Seon Ki Park · Liang Xu *Editors*

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)

Springer

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)

Seon Ki Park · Liang Xu
Editors

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)

Springer

Editors
Seon Ki Park
Climate and Energy Systems Engineering
Ewha Womans University
Seoul, Korea (Republic of)

Liang Xu
Marine Meteorology Division
Naval Research Laboratory
Monterey, CA, USA

ISBN 978-3-030-77721-0 ISBN 978-3-030-77722-7 (eBook)
https://doi.org/10.1007/978-3-030-77722-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To
Yoshi K. Sasaki
and
Anna Trevisan

Preface

Data assimilation (DA) seeks to find the best available estimation of a state by combining all the available information, such as the dynamic models, the observations, and the associated error statistics, using one of the many algorithms based on modern estimation theory. DA has been used to provide the numerical model with the best initial conditions in the early days of numerical weather prediction (NWP). It has become one of the cornerstones in giving state estimations and numerical forecasts in various earth science disciplines such as oceanography and hydrology.

Since June 2005, we have convened a recurring session, titled "Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications," at the Asia Oceania Geosciences Society (AOGS) annual meetings. We added the "Sasaki Symposium" to the session title in June 2007 to honor Prof. Yoshi K. Sasaki of the University of Oklahoma for his life-long contributions to DA in geosciences. At the 5th AOGS Annual Meeting held in Busan, Korea, in June 2008, we realized the need for a book in DA that can include the fundamentals of DA theory, development, applications, and notable recent advances in the field. Right after the AOGS meeting in Busan, we prepared the book; at that time, we had never thought or planned to publish the books in a series. Since then, we have successfully published three volumes of the book.

Volume I was published in March 2009 with 27 chapters, including notable invited ones and some selected ones from the previous symposiums. Among them, "Data Assimilation for Numerical Weather Prediction: A Review" by Ionel M. Navon provided a thorough review on 4DVar and recorded the highest citations among all the volumes in this series (171 times as of February 2021 from Google Scholar). "Real Challenge of Data Assimilation for Tornadogenesis" was authored by Yoshi himself, proposing a new theory based on the entropic balance.

Volume II was published in May 2013, again with 27 chapters, by collecting both invited papers and selected papers from the previous symposiums held in 2009 (Singapore, Singapore), 2010 (Hyderabad, India), and 2011 (Taipei, Taiwan). The volume included excellent overviews on estimation theory, nudging and variational methods, and Markov chain Monte Carlo methods. Most prominently, Yoshi extended his entropy balance theory for tornado DA and contributed a chapter titled "Entropic

Balance Theory and Radar Observation for Prospective Tornado Data Assimilation," which became his last contribution to this series.

Volume III was published in December 2016 with 24 chapters with a special dedication section, titled "In Memory of Yoshi," by collecting memories on and photos of Yoshi from several authors. The volume included excellent overviews on variational DA (Le Dimet et al.), DA for coupled modeling systems (Milija Županski), representer-based variational DA (Chua and Xu), soil moisture DA (Maggioni and Houser), and surface DA over complex terrain (Zhaoxia Pu), to mention a few, as well as updates on a wide range of other DA topics.

Since Volume III's publishing, many advances have been made in various parts of geoscientific data assimilation. In January 2020, we thought it was the right time to publish Volume VI to provide our readers with the needed updates in the recent advances of DA theories and applications. We surveyed many well-known leading data assimilation experts in their respective specialty areas regarding their willingness to contribute to Volume IV. Encouraged by the overwhelmingly positive responses from the survey, we officially started the process of publishing Volume IV in February 2020. Despite all the challenges associated with the COVID-19 pandemic during the past year, we completed the new volume thanks to the authors' and anonymous reviewers' tireless efforts.

Volume IV contains 26 chapters contributed by experts from universities and operational centers around the world. It includes an excellent overview of the fundamentals and the applications of the GNSS-RO sounding in the troposphere and stratosphere (Ruston et al.), a must-have resource on all things associated with the assimilation of in situ observations in numerical weather predictions at ECMWF and NRL (Pauley and Ingley), some theoretical considerations regarding the observation network design (Lewis et al. and Lakshmivarahan et al.), a sparsity-based Kalman filter algorithm to reduce computational requirement (Kang and Xu), modeling observation error through statistical parameter estimation (Satterfield et al.), an excellent review of the second-order methods in variational data assimilation (Le Dimet and Shutyaev), the strongly coupled data assimilation (Wu et al. and Zhaoxia Pu), sensitivity analysis in ocean acoustic propagation (Ngodock et al.), and recent progress at the operational centers (Baker et al., Bouyssel et al., Lee and Huang, Martet et al., and Shin et al.) among other progresses in the DA community.

This book will be useful to individual researchers and graduate students as a reference to the most recent data assimilation progress in geosciences. We dedicate this volume to Yoshi Sasaki and Anna Trevisan for their monumental contributions to data assimilation. Yoshi's journey into and contribution to DA was well described in "Sasaki's Pathway to Deterministic Data Assimilation" in Volume I, authored by John M. Lewis. Anna was a pioneer working at the crossroad between the dynamical system and data assimilation: she invented a new DA concept that was based on the dynamical properties of underlying systems, paving the way toward efficient DA and deployment of adaptive observations. "Data Assimilation for Chaotic Dynamics" in this volume, by Carrassi et al., can be considered as Anna's legacy.

Lastly, we are deeply saddened by the passing of Prof. François-Xavier Le Dimet, an eminent scientist who made significant contributions to the data assimilation

community. He had also served as a co-convener of the Sasaki symposiums and contributed several chapters to our book series, including two chapters in this volume. He will be greatly missed.

Seoul, Korea (Republic of) Seon Ki Park
Monterey, USA Liang Xu
March 2021

Contents

Contributors

Kwang Deuk Ahn Korea Meteorological Administration, Seoul, Korea

Boujemaa Ait-El-Fquih King Abdullah University of Science and Technology (KAUST), PSE, Thuwal, Saudi Arabia

Clark Amerault Naval Research Laboratory, Monterey, CA, USA

Nancy L. Baker Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

Paulo Batista Instituto Nacional de Pesquisas Esaciais, São José dos Campos, SP, Brazil

L. Berre CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Marc Bocquet CEREA, Joint Laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-Sur-Marne, France

F. Bouyssel CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Neill Bowler Met Office, Exeter, UK

Pierre Brousseau CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France

Peter Brown Department of Physics and Astronomy, University of Western Ontario, Ontario, CA, USA

H. Bénichou Direction des opérations, Météo-France, Toulouse, France

Alberto Carrassi Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, UK;
Mathematical Institute, University of Utrecht, Utrecht, Netherlands

Matthew J. Carrier Oceanography Division, U.S. Naval Research Laboratory, Stennis Space Center, MS, USA

P. Chambon CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Hsiang-Wen Cheng National Space Organization, Hsinchu, Taiwan

Adam Clayton Korea Institute of Atmospheric Prediction Systems, Seoul, South Korea

Jonathan Demaeyer Royal Meteorological Institute of Belgium, Brussels, Belgium

James D. Doyle Naval Research Laboratory, Monterey, CA, USA

Le Duc Meteorological Research Institute, Tsukuba, Japan;
Japan Meteorological Business Support Center, Tokyo, Japan

Stephen D. Eckermann U.S. Naval Research Laboratory, Washington, DC, USA

Patrick J. Espy Atmospheric and Environmental Physics at the Norwegian University of Science and Technology, Trondheim, Norway

Josette Fabre The Naval Research Laboratory Code 7180, 1005, Stennis Space Center, Balch Blvd, MS, USA

David C. Fritts GATS, Inc., Boulder, CO, USA

N. Girardot CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Leah D. Grant Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

Lewis Grasso Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Colin Grudzien Department of Mathematics and Statistics, University of Nevada, Reno, Reno, NV, USA

V. Guidard CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Frank Guillaume CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France

Ji-Hyun Ha Korea Meteorological Administration, Seoul, Korea

Chris Hall Tromsø Geophysical Observatory, University of Tromsø, Tromsø, Norway

Sean Healy European Centre for Medium Range Weather Forecasts (ECMWF), Reading, UK

Dan Hodyss U.S. Naval Research Laboratory, Washington, DC, USA

Xiaodong Hong Naval Research Laboratory, Monterey, CA, USA

Karl W. Hoppel U.S. Naval Research Laboratory, Washington, DC, USA

Ibrahim Hoteit King Abdullah University of Science and Technology (KAUST), PSE, Thuwal, Saudi Arabia

Xiang-Yu Huang Centre for Climate Research Singapore, Singapore, Singapore

Zih-Mao Huang Central Weather Bureau, Taipei, Taiwan

Hiroyuki Iimura GATS, Inc., Boulder, CO, USA

Bruce Ingleby Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Christoph Jacobi Institute for Meteorology, University of Leipzig, Leipzig, Germany

Wei Kang Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, USA

Chang Hwan Kim Korea Institute of Atmospheric Prediction Systems, Seoul, South Korea

Anton Kliewer Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA;
NOAA Earth System Research Laboratory, Global System Laboratory, Boulder, CO, USA

David D. Kuhl U.S. Naval Research Laboratory, Washington, DC, USA

Takuya Kurihana Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan;
Department of Computer Science, University of Chicago, Chicago, USA

S. Lakshmivarahan National Severe Storms Laboratory (NSSL), Desert Research Institute (DRI), Reno, NV, USA;
School of Computer Science, University of Oklahoma Norman, Norman, OK, USA

Rolf H. Langland Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

Ralph Latteck Leibniz Institute of Atmospheric Physics at the University of Rostock, Kuhlungsborn, Germany

François-Xavier Le Dimet Laboratoire Jean-Kuntzman, Université Grenoble-Alpes, Grenoble, France

Eun Hee Lee Korea Meteorological Administration, Seoul, Korea

Joshua Chun Kwang Lee Centre for Climate Research Singapore, Singapore, Singapore

Yong Hee Lee Korea Institute of Atmospheric Prediction Systems, Seoul, South Korea

John M. Lewis National Severe Storms Laboratory (NSSL), Desert Research Institute (DRI), Reno, NV, USA

Guozhu Li Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, CN, China

Na Li National Key Laboratory of Electromagnetic Environment, China Research Institute of Radiowave Propagation, Qingdao, CN, China

Sujeong Lim Department of Climate and Energy Systems Engineering, Ewha Womans University, Seoul, Republic of Korea

C. Loo CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Jun Ma CPI, Springfield, VA, USA

Jean-François Mahfouf CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Jean-François Mahfouf CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France

John Marino Colorado Center for Astrodynamics Research, Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, USA

Maud Martet CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France

S. K. R. Maryada National Severe Storms Laboratory (NSSL), Desert Research Institute (DRI), Reno, NV, USA

Takumi Matsunobu Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan;
Meteorologisches Institut, Ludwig-Maximilians-Universität, Munich, Germany

John P. McCormack U.S. Naval Research Laboratory, Washington, DC, USA;
ATRAD Pty Ltd, 24 Now at Heliophysics Division, NASA Headquarters, Washington, DC, USA

Chris E. Meek Institute of Space and Atmospheric Studies, University of Saskatchewan, Saskatoon, CA, USA

Nicholas J. Mitchell Centre for Space Atmosphere and Oceanic Science, Department of Electronic and Electrical Engineering, University of Bath, Bath, UK

Takemasa Miyoshi RIKEN Center for Computational Science, Kobe, Japan;
RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program, Kobe, Japan;
Prediction Science Laboratory, RIKEN Cluster for Pioneering Research, Kobe, Japan;

University of Maryland, College Park, Maryland, USA;
Japan Agency for Marine–Earth Science and Technology, Yokohama, Japan

Tracy Moffat-Griffin British Antarctic Survery, Cambridge, UK

P. Moll CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Damian J. Murphy Environment, Water, Population and Communities, Australian Antarctic Division of Sustainability, Kingston, TAS, Australia

Hans E. Ngodock The Naval Research Laboratory Code 7320, 1009, Stennis Space Center, Balch Blvd, MS, USA

Hans E. Ngodock The US Naval Research Laboratory, Stennis Space Center, MS, USA;
Oceanography Division, U.S. Naval Research Laboratory, Stennis Space Center, MS, USA

John J. Osborne Oceanography Division, U.S. Naval Research Laboratory, Stennis Space Center, MS, USA

Seon Ki Park Department of Climate and Energy Systems Engineering, Ewha Womans University, Seoul, Republic of Korea

Patricia M. Pauley Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

C. Payan CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Andrey A. Popov Virginia Tech, Blacksburg, VA, USA

Zhaoxia Pu Department of Atmospheric Sciences, University of Utah, Salt Lake City, UT, USA

Patrick Raanes NORCE, Bergen, Norway

D. Raspaud CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Sai Kiran Reddy Maryada School of Computer Science, University of Oklahoma Norman, Norman, OK, USA

Iain M. Reid School of Physical Sciences, The University of Adelaide, Adelaide, Australia;
ATRAD Pty Ltd, 24 Now at Heliophysics Division, NASA Headquarters, Washington, DC, USA

Benjamin Ruston Naval Research Laboratory (NRL), Monterey, CA, USA

Kazuo Saito Atmosphere and Ocean Research Institute, University of Tokyo, Chiba, Japan;
Meteorological Research Institute, Tsukuba, Japan;
Japan Meteorological Business Support Center, Tokyo, Japan

Adrian Sandu Virginia Tech, Blacksburg, VA, USA

Elizabeth A. Satterfield Marine Meteorology Division, U.S. Naval Research Laboratory, Monterey, CA, USA

Hui Shao Joint Center for Satellite Data Assimilation (JCSDA)/University Corporation for Atmospheric Research (UCAR), College Park, MD, USA

Hyun-Cheol Shin Korea Meteorological Administration, Seoul, Korea

Victor Shutyaev Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow, Russia

Scott Smith The Naval Research Laboratory Code 7320, 1009, Stennis Space Center, Balch Blvd, MS, USA

Innocent Souopgui Department of Physics, The University of New Orleans, New Orleans, LA, USA;
Department of Marine Sciences, The University of Southern Mississippi, Stennis Space Center, Balch Blvd, MS, USA

Gunter Stober Institute of Applied Physics & Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Rebecca E. Stone SAIC, Monterey, CA, USA

Eswaraiah Sunkara Space Science Laboratory, Department of Astronomy and Space Science, Chungnam National University, Daejeon, Korea

Koji Terasaki RIKEN Center for Computational Science, Kobe, Japan

Chih-Chien Tsai National Science and Technology Center for Disaster Reduction, New Taipei, Taiwan

Masaki Tsutsumi National Institute of Polar Research, Tokyo, Japan

Daniel P. Tyndall Naval Research Laboratory, Monterey, CA, USA

Stéphane Vannitsem Royal Meteorological Institute of Belgium, Brussels, Belgium

Joanne A. Waller Met Office, Reading, UK

Eric Wattrelot CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France

Pin-Ying Wu Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan

Ting-Chi Wu Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Liang Xu Naval Research Laboratory, Monterey, CA, USA

Shu-Chih Yang Department of Atmospheric Sciences, National Central University, Taoyuan, Taiwan

Wen Yi CAS Key Laboratory of Geospace Environment, Department of Geophysics and Planetary Sciences, University of Science and Technology of China, Hefei, CN, China

Robert Zingarelli The Naval Research Laboratory Code 7180, 1005, Stennis Space Center, Balch Blvd, MS, USA

Milija Zupanski Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Data Assimilation for Chaotic Dynamics

Alberto Carrassi, Marc Bocquet, Jonathan Demaeyer, Colin Grudzien, Patrick Raanes, and Stéphane Vannitsem

Abstract Chaos is ubiquitous in physical systems. The associated sensitivity to initial conditions is a significant obstacle in forecasting the weather and other geophysical fluid flows. Data assimilation is the process whereby the uncertainty in initial conditions is reduced by the astute combination of model predictions and real-time data. This chapter reviews recent findings from investigations on the impact of chaos on data assimilation methods: for the Kalman filter and smoother in linear systems, analytic results are derived; for their ensemble-based versions and nonlinear dynamics, numerical results provide insights. The focus is on characterizing the asymptotic statistics of the Bayesian posterior in terms of the dynamical instabilities, differentiating between deterministic and stochastic dynamics. We also present two novel results. Firstly, we study the functioning of the ensemble Kalman filter in the context of a chaotic, coupled, atmosphere-ocean model with a quasi-degenerate spectrum of Lyapunov exponents, showing the importance of having sufficient ensemble mem-

A. Carrassi (✉)
Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, UK
e-mail: n.a.carrassi@reading.ac.uk

Mathematical Institute, University of Utrecht, Utrecht, Netherlands

M. Bocquet
CEREA, Joint Laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-Sur-Marne, France
e-mail: marc.bocquet@enpc.fr

J. Demaeyer · S. Vannitsem
Royal Meteorological Institute of Belgium, Brussels, Belgium
e-mail: jodemaey@meteo.be

S. Vannitsem
e-mail: svn@meteo.be

C. Grudzien
Department of Mathematics and Statistics, University of Nevada, Reno, Reno, NV, USA
e-mail: cgrudzien@unr.edu

P. Raanes
NORCE, Bergen, Norway
e-mail: para@norceresearch.no

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_1

bers to track all of the near-null modes. Secondly, for the fully non-Gaussian method of the particle filter, numerical experiments are conducted to test whether the curse of dimensionality can be mitigated by discarding observations in the directions of little dynamical growth of uncertainty. The results refute this option, most likely because the particles already embody this information on the chaotic system. The results also suggest that it is the rank of the unstable-neutral subspace of the dynamics, and not that of the observation operator, that determines the required number of particles. We finally discuss how knowledge of the random attractor can play a role in the development of future data assimilation schemes for chaotic multiscale systems with large scale separation.

1 Introduction

This chapter attempts a unified and comprehensive discussion of a number of studies that in about a decade have contributed to shape our nowadays's understanding of the implications, impacts and consequences for data assimilation when is applied to chaotic dynamics. The chapter presents a review of the essential results appeared in different studies, but for the first time together in a coherent treatment. In addition we present new original findings addressing two key aspects that were not covered in previous studies. The first treats the impact of data assimilation on a chaotic and multiscale system, the second concerns the consequences for nonlinear, non-Gaussian, data assimilation (a particle filter) in face of the chaotic nature of the underlying dynamics.

The exposition is organised as follows. We first discuss, in Sect. 2, the chaotic character of atmospheric and oceanic flows, and provide a treatment of the key mathematical concepts and tools that allow for characterising chaos and to "measure" the degree of instabilities. In doing so, we review classical results from dynamical system theory, including the *multiplicative ergodic theorem* and associated definition of Lyapunov (forward, backward and covariant) vectors and exponents.

Section 3 analyses how a chaotic dynamics impacts data assimilation. Our focus is on Kalman filter (KF) and smoother (KS) in Sect. 3.1, and on their ensemble-based formulations, the ensemble Kalman filter (EnKF) and smoother (EnKS), in Sect. 3.2. Section 3.1 contains primarily analytic results and treats KF and KS in linear systems, either purely deterministic (Sect. 3.1.1) or with stochastic additive noise (Sect. 3.1.2). Section 3.2 is dedicated to nonlinear systems (deterministic, in Sect. 3.2.1 and stochastic in Sect. 3.2.2) and on how the EnKF (Evensen 2009a) and the extended Kalman filter (EKF, Ghil and Malanotte-Rizzoli 1991) works in this scenario. In Sect. 3.2.1 we present original results on the impact of chaos on the performance of the EnKF in a coupled atmosphere-ocean model which possesses a degenerate-like spectrum of Lyapunov exponents, disentangling on the role of the quasi-null exponents.

Section 4 reverses the perspective and instead of studying the effect of chaos on data assimilation, reports on how properties of the dynamics have been used to devise

adaptive observation strategies and ad-hoc data assimilation methods. In particular, in Sect. 4.2, we succinctly review the assimilation in the unstable subspace (AUS, Palatella et al. 2013), a known approach that exploits the unstable-neutral subspace of the dynamics to perform the analysis. Remarkably, AUS was conceived before the findings reviewed in Sect. 3. The latter work was inspired by these early studies, and the need to provide mathematical rigor to clarify the mechanisms that made these early studies successful. In turn, the later work has furthermore provided the framework to generalize these early ideas to a variety of other types of dynamics.

Section 5 presents what we consider two main areas of future developments. In Sect. 5 the paradigm of AUS is incorporated within a fully nonlinear data assimilation scheme, the particle filter (Van Leeuwen et al. 2019). It is shown that observing in the directions of instabilities is effective but also that it is not deleterious to observe the stable directions. As opposed to the data or model sizes, our results suggest that the number of particles required to achieve good results scales with the size of the unstable-neutral subspace. Section 5.2 discusses how the concept of random attractor can offer novel ways to handle the data assimilation problem in stochastic chaotic multi-scale systems with large scale separation. It also treats the implications of the numerical scheme on the output of the data assimilation cycle and ensemble-based forecasts, as well as poses some key questions for future studies.

Final conclusions and a summary are drawn in Sect. 6.

2 Chaos in Atmospheric and Oceanic Flows

The atmosphere and the ocean are fluids that are described by the set of classical conservation laws of hydrodynamics, including the conservation of mass, momentum and energy (Vallis 2017). For the atmosphere, these are often complemented by the conservation of moisture present in the air. These laws lead to a set of local dynamical equations describing the motion of each parcel of fluid. Given that these equations are nonlinear with complex interactions with the boundaries, realistic solutions cannot be obtained analytically and one must rely on numerical simulations starting from appropriate initial conditions.

Numerical simulations are based on discrete approximations in space and time of the dynamical equations, and are often accompanied with simplifications of the equations in order to describe appropriately the scales of interest. One of the most famous approximations is the geostrophic approximation which assumes a balance between the horizontal velocity field and the horizontal pressure gradient. Geostrophic balance is a good approximation for both the ocean and the atmosphere albeit at different spatial scales. The geostrophic approximation is at the basis of the models that will be used later in Sect. 3.2.1.

Whatever the scale at which the atmospheric or the ocean fluids are observed, they display an apparently erratic evolution. This erratic behavior is also present in atmospheric, ocean and climate models when appropriate forcing are imposed. This feature should not be confused with randomness as most of the models used

since the start of numerical modelling were purely deterministic. Lorenz (1963) showed in a simple low-order model that this erratic behavior is concomitant with the property of sensitivity to initial conditions, by which whatever small an error in the initial condition is, it will increase rapidly in time. This property implies that given the inevitable error in the initial conditions, any forecast will ultimately become useless as the error finally reaches an amplitude of the same order as the natural variability of the variable considered. The sensitivity to initial conditions and the consequent erratic-like evolution are the key properties of deterministic dynamical systems displaying *chaos*.

This behaviour has been found in many atmospheric, oceanic and climate models (see e.g., Vannitsem 2017, for a review). In particular, a detailed investigation of the chaotic nature of the coupled ocean-atmosphere system that will be used in Sect. 3.2.1 has been performed by Vannitsem et al. (2015).

2.1 Measuring Sensitivity to Initial Conditions

The notion of sensitivity to initial conditions of dynamical solutions was already discovered and studied in a mathematical context by Poincaré (1899). In the second half of the 20th century, this property was discovered in models of atmospheric and climate relevance (Thompson 1957; Lorenz 1963). Its important practical implications drove the regain of interest in developing the appropriate mathematics in support of its description and understanding. These efforts culminated with the development of the ergodic theory of deterministic dynamical systems and chaos theory. In the following we shall briefly summarise what we consider to be the key developments that led to the definition of Lyapunov exponents and vectors.

The importance of these mathematical objects stands on their ability to "measure" the degree of instabilities and thus to quantify the aforementioned sensitivity to initial conditions. While more rigorous mathematical treatments can be found in appropriate mathematical literature (see e.g., Pikovsky and Politi 2016, and references therein), and we will invoke that rigour to a certain extent in the following sections, here we approach the discussion with a physical and intuitive angle. Further details can also be found in Legras and Vautard (1996), Barreira and Pesin (2002), Kuptsov and Parlitz (2012).

Let us write the evolution laws of a deterministic dynamical system in the form of a set of ordinary differential equations (ODEs),

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \boldsymbol{\sigma}), \tag{1}$$

where $\mathbf{x}$ is a vector containing the entire set of relevant variables $\mathbf{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{\sigma}$ represents a set of parameters. The discussion that follows holds also when system (1) explicitly depends on time, thus being non-autonomous, provided it remains ergodic.

Since the process of measurement is always subject to finite precision, the initial state is never known exactly. To study the evolution and the implications of such an error, let us consider an initial state displaced slightly from $\mathbf{x}_0$ by an initial error $\delta\mathbf{x}_0$. The perturbed initial state, $\mathbf{x}_0 + \delta\mathbf{x}_0$, generates a new trajectory in phase space and one can define the instantaneous error vector as the vector joining the representative points of the reference trajectory and the perturbed one at a given time, $\delta\mathbf{x}(t)$. Provided that this perturbation is sufficiently small and smooth, its dynamics can be described by the linearized equation,

$$\frac{\mathrm{d}\delta\mathbf{x}}{\mathrm{d}t} \approx \frac{\partial\mathbf{F}}{\partial\mathbf{x}_{|\mathbf{x}(t)}}\delta\mathbf{x}, \tag{2}$$

with a formal solution,

$$\delta\mathbf{x}(t) \approx \mathbf{M}(t, \mathbf{x}(t_0))\delta\mathbf{x}(t_0). \tag{3}$$

The matrix $\mathbf{M}$ is referred to as the "fundamental matrix" and is the resolvent of Eq. (2), i.e. $\mathbf{M}(t, \mathbf{x}(t_0)) = e^{\int_{t_0}^{t} \frac{\partial\mathbf{F}}{\partial\mathbf{x}}_{|\mathbf{x}(t)} \mathrm{d}t}$. The fundamental matrix contains the information on the amplification of infinitesimally small perturbations.

In the context of the ergodic theory of deterministic dynamical systems, the Oseledets theorem (Kuptsov and Parlitz 2012) shows that the limit of the matrix $(\mathbf{M}^\top\mathbf{M})^{1/[2(t-t_0)]}$, for time going to infinity, exists; let us refer to this limiting matrix as $\mathbf{S}$. The logarithm of its eigenvalues are called the *Lyapunov exponents* (LEs), whereas the full set of LEs is called the Lyapunov spectrum, and is usually represented in decreasing order. The eigenvectors of $\mathbf{S}$, which are local properties of the flow (they change along the trajectory thus being time-dependent) and depend on the initial time t_0, are called the *forward Lyapunov vectors* (FLVs) (Legras and Vautard 1996).

The LEs are a powerful tool to "measure" chaos and to characterise the degree of instability of a system. For instance, LEs are averaged (asymptotic) indicators of exponential growth (LE > 0) or decay (LE < 0) of perturbations under the tangent-linear model. A deterministic chaotic system is uniquely characterised by having at least its leading LE larger than zero, i.e. $\mathrm{LE}_1 > 0$. The sum of the LEs is equal to the average divergence of flow generated by Eq. (1) (see e.g. Pikovsky and Politi 2016). This means that in dissipative (conservative, e.g. Hamiltonian) systems the sum of the LEs is negative (zero): volumes in the phase space of a dissipative (conservative) dynamics reduce (are conserved) on average with time. Furthermore, autonomous continuous-in-time systems generally possess at least one LE $= 0$, unless they converge to a motion-less state. This null exponent is related to perturbations aligned to the system's velocity vector (i.e. $\delta\mathbf{x} = \mathbf{f}$), so although they shall fluctuate depending on the local flow, they will not on average decay nor growth. These last two properties can also be used as a check for numerical accuracy when computing the LEs.

The multiplicative ergodic theorem (MET) (Barreira and Pesin 2002, Theorem 2.1.2) guarantees that, under general hypotheses, the eigenvalues of the matrix $(\mathbf{M}^\top\mathbf{M})^{1/[2(t-t_0)]}$ obtained for $t \to \infty$ are equivalent to the ones of the matrix $\mathbf{S}' = (\mathbf{M}\mathbf{M}^\top)^{1/[2(t-t_0)]}$ when $t_0 \to -\infty$. The equivalence of the spectrum of $\mathbf{S}$ and $\mathbf{S}'$

is not generically true for the fundamental matrix of an arbitrary, linear dynamical system; however, the MET guarantees that this is a fairly generic property of the resolvent of the *tangent-linear model of a nonlinear dynamical system*. If the flow of the time derivative $\mathbf{f}$ is a $\mathcal{C}^1$ diffeomorphism of a compact, smooth, Riemannian manifold M, then the MET assures that the LEs are defined equivalently by the log-eigenvalues of $\mathbf{S}$ or $\mathbf{S}'$ for any initial condition $\mathbf{x}(t_0)$ and that the LEs are unique on a subset of M of full measure with respect to any ergodic invariant measure μ of the flow. Note that, without the condition of ergodicity of μ, the LEs may be well-defined point-wise, but the specific values and their multiplicity may depend on the initial condition $\mathbf{x}(t_0)$.

The matrices $\mathbf{S}$ and $\mathbf{S}'$ are symmetric. However, contrary to the eigenvalues, the eigenvectors of these two matrices are not equivalent due to the asymmetric character of the fundamental matrix $\mathbf{M}$ in forward- and reverse-time. The eigenvectors of the latter are called the *backward Lyapunov vectors* (BLVs). Theoretically, each matrix can be evaluated at the same place along the reference trajectory $\mathbf{x}(t)$ and their orthogonal eigenvectors can be computed as $\mathbf{L}_t^{\mathrm{f},i}$ and $\mathbf{L}_t^{\mathrm{b},i}$ for $\mathbf{S}$ and $\mathbf{S}'$, respectively, where it is understood that the time-dependence on t is with respect to the linearization of the dynamics at $\mathbf{x}(t)$. There exist *Oseledet* subspaces W_t^i,

$$W_t^i = \mathbf{L}_t^{\mathrm{b},1} \oplus \cdots \oplus \mathbf{L}_t^{\mathrm{b},i} \cap \mathbf{L}_t^{\mathrm{f},i} \oplus \cdots \oplus \mathbf{L}_t^{\mathrm{f},N}, \tag{4}$$

with $\oplus$ being the direct product (Ruelle 1979), that have the important properties of being invariant under the effect of the fundamental matrix, such that

$$\mathbf{M}(\tau, \mathbf{x}(t))W_t^i = W_\tau^i. \tag{5}$$

Due to their orthogonal nature, the FLVs and the BLVs require by definition the choice of a norm and of an inner product. Nevertheless, the Oseledet subspaces themselves do not have this dependence; in this way, they can be considered to embed more invariant information about the dynamics. The decomposition of the tangent-linear space into these covariant subspaces is commonly known as the "Oseledet splitting" or decomposition. The classical form of the MET thus guarantees that the Oseledet splitting is well-defined and consistent with probability one over all initial conditions of the attractor, with respect to the invariant, ergodic measure. Other covariant splittings of the tangent-linear model, such as by exponential dichotomy, exist under more general forms of the MET (Froyland et al. 2013).

When the Lyapunov spectrum is non-degenerate, one can define a time-varying basis, subordinate to the Oseledet spaces, $W_t^i = \text{span}\left\{\mathbf{L}_t^{\mathrm{c},i}\right\}$, such that

$$\mathbf{M}(\tau, \mathbf{x}(t))\mathbf{L}_t^{\mathrm{c},i} = \alpha_i(\tau, \mathbf{x}(t))\mathbf{L}_\tau^{\mathrm{c},i}, \tag{6}$$

where $\| \mathbf{L}_t^{\mathrm{c},i} \| = \| \mathbf{L}_\tau^{\mathrm{c},i} \|$, and $\alpha_i(\tau, \mathbf{x}(t)) \in \mathbb{R}$ describes an amplification factor. The vectors $\mathbf{L}_t^{\mathrm{c},i}$ are known as the *covariant Lyapunov vectors* (CLVs). In the long-time limit, the amplifications $\alpha_i(\tau, \mathbf{x}(t))$ can be associated to the LEs as,

$$\pm\lambda_i = \lim_{\tau\to\pm\infty} \frac{1}{\tau} \ln\ |\alpha_i(\tau, \mathbf{x}(t))| = \lim_{\tau\to\pm\infty} \lambda_i^\tau(\mathbf{x}(t)) \tag{7}$$

where we indicate by $\lambda_i^\tau(\mathbf{x}(t))$ the average of the growth rate taken over a time window τ starting at t at position $\mathbf{x}(t)$), and are commonly known as the local Lyapunov exponents (LLEs) (Pikovsky and Politi 2016, see Chap. 5).

Throughout the chapter we will denote the matrix with columns corresponding to the full tangent linear space ordered basis of BLVs/FLVs/CLVs at time t as $\mathbf{L}_t^{\mathrm{e}}$ for $\mathrm{e} \in \{\mathrm{b}, \mathrm{f}, \mathrm{c}\}$ respectively. A sub-slice of this matrix of Lyapunov vectors corresponding, inclusively, to columns i through j will be denoted $\mathbf{L}_t^{\mathrm{e},i:j}$. In the following, the relationships between these Lyapunov vectors and their asymptotic dynamics will be key to understanding the predictability of chaotic systems. For this purpose, we will largely use the BLVs and the CLVs, and to a lesser extent the FLVs.

Note that the basis $\left\{\mathbf{L}_t^{\mathrm{c},i}\right\}_{i=1}^{N_x}$ is not orthogonal in general and, in fact, the angles between any two Oseledec subspaces W_i and W_j are not in general bounded away from zero in their limits in forward- or reverse-time. For this reason, coordinate transformations to covariant Oseledet bases are not generally numerically well-conditioned asymptotically, and therefore hard to compute. The property of *integral separation* (Dieci and Van Vleck 2002), describing the stability of the Lyapunov spectrum under bounded perturbations of the tangent-linear equations, ensures that the angles between the covariant subspaces will remain bounded away from zero, but this is a strong condition and it is not as generic of a property as the existence of the Oseledet decomposition under the MET. However, if a dynamical system is integrally separated as above, there exists a well-defined, numerically well-conditioned transformation of coordinates of the tangent-linear model for which the action of the resolvent $\mathbf{M}$ can be expressed as a block-diagonal matrix with each block describing the invariant dynamics of a single Oseledet space. For degenerate spectrum, these blocks may be upper-triangular, but for non-degenerate spectrum this representation of the resovlent becomes a strictly diagonal matrix; see Theorem 5.4.9 of Adrianova (1995) for the classical result, or Theorem 5.1 of Dieci and Van Vleck (2007) and Froyland et al. (2013) for more recent extensions.

3 Data Assimilation in Chaotic Systems—How the Dynamics Impacts the Way We Assimilate Data

The high sensitivity of a chaotic dynamical system to the initial condition makes it hard to forecast it, even when their evolution equations are perfectly known. Indeed the typical error grows exponentially over time with a rate given by the largest positive LE. With a view to forecasting, one has no choice but to regularly correct its trajectory using information on the system state obtained through observations. This is the primary goal of data assimilation (DA) and has been key to the success

of numerical weather forecasting. We refer to Kalnay (2003), Asch et al. (2016) and references therein for textbooks and reviews on geophysical DA.

When framed in a Bayesian formalism, the goal of DA is to estimate the conditional probability density function (pdf) of the system state knowing observations of that system. For high-dimensional systems, only approximations of this pdf can be obtained, among which the Gaussian approximation is the most practical and common (Carrassi et al. 2018). In the case of Gaussian statistics of the error and linear dynamics, the conditional pdf can be obtained analytically: it is Gaussian, and can be sequentially computed using the Kalman filter (KF) (Kalman 1960). Directly inspired from the Kalman filter, the extended Kalman filter (EKF) offers an approximate DA scheme when the operators are nonlinear (Ghil and Malanotte-Rizzoli 1991). However, it can hardly be used in the context of high-dimensional systems, where it has to be replaced with the ensemble Kalman filter (EnKF) (Evensen 2009b). In the EnKF, the covariance matrices are represented by state perturbations which are representative of the errors and, together with the state mean, form a limited-size ensemble of state vectors.

Understanding how the ensemble of the EnKF evolves under the action of the DA scheme and the forecast model dynamics is important. Several numerical results suggest that the skills of ensemble-based DA methods in chaotic systems are related to the instabilities of the underlying dynamics (Ng et al. 2011). Numerical evidences exist that some asymptotic properties of the ensemble-based covariance matrices (rank, span, range) relate to the unstable modes of the dynamics (Sakov and Oke 2008a; Carrassi et al. 2009). Nevertheless, a better, more profound, understanding of these results was needed to aim at designing reduced rank, computationally cheap, formulations of the filters.

Analytic results that have shed lights on the behaviour of filters and smoothers on chaotic dynamics, and explained the numerically observed properties, have been obtained for linear dynamics. Section 3.1.1 reviews those findings in the case of deterministic systems without model error, based on the work by Gurumoorthy et al. (2017) and Bocquet et al. (2017). Section 3.1.2 treats their extension to the case of stochastic systems, following the work by Grudzien et al. (2018a), Grudzien et al. (2018b). In the case of nonlinear dynamics, robust numerical evidences will be described by the original experiments in Sect. 3.2, that extent the previous findings by Bocquet and Carrassi (2017).

3.1 Linear Dynamics: The Effect of Chaos on the Kalman Filter and Smoother

3.1.1 Perfect and Deterministic Dynamics

At time t_k, let $\mathbf{x}_k \in \mathbb{R}^{N_x}$ and $\mathbf{y}_k \in \mathbb{R}^{N_y}$ be the state and observation vector, respectively. Let us assume linear evolution model dynamics $\mathbf{M}_k$ and observation model $\mathbf{H}_k$ such

that

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \mathbf{w}_k, \tag{8a}$$
$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k. \tag{8b}$$

The model and observation noises, $\mathbf{w}_k$ and $\mathbf{v}_k$, are assumed mutually independent, zero-mean Gaussian white sequences with statistics

$$\mathrm{E}[\mathbf{v}_k \mathbf{v}_l^\top] = \delta_{k,l} \mathbf{R}_k, \quad \mathrm{E}[\mathbf{w}_k \mathbf{w}_l^\top] = \delta_{k,l} \mathbf{Q}_k, \quad \mathrm{E}[\mathbf{v}_k \mathbf{w}_l^\top] = \mathbf{0}\,. \tag{9}$$

In the Kalman filter, which yields the optimal DA solution with such assumptions, the forecast error covariance matrix $\mathbf{P}_k$ satisfies the recurrence

$$\mathbf{P}_{k+1} = \mathbf{M}_{k+1} \left(\mathbf{I} + \mathbf{P}_k \mathbf{\Omega}_k\right)^{-1} \mathbf{P}_k \mathbf{M}_{k+1}^\top + \mathbf{Q}_{k+1}, \tag{10}$$

where

$$\mathbf{\Omega}_k \equiv \mathbf{H}_k^\top \mathbf{R}_k^{-1} \mathbf{H}_k, \tag{11}$$

is the *precision matrix* of the observations mapped into state space. In the absence of model error, i.e. $\mathbf{Q}_k \equiv \mathbf{0}$, Gurumoorthy et al. (2017) proved rigorously that in the full-rank KF (in particular $\mathbf{P}_0$ is full rank), $\mathbf{P}_k$ collapses onto the unstable-neutral subspace.

We shall summarise in the following some key results that apply to the full-rank but also to the degenerate case, i.e. even if $\mathbf{P}_0$ is of arbitrary rank and the initial errors of the DA scheme only lie in a subspace of $\mathbb{R}^{N_x}$. In particular they apply to the EnKF if the dynamical and observations operators are both linear. We recall them here, but readers can find full details in Bocquet et al. (2017), Bocquet and Carrassi (2017).

Result 1: Bound on the free forecast error covariance matrix

Let us define the resolvent of the dynamics from t_l to t_k as $\mathbf{M}_{k:l} = \mathbf{M}_k \mathbf{M}_{k-1} \cdots \mathbf{M}_{l+1}$, with the convention that $\mathbf{M}_{k,k} = \mathbf{I}$. The first key result is the following inequality in the set of the semi-definite symmetric matrices:

$$\mathbf{P}_k \leq \mathbf{M}_{k:0} \mathbf{P}_0 \mathbf{M}_{k:0}^\top + \mathbf{\Xi}_k =: \mathbf{P}_k^{\mathrm{free}}, \tag{12}$$

where

$$\mathbf{\Xi}_0 \equiv \mathbf{0} \quad \text{and for } k \geq 1 \quad \mathbf{\Xi}_k \equiv \sum_{l=1}^{k} \mathbf{M}_{k:l} \mathbf{Q}_l \mathbf{M}_{k:l}^\top, \tag{13}$$

is known as the *controllability* matrix (Jazwinski 1970), and the suffix "free" is used to refer to the forecast error of the system unconstrained by data. In the absence of model noise ($\mathbf{Q}_k \equiv \mathbf{0}$ for the rest of this section), it reads

$$\mathbf{P}_k \leq \mathbf{M}_{k:0} \mathbf{P}_0 \mathbf{M}_{k:0}^\top. \tag{14}$$

Assuming the dynamics to be non-singular, the column subspace of the forecast error covariance matrix satisfies

$$\mathrm{Im}(\mathbf{P}_k) = \mathbf{M}_{k:0}\left(\mathrm{Im}(\mathbf{P}_0)\right). \tag{15}$$

If n_0 is the dimension of the unstable-neutral subspace of the dynamics, it can further be shown that

$$\lim_{k\to\infty} \mathrm{rank}(\mathbf{P}_k) \leq \min\left\{\mathrm{rank}(\mathbf{P}_0), n_0\right\}, \tag{16}$$

which is a first proof of the collapse of the error covariance matrix (actually its column space) onto the unstable and neutral subspace of the dynamics.

Result 2: Collapse onto the unstable subspace

Let σ_i^k, for $i = 1, \ldots, N_\mathrm{x}$ denote the eigenvalues of $\mathbf{P}_k$ ordered as $\sigma_1^k \geq \sigma_2^k \cdots \geq \sigma_{N_\mathrm{x}}^k$. It was shown that

$$\sigma_i^k \leq \beta_i \exp\left(2\lambda_i^k k\right), \tag{17}$$

for some $\beta_i > 0$, where λ_i^k is a log-singular value of $\mathbf{M}_{k:0}$ that converges to the LE λ_i. This gives an upper bound for all eigenvalues of $\mathbf{P}_k$ and a rate of convergence for the $N_\mathrm{x} - n_0$ smallest ones. Moreover, if $\mathbf{P}_k$ is uniformly bounded, it can further be shown that the stable subspace of the dynamics is asymptotically in the null space of $\mathbf{P}_k$, i.e.,

$$\lim_{k\to\infty} \left\| \mathbf{P}_k \mathbf{L}_k^{\mathrm{b},i} \right\| = 0 \tag{18}$$

for all $i > n_0$; this extends to any norm and linear combination of these vectors.

Result 3: Explicit dependence of $\mathbf{P}_k$ on $\mathbf{P}_0$

To study the dependence of $\mathbf{P}_k$ on $\mathbf{P}_0$, it has been shown that the forecast error covariance matrix can be written as

$$\mathbf{P}_k = \mathbf{M}_{k:0}\mathbf{P}_0\mathbf{M}_{k:0}^\top \left(\mathbf{I} + \mathbf{\Gamma}_k \mathbf{M}_{k:0}\mathbf{P}_0\mathbf{M}_{k:0}^\top\right)^{-1}, \tag{19}$$

where the matrix

$$\mathbf{\Gamma}_k \equiv \sum_{l=0}^{k-1} \mathbf{M}_{k:l}^{-\top} \mathbf{\Omega}_l \mathbf{M}_{k:l}^{-1}, \tag{20}$$

is known as the *information matrix* and it measures the *observability* of the system by propagating the precision matrices $\mathbf{\Omega}_l$ up to t_k.

An alternative formulation of Eq. (19) is

$$\mathbf{P}_k = \mathbf{M}_{k:0}\mathbf{P}_0 \left[\mathbf{I} + \mathbf{\Theta}_k \mathbf{P}_0\right]^{-1} \mathbf{M}_{k:0}^\top, \tag{21}$$

where

$$\mathbf{\Theta}_k \equiv \mathbf{M}_{k:0}^{\top}\mathbf{\Gamma}_k\mathbf{M}_{k:0} = \sum_{l=0}^{k-1}\mathbf{M}_{l:0}^{\top}\mathbf{\Omega}_l\mathbf{M}_{l:0}, \tag{22}$$

is also an *information* matrix, directly related to the *observability* of the DA system, but pulled back at the initial time t_0. Equation (21) is of key importance because it allows to study the asymptotic behaviour of $\mathbf{P}_k$, i.e. the filter "believed" error, using the asymptotic properties of the dynamics.

Result 4: Asymptotics of $\mathbf{P}_k$

For $\mathbf{P}_k$ to forget about $\mathbf{P}_0$ as k tends to infinity, it was shown that one can impose the following sufficient conditions:

- **Condition 1**: Recall that the FLVs at t_0 associated to the unstable and neutral exponents are the columns of $\mathbf{L}_0^{\mathrm{f},1:n_0} \in \mathbb{R}^{N_\mathrm{x}\times n_0}$. Moreover, let us define the anomaly matrix $\mathbf{X} \in \mathbb{R}^{N_\mathrm{x}\times n_0}$ such that $\mathbf{P} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. The condition reads

$$\mathrm{rank}\left(\mathbf{X}_0^{\top}\mathbf{L}_0^{\mathrm{f},1:n_0}\right) = n_0. \tag{23}$$

  In practice the initial ensemble anomalies $\mathbf{X}_0$ projects onto the first n_0 FLVs at t_0.
- **Condition 2**: The model is sufficiently observed so that the unstable and neutral directions remain under control, i.e., there exits $\varepsilon > 0$ such that

$$\left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^{\top}\mathbf{\Gamma}_k\mathbf{L}_k^{\mathrm{b},1:n_0} > \varepsilon\mathbf{I}. \tag{24}$$

- **Condition 3**: Furthermore, for any neutral BLV we have

$$\lim_{k\to\infty}\left(\mathbf{L}_k^{\mathrm{b},n_0}\right)^{\top}\mathbf{\Gamma}_k\mathbf{L}_k^{\mathrm{b},n_0} = \infty, \tag{25}$$

  implying that the neutral direction is well observed and controlled.

Under these three conditions, we obtain

$$\lim_{k\to\infty}\left\{\mathbf{P}_k - \mathbf{L}_k^{\mathrm{b},1:n_0}\left[\left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^{\top}\mathbf{\Gamma}_k\mathbf{L}_k^{\mathrm{b},1:n_0}\right]^{-1}\left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^{\top}\right\} = \mathbf{0}. \tag{26}$$

Hence, the asymptotic sequence does not depend on $\mathbf{P}_0$, but only on $\mathbf{\Gamma}_k$, i.e. on the dynamics and observations. It can also be shown that the neutral modes have a peculiar role and a long lasting influence: their influence on the current estimate decreases sub-exponentially.

Result 5: From the degenerate KF to the square-root EnKF and EnKS

The recurrence Eq. (21) can be reformulated in a factorised form which is suited to the square-root EnKF. The standard perturbation decomposition of the forecast error

covariance is

$$\mathbf{P}_k = \mathbf{X}_k \mathbf{X}_k^\top, \tag{27}$$

where $\mathbf{X}_k$ is the matrix of centred perturbations (the anomaly matrix aforementioned). A right-transform update formula can then be obtained from (21):

$$\mathbf{X}_k = \mathbf{M}_{k:0}\mathbf{X}_0 \left[\mathbf{I} + \mathbf{X}_0^\top \boldsymbol{\Theta}_k \mathbf{X}_0\right]^{-1/2} \boldsymbol{\Upsilon}_k, \tag{28}$$

where $\boldsymbol{\Upsilon}_k$ is an arbitrary orthogonal matrix such that $\boldsymbol{\Upsilon}_k \mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the vector $[1 \ldots 1]^\top$ defined in the ensemble subspace. It is equivalent to the left-transform update formula

$$\mathbf{X}_k = \left[\mathbf{I} + \mathbf{M}_{k:0}\mathbf{P}_0\mathbf{M}_{k:0}^\top \boldsymbol{\Gamma}_k\right]^{-1/2} \mathbf{M}_{k:0}\mathbf{X}_0 \boldsymbol{\Upsilon}_k. \tag{29}$$

The importance of Eqs. (28) and (29) stands on the fact that with linear models, Gaussian observation and initial errors, the (square-root) degenerate KF (with $\mathbf{X} \in \mathbb{R}^{N_x \times n}$ and $n < N_x$) is equivalent to the square-root EnKF and can serve as a proxy to the EnKF applied to nonlinear models.

All of these results can be generalised to linear smoothers. In particular, the smoother forecast error covariance matrix is similar to that of the filter, i.e. given by (19), or (21) but with the following modified information matrix:

$$\widehat{\boldsymbol{\Gamma}}_k = \boldsymbol{\Gamma}_k + \sum_{l=k}^{k+L-S} \mathbf{M}_{k:l}^{-\mathrm{T}} \boldsymbol{\Omega}_k \mathbf{M}_{k:l}^{-1}, \tag{30}$$

where L is the lag of the smoother (how far in the past observations are accounted for) and S tells by how many time steps the smoother's window is shifted between two consecutive updates. Note that $\widehat{\boldsymbol{\Gamma}}_k \geq \boldsymbol{\Gamma}_k$ (using the Loewner order on the set of semi-definite positive matrices) reflecting the general higher amount of information incorporated within a smoother update relative to a filter. Therefore the asymptotic sequences for the filter (right-hand side) and smoother (left-hand-side) follow the inequality:

$$\begin{aligned} &\mathbf{L}_k^{\mathrm{b},1:n_0} \left[\left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^\top \widehat{\boldsymbol{\Gamma}}_k \mathbf{L}_k^{\mathrm{b},1:n_0} \right]^{-1} \left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^\top \leq \\ &\mathbf{L}_k^{\mathrm{b},1:n_0} \left[\left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^\top \boldsymbol{\Gamma}_k \mathbf{L}_k^{\mathrm{b},1:n_0} \right]^{-1} \left(\mathbf{L}_k^{\mathrm{b},1:n_0}\right)^\top. \end{aligned} \tag{31}$$

The linear smoothers can serve as a proxy to the ensemble Kalman smoother (Evensen 2009b) or the iterative ensemble Kalman smoother (IEnKS) (Bocquet and Sakov 2014) applied to nonlinear models (Bocquet and Carrassi 2017). With the IEnKS, where the true state trajectory is even better estimated and the errors are reduced, the collapse of the perturbations onto the unstable and neutral subspace is expected to be even faster.

3.1.2 Stochastic Dynamics

The above perfect-deterministic model configuration demonstrates how chaos shapes the inferences of the posterior. Asymptotic statistics are determined by the ability of the filter to control the growth of initial error in the unstable-neutral subspace with respect to its sensitivity to observations therein. However, in realistic DA, additional forecast errors are introduced throughout the forecast cycle due to the inadequacy of numerical models in representing reality. One classical approach to treat these model errors is to represent them as additive or multiplicative noise (Jazwinski 1970). Oseledet's theorem and the Lyapunov spectrum are also formulated for such systems of stochastic differential equations (SDEs) and discrete maps.

Suppose $\{\mathbf{v}^i\}_{i=0}^d$ is a collection of $\mathcal{C}^3$ vector fields on $M \subset \mathbb{R}^{N_x}$, a smooth manifold without boundary. Define $\{W_t^i\}_{i=1}^d$ where each W_t^i is an independent Wiener process defined on the probability space $(\Omega, \mathcal{F}, P)$. This describes a generic Stratonovich SDE,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \boldsymbol{\sigma}, \omega) \triangleq \mathbf{v}^0(\mathbf{x}, \boldsymbol{\sigma}) + \sum_{i=1}^{d} \mathbf{v}^i(\mathbf{x}, \boldsymbol{\sigma}) \circ W_t^i(\omega). \tag{32}$$

For fixed $\boldsymbol{\sigma}$, if $\{\mathbf{v}^i\}_{i=1}^d$ span the tangent space $T_{\mathbf{x}}M$ for each $\mathbf{x} \in M$, then the system of SDEs gives rise to a unique probability measure μ on M that is invariant with respect to the random flow induced by the system of SDEs. For $\mu \times P$ almost every $(\mathbf{x}, \omega) \in M \times \Omega$, the Lyapunov exponents and their multiplicity are defined and only depend on $\mathbf{x}$ (Liu and Qian 2006, Theorem 2.1). For the SDE, $\mathbf{f}(\mathbf{x}, \boldsymbol{\sigma}, \omega)$, the tangent linear model is once again defined as in Eq. (2) from which the exponents and vectors can be computed as usual (Pikovsky and Politi 2016, chaps. 2 and 8).

The analysis from perfect-deterministic models thus extends to stochastically forced models as in Eq. 8a, but with key differences. Firstly, the forecast error covariance can generally be considered to be of full rank due to the injection of the stochastic forcing $\mathbf{w}_k$ into arbitrary subspaces; the standard controllability assumption actually guarantees that it is of full rank after a sufficient lead time (Jazwinski 1970, Lemma 7.3). Stochastic perturbations $\mathbf{w}_k$ are, moreover, subject to growth and decay rates of the LLEs, λ_i^τ of Eq. (7). These LLEs are distributed about the LE such that even asymptotically stable modes may have transient periods of rapid growth.

To make the analysis tractable, assume that the dynamics are stationary in the sense that the recursive QR algorithm (Shimada and Nagashima 1979; Benettin et al. 1980) converges uniformly to the theoretical LEs uniformly in number of iterations from any initial time. While the forecast error covariance is not generally reduced-rank, the stable dynamics is actually sufficient to uniformly bound the forecast error variances in the stable subspace *without any assimilation*.

Result 6: Uniform bounds on the variance of errors in the stable subspace

Recall that $\mathbf{L}_k^{\mathrm{b},i}$ is the i-th BLV and that $\mathbf{P}_k^{\mathrm{free}}$ is the free forecast error covariance (e.g. without DA). For $\lambda_i < 0$, define $\epsilon > 0$ such that $\exp\{2\lambda_i + \epsilon\} < 1$. By the

stationarity assumed above, there exists N_ϵ such that if $k - l > N_\epsilon$, then

$$-\epsilon < \frac{1}{k-l} \log \left(\| \mathbf{M}_{k:l}^{\top} \mathbf{L}_k^{\mathrm{b},i} \| \right) - \lambda_i < \epsilon. \tag{33}$$

Assuming that the system is uniformly completely controllable and $\mathbf{Q}_k \leq q_{\sup} \mathbf{I}_n < \infty$ for all k, the controllability matrix, Eq. (13), is uniformly bounded above by $C_{N_\epsilon} \mathbf{I}_n$. A variation on Eq. (12) can be used to obtain a uniform bound on the i-th variance of $\mathbf{P}_k^{\mathrm{free}}$ in the basis of the BLVs as

$$\limsup_{k \to \infty} \left(\mathbf{L}_k^{\mathrm{b},i} \right)^{\top} \mathbf{P}_k^{\mathrm{free}} \mathbf{L}_k^{\mathrm{b},i} \leq C_{N_\epsilon} + \frac{q_{\sup} \exp\{2(\lambda_i + \epsilon) N_\epsilon + 1\}}{1 - \exp\{2(\lambda_i + \epsilon)\}}. \tag{34}$$

This bound represents the competing forces of the transient growth rates of recently introduced perturbations in the controllability matrix, and perturbations that adhere to their aysmptotic log-average rate of decay in the stable subspace within a margin of ϵ (Grudzien et al. 2018a, Proposition 2 and Corollary 3). This demonstrates that, if assimilation prevents error growth in the span of the unstable-neutral BLVs, errors in the span of the stable BLVs can be neglected without relinquishing filter boundedness. However, this does not state whether the error in the span of the stable subspace will remain within tolerable bounds.

Redefine $q_{\sup}$ such that $\mathbf{P}_0^{\mathrm{free}}, \mathbf{Q}_k \leq q_{\sup} \mathbf{I}_n$ for all k. The variance of the free forecast in the i-th BLV is bounded directly as

$$\begin{aligned} \left(\mathbf{L}_k^{\mathrm{b},i} \right)^{\top} \mathbf{P}_k^{\mathrm{free}} \mathbf{L}_k^{\mathrm{b},i} &= \left(\mathbf{L}_k^{\mathrm{b},i} \right)^{\top} \mathbf{M}_{k:0} \mathbf{P}_0^{\mathrm{free}} \mathbf{M}_{k:0}^{\top} \mathbf{L}_k^{\mathrm{b},i} + \sum_{l=1}^{k} \left(\mathbf{L}_k^{\mathrm{b},i} \right)^{\top} \mathbf{M}_{k:l} \mathbf{Q}_l \mathbf{M}_{k:l}^{\top} \mathbf{L}_k^{\mathrm{b},i} \\ &\leq q_{\sup} \sum_{l=0}^{k} \left(\mathbf{L}_k^{\mathrm{b},i} \right)^{\top} \mathbf{M}_{k:l} \mathbf{M}_{k:l}^{\top} \mathbf{L}_k^{\mathrm{b},i} = q_{\sup} \sum_{l=0}^{k} \| \left(\mathbf{T}_{k:l}^{\top} \right)^i \|^2, \end{aligned} \tag{35}$$

where $\| \left(\mathbf{T}_{k:l}^{\top} \right)^i \|$ is the norm of the i-th row of $\mathbf{T}_{k:l}$ in the recursive QR decomposition of the propagator, $\mathbf{M}_{k:l} = \mathbf{L}_k^{\mathrm{b}} \mathbf{T}_{k:l} \left(\mathbf{L}_l^{\mathrm{b}} \right)^{\top}$. The sum

$$\Psi_k^i \triangleq \sum_{l=0}^{k} \| \left(\mathbf{T}_{k:l}^{\top} \right)^i \|^2 \tag{36}$$

describes the invariant evolution for the i-th variance of the free forecast error covariance matrix in the basis of BLVs. In the case that $\mathbf{Q}_k = q \mathbf{I}_n$ for all k, Eq. (35) becomes an equality and Ψ_k^i can be interpreted as the evolution of the i-th variance when $q = 1$. As $k - l$ grows $\| \left(\mathbf{T}_{k:l}^{\top} \right)^i \|$ converges exponentially to zero while for $k - l$ close to one this describes transient dynamics in the basis of the BLVs. Although Ψ_k^i is guar-

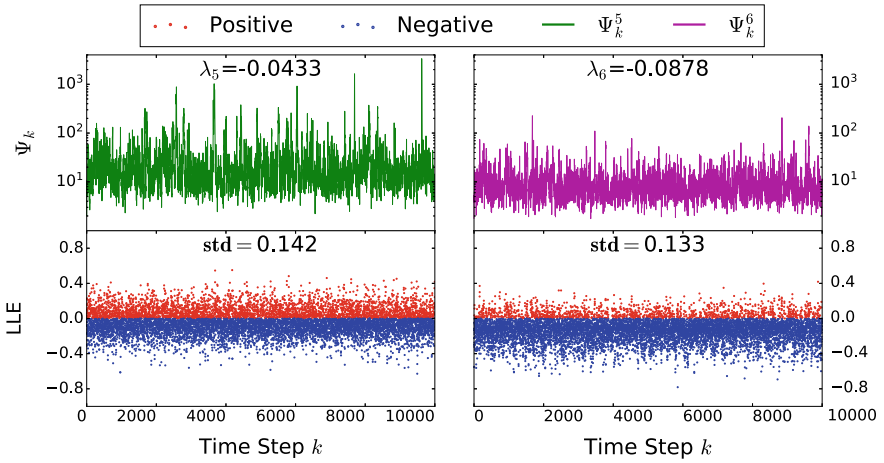

Fig. 1 **Upper**: time series of Ψ_k^5 and Ψ_k^6 as defined in Eq. (36). **Lower**: LLEs of $\mathbf{L}^{\mathrm{b},5}$ and $\mathbf{L}^{\mathrm{b},6}$. Adapted from Grudzien et al. (2018a)

anteed to be uniformly bounded in k (Grudzien et al. 2018a, Corollary 3), numerical simulations demonstrate how this uniform bound can be extremely large.

Figure 1, presents an example from Grudzien et al. (2018a) of the free forecast error variance, Ψ_k^i, over 10^4 forecast cycles where the model propagator $\mathbf{M}_k$ is defined by the evolution of the tangent linear model of the Lorenz-96 system (Lorenz 1996) in $N_x = 10$ dimensions, with an interval between observations of 0.1, and $\mathbf{Q}_k \triangleq \mathbf{I}_{N_x}$. Although this model is generated from the underlying nonlinear Lorenz system, the state model for the experiment is treated as a discrete linear model using only the tangen-linear resolvent as above. This model has three unstable, one neutral and six stable LEs. While λ_5 and λ_6 are negative, $\mathbf{L}_k^{\mathrm{b},5}$ and $\mathbf{L}_k^{\mathrm{b},6}$ experience frequent transient instabilities in the timeseries of their LLEs (see bottom row in Fig. 1). The LLEs of $\mathbf{L}_k^{\mathrm{b},5}$ have more intense growth, reflected in the differences between Ψ_k^5 and Ψ_k^6 (top row): the maximum of Ψ_k^6 is on the order of $\mathcal{O}(10^2)$ and the mean is approximately 28; for Ψ_k^5 the max is of $\mathcal{O}\left(10^3\right)$ and the mean is approximately 808. This suggests that, as opposed to the case of deterministic dynamics, for successful DA in stochastic systems with additive model error, it is necessary to control the growth of forecast errors also in the span of weakly stable BLVs. While errors in this span will not grow indefinitely and remain bounded, if those directions are left uncontrolled their error bounds can be practically too large for any meaningful state estimation purposes.

Result 7: The unfiltered-to-filtered error upwell and the need for inflation

Motivated by the results above, suppose that an approximate, reduced-rank Kalman estimator is defined such that the resulting forecast error covariance and the Kalman gain have image (column) spaces constrained to the span of the leading $r \geq n_0$ BLVs, $\mathbf{L}_k^{\mathrm{b},1:r}$. For perfect and deterministic dynamics, this estimator is asymptotically

equivalent to the optimal Kalman filter by the results of Sect. 3.1.1. On the other hand, Result 6 establishes that, in the presence of additive model errors, the variance of the unfiltered error in the span of the trailing $N_x - n_0$ BLVs will remain finite, albeit bounded as per Eq. (35). When $r > n_0$ this reduced-rank Kalman filter will correct the $r - n_0$ stable modes in addition to unstable-neutral modes, thereby reducing the variance of the free forecast errors below the bounds pictured in Fig. 1.

Grudzien et al. (2018b) derive the full forecast error covariance dynamics for the reduced-rank Kalman estimator described above. Note that, as opposed to the free forecast error covariance described in relation to Result 6, the discussion pertains now to the forecast error covariance cycled within the KF. While the standard reduced-rank KF formalism would write the recursion for the forecast error covariance entirely within the span of $\mathbf{L}^{\mathrm{b},1:r}$, it is proven in Grudzien et al. (2018b) that forecast errors in the column span of trailing $\mathbf{L}_k^{\mathrm{b},r+1:N_x}$ (those left "uncorrected" by DA) are transmitted into the column span of $\mathbf{L}_k^{\mathrm{b},1:r}$. This "error upwell" is a consequence of the KF rank-reduction within the first r BLVs and is driven by the upper triangular dynamics of the BLVs in the recursive QR algorithm. Therefore, neglecting the contribution of the "upwelling" of error from the trailing to the leading BLVs, as in the standard recursion, leads to a systematic underestimation of the true forecast error in the presence of additive noise. Furthermore, because the leading r BLVs share the same span as the leading r Oseledet spaces, Eq. (4), the upwelling of errors from the span of the trailing BLVs to the leading BLVs holds for any estimator that is restricted to the span of the leading r covariant subspaces.

Figure 2 presents an example from Grudzien et al. (2018b), using the same tangent linear model from the 10-dimensional Lorenz-96 system as in Fig. 1, and fixing each of $\mathbf{H}_k = \mathbf{R}_k = \mathbf{Q}_k = \mathbf{I}$. In each window, the eigenvalues of the forecast error covariance matrix of the optimal full-rank KF (yellow) and of the "exact", reduced-rank estimator (red) are averaged over 10^5 analysis cycles and plotted with triangles. By exact it is meant here that the full covariance equation is evolved analytically, including the covariance within the unfiltered trailing BLVs and their cross covariances with the leading filtered modes. The rank, r, of the reduced-rank estimator is varied to examine the differences between the forecast error covariances arising from the optimal KF and the reduced-rank one when only the unstable-neutral subspace is corrected by the gain ($r = n_0 = 4$), the first stable mode is corrected by the gain ($r = 5$) and so on. As suggested by Fig. 1, correcting the first stable mode reduces the leading eigenvalue of reduced-rank estimator's forecast error covariance by an order of magnitude versus the case when only the unstable-neutral subspace has been corrected (cf the red curves between the two top windows).

In each window the projection coefficients of the reduced-rank estimator's forecast error covariance into the basis of BLVs are also plotted (green line). For the full-rank optimal KF, the projection coefficients closely follow the eigenvalues and this is not shown due to redundancy. By contrast, it is clear that for the reduced-rank estimator the leading eigenvector is typically close to the first BLV that is contained in the null space of the reduced-rank gain, i.e. the first among the unconstrained directions. Figure 2 demonstrates thus a fundamental difference between the perfect model and

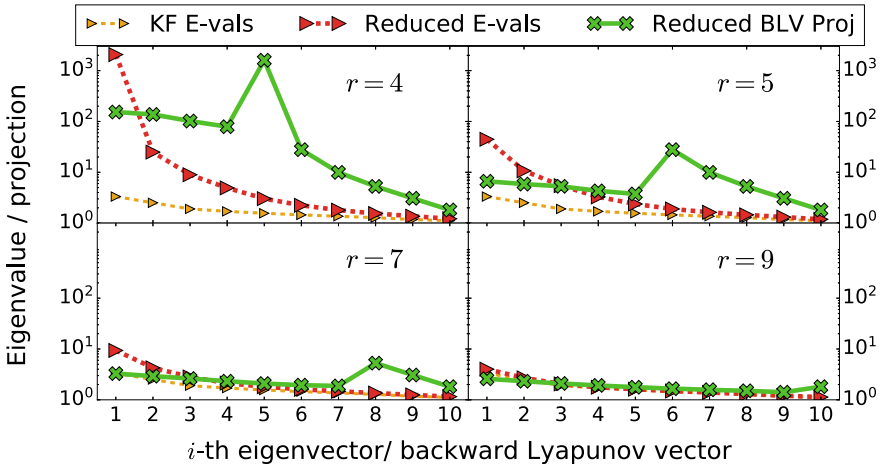

Fig. 2 Eigenvalues of the KF and the reduced-rank estimator covariances plotted with triangles. Projection coefficients of the reduced-rank estimator covariance plotted with X's

stochastically forced model configurations: in stochastic systems, using a reduced-rank $n_0 \leq r < N_x$ Kalman estimator, the leading order forecast errors can actually lie in the stable, unfiltered modes. Those modes must be taken thus into account in the DA procedure by, for instance, appropriately enlarging the ensemble size or otherwise augmenting the span of the ensemble-based gain.

The exact recursion for the reduced-rank gain (red lines in Fig. 2) requires resolving the full error covariance dynamics and therefore cannot be used practically for DA. Nevertheless, it was used here to render apparent, the upwelling phenomena: a fundamental source of uncertainty that is not captured by the standard reduced-rank KF (and thus almost all EnKF) recursion. This mechanism is ubiquitous whenever one solves for a reduced rank estimator and is present whenever the forecast error evolution can be well approximated by the tangent-linear dynamics (see Sect. 3.2.2). Notably, it also provides one basic, mathematically grounded, justification for using covariance inflation (Grudzien et al. 2018b), a powerful common fix used in ensemble-based DA (that are commonly rank-deficient by construction) to mitigate for sampling, and sometimes model, error (see e.g., Carrassi et al. 2018, their Sect. 4.4.2 and references therein). Finally, the exact recursion also demonstrates the asymptotic characteristics of the forecast error covariance when using a reduced-rank gain, in the absence of sampling error. It is extremely important to note that in the exact error dynamics for the reduced rank estimator, the leading order forecast errors lie in the directions that are asymptotically stable but unfiltered. This also highlights the importance of localization (Sakov and Bertino 2011) and gain-hybridization (Penny 2017) as effective means for preventing the growth of forecast errors that lie outside of the ensemble span in the EnKF.

3.2 Nonlinear Dynamics: The Effect of Chaos on the Ensemble Kalman Filter

3.2.1 Perfect and Deterministic Dynamics

The performance and the functioning mechanisms of the EnKF in nonlinear systems are studied with the aid of numerical simulations performed using the `qgs` code platform (Demaeyer et al. 2020; Demaeyer and De Cruz 2020).

We consider first a spectral 2-layer channel quasi-geostrophic atmospheric model. The Fourier modes decomposition is truncated at wavenumber 2 in both meridional and zonal directions on a beta-plane, leading to a set of 20 ODEs for the time evolution of the first 10 components of the atmospheric streamfunction $\boldsymbol{\psi}$, and temperature $\boldsymbol{\theta}$ (Reinhold and Pierrehumbert 1982). The model dimension is $N_x = 20$ and its spectrum of LEs includes 3 positive and one neutral exponents, so that $n_0 = 4$.

The model is integrated with a time step of approximately 15 min and is spun-up for 3 years to ensure the solution has reached the model attractor. Afterward we initialise the DA experiments with the following protocol: a "true" trajectory is computed for 4, 6 years, and synthetic observations are generated by first sampling this trajectory at regular analysis time t_k, each separated by a time interval Δt, $t_{k+1} = t_k + \Delta t$. The observations are then obtained by adding a zero mean Gaussian random error sampled from $\mathcal{N}(\mathbf{0}, \mathbf{R})$, with $\mathbf{R}$ being the (assumed to be known) observation error covariance matrix. It is assumed that we observe the spectral components directly and that the full system is observed, implying that the observation operator is the identity matrix, $\mathbf{H} = \mathbf{I}_{N_x} \in \mathbb{R}^{N_x \times N_x}$. Although the former hypothesis cannot be met in practice (instrumental devices do not observe the spectral modes) and the latter rarely holds in high-dimensional applications, they are done here for the sake of clarity and will facilitate the study of the dynamical behaviour we intend to discuss. Furthermore, observational error is supposed to be spatially (in spectral space) uncorrelated with an amplitude proportional (with factor $\sigma^{\%}$) to the corresponding model variable's standard deviation, σ^i_{md}, $i = 1, \ldots, N_x$. These imply $\mathbf{R} = \sigma^{\%}\text{diag}(\sigma^1_{\text{md}}, \ldots, \sigma^{N_x}_{\text{md}})$.

Data assimilation is performed using the EnKF-N, hereafter simply referred to as EnKF (Bocquet 2011). This EnKF belongs to the family of deterministic filters but it furthermore possesses the appealing property that the ensemble covariance multiplicative inflation is computed automatically as part of the DA process. Inflation is one of the unavoidable feature making the EnKF methods suitable for high dimensional problems (Carrassi et al. 2018). It comes under two ways known as multiplicative and additive inflation, that are often used together. We shall use multiplicative inflation alone because, as opposed to the additive version, it does not change the rank and span of the ensemble error covariance but it only inflates the matrix entries amplitude. This allows us to study the effect (if any) on the ensemble subspace (reflected into the ensemble covariance rank and span) that comes from the dynamics, without artifacts from the DA procedure. At each analysis step, the forecast anomaly matrix is inflated as $\alpha \mathbf{X}^{\text{f}} \leftarrow \mathbf{X}^{\text{f}}$, $\alpha \geq 1$.

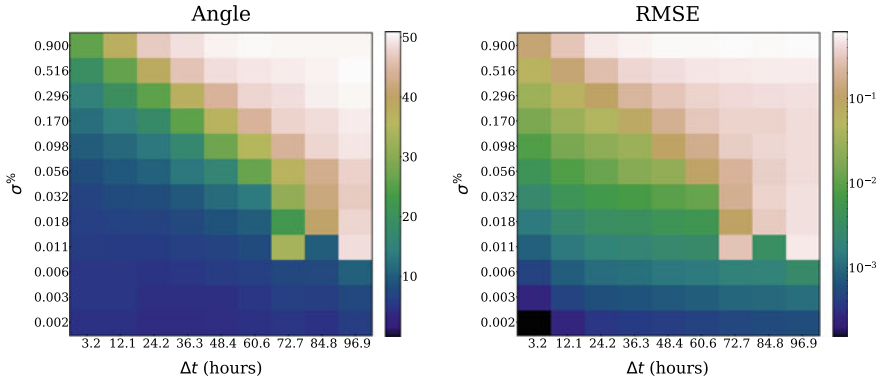

Fig. 3 Atmospheric quasi-geostrophic model. Time- and ensemble-averaged angle, from Eq. (37) (in degree), between the anomalies of the EnKF and the unstable-neutral subspace (left panel, shadow colours, in degree), and time averaged normalised RMSE of the EnKF analysis (right panel, shadow colours), both on the plane $(x, y) = (\Delta t, \sigma^{\%})$. The set-up is $\mathbf{H} = \mathbf{I}_d$, $\mathbf{R} = \sigma^{\%}\mathrm{diag}(\sigma_{\mathrm{md}}^1, \ldots, \sigma_{\mathrm{md}}^{N_x})$ and $N = 10$. Note that the logarithmic scale is used on the right panel

We will study the properties of the EnKF ensemble subspace, its dimension and alignment to the unstable subspace of the underlying dynamics, and will investigate how those will relate to the skill of the EnKF. Following Bocquet and Carrassi (2017), at each analysis time t_k, the alignment between the ensemble and the unstable-neutral subspaces, $\mathcal{U}_k$, is computed as

$$\cos^2(\theta_k^i) = \sum_{p=1}^{n_0} \cos^2(\theta_k^{i,p}) = \sum_{p=1}^{n_0} \frac{\left\{(\mathbf{u}_k^p)^{\mathrm{T}}\mathbf{v}_k^i\right\}^2}{\left\|\mathbf{v}_k^i\right\|^2} \quad 1 \leq i \leq N, \tag{37}$$

where $\theta_k^i \in [0, \pi]$ is the angle between the anomaly $\mathbf{v}_k^i$ and $\mathcal{U}_k$, $\mathbf{u}_k^p$ is the p-th BLV, and N is the number of ensemble members.

The RMSE of the EnKF analyses and the angle θ_k^i are shown on the plane $(x, y) = (\Delta t, \sigma^{\%})$ in Fig. 3. Values are averaged over 4 years of simulated time after discarding the first 200 analyses. The RMSE of the analyses is also averaged over all model variables, once the error on each variable is normalised with respect to the corresponding standard deviation σ_{md}. The number of ensemble members is set to $N = 10$.

Figure 3 immediately shows the strong resemblance of patterns between the angle and the RMSE. In practice, whenever the observation time interval Δt and error $\sigma^{\%}$ are small enough then the angle between the two subspaces get smaller and the RMSE decreases. Similarly to what is discussed in Bocquet and Carrassi (2017), we also see here how increasing observation frequency is more effective in reducing the RMSE than reducing the observation error (see discussion in the Appendix of Bocquet and Carrassi 2017). This result also suggests that the use of a large number and frequent observations, like the ones produced by non-conventional systems (e.g.

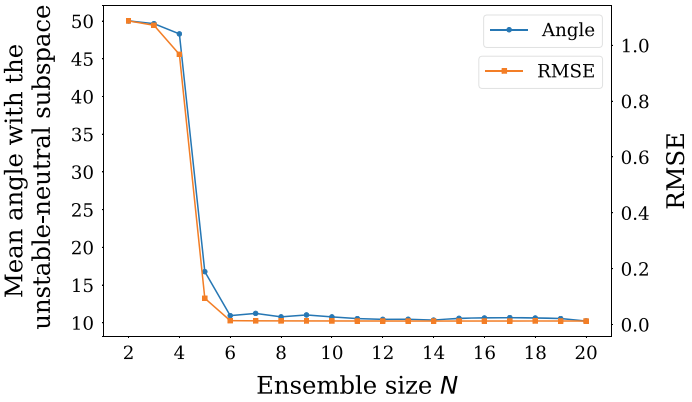

Fig. 4 Atmospheric quasi-geostrophic model. Time- and ensemble-averaged angle, Eq. (37) (in degree), between the ensemble anomalies of the EnKF and the unstable-neutral subspace as a function of the ensemble size N (left y axis), and the corresponding time-averaged RMSE of the EnKF (right y axis). The set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$ with $\sigma^{\%} = 0.08$ and $\Delta t = 12.11$ h

crowdsourcing) has the potential for improving analyses, despite that observational errors may be larger than in conventional measurement systems (Nipen et al. 2019).

A complementary picture of the relation between the filter skill and the subspaces alignment is provided in Fig. 4, that displays the angle (left y-axis) and RMSE of the EnKF analysis (right y-axis) both against the ensemble size N. The remaining experimental set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$ with $\sigma^{\%} = 0.08$ and $\Delta t = 12.11$ h. Recalling that the unstable-neutral subspace has dimension $n_0 = 4$, Fig. 4 shows that as soon as $N \geq n_0 + 1$, i.e. the ensemble fully spans the unstable-neutral subspace, the RMSE suddenly reduces to very small values and it does not further decrease when N is increased beyond $n_0 + 1$. The fact that the convergence occurs for $n_0 + 1$ instead of n_0 is due to the ensemble anomalies subspace to be at most of rank $N - 1$ because one degree of freedom is removed when computing the ensemble mean. The behavior depicted in Fig. 4 is peculiar of the EnKF applied to chaotic dynamics and suggests a natural way to reduce computational cost by applying the EnKF with "only" $N = n_0 + 1$ members.

At the convergence, the angle between the ensemble and unstable-neutral subspace (of dimension $n_0 = 4$) is about 10 degrees (see Fig. 4): a remaining small portion of the ensemble subspace is projecting outside the unstable-neutral space. To investigate how large is such a portion, we compute the angle between an ensemble subspace with fixed $N = 10$ and a Lyapunov/Oseledets subspace of increasing dimension beyond $n_0 + 1$; results are shown in Fig. 5. The angle between the two subspaces decreases monotonically with the size of the subspace, eventually reaching zero once the full phase space ($n = 20$) is considered. Interestingly, the rate of convergence is initially faster until approximately dimension 10, and slower afterwards. This indicates that the additional projection beyond the asymptotic unstable-neutral subspace is largely confined to the less stable (i.e. close to neutral) directions, those that can often be

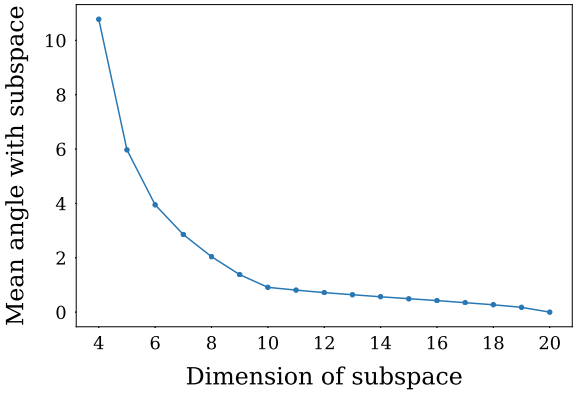

Fig. 5 Atmospheric quasi-geostrophic model. Time- and ensemble-averaged angle (in degree) between the ensemble anomalies of the EnKF with $N = 10$ members, and subspaces spanned by the BLVs of increasing dimensions. These subspaces are constructed by starting from the unstable-neutral subspace ($n_0 = 4$) and adding one by one stable directions ordered decreasingly by their Lyapunov exponent. The set-up is $\mathbf{H} = \mathbf{I}_d$, $\mathbf{R} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$ with $\sigma^{\%} = 0.08$ and $\Delta t = 12.11$ h

locally unstable. This mechanism is a reminiscent of what was shown by Grudzien et al. (2018b) for stochastic systems and reviewed in Sect. 3.1.2. It suggests that the addition of few ensemble members beyond $n_0 + 1$ may lead to modest improvements in the filter skill, even though the long term performance is not much impacted.

To discuss the impact of multiple timescales on the ensemble Kalman filtering, we consider now the addition of a shallow-water ocean component to the previous atmospheric model. The atmospheric and oceanic models are coupled together through wind and radiative forcing as well as through heat exchanges, yielding the Modular Arbitrary-Order Ocean-Atmosphere Model MAOOAM (De Cruz et al. 2016). MAOOAM has the same 20 ODEs of the atmospheric quasi-geostrophic model, and additional 16 ODEs for the ocean. Amongst these 16 equations, the first 8 govern the time evolution of the first eights components of the oceanic streamfunction $\boldsymbol{\psi}_{\mathrm{o}}$, while last 8 are relative to the first eights components of the oceanic temperature anomaly $\delta\boldsymbol{T}_{\mathrm{o}}$. The model dimension is thus $N_x = 36$ and its parameters are chosen such that a decadal low-frequency variability appears as a consequence of the coupling (Vannitsem et al. 2015; De Cruz et al. 2016).

The model is integrated with a time step of approximately 15 min and is spun-up for as many as 18,500 years to ensure the solution has reached the model attractor, given the low-frequency (i.e. slow time-scale) of the model. From this spun-up trajectory we start DA experiments using the EnKF-N with the same protocol used for the atmospheric model detailed above. In this case the true trajectory lasts 185 years, and it is again assumed that we observe the spectral components directly and completely. The EnKF is used in a strongly-coupled DA mode (see e.g. Penny and Hamill 2017) so that, at analysis steps, atmospheric data can impact the ocean and vice-versa. MAOOAM has been already used as a prototypical coupled model to study cou-

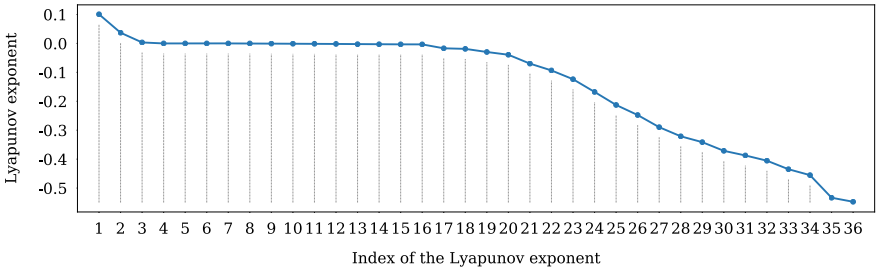

Fig. 6 MAOOAM coupled Ocean-Atmosphere model. Time-averaged Lyapunov exponents in days^{-1} computed along the true trajectory of the DA experiments

pled DA by Penny et al. (2019), Tondeur et al. (2020). Again, observational error is assumed to be spatially (in spectral space) uncorrelated with an amplitude proportional to the corresponding model variable's standard deviation, $\sigma^i_{\rm md}$, $i = 1, \ldots, N_x$. These imply $\mathbf{R} = \sigma^{\%}\mathrm{diag}(\sigma^1_{\rm md}, \ldots, \sigma^{N_x}_{\rm md})$.

The spectrum of LEs of MAOOAM is displayed in Fig. 6 and presents some key features.

With some arbitrariness, it can be decomposed into three subsets. A first subset of LEs corresponds to the unstable ($\lambda_i > 0$) and neutral ($\lambda_{n_0} = 0$) directions. The dimension of the subspace spanned by these directions is $n_0 = 6$. A second subset of small negative LEs ($\lambda_i \in [-5 \times 10^{-3}, 0[\ \mathrm{day}^{-1}$) corresponds to *nearly neutral directions*. The subspace they span has dimension $n_1 = 10$ and we hereafter define the *unstable-near–neutral subspace* as the one spanned by the first $n_0 + n_1$ directions. The presence of these nearly-neutral directions amounts as a form of degeneracy of the neutral direction. Although the model is theoretically possessing only one single neutral mode (cf Sect. 2), it is computationally extremely difficult to distinguish it from other nearly neutral ones, and the model can be said to degenerate in the neutral direction in any practical sense. We shall see how this feature has important consequences on the functioning and performance of DA. Finally, after a clear gap, the nearly neutral Lyapunov exponents are followed by the remaining stable directions which form the last subset of the spectrum.

Similarly to Fig. 3, we study the filter performance along with its ensemble subspace alignment to the unstable subspace in Fig. 7. The figure shows the RMSE of the EnKF analyses and the angle θ^i_k given by Eq. (37), between the ensemble subspace and the unstable-neutral (left) and unstable-near–neutral (middle) subspace. Values are averaged over 185 years of simulated time after discarding the first 100 analyses. RMSE of the analysis is also averaged over all model variables, once the error on each variable is normalised with respect to the corresponding standard deviation $\sigma_{\rm md}$. The number of ensemble members is set to $N = 20$.

In contrast to what was observed in Fig. 3, we see here that the pattern of the RMSE does not longer resemble the pattern of the angle between the ensemble and unstable-neutral subspace (cf left and right panels in Fig. 7). Nevertheless, it bears great similarities with the pattern of the angle between the ensemble and unstable-

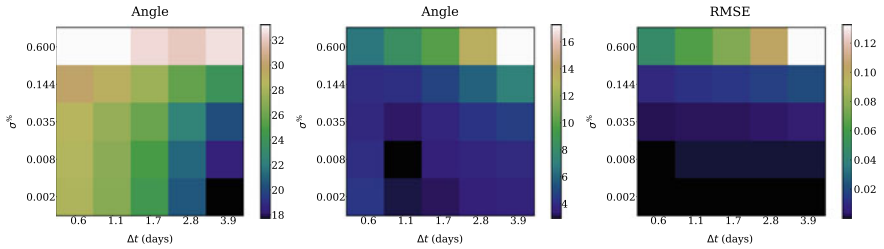

Fig. 7 MAOOAM coupled Ocean-Atmosphere model. Time- and ensemble-averaged angle given by Eq. (37) between the anomalies of the EnKF and the unstable-neutral subspace (left panel, shadow colours, in degree) or the unstable-near–neutral subspace (mid panel, shadow colours, in degree). The time averaged normalised RMSE of the EnKF analysis is also depicted (right panel, shadow colours). All the figures are on the plane $(x, y) = (\Delta t, \sigma^{\%})$. The set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \sigma^{\%}\mathrm{diag}(\sigma_{\mathrm{md}}^1, \ldots, \sigma_{\mathrm{md}}^{N_x})$ and $N = 20$. The unstable-near–neutral subspace is defined as follows: it includes the subspace spanned by the unstable and neutral $n_0 = 6$ directions, but also an additional $n_1 = 10$ stable but near–neutral directions with LEs $\lambda_i \in [-5 \times 10^{-3}, 0[$ day^{-1}

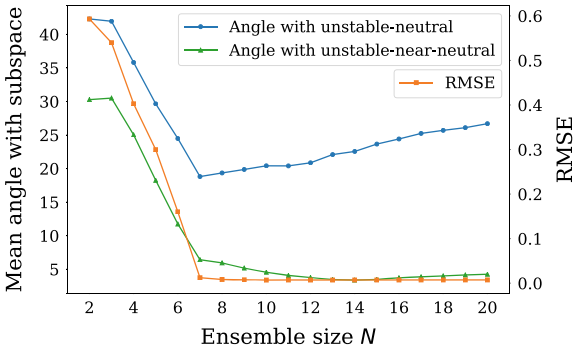

Fig. 8 MAOOAM coupled Ocean-Atmosphere model. Time- and ensemble-averaged angle (in degree) between the ensemble anomalies of the EnKF and the unstable-neutral subspace or the unstable-near–neutral subspace as a function of the ensemble size N (left y axis), and the corresponding time-averaged RMSE of the EnKF (right y axis). The set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \sigma^{\%}\mathrm{diag}(\sigma_{\mathrm{md}}^1, \ldots, \sigma_{\mathrm{md}}^{N_x})$ with $\sigma^{\%} = 0.08$ and $\Delta t = 1.68$ days

near–neutral subspace (cf mid and right panels in Fig. 7), suggesting that it is now this larger subspace (that includes the n_1 weakly stable modes) that contains most of the error.

This is further confirmed by looking at Fig. 8 that, similarly to Fig. 4, shows the angle (left y-axis) and RMSE of the analysis (right y-axis) against the ensemble size N. The set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \sigma^{\%}\mathrm{diag}(\sigma_{\mathrm{md}}^1, \ldots, \sigma_{\mathrm{md}}^{N_x})$ with $\sigma^{\%} = 0.08$ and $\Delta t = 1.68$ days.

The critical role of the n_1 weakly stable modes appears now evident when looking at Fig. 8 (cf green and blue lines). As opposed to the behaviour of the EnKF applied to the atmospheric model alone, or to the single scale Lorenz96 system used by Bocquet and Carrassi (2017), we see that the ensemble subspace does not longer project much

onto the unstable-neutral subspace (blue line with solid circles markers): when $N = n_0 + 1$ the angle reaches its minimum at around 20 degrees (recall that it was about 10 degrees for the atmospheric model, Fig. 4), and it even further increases when $N > n_0 + 1$, indicating the presence of non-negligible projections outside the $n_0 = 6$ unstable-neutral directions. However, when the angle is computed with respect to the larger unstable-near–neutral subspace (green line with solid triangle markers), we retrieve the match with the RMSE curve (orange line with solid squares markers). A closer inspection further reveals that the angle decreases (the projection grows) fast until $N = n_0 + 1$: the first unstable-neutral $n_0 = 6$ modes still span most of the error. After this initial fast decrease, the angle keeps decreasing at a slower, yet monotonic, rate until approximately $N = 15$ and stays almost constant afterwards. This result demonstrates undoubtedly the importance of the n_1 near–neutral modes. Although possibly asymptotically weakly stable, these directions span a small portion of the filter error that, if included in the ensemble subspace (by properly enlarging the ensemble size) leads to further improvement of the filter performance. This is finally emphasized in Fig. 9, which shows the decrease of the ensemble-averaged angle between an ensemble of $N = 10$ members and the subspaces of increasing dimension beyond the unstable-neutral one. In contrast to Fig. 5, it indicates that the addition of n_1 extra ensemble members may lead to RMSE improvement that is far from negligible. Also, the gap observed around the value $n_0 + n_1 + 1 = 17$ clearly shows that adding more members beyond the range of the near-neutral stability does not bring any benefit. This is a strong indication in favor of a cautious assessment of the ensemble size when working with coupled multi-scale dynamics and performing strongly-coupled DA. Furthermore, as elucidate by Vannitsem and Lucarini (2016) and Tondeur et al. (2020), the near-neutral part of the spectrum in MAOOAM is directly connected to the coupling: including those directions within the ensemble subspace is paramount to propagate the data information content between ocean and atmosphere. This subspace also proved to be key in producing reliable ensemble forecasts in coupled ocean-atmosphere systems (Vannitsem and Duan 2020).

3.2.2 Stochastic Dynamics

Results in Sect. 3.1.1 show that the optimal KF and the reduced-rank KF with gain restricted to the leading BLVs are asymptotically equivalent for linear, perfect-deterministic models. Moreover, results in Sect. 3.2.1 demonstrate that in weakly non-linear error dynamics, with a perfect-deterministic model, these conclusions largely extend to the EnKF. Yet the exact reduced-rank recursion introduced in Sect. 3.1.2 evidenced important differences between the optimal and the reduced-rank formulations in the presence of model error for linear models. In this case, stochastic noise injected in asymptotically stable directions is not entirely damped out and remains finitely bounded. In addition, the upwelling process, albeit inherent to the reduced-rank formulation and not a consequence of model error, will move the constantly injected model noise from unfiltered to filtered directions.

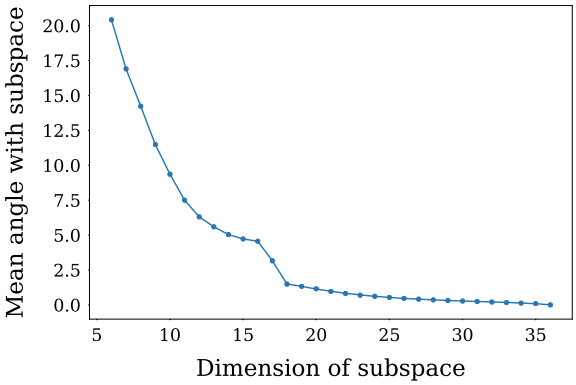

Fig. 9 MAOOAM coupled Ocean-Atmosphere model. Time- and ensemble-averaged angle (in degree) between the ensemble anomalies of the EnKF with $N = 10$ members, and subspaces spanned by the BLVs of increasing dimensions. These subspaces are constructed by starting from the unstable-neutral subspace ($n_0 = 6$) and adding one by one stable directions ordered decreasingly by their LE. The set-up is $\mathbf{H} = \mathbf{I}_{N_x}$, $\mathbf{R} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$ with $\sigma^{\%} = 0.08$ and $\Delta t = 1.68$ days

It is of interest thus to compare the differences between the full rank Kalman estimator, and both the standard and the exact reduced-rank KF recursions in a stochastically forced, nonlinear models. As a prototype, we use the model defined by the nonlinear flow of the Lorenz-96 (Lorenz 1996) system with additive noise, and $N_x = 40$. Suppose that $t_{k+1} = t_k + \Delta t$, so that the flow map taking all initial conditions to time $+\Delta t$ is defined $\phi_{\Delta t}(\mathbf{x}_k) = \mathbf{x}_{k+1}$. If $\mathbf{x}_k^{\mathrm{t}}$ represents the true physical state at time t_k, we define the dynamical model analogously to Eq. (8a) as a discrete, nonlinear map, $\mathbf{x}_{k+1}^{\mathrm{t}} = \phi_{\Delta t}\left(\mathbf{x}_k^{\mathrm{t}}\right) + \mathbf{w}_k$.

To avoid the interplay and superposition between sampling and model errors, and to be able to focus on the latter alone, instead of the EnKF we use here the EKF (see Sect. 3 and Jazwinski (1970)). The EKF estimates the forecast distribution for $\mathbf{x}_k^{\mathrm{t}}$ via the equation, $\mathbf{x}_{k+1}^{\mathrm{f}} = \phi_{\Delta t}\left(\mathbf{x}_k^{\mathrm{a}}\right)$, taking the analysis mean at time t_k to the forecast mean at time t_{k+1}, and by the linearized forecast equation for the covariance. In the following, the EKF propagates the full-rank covariance equation via the tangent-linear model defined along the mean equation and assimilates observations in all state components. On the other hand, both the standard and the exact reduced-rank formulations restrict the assimilation so that the image space of the gain is equal to the span of the leading r BLVs defined along the tangent-linear model of the mean equation. The difference between the standard and the exact formulation is as follows: the standard formulation only estimates the error covariance in the span of the leading r BLVs while the exact formulation simulates the entire covariance equation, estimating the free forecast covariance in the trailing modes and including its upwelling in the estimate of the uncertainty in the span of leading BLVs (Grudzien et al. 2018b, see Proposition 1)

Figure 10, adapted from Grudzien et al. (2018b), illustrates the differences in performance between the above described schemes. On the left, the average RMSE of

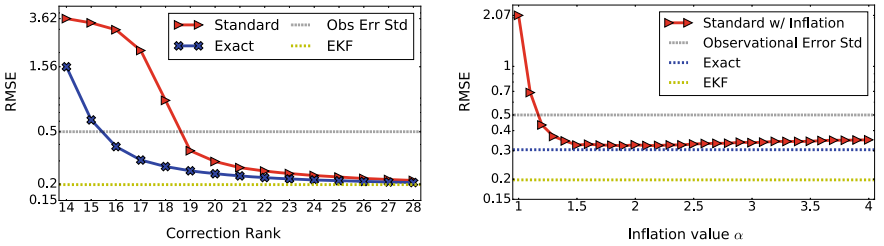

Fig. 10 **Left**: RMSE of the full-rank, standard reduced-rank and exact reduced-rank EKF. The correction rank of the reduced-rank estimators is varied in the horizontal axis. **Right**: RMSE of the full-rank, standard reduced-rank and exact reduced-rank EKF. Rank is fixed at $r = 17$ and multiplicative inflation in the standard reduced-rank recursion is varied in the horizontal axis

the (i) *full-rank*, the (ii) *standard reduced-rank* and the *exact reduced-rank EKF* formulations are plotted versus the rank of the reduced-rank gain over 10^5 analyses. The model possesses 13 positive LEs, therefore the reduced-rank EKFs under consideration have at least $r \geq n_0 + 1 = 14$. The system is fully observed with $\mathbf{R} = 0.25\mathbf{I}_{N_x}$. As r approaches $N_x = 40$, the two reduced rank formulations converge to the full rank EKF, with performance considered optimal for a filter in this system. However, for correction rank $r \leq 20$, there are substantial differences in performance between the standard formulation and that which includes the effect of the upwelling of error from the trailing BLVs: the exact formulation reaches both adequate and near-optimal performance with smaller r than the standard recursion.

On the right, Fig. 10 demonstrates the effect of multiplicative inflation on the standard recursion when the correction rank is fixed at $r = 17$. While multiplicative inflation greatly improves the performance of the standard recursion, this performance is actually bounded below by the RMSE of the exact formulation. This example shows that, in addition to sampling error, multiplicative inflation can be used to remedy the inadequacies of the standard reduced-rank formalism which neglects the effects of dynamic upwelling. This dynamic upwelling is a direct byproduct of the estimator being restricted to the span of the leading BLVs. Such an estimator arises when, e.g., the ensemble span aligns with the span of the leading BLVs in the EnKF, as demonstrated in Sect. 3.2.1. The efficacy of dynamic, multiplicative covariance inflation for treating the effect of model error separately from sampling error has also been demonstrated with statistical and optimization methods by, e.g., Mitchell and Carrassi (2015), Raanes et al. (2015), Raanes et al. (2019), Sakov et al. (2018), Fillion et al. (2020). The dynamical upwelling derived in Grudzien et al. (2018b) provides an explanation of one of the mechanisms responsible for the need for covariance inflation.

4 Data Assimilation for Chaotic Systems—How Chaos Becomes an Opportunity

We present here two areas where the knowledge about the chaotic nature and properties of the model dynamics, have been used pro-actively to achieve a better track of the system of interest and to reduce the forecast error. There have been two ways how this has been accomplished. The former has to do with the design of methods to inform where and when additional fewer data would lead to a major improvement in the analysis and forecast skill. This is usually referred to as "adaptive observation" or "target observation", and is the content of Sect. 4.1. The second has to do with incorporating, within the DA process itself, the information about the unstable subspace, with the goal of achieving a computational economy while maximising the error reduction. This gave raise to a family of DA methods known as "assimilation in the unstable subspace", summarised in Sect. 4.2.

Our treatments here are intentionally succinct and have the character of a review, but interested readers will find all the appropriate references to the original works.

4.1 Targeting Observations Using the Unstable Subspace

Historically, one very important development in this type of dynamical analysis of the climate predictability/DA problem arose from the question of how to generate adaptive observation systems. By "adaptive observations" is meant here observations whose locations, time and type is chosen such that their impact on the state estimate or on the forecast skill is the largest. It was well understood at the time that the growth of forecast errors was confined to a lower dimensional subspace of rapidly growing perturbations (Trevisan and Legnani 1995). The Fronts and Atlantic Storm Track (FASTEX) program (Snyder 1996), a multinational collaboration to investigate the growth and development of frontal cyclones, in particular was motivated by these dynamical approaches to generate adaptive observation schemes that would target regions of rapid forecast error growth. Other similar international efforts have followed, some of them including actual field campaign of measurements: the THORPEX (Fourrié et al. 2006) and the Winter Storm Reconnaissance programs (Szunyogh et al. 2002; Hamill et al. 2013). Two main approaches were considered, the *forced singular vectors* (Buizza et al. 1993; Palmer et al. 1998), and the *bred vectors* (Toth and Kalnay 1993, 1997), to identify the sensitivity areas of high forecast uncertainty.

Forced singular vectors are generated by the right singular vectors of the forward-in-time, tangent-linear model resolvent $\mathbf{M}$. It can be seen from the earlier discussions in Sect. 2 that the singular vectors can be interpreted as a finite-time approximation of the FLVs along a model forecast. They therefore indicate region where error will grow. On the other hand, the bred vectors are an ensemble-based approach to identify sensitivity regions. Particularly, the breeding scheme simulates how the modes of

fast growing error are maintained and propagated through the successive use of short range forecasts in weather prediction. The bred vectors are formed by initializing small perturbations of a control trajectory and forecasting these in parallel along the control. By successive rescaling of the perturbations amplitude back to a small value, this mimics the evolution of small perturbations under the tangent-linear model, and the span of these perturbations generically converges to the leading BLVs. Both of these approaches represent an early and intuitive way to utilizing the ergodic theory of chaotic dynamical systems in designing an effective adaptive observation scheme by targeting an unstable subspace in some form.

Trevisan and Uboldi (2004a), Uboldi and Trevisan (2006) utilized the bred vectors/BLV analysis of Toth and Kalnay (1993, 1997) and explicitly linked the methodology to Lyapunov stability theory. Importantly it was recognised that, instead of mimicking the error growth of the unforced (free) forecast, it was important to track the errors that develop within the DA cycle. This led to a modified version of the breeding approach known as Breeding on the Data Assimilation System (BDAS) Uboldi et al. (2005), Carrassi et al. (2007). The BDAS scheme for adaptive observations is based on the principle of the support of the forecast error lying primarily in the span of the BLVs, with the bred vectors acting as a proxy for the explicit decomposition. In practice the locations of few adaptive observations were selected to be in the areas where the leading BDAS modes attained their local maxima. In experiments with an atmospheric quasi-geostrophic model, BDAS was used successfully to locate one additional observation at each analysis time of a 3DVar cycle, leading to a dramatic improvement of the analysis skill, compared to cases where either a fixed or a randomly located adaptive observation was assimilated Carrassi et al. (2007).

4.2 Assimilation in the Unstable Subspace

Data assimilation has long been studied with the trade-off between accuracy and numerical efficiency as a goal. To this end a natural choice has been that of devising reduced-order schemes that focus the observational constraint on smaller, albeit crucial, part of the full dynamics. One of the most celebrated among those approaches is the assimilation in the unstable subspace (AUS) where the DA procedure is explicitly designed to track and control the unstable manifold of the dynamics, usually of much smaller dimension of the full phase space, thus aiming to a reduction in computational cost (Palatella et al. 2013).

We have seen in Sect. 2 that the full phase space of a chaotic dissipative dynamical system can be seen as split in a (usually much smaller) unstable-neutral subspace and a stable one (Kuptsov and Parlitz 2012). For instance, Carrassi et al. (2007) have shown how a quasi-geostrophic atmospheric model of $\mathcal{O}(10^5)$ degrees of freedom possesses an unstable-neutral subspace of dimension as small as $n_0 = 24$.

We have furthermore seen that, in deterministic chaotic systems and under the linear regime of error evolution, the uncertainty in the state estimate converges to zero outside of the unstable-neutral subspace. This phenomenon was at the core

of the idea of AUS, whereby the unstable-neutral subspace (or a suitable numerical approximation of it) is explicitly used in the DA scheme to parametrise the description (both temporally and spatially) of the uncertainty in the state estimate (Trevisan and Uboldi 2004b; Uboldi and Trevisan 2006; Carrassi et al. 2008a).

The AUS concept has been since then applied to different model scenarios and embedded within either KF-like or variational methods (Palatella et al. 2013). Carrassi et al. (2008b) plugged AUS into a 3DVar cycle in such a way that the observations in the proximity of the leading BDAS mode's maximum were assimilated by imposing that the analysis increment follows the BDAS mode. This implied that the larger the estimated error growth with BDAS, the larger the analysis correction. The combined 3DVar-AUS was extraordinarily more accurate than the 3DVar when the same amount of data were assimilated. AUS was subsequently generalised and embedded into 4DVar (4DVar-AUS, Trevisan et al. 2010), and in an EKF, (EKF-AUS, Trevisan and Palatella 2011). The forecast error covariance was projected so as to confine the analysis correction to the unstable-neutral subspace. Remarkably both reduced-rank formulations, 4DVar-AUS and EKF-AUS, showed superior skills than their full-rank counterparts.

AUS relied on the assumption that errors evolve linearly. Going beyond this, Palatella and Trevisan (2015) presented an original way of mixing the contributions from the various Lyapunov vectors such that a quadratic expansions of the error is considered. This improved the performance of the standard EKF-AUS particularly in regimes of increasing nonlinearities. It remains however to be seen to which extent AUS concept could be used within fully nonlinear DA schemes. This is the subject of Sect. 5.1 and of the references therein. Although AUS has been largely used in a perfect model scenario, Palatella and Grasso (2018) proposed a suitable modification that allows for incorporating parametric model errors. As proved in Grudzien et al. (2018b) and detailed in Sects. 3.1.2 and 3.2.2 the use of AUS in chaotic systems forced by additive stochastic noise would necessarily require the additional inclusion of the asymptotically weakly stable modes.

A key caveat in all of the aforementioned applications of AUS is that one needs to compute in real time the BLVs to be used in the analysis. Therefore, while AUS proved capable to improve accuracy, it did not accomplishes a computational economy, unless the BLVs were all pre-computed and stored. Note however that the latter is not just a technological challenge given that the BLVs depends on the system's state and vice-versa if BLVs are to be used in the analysis update. Thus it is not straightforward to decouple their online estimation and use within the analysis. At the same time however, as we have seen in Sect. 3, AUS concept proved to be very powerful to understand, design and interpret the functioning of the KF and EnKF in chaotic systems.

5 Forward Looking

5.1 AUS in a Non-Gaussian Filter?

In this subsection we attempt to improve the performance of the (bootstrap) particle filter (PF, Farchi and Bocquet 2018) by AUS. The underlying hypothesis is that observational components in the stable subspace contribute little in the way of precision (since nearby orbits within the stable subspace converge), but a lot of noise. Therefore, the investigation will explore whether discarding observational information outside of the unstable subspace can mitigate the acute collapse of weights experienced by PFs in high-dimensional systems, manifesting the "curse of dimensionality". If so, this could be used to reduce the number of particles required, which scales exponentially with some measure of the system size (Snyder et al. 2008). A secondary objective is to investigate the effectiveness of a few different methods of targeting observing systems to the unstable subspace, potentially also reducing costs.

We perform synthetic DA experiments with the Lorenz-96 system. The state size is set to $N_x = 10$ and there is no dynamical noise ($\mathbf{Q} = \mathbf{0}$). Four different observation configurations targeting the unstable subspace are tested. For each of them, observations are taken 0.2 apart in time, with independent noise of variance 1.5. Each experiment lasts for 10^5 analysis cycles. The RMSE averages of each method are tabulated for a range of ensemble sizes and observation operator ranks, N_y, and plotted as curves in Fig. 11. The plotted scores represent the lowest obtained among a large number of tuning settings, selected for optimality at each point. For the PF the tuning parameters are: (i) the threshold for (universal) resampling, which is triggered if the threshold is larger than the effective ensemble size, $\|\mathbf{w}\|^{-2}$, where $\mathbf{w}$ is the vector of weights, and (ii) the bandwidth (scaling) of the regularizing post-resample jitter, whose covariance is computed from the weighted ensemble. For comparison the (symmetric square-root) EnKF (Hunt et al. 2004) is also tested. Its tuning parameters are (i) the post-analysis inflation factor and (ii) whether or not to apply random, covariance-preserving rotations (Sakov and Oke 2008b).

The top-left panel Fig. 11 shows that the error decreases monotonically in N_y when the observation operator are rows of the identity. The top-right panel shows, by contrast, that when observing the BLVs, $\mathbf{L}^{\mathrm{b}}_k$, the PF with the largest ensemble levels off at near-optimal performance with as few as $N_y = n_0 = 4$ observations, only improving marginally thereafter. The same trait can also be noted for lower N, albeit less clearly. Also recall that a similar "levelling off" of RMSE around n_0 occurred in Fig. 4 of Sect. 3.2.1, whose x-axis is N (rather than N_y, as here). This result demonstrates that targeting observations to the directions of dynamical growth of the uncertainty is efficacious. Interestingly, the performance of any given DA method is nearly independent of the observing system (i.e. panel) for $N_y = 10$. This makes sense considering that all of $\mathbf{I}$, $\mathbf{L}^{\mathrm{b}}_k$, $\mathbf{V}_{k+1}$, and $\mathbf{U}_k$ are orthogonal, i.e. equal up to a rotation.

In all panels of Fig. 11, the RMSE score of the PF (as well as that of the EnKF), for any ensemble size, never degrades by the inclusion of more observations (except for

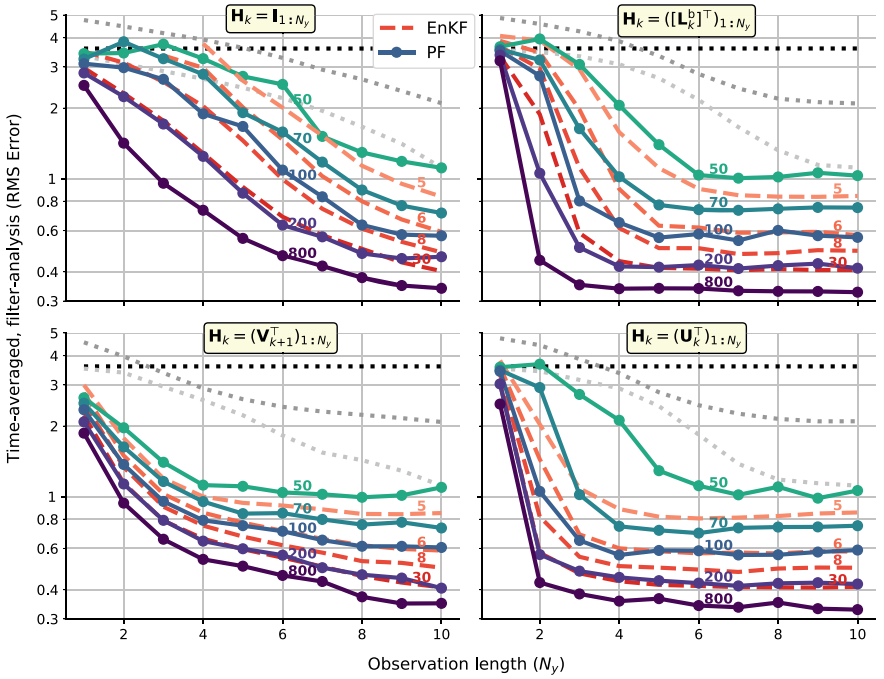

Fig. 11 Benchmarks of filter accuracy (RMSE) from synthetic DA experiments on the Lorenz-96 system, plotted as functions of the observation dimension (N_y). Specifically, the observation operator consists of the N_y leading rows of the transpose of the identity matrix (*top-left*), the BLV matrix, $\mathbf{L}_k^b$, computed by recursive QR decompositions (*top-right*), the 1-cycle forward singular vectors, $\mathbf{V}_{k+1}$ (*bottom-left*), and 1-cycle backward singular vectors, $\mathbf{U}_k$ (*bottom-right*), all of which are defined via the fundamental matrix of the orbit of the (supposedly unknown) truth. The number of members/particles used for the ensemble Kalman filter (EnKF, *dashed*) and Particle filter (PF, *solid*) is tagged above each line. No further improvement is obtained by increasing the ensemble size beyond $N = 30$ for the EnKF and $N = 800$ for the PF, for which the PF achieves better accuracy than the EnKF, as expected for nonlinear systems. The greyscale, dotted lines, included for context, show the performance of baseline methods, whose analysis estimates, $\mathbf{x}^a$, are given by $\overline{\mathbf{x}}$ for Climatology (*black*), $\overline{\mathbf{x}} + \mathbf{K}(\overline{\mathbf{C}})\,[\mathbf{y} - \overline{\mathbf{x}}]$ for Optimal interp. (*dark grey*), $\mathbf{x}^f + \mathbf{K}(c\mathbf{I})\,[\mathbf{y} - \mathbf{x}^f]$ for 3D-Var (*light grey*). Here, $\overline{\mathbf{x}}$ and $\overline{\mathbf{C}}$ are the mean and covariance of the (invariant measure of the) system dynamics, $\mathbf{K}(\mathbf{C}) = \mathbf{C}\mathbf{H}^\top(\mathbf{H}\mathbf{C}\mathbf{H}^\top + \mathbf{R})^{-1}$ is a gain matrix, $\mathbf{x}^f$ is the forecast of the previous $\mathbf{x}^a$, and c is a scaling factor subject to tuning. The plots show that targeting observations to the directions of dynamical growth of the uncertainty is efficacious and that, for this task, $\mathbf{L}_k^b$ and $\mathbf{U}_k$ are similarly effective, and superior to $\mathbf{V}_{k+1}$. Moreover, the RMSE performance of any method, also for the PF (which is our focus), never degrades with the inclusion of more observations

what we adjudge to be noise). This feature was also found in experiments with even smaller ensembles than shown, and experiments with larger observation errors. Thus it seems that the curse of dimensionality is not mitigated by discarding observations. Hence there is little to gain by eliminating observation components outside of the unstable-neutral subspace, apart for the potential computational efficiency in the case that the unstable-neutral has been pre-computed.

It should be noted that this finding runs counter to that of Maclean and Vleck (2019), Beeson and Namachchivaya (2020), Potthast et al. (2019), all of which report success in mitigating weight collapse by reducing the observations to the leading components of $\mathbf{L}_k^{\mathrm{b}}$ or some related matrix. Beeson and Namachchivaya (2020) also tested observation operators given by time-local backward and forward Lyapunov vectors, defined through the singular value decomposition of the resolvent: $\mathbf{M}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\top$. They reported better targeting results with $\mathbf{V}_{k+1}$ than with $\mathbf{U}_k$, which in turn was more effective than $\mathbf{L}_k^{\mathrm{b}}$. Our results similarly show $\mathbf{U}_k$ (bottom-right panel) to be slightly more effective for targeting observations than $\mathbf{L}_k^{\mathrm{b}}$ (top-right panel), albeit only for intermediate ensemble sizes. Both, however, are more effective than $\mathbf{V}_{k+1}$ (bottom-left). Moreover, also for $\mathbf{V}_{k+1}$ and $\mathbf{U}_k$, we never observe lower RMSE scores when using fewer observation components. There are some differences in the experimental setting; notably our system is deterministic, and the jitter we apply is restricted to the ensemble subspace. It is unclear if these differences can account for the disparity in conclusions.

Why does PF performance not suffer from the inclusion of further, visibly redundant, observations (contrary to our initial expectation)? Consider the likelihood (i.e. weighting) of particle $n \in [1, \ldots, N]$, at a given (implicit) time:

$$p(\mathbf{y}|\mathbf{x}_n) = \phi\big(\|\mathbf{y} - \mathbf{H}\mathbf{x}_n\|_{\mathbf{R}}^2\big)\,, \tag{38}$$

where the norm is defined as $\|\mathbf{y}\|_{\mathbf{R}}^2 = \mathbf{y}^\top \mathbf{R}^{-1}\mathbf{y}$, and ϕ is a radial density such that Eq. (38) represents an elliptical distribution, for example a Gaussian. Suppose for the sake of simplicity that both the observation operator and its error covariance are identity, i.e. $\mathbf{H} = \mathbf{R} = \mathbf{I}$, and let $\boldsymbol{\Pi}$ be the matrix of orthogonal projection onto the unstable-neutral subspace. Now, assuming the PF controls the error in the unstable-neutral subspace, it seems reasonable to assume, given the results for linear systems of Sect. 3, that the particles will converge onto the unstable-neutral subspace: $(\mathbf{I} - \boldsymbol{\Pi})(\mathbf{x}_n - \mathbf{x}) \to \mathbf{0}$, at least as a first-order approximation. Supposing this, and by decomposing the norm into two orthogonal components, it can be shown that

$$\|\mathbf{y} - \mathbf{H}\mathbf{x}_n\|^2 = \|\boldsymbol{\Pi}(\mathbf{y} - \mathbf{H}\mathbf{x}_n)\|^2 + \|(\mathbf{I} - \boldsymbol{\Pi})\mathbf{v}\|^2\,, \tag{39}$$

where $\mathbf{v}$ is the observation noise. Thus, projecting the observations onto the unstable-neutral subspace, which would eliminate the last term of Eq. (39), merely reduces the data mismatch by a *constant* (with respect to n, the particle index). Thus

$$p(\mathbf{\Pi}\mathbf{y}|\mathbf{x}_n) = c\, p(\mathbf{y}|\mathbf{x}_n)\,, \tag{40}$$

for some $c > 0$ which is rendered inconsequential by the subsequent weight normalization. Hence, the inclusion of the remaining observation (outside the unstable-neutral subspace), does not cause a higher variance in the weights, nor the associated tendency to collapse/degeneracy.

The essence of the above reasoning is that the PF prior has zero support, i.e. uncertainty, in the stable subspace, and therefore ignores observational components, *including the noise*, in that subspace. This should be contrasted with the situation for dynamical noise (model error), which is "up-welled" from the stable to the unstable subspace, as per Sect. 3.1.2. This distinction exemplifies the difference between uncertainty addition (by dynamical noise) and subtraction (by likelihood updates).

For nonlinear dynamical systems, as illustrated in Sect. 3.2, the particles do not neatly align with the unstable-neutral subspace. Capturing this nonlinear aspect of the flow is generally seen as an advantage of the PF. It might be argued, though, that the observation noise is likely large compared to the spread of the prior particles in the stable subspace, and therefore the observations should be reduced by discarding the corresponding components. However, as highlighted above, the observational noise is constant in the particle index, and therefore its amplitude does not constitute a particular source of weight variability. Instead, the weight variability originates in the prior, which was assumed to have low variability in this scenario.

The analysis resulting in Eq. (40) assumed that $\mathbf{H} = \mathbf{R} = \mathbf{I}$ followed by $\mathbf{H} = \mathbf{\Pi}$. In the case of the observation operator consisting of the leading $N_y \geq n_0$ rows of the transposed BLV matrix, $\mathbf{H}_k = (\mathbf{L}_k^{\mathrm{b}})^{\top}_{1:N_y}$, the same conclusion can be derived along similar lines. In the general case, for any $\mathbf{H}$, it is not obvious how to reduce the observations so as to only measure the unstable-neutral subspace. To accomplish this, both Maclean and Vleck (2019) and Beeson and Namachchivaya (2020) apply the pseudo-inverse $\mathbf{H}^{+}$ before their reduction. This can be costly if $\mathbf{H}$ is large. A more practical approach is to reduce the observations as $\hat{\mathbf{y}} = (\mathbf{H}\hat{\mathbf{Q}})^{+}\mathbf{y}$, or $\hat{\mathbf{y}} = \mathbf{\Pi}\mathbf{y}$ with $\mathbf{\Pi} = (\mathbf{H}\hat{\mathbf{Q}})(\mathbf{H}\hat{\mathbf{Q}})^{+}$, with $\hat{\mathbf{Q}} = (\mathbf{L}_k^{\mathrm{b}})^{\top}_{1:N_y}$ In any case, re-doing the same derivation, Eq. (40) again follows, including the same implications for the weights.

In summary, discarding observational information outside of the unstable subspace does not yield improvements in the PF because it already embodies this "flow-dependent" information. A similar conclusion was also drawn for the iterative ensemble Kalman smoother by Bocquet and Carrassi (2017). Thus, while AUS is a powerful explanatory and diagnostics tool, it is not obvious if it can be used to combat the curse of dimensionality for PFs.

Lastly, this section adds some clarification to the influential paper by Snyder et al. (2008), whose conclusion is sometimes taken to be that the required ensemble size for PFs scales exponentially with *observation* size. Our results rather indicate that the performance of a well-tuned PF will not deteriorate with the inclusion of more observations (even if they are redundant). In other words, that the required ensemble size depends on the rank of the *state* space, or more precisely for chaotic dynamics, the rank of the unstable-neutral subspace. Snyder et al. (2015) points out that the "effective dimension" may be limited by the observation size if this is smaller than

the state size. However, in case $N_y < n_0$, no filtering system using flow-dependent priors will be able to achieve satisfactory performance, because the system is not sufficiently observed. Of course, the question of observability is complicated by considering time-dependent observation networks, while localization can also be applied to alleviate the curse of dimensionality (Farchi and Bocquet 2018).

5.2 Data Assimilation and Random Attractors

AUS and its theoretical extensions provide a framework to interpret the asymptotic inferences of the EnKF. Provided that the forecast anomalies can be considered to be perturbations of the true physical state, and if their evolution can be approximated by the tangent-linear model along the true trajectory, the dynamics of the EnKF anomalies can be decomposed along the Oseledec spaces of the true physical state. The stability and accuracy of the EnKF is largely determined by the ability of the ensemble-based gain to correct the growth of forecast errors in the unstable-neutral subspace, with respect to the uniform-complete observability of these modes. Additive noise complicates the description of the asymptotic forecast uncertainty as uncorrected forecast errors in the span of the stable BLVs may be bounded but impractically large. The upwelling of such errors into the ensemble span furthermore necessitates covariance inflation to rectify the systematic underestimation of the forecast uncertainty in the standard KF-AUS recursion.

Model stochasticity arises systematically in multiscale climate dynamics where there are large scale-separations between resolved and unresolved dynamic processes. In the asymptotic limit of scale-separation, unresolved dynamics can be reduced to additive Gaussian noise due to the Central Limit Theorem; finite scale-separation in reality leads to non-Markovian memory terms in addition to additive stochastic forcing in the exact model reduction of a multiscale model, as in Mori-Zwansig formalism (Gottwald et al. 2015). Several mathematically rigorous frameworks have been developed to model and simulate the effect of small-scale dynamics on the large-scale dynamics with stochastic parameterization, including averaging methods, perturbation methods and combinations of the two—see, e.g., the survey of approaches by Demaeyer and Vannitsem (2018).

The theory of random dynamical systems offers a novel means of analysis of the DA cycle for multi-scale chaotic systems with large scale separations and model reduction error. Characterizing the asymptotic Bayesian posterior in terms of the properties of a random, nonlinear and ergodic attractor is a natural step forward in the philosopy of AUS. Recent work suggests that the support of the posterior measure of the DA cycle can be asymptotically bounded by the support of the SRB measure in deterministic, nonlinear dynamical systems (Oljača et al. 2018). While this is an intuitively appealing result, the existence of an SRB measure in deterministic dynamics usually requires a hyperbolicity assumption which may not be appropriate in weather and climate, e.g. Vannitsem and Lucarini (2016). However, many of the theoretical challenges in showing the existence of SRB measures in deterministic dynamics are

relaxed in a random dynamical systems setting. Indeed, the Pesin entropy formula holds under very general assumptions for stochastic flows of diffeomorphisms (Liu and Qian 2006, Theorem 3.1 and discussion on page 127), establishing the link once again between the observed instability in the dynamics and the statistical properties of the invariant measure.

For such a dynamical interpretation of the DA cycle to be credible, the correct specification of random models in stochastic-physical systems is a primary concern; stochastically reduced models should be specified to preserve conservation laws and the original model's dynamics (Cotter et al. 2019). In addition to the correct stochastic model specification, important differences in the statistical properties of model forecasts of stochastic dynamical systems have been observed due to the discretization errors of certain low-order schemes. For example, Frank and Gottwald (2018) develop an order 2.0 Taylor scheme to correct the bias in the drift term induced by the Euler-Maruyama scheme in their study system. Grudzien et al. (2020) likewise find that the bias due to discritization error of the Euler-Maruyama scheme can be sufficient to cause filter divergence of the EnKF in the stochastically forced Lorenz-96 model. Grudzien et al. (2020) emphasize the important role of efficient, weak numerical schemes for the simulation of ensemble-based forecasts. Unlike strong convergent schemes, numerical schemes that converge in the weak sense can make reductions in the complexity of simulation by emphasizing the accuracy of the convergence of the ensemble to the forecast distribution rather than the accuracy of any individual ensemble member.

6 Summary and Conclusion

Chaos is ubiquitous in natural, physical and laboratory systems. Scientists have long coped with this whenever attempting to model, predict or control such systems. Combining and confronting models with data is common in science and *data assimilation* (DA) is the term coined in the context of numerical weather prediction science to encompass the methods that perform such a combination. The outputs of DA is the improved representation of the system under study, and an estimate of the associated uncertainty.

Inevitably, the sensitivity to initial conditions of chaotic systems, including state-dependence of the directions of error growth, is a challenge for DA. On the other hand, the energy dissipation typical of real systems implies a "dimension reduction" in that errors are confined within a subspace of the full system's phase space.

The DA process requires furnishing a prior distribution, whose specification is a severe difficulty in high dimension. Gaussian methods reduce the complexity to that of estimating the prior mean and covariance. Yet, with the exceptionally high-dimensions of geophysical problems, a proper estimate, and storage of the prior covariance matrices is still challenging. Thus, suitable reduced-rank formulations should be used to lower the computational load while maintaining a good description of the errors about the full system.

In chaotic dissipative systems this goal can be achieved by monitoring the unstable-neutral subspace of the model dynamics and performing DA only within that subspace. This is the idea at the basis of the class of methods known as *assimilation in the unstable subspace* (Palatella et al. 2013), that were developed mainly in the years between 2004 and 2015; we reviewed them in Sect. 4. Despite reducing the problem size to that of the unstable-neutral subspace (of dimension $n_0 \ll N_x$), AUS methods proved aptly skillful, very close to those of their more costly full-rank competitors.

However, the reduction of cost achieved at the analysis steps is offset by the additional cost of monitoring and tracking the unstable-neutral subspace. This requires computing the Lyapunov vectors (usually the backward Lyapunov vectors), which implies computing the tangent linear model (or alternatively evolving an ensemble of bred perturbations mimicking the evolution under the tangent linear model) and a repeated QR matrix decomposition. For specific purposes, one can opt for tracking only a few dominant unstable modes. This was the case for early adaptive (targeted) observation results (Carrassi et al. 2007) or when AUS was used to complement a classical DA method (Carrassi et al. 2008b).

In parallel to the early developments of AUS, several studies with ensemble DA methods in chaotic dissipative systems were suggesting that a number of their key properties were related to the model instabilities, including the rank and span of the ensemble-based forecast error covariance, as well as the skill of the analysis (Sakov and Oke 2008b; Carrassi et al. 2009; Ng et al. 2011; Bocquet and Sakov 2014). These results indicate that the ensemble anomalies automatically align with the unstable-neutral subspace, thus resulting in the analysis to be confined to it.

To put this mechanism on a more rigorous theoretical footing, a stream of works have studied the relation between the unstable-neutral subspace in linear systems using the Kalman filter (KF) and Kalman smoother (KS) (Gurumoorthy et al. 2017; Bocquet et al. 2017; Bocquet and Carrassi 2017). These works, reviewed in Sect. 3.1, have provided analytic proofs that the span of the error covariance matrices of the KF and KS tends asymptotically to the unstable-neutral subspace, independent of the initial condition (i.e. no matter the number of the ensemble members, provided it exceeds the size of the unstable-neutral subspace). For stochastic systems with additive noise it was proved that asymptotically weakly stable modes, that one might discard in deterministic systems, must be included and analytic bounds for the error were provided (Grudzien et al. 2018a).

How do these results hold for nonlinear systems? In the case of chaotic deterministic systems, this was studied in Bocquet and Carrassi (2017) and further in Sect. 3.2.1 of this chapter. It was numerically showed that an ensemble comprising at least as many members as the size of the unstable-neutral subspace plus one ($N \geq n_0 + 1$) is needed to achieve satisfactorily skill with the ensemble Kalman filter (EnKF). Section 3.2.1 also considered the case of a coupled multiscale system with a quasi-degenerate spectrum of Lyapunov exponents. This originates in the presence of many close-to-zero exponents that are related to the coupling mechanisms (Vannitsem and Lucarini 2016). It is shown that their full inclusion in the ensemble design is needed to reduce the EnKF analysis error to a satisfactorily low level.

The case of nonlinear stochastic chaotic systems with additive noise was studied in Grudzien et al. (2018b) and reviewed in Sect. 3.2.2. The section explains the role of the weakly stable modes already identified in linear systems, but also discovered the *upwelling mechanism* for which uncertainty is upwelled from unfiltered (stable) modes to filtered (unstable) ones. The upwelling phenomenon is not exclusive of nonlinear systems and it is in fact present in linear systems too. It provides an additional rationale to the use of multiplicative inflation, otherwise known by numerically evidence to be needed for a proper functioning of reduced-rank filters even in perfect model scenarios (see e.g., Raanes et al. 2019).

An outlook at how this research may evolve is given in Sect. 5. In particular in Sect. 5.1 we provide original results on the use of the AUS approach, i.e. exploiting the unstable-neutral subspace, in particle filters (PFs), a fully non-Gaussian DA method. Results indicate that targeting observations within the unstable-neutral subspace is very effective. However, by analogy with what was proved for the EnKF in Sect. 3.2.1, adding observations along the stable modes does not deteriorate the analysis. In the particle filter too, the particles automatically align along the unstable-neutral subspace so that the contribution from observations in its complement stable subspace is negligible. Our results shed also new insight on the scaling of the particle numbers needed to reach convergence. It is shown to depend on the size of the unstable-neutral subspace rather than the observation vector size. In Sect. 5.2 we surveyed how the novel concept of random attractors could offer new ways to further exploit the idea behind AUS on stochastic multi-scale systems with large scale separation. Moreover, we also described how to amend numerical integration schemes when doing ensemble DA on such systems, as a trade-off between accuracy and computational cost.

It is important to recall that all of the results with the EnKF and PF that we have presented are obtained without the use of *localization* (see e.g. Carrassi et al. 2018, their Sect. 4.4, and Farchi and Bocquet 2018 for localisation in the EnKF and PF, respectively). While we are well aware of the dramatic positive impact of localization on the filters' skill, we intentionally did not use it as it artificially changes the ensemble-covariance rank and span, thus making it impossible to link them exclusively to the model instabilities.

The use of the time-dependent unstable-neutral subspace to represent uncertainty in dynamical systems is still very appealing and potentially prone to success in a wider area than explored so far. For instance, in a recent work by Bocquet et al. (2021) model error arise from parameter mispecification and the EnKF is applied to estimate simultaneously the (chaotic) model state and N_p parameters. The EnKF is used in the state-augmentation formulation and the standard persistence model is adopted for the parameter dynamics. It was shown that the bound for the necessary ensemble size becomes $N \geq n_0 + N_p + 1$: N_p additional members are required to infer the N_p parameters. While the linear one-to-one relation between the number of parameters and that of the additional members is a consequence of the choice of a persistence model and will change if a different parameters dynamics is in place, this result further highlights how much the design of the EnKF is, and can be tied to the properties of the dynamical model.

Future developments along these lines will unavoidably have to tackle the obstacle of the computing cost of the unstable-neutral subspace. We speculate that recent progress in the area of machine learning (Goodfellow et al. 2016) may help. Neural network surrogate models of chaotic systems have shown capabilities to reproduce the spectrum of the asymptotic Lyapunov exponents fairly well (Pathak et al. 2017; Brajard et al. 2020). It is matter of future investigations to explore the possibilities of machine learning algorithms that learn about the time-dependent instabilities from offline long model simulations and then assist the model in the prediction mode by providing a proxy of the unstable-subspace at each analysis time.

Acknowledgements AC has been funded by the UK Natural Environment Research Council award NCEO02004. CEREA is member of Institut Pierre-Simon Laplace (IPSL). PNR has been partly funded by DIGIRES, a project sponsored by industry partners and the PETROMAKS2 programme of the Research Council of Norway.

References

Adrianova LY (1995) Introduction to linear systems of differential equations. Am Math Soc

Asch M, Bocquet M, Nodet M (2016) Data assimilation: methods, algorithms, and applications. SIAM

Barreira L, Pesin YB (2002) Lyapunov exponents and smooth ergodic theory. American Mathematical Society, Student Mathematical Library. 9780821829219

Beeson R, Namachchivaya NS (2020) Particle filtering for chaotic dynamical systems using future right-singular vectors. Nonlinear Dynamics. https://doi.org/10.1007/s11071-020-05727-y

Benettin G, Galgani L, Giorgilli A, Strelcyn J (1980) Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 1: Theory. Meccanica 15(1):9–20

Bocquet M (2011) Ensemble Kalman filtering without the intrinsic need for inflation. Nonlinear Process Geophys 18(5):735–750

Bocquet M, Carrassi A (2017) Four-dimensional ensemble variational data assimilation and the unstable subspace. Tellus A 69(1):1304504

Bocquet M, Sakov P (2014) An iterative ensemble Kalman smoother. Q J R Meteorol Soc 140:1521–1535

Bocquet M, Gurumoorthy KS, Apte A, Carrassi A, Grudzien C, Jones CKRT (2017) Degenerate Kalman filter error covariances and their convergence onto the unstable subspace. SIAM/ASA J Uncertain Quant 5(1):304–333

Bocquet M, Farchi A, Malartic Q (2021) Online learning of both state and dynamics using ensemble Kalman filters. Foundation of Data Science. https://doi.org/10.3934/fods.2020015

Brajard J, Carrassi A, Bocquet M, Bertino L (2020) Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. J Comput Sci 44

Buizza R, Tribbia J, Molteni F, Palmer TN (1993) Computation of optimal unstable structures for a numerical weather prediction model. Tellus A 45(5):388–407

Carrassi A, Trevisan A, Uboldi F (2007) Adaptive observations and assimilation in the unstable subspace by breeding on the data-assimilation system. Tellus A 59(1):101–113

Carrassi A, Ghil M, Trevisan A, Uboldi F (2008a) Data assimilation as a nonlinear dynamical systems problem: stability and convergence of the prediction-assimilation system. Chaos 18

Carrassi A, Trevisan A, Descamps L, Talagrand O, Uboldi F (2008b) Controlling instabilities along a 3DVar analysis cycle by assimilating in the unstable subspace: a comparison with the EnKF. Nonlinear Process Geophys 15:503–521
Carrassi A, Vannitsem S, Zupanski D, Zupanski M (2009) The maximum likelihood ensemble filter performances in chaotic systems. Tellus A 61:587–600
Carrassi A, Bocquet M, Bertino L, Evensen G (2018) Data assimilation in the geosciences-an overview on methods, issues and perspectives. WIREs Clim Change e535. https://doi.org/10.1002/wcc.535
Cotter C, Crisan D, Holm DD, Pan W, Shevchenko I (2019) Numerically modeling stochastic lie transport in fluid dynamics. Multiscale Model Simul 17(1):192–232
De Cruz L, Demaeyer J, Vannitsem S (2016) The modular arbitrary-order ocean-atmosphere model: MAOOAM v1.0. Geosci Model Dev 9(8):2793–2808. https://doi.org/10.5194/gmd-9-2793-2016
Demaeyer J, De Cruz L (2020) Climdyn/QGS: QGS version 0.2.0 release, July 2020. https://doi.org/10.5281/zenodo.3941877
Demaeyer J, Vannitsem S (2018) Stochastic parameterization of subgrid-scale processes: a review of recent physically based approaches. In: Advances in nonlinear geosciences. Springer, pp 55–85
Demaeyer J, De Cruz L, Vannitsem S (2020) QGS: a flexible python framework of reduced-order multiscale climate models. J Open Source Softw 5(56):2597. https://doi.org/10.21105/joss.02597
Dieci L, Van Vleck ES (2002) Lyapunov spectral intervals: theory and computation. SIAM J Numer Anal 40(2):516–542
Dieci L, Van Vleck ES (2007) Lyapunov and Sacker-sell spectral intervals. J Dyn Diff Equat 19(2):265–293
Evensen G (2009a) Data assimilation: the ensemble Kalman filter, 2nd edn. Springer, Berlin, Heildelberg
Evensen G (2009b) Data assimilation: the ensemble Kalman filter. Springer Science & Business Media
Farchi A, Bocquet M (2018) Review article: comparison of local particle filters and new implementations. Nonlinear Process Geophys 25(4):765–807. https://doi.org/10.5194/npg-25-765-2018
Fillion A, Bocquet M, Gratton S, Gürol S, Sakov P (2020) An iterative ensemble Kalman smoother in presence of additive model error. SIAM/ASA J Uncertain Quant 8(1):198–228. https://doi.org/10.1137/19M1244147
Fourrié N, Marchal D, Rabier F, Chapnik B, Desroziers G (2006) Impact study of the 2003 north Atlantic thorpex regional campaign. Q J R Meteorol Soc 132(615):275–295
Frank J, Gottwald GA (2018) A note on statistical consistency of numerical integrators for multiscale dynamics. Multiscale Model Simul 16(2):1017–1033
Froyland G, Hüls T, Morriss GP, Watson TM (2013) Computing covariant lyapunov vectors, oseledets vectors, and dichotomy projectors: a comparative numerical study. Physica D 247(1):18–39
Ghil M, Malanotte-Rizzoli P (1991) Data assimilation in meteorology and oceanography. In: Advances in geophysics, vol 33. Elsevier, pp 141–266
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press
Gottwald GA, Crommelin D, Franzke C (2015) Stochastic climate theory. In: Nonlinear and stochastic climate dynamics, pp 209–240
Grudzien C, Carrassi A, Bocquet M (2018a) Asymptotic forecast uncertainty and the unstable subspace in the presence of additive model error. SIAM/ASA J Uncertain Quant 6(4):1335–1363
Grudzien C, Carrassi A, Bocquet M (2018b) Chaotic dynamics and the role of covariance inflation for reduced rank Kalman filters with model error. Nonlinear Process Geophys 25(3):633–648
Grudzien C, Bocquet M, Carrassi A (2020) On the numerical integration of the Lorenz-96 model, with scalar additive noise, for benchmark twin experiments. Geosci Model Dev 13(4):1903–1924
Gurumoorthy KS, Grudzien C, Apte A, Carrassi A, Jones CKRT (2017) Rank deficiency of Kalman error covariance matrices in linear time-varying system with deterministic evolution. SIAM J Control Optim 55(2):741–759

Hamill TM, Yang F, Cardinali C, Majumdar SJ (2013) Impact of targeted winter storm reconnaissance dropwindsonde data on midlatitude numerical weather predictions. Mon Weather Rev 141(6):2058–2065

Hunt BR, Kalnay E, Kostelich EJ, Ott E, Patil DJ, Sauer T, Szunyogh I, Yorke JA, Zimin AV (2004) Four-dimensional ensemble Kalman filtering. Tellus A 56(4):273–277

Jazwinski AH (1970) Stochastic processes and filtering theory. Academic Press, New-York

Kalman RE (1960) A new approach to linear filtering and prediction problems. J Fluids Eng 82:35–45

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press

Kuptsov PV, Parlitz U (2012) Theory and computation of covariant Lyapunov vectors. J Nonlinear Sci 22(5):727–762

Legras B, Vautard R (1996) A guide to Lyapunov vectors. In: ECMWF workshop on predictability. Reading, United-Kingdom, pp 135–146

Liu PD, Qian M (2006) Smooth ergodic theory of random dynamical systems. Springer

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141

Lorenz EN (1996) Predictability: a problem partly solved. In: Proceedings of seminar on predictability, vol 1

Maclean J, Vleck ESV (2019) Particle filters for data assimilation based on reduced order data models

Mitchell L, Carrassi A (2015) Accounting for model error due to unresolved scales within ensemble Kalman filtering. Q J R Meteorol Soc 141(689):1417–1428

Ng GHC, McLaughlin D, Entekhabi D, Ahanin A (2011) The role of model dynamics in ensemble Kalman filter performance for chaotic systems. Tellus A 63(5):958–977

Nipen TN, Seierstad IA, Lussana C, Kristiansen J, Hov O (2019) Adopting citizen observations in operational weather prediction. Bull Am Meteorol Soc 101(1):E43–E57. ISSN: 0003-0007. https://doi.org/10.1175/BAMS-D-18-0237.1

Oljača L, Bröcker J, Kuna T (2018) Almost sure error bounds for data assimilation in dissipative systems with unbounded observation noise. SIAM J Appl Dyn Syst 17(4):2882–2914. https://doi.org/10.1137/17M1162305

Palatella L, Grasso F (2018) The EKF-AUS-NL algorithm implemented without the linear tangent model and in presence of parametric model error. SoftwareX 7:28–33. ISSN 2352-7110

Palatella L, Trevisan A (2015) Interaction of Lyapunov vectors in the formulation of the nonlinear extension of the Kalman filter. Phys Rev E 91

Palatella L, Carrassi A, Trevisan A (2013) Lyapunov vectors and assimilation in the unstable subspace: theory and applications. J Phys A: Math Theor 46

Palmer T, Gelaro R, Barkmeijer J, Buizza R (1998) Singular vectors, metrics, and adaptive observations. J Atmos Sci 55(4):633–653

Pathak J, Lu Z, Hunt BR, Girvan M, Ott E (2017) Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. Chaos Interdiscip J Nonlinear Sci 27(12):121102

Penny S, Bach E, Bhargava K, Chang C-C, Da C, Sun L, Yoshida T (2019) Strongly coupled data assimilation in multiscale media: experiments using a quasi-geostrophic coupled model. J Adv Model Earth Syst 11(6):1803–1829

Penny SG (2017) Mathematical foundations of hybrid data assimilation from a synchronization perspective. Chaos Interdiscip J Nonlinear Sci 27(12):126801

Penny SG, Hamill TM (2017) Coupled data assimilation for integrated earth system analysis and prediction. Bull Am Meteorol Soc 98(7):ES169–ES172

Pikovsky A, Politi A (2016) Lyapunov exponents: a tool to explore complex dynamics. Cambridge University Press

Poincaré H (1899) Les méthodes nouvelles de la mécanique céleste. Tome III, GAUTHIER-VILLARS

Potthast R, Walter A, Rhodin A (2019) A localized adaptive particle filter within an operational NWP framework. Mon Weather Rev 147(1):345–362
Raanes PN, Carrassi A, Bertino L (2015) Extending the square root method to account for additive forecast noise in ensemble methods. Mon Weather Rev 143(10):3857–3873
Raanes PN, Bocquet M, Carrassi A (2019) Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. Q J R Meteorol Soc 145(718):53–75
Reinhold BB, Pierrehumbert RT (1982) Dynamics of weather regimes: quasi-stationary waves and blocking. Mon Weather Rev 110(9):1105–1145
Ruelle D (1979) Ergodic theory of differentiable dynamical systems. Inst Hautes Études Sci Publ Math 50(50):27–58. ISSN 0073-8301
Sakov P, Bertino L (2011) Relation between two common localisation methods for the ENKF. Comput Geosci 15(2):225–237
Sakov P, Oke PR (2008a) A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. Tellus A 60(2):361–371
Sakov P, Oke PR (2008b) Implications of the form of the ensemble transformation in the ensemble square root filters. Mon Weather Rev 136(3):1042–1053
Sakov P, Haussaire JM, Bocquet M (2018) An iterative ensemble Kalman filter in presence of additive model error. Quart J R Meteorol Soc
Shimada I, Nagashima T (1979) A numerical approach to ergodic problem of dissipative dynamical systems. Progress Theoret Phys 61(6):1605–1616
Snyder C (1996) Summary of an informal workshop on adaptive observations and fastex. Bull Am Meteor Soc 77(5):953–961
Snyder C, Bengtsson T, Bickel P, Anderson J (2008) Obstacles to high-dimensional particle filtering. Mon Weather Rev 136(12):4629–4640
Snyder C, Bengtsson T, Morzfeld M (2015) Performance bounds for particle filters using the optimal proposal. Mon Weather Rev 143(11):4750–4761
Szunyogh I, Toth Z, Zimin AV, Majumdar SJ, Persson A (2002) Propagation of the effect of targeted observations: the 2000 winter storm reconnaissance program. Mon Weather Rev 130(5):1144–1165
Thompson PD (1957) Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. Tellus 9(3):275–295. https://doi.org/10.1111/j.2153-3490.1957.tb01885.x
Tondeur M, Carrassi A, Vannitsem S, Bocquet M (2020) On temporal scale separation in coupled data assimilation with the ensemble Kalman filter. J Stat Phys 179:1161–1185
Toth Z, Kalnay E (1993) Ensemble forecasting at NMC: the generation of perturbations. Bull Am Meteor Soc 74(12):2317–2330
Toth Z, Kalnay E (1997) Ensemble forecasting at NCEP and the breeding method. Mon Weather Rev 125(12):3297–3319
Trevisan A, Legnani R (1995) Transient error growth and local predictability: a study in the Lorenz system. Tellus A 47(1):103–117
Trevisan A, Palatella L (2011) On the Kalman filter error covariance collapse into the unstable subspace. Nonlinear Process Geophys 18(2):243–250
Trevisan A, Uboldi F (2004a) Assimilation of standard and targeted observations within the unstable subspace of the observation-analysis-forecast cycle. J Atmos Sci 61:103–113
Trevisan A, Uboldi F (2004b) Assimilation of standard and targeted observations within the unstable subspace of the observation-analysis-forecast cycle system. J Atmos Sci 61(1):103–113
Trevisan A, D'Isidoro M, Talagrand O (2010) Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. Q J R Meteorol Soc 136:487–496
Uboldi F, Trevisan A (2006) Detecting unstable structures and controlling error growth by assimilation of standard and adaptive observations in a primitive equation ocean model. Nonlinear Process Geophys 16:67–81
Uboldi F, Trevisan A, Carrassi A (2005) Developing a dynamically based assimilation method for targeted and standard observations. Nonlinear Process Geophys 12(1):149–156. https://doi.org/10.5194/npg-12-149-2005

Vallis GK (2017) Atmospheric and oceanic fluid dynamics: fundamentals and large-scale circulation, 2nd ed. Cambridge University Press. https://doi.org/10.1017/9781107588417

Van Leeuwen PJ, Künsch HR, Nerger L, Potthast R, Reich S (2019) Particle filters for high-dimensional geoscience applications: a review. Q J R Meteorol Soc 145(723):2335–2365

Vannitsem S (2017) Predictability of large-scale atmospheric motions: Lyapunov exponents and error dynamics. Chaos Interdiscip J Nonlinear Sci 27(3):032101

Vannitsem S, Duan W (2020) On the use of near-neutral backward Lyapunov vectors to get reliable ensemble forecasts in coupled ocean-atmosphere systems. Clim Dyn 55:1125–1139. https://doi.org/10.1007/s00382-020-05313-3

Vannitsem S, Lucarini V (2016) Statistical and dynamical properties of covariant Lyapunov vectors in a coupled atmosphere-ocean model-multiscale effects, geometric degeneracy, and error dynamics. J Phys A: Math Theor 49(22)

Vannitsem S, Demaeyer J, De Cruz L, Ghil M (2015) Low-frequency variability and heat transport in a low-order nonlinear coupled ocean-atmosphere model. Phys D: Nonlinear Phenom 309:71–85. ISSN 0167-2789

Multifidelity Data Assimilation for Physical Systems

Andrey A. Popov and Adrian Sandu

Abstract Multifidelity methods aim to leverage the availability of models at different levels of fidelity describing the same physical phenomena and are receiving growing attention in computational science. One field that can considerably benefits from statistical multifidelity approaches is data assimilation. This chapter presents a broad overview of multifidelity methods in data assimilation for hierarchies of models and hierarchies of observations. We introduce the theoretical multifidelity Kalman filter, and discuss its practical implementation using an ensemble-based framework as the multifidelity ensemble Kalman filter (MFEnKF). The discussion builds upon the theory of linear and nonlinear control variates. Numerical examples compare the multifidelity and the traditional EnKF.

1 Introduction

An often ignored principle in Bayesian inference is that the inference requires the utilization of all available knowledge and all the relevant information available (Jaynes 2003). In the context of data assimilation, especially for physical systems, one often has access to hierarchies of multiple models, each one more accurate than its predecessor in the hierarchy; higher resolution models can be obtained by simply refining the simulation grid or step-size, or through the ability to more accurately capture the underlying physical phenomena. In addition, one frequently has access to observations of the same physical variable through different types of sensors. The principle of Bayesian inference asks to not discard this information, but to incorporate it whenever possible.

Multifidelity data assimilation refers to methods that merge information about the same underlying natural truth obtained through the use of multiple models or obser-

A. A. Popov (✉) · A. Sandu
Virginia Tech, Blacksburg, VA, USA
e-mail: apopov@vt.edu

A. Sandu
e-mail: sandu@cs.vt.edu

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_2

vation operators at different levels of fidelity. For a survey of general multifidelity methods and types of multifidelity models see Peherstorfer et al. (2018).

First introduced by Giles (2008) and then later more formally defined in Giles (2015) the 'multilevel' Monte Carlo filter describes the optimal linear coupling between different 'levels' of simulations (which we generalize to calling fidelities). Here we aim to generalize the notion of levels to fidelities. We introduce a formal notion of structure in the relation between fidelities, and also within the structure that optimally combines the information contained therein.

This chapter is organized as follows. The rest of the introduction is concerned with describing the data assimilation problem, and the idea of multifidelity models. Control variate theory is introduced in Sect. 2. The problem of multifidelity inference is introduced in Sect. 3, with the multifidelity Kalman filter discussed in Sect. 3.1 and the multifidelity ensemble Kalman filter in Sect. 3.2. We then introduce the concept of multifidelity observations in Sect. 4. A trivial numerical example with the Lorenz '96 system is shown in Sect. 5. We conclude with some remarks in Sect. 6.

1.1 Notation

Consider a random variable χ. The distribution of χ is denoted by π_χ, and an ensemble representing N samples from the distribution by $\mathbf{E}_\chi = [\boldsymbol{\chi}^{(1)}, \ldots, \boldsymbol{\chi}^{(N)}]$. We (exact) denote the mean by $\boldsymbol{\mu}_\chi$, and the empirical sample mean by $\widetilde{\boldsymbol{\mu}}_\chi$. The covariance between random variables χ and υ is denoted by $\Sigma_{\chi,\upsilon}$, and the empirical sample covariance by $\widetilde{\Sigma}_{\chi,\upsilon}$.

1.2 The Data Assimilation Problem

We seek to model the state X of a dynamical system with an imperfect model,

$$X_i = \mathcal{M}_i(X_{i-1}) + \Xi_i, \tag{1}$$

where the model errors at different times Ξ_i are independent of each other. We assume the errors have mean zero, $\boldsymbol{\mu}_{\Xi_i} = 0$, and covariances Σ_{Ξ_i,Ξ_i}.

Observations of the true state X_i^t are available at discrete time moments i:

$$Y_i = \mathcal{H}_i(X_i^t) + \eta_i, \tag{2}$$

where we again assume that the observation errors η_i at different times are independent of each other, have mean zero, $\boldsymbol{\mu}_{\eta_i} = 0$, and covariances Σ_{Y_i,Y_i}.

Given prior information X_i^b about the state at time i, and noisy observations of the truth, the filtering problem consists of sequentially computing the posterior, X_i^a in some (usually Bayesian) inference sense.

Explicitly, the Bayesian formulation (Reich and Cotter 2015) aims to find

$$\pi(X_i^a) = \pi(X_i^b|Y_i) = \frac{\pi(Y_i|X_i^b)\pi(X_i^b)}{\pi(Y_i)}, \tag{3}$$

in some approximate way, as the problem, more often than not, is computationally intractable.

1.3 Multifidelity Models

The focus of multifidelity data assimilation is to compute the inference (3) using not a single model (1), but leveraging a hierarchy of models at different fidelities. Consider two levels of fidelity, and assume that our high fidelity state variable is X and coarse fidelity variable is U. The two models that propagate these quantities in time are:

$$X_i = \mathcal{M}_i^X(X_{i-1}), \tag{4}$$
$$U_i = \mathcal{M}_i^U(U_{i-1}). \tag{5}$$

The goal of multifidelity data assimilation is to make use of these different models to incorporate as much information as possible.

An important aspect of multifidelity models, which is a generalization of multilevel hierarchies, is that the the state spaces of the different models do not necessarily have to be the same. In fact, we will assume that the fine fidelity model state can be embedded into $\mathbb{R}^n$ and that the coarse fidelity model state can be embedded into $\mathbb{R}^r$, where typically $r < n$, though this is not necessarily the case.

Alternatively, we can think of the word 'model' as describing the operator about which we are optimizing. In a data assimilation context this will often be our observations. Assume that there exist two ways of obtaining observations (2) of the same fundamental phenomenon, one defined by a fine fidelity operator $\mathcal{H}^\chi$, and the other defined by a coarse fidelity observation operator, $\mathcal{H}^\upsilon$, such that,

$$Y^\chi = \mathcal{H}^\chi(\mathbf{X}^t) + \eta^\chi, \tag{6}$$
$$Y^\upsilon = \mathcal{H}^\upsilon(\mathbf{X}^t) + \eta^\upsilon, \tag{7}$$

wherein the goal would shift to either combining and utilizing the observations in some optimal way without loss of information, but also without duplication of information.

2 Control Variates

The linear control variate technique (Rubinstein and Marcus 1985) is a method for reducing the variance of an estimator by making use of highly correlated data about which additional information is known. Assume our quantity of interest is described by the distribution of the *principal variate* χ. The distribution of the highly correlated *control variate* $\hat{\upsilon}$ describes information in an alternate way (such as in a different space), and the distribution of an independent (or more weaker, uncorrelated) *ancillary variate* υ describes information related to that of the control variate and shares the same mean, $\boldsymbol{\mu}_{\hat{\upsilon}} = \boldsymbol{\mu}_{\upsilon}$. The linear control variate approach builds a *total variate* ζ

$$\zeta = \chi - \mathbf{S}(\hat{\upsilon} - \upsilon), \tag{8}$$

where the free parameter $\mathbf{S}$, known as the *gain operator*, is chosen to minimize the generalized variance of ζ. The three variates that make up the total variate will be collectively called the *constituent variates*.

Theorem 1 (Unbiased nature of linear control variates) *Without proof, the mean of the total variate equals the mean of the principal variate,*

$$\boldsymbol{\mu}_{\zeta} = \boldsymbol{\mu}_{\chi}. \tag{9}$$

Theorem 2 (Optimal gain for linear control variates) *The optimal gain matrix* $\mathbf{S}$ *that minimizes the trace of the covariance of* (8) *is*

$$\mathbf{S} = \Sigma_{\chi,\hat{\upsilon}}\left(\Sigma_{\hat{\upsilon},\hat{\upsilon}} + \Sigma_{\upsilon,\upsilon}\right)^{-1}. \tag{10}$$

Proof Observe that the covariance of (8) is

$$\Sigma_{\zeta,\zeta} = \Sigma_{\chi,\chi} - \mathbf{S}\Sigma_{\hat{\upsilon},\chi} - \Sigma_{\chi,\hat{\upsilon}}\mathbf{S}^T + \mathbf{S}\Sigma_{\hat{\upsilon},\hat{\upsilon}}\mathbf{S}^T + \mathbf{S}\Sigma_{\upsilon,\upsilon}\mathbf{S}^T. \tag{11}$$

Taking the derivative with respect to $\mathbf{S}$ of the trace of (11),

$$\frac{\partial}{\partial \mathbf{S}}\mathrm{tr}\left(\Sigma_{\zeta,\zeta}\right) = -2\Sigma_{\chi,\hat{\upsilon}} + 2\mathbf{S}\left(\Sigma_{\hat{\upsilon},\hat{\upsilon}} + \Sigma_{\upsilon,\upsilon}\right), \tag{12}$$

the local minimum is found at (10), as required.

Corollary 1 *By simple manipulation, the covariance of the total variate under the optimal gain from Theorem 2 is:*

$$\Sigma_{\zeta,\zeta} = \Sigma_{\chi,\chi} - \Sigma_{\chi,\hat{\upsilon}}\left(\Sigma_{\hat{\upsilon},\hat{\upsilon}} + \Sigma_{\upsilon,\upsilon}\right)^{-1}\Sigma_{\hat{\upsilon},\chi}, \tag{13}$$

from which it is clear that $\Sigma_{\zeta,\zeta} \leq \Sigma_{\chi,\chi}$ *in the symmetric semi-positive definite sense.*

The linear control variate technique can be derived in a parallel but completely alternate way. Taking the principal, control, and ancillary variates as Gaussian random variables, the total variate is the solution to the Bayesian inference problem,

$$\pi(\zeta) = \pi(\chi|\hat{\upsilon}, \upsilon). \tag{14}$$

which is a well-known result due to Kalman (1960).

Following Nelson (1987) we now attempt to introduce the idea of non-linear control variates. Instead of searching for function approximations that follow a set of rules, we will instead view the problem of finding the total variate ζ in terms of the principal, control, and ancillary variates as an inference problem, generalizing (14) to arbitrary distributed variables. Specifically, we seek to cast the general inference problem (and specific approximations thereof) into an application of some problem-specific transform,

$$\zeta = \mathcal{T}(\chi, \hat{\upsilon}, \upsilon), \tag{15}$$

with the function $\mathcal{T}$ represents a distribution transformation on the principal variate, built making use of the information given by the control and ancillary variates.

For the remainder of this chapter we will assume that the control variate is related to the principal variate through a deterministic function (coupling),

$$\hat{\upsilon} = \theta(\chi), \tag{16}$$

which implies that there is necessarily some loss of information from the space of the total and principal variates to the space of the control and ancillary variates.

An important generalization of the control variate concept is its ability to be applied in a nested form. This means that the total variate ζ can itself be an ancillary variate for a finer fidelity principal variate. Assume that we have to have $\mathcal{F}$ fidelities, with $\upsilon_{\mathcal{F}}$ being the coarsest fidelity ancillary variate. Its corresponding control variate is $\hat{\upsilon}_{\mathcal{F}}$, and its principal and total variates are on level $\mathcal{F}-1$: $\chi_{\mathcal{F}-1}$ and $\zeta_{\mathcal{F}-1}$. The total variate is then also the ancillary variate for the next set,

$$\upsilon_{\mathcal{F}-1} \leftarrow \zeta_{\mathcal{F}-1}, \tag{17}$$

which can be generalized all the way up the chain, until we reach the the constituent variates χ_1, $\hat{\upsilon}_2$, and υ_2, that represent the full information content through the total variate ζ_1. Explicitly, from Popov et al. (2020), the total variate for $\mathcal{F}$ fidelities and the corresponding optimal gain matrices can be written as,

$$\zeta = \chi - \sum_{f=1}^{\mathcal{F}} \overline{\mathbf{S}}_f \left(\hat{\upsilon}_f - \upsilon_f\right), \quad \overline{\mathbf{S}}_f = \prod_{\psi=1}^{f} \mathbf{S}_\psi. \tag{18}$$

2.1 Ensemble Control Variates

Instead of employing the exact distribution of a random variable, which usually is considered to be an intractable task, an ensemble of samples is typically used.

We will discuss the ways in which ensemble multifidelity inference is performed in a later section. Here we concern ourselves with the problem of finding an ensemble representation of the total variate (8) given an ensemble of N_χ samples of the principal variate $\mathbf{E}_\chi = [\boldsymbol{\chi}^{(1)}, \dots \boldsymbol{\chi}^{(N_\chi)}]$ and corresponding pairwise samples of the control variate $\mathbf{E}_{\hat{\upsilon}} = [\hat{\boldsymbol{\upsilon}}^{(1)}, \dots \hat{\boldsymbol{\upsilon}}^{(N_\chi)}]$. We seek to find an ensemble of the total variate $\mathbf{E}_\zeta = [\boldsymbol{\zeta}^{(1)}, \dots \boldsymbol{\zeta}^{(N_\chi)}]$.

We will define the ensemble means as

$$\widetilde{\boldsymbol{\mu}}_\chi = \frac{1}{N_\chi}\mathbf{E}_\chi \mathbf{1}_{N_\chi} \in \Re^n, \quad \widetilde{\boldsymbol{\mu}}_{\hat{\upsilon}} = \frac{1}{N_\chi}\mathbf{E}_{\hat{\upsilon}} \mathbf{1}_{N_\chi} \in \Re^r, \tag{19}$$

and the anomalies as,

$$\mathbf{A}_\chi = \mathbf{E}_\chi - \widetilde{\boldsymbol{\mu}}_\chi \mathbf{1}_{N_\chi}^T \in \Re^{n\times N_\chi}, \quad \mathbf{A}_{\hat{\upsilon}} = \mathbf{E}_{\hat{\upsilon}} - \widetilde{\boldsymbol{\mu}}_{\hat{\upsilon}} \mathbf{1}_{N_\chi}^T \in \Re^{r\times N_\chi}. \tag{20}$$

Assume that we are given either the mean and covariance of the ancillary variate ($\boldsymbol{\mu}_\upsilon$ and $\Sigma_{\upsilon,\upsilon}$), or that we are able to derive empirical approximations $\widetilde{\boldsymbol{\mu}}_\upsilon$ and $\widetilde{\Sigma}_{\upsilon,\upsilon}$ an ensemble of N_υ samples of υ, $\mathbf{E}_\upsilon$. In the first approach we utilize the linear control variate framework (8).

There are numerous ways in which to derive the ensemble of the total variate. One way is to create a synthetic ensemble of N_χ samples of the ancillary variate sampled from its known distribution. Denote this ensemble $\widetilde{\mathbf{E}}_\upsilon$. Under the linear control variate approach,

$$\mathbf{E}_\zeta = \mathbf{E}_\chi - \mathbf{S}(\mathbf{E}_{\hat{\upsilon}} - \widetilde{\mathbf{E}}_\upsilon), \tag{21}$$

where the optimal gain is approximated by

$$\widetilde{\mathbf{S}} = \widetilde{\Sigma}_{\chi,\hat{\upsilon}}\left(\widetilde{\Sigma}_{\hat{\upsilon},\hat{\upsilon}} + \Sigma_{\upsilon,\upsilon}\right)^{-1}. \tag{22}$$

The astute reader will recognize this as the 'perturbed observations' ensemble Kalman filter (Houtekamer and Mitchell 1998).

An alternate formulation that does away with the synthetic ensemble assumes a Gaussian prior on the ancillary variate, and uses the optimal empirical gain,

$$\mathbf{E}_\zeta = \widetilde{\boldsymbol{\mu}}_\chi \mathbf{1}_{N_\chi}^T + \mathbf{A}_\chi \mathbf{T}\left(\mathbf{I}_{N_\chi} - \frac{1}{N_\chi - 1}\mathbf{T}^T \mathbf{A}_{\hat{\upsilon}}^T \Sigma_{\upsilon,\upsilon}^{-1}(\widetilde{\boldsymbol{\mu}}_{\hat{\upsilon}} - \boldsymbol{\mu}_\upsilon)\mathbf{1}_{N_\chi}^T\right), \tag{23}$$

$$\mathbf{T} = \left(\mathbf{I}_{N_\chi} - \frac{1}{N_\chi - 1}\mathbf{A}_{\hat{\upsilon}}^T\left(\widetilde{\Sigma}_{\hat{\upsilon},\hat{\upsilon}} + \Sigma_{\upsilon,\upsilon}\right)^{-1}\mathbf{A}_{\hat{\upsilon}}\right)^{\frac{1}{2}}, \tag{24}$$

which the astute reader will recognize as the ensemble transform Kalman filter (ETKF) (Bishop et al. 2001).

Note that if it is not possible to represent the covariance of the ancillary variate exactly, then one needs to compute the optimal gain in alternate ways.

Another interesting approach to ensemble inference is the importance sampling optimal transport procedure (Reich 2013). In essence, one constructs the posterior mean from the importance sampling procedure,

$$\boldsymbol{\mu}_\zeta = \mathbf{E}_\chi \mathbf{w}, \tag{25}$$

$$[\mathbf{w}]_i \propto \pi_\upsilon(\hat{\boldsymbol{\upsilon}}^{(i)}). \tag{26}$$

The anzatz is made that the optimal transportation into an equally weighted ensemble with the same mean defines an ensemble with the same empirical moments as those defined by the importance sampling weights,

$$\mathbf{E}_\zeta = \mathbf{E}_\chi \mathbf{T}^*, \tag{27}$$

where the optimal transport is defined in the Monge-Kantorovich sense,

$$\begin{aligned} \mathbf{T}^* = \arg\min_{\mathbf{T}} & \sum_{1 \le i,j \le N_\chi} [\mathbf{T}]_{ij} \left\| \boldsymbol{\chi}^{(i)} - \boldsymbol{\chi}^{(j)} \right\|_2^2 \\ \text{subject to: } & \mathbf{T}\mathbf{1}_{N_\chi} = N_\chi \mathbf{w},\ \mathbf{T}^T \mathbf{1}_{N_\chi} = \mathbf{1}_{N_\chi},\ [\mathbf{T}]_{ij} \ge 0, \end{aligned} \tag{28}$$

which ensures that the weights of the new posterior ensemble are equal.

Second order accurate (preserving the weighted ensemble covariance) extensions to this formulation exist (Acevedo et al. 2017) and should be used if this methodology is to be attempted operationally.

3 Multifidelity Filtering

For ease of exposition we primarily focus on the case of two fidelities; multifidelity extensions will be described separately.

Assume now that the state of a dynamical system is our quantity of interest, and that there are two different fidelities in which we can represent it: fine and coarse. Let the distribution of the principal variate X^b represent the prior information about the state at fine fidelity. Let $\hat{U}^b$ be its corresponding control variate, and U^b be the ancillary variate, the distributions of which describe information about the state at coarse fidelity.

Assume that the prior total variate Z^b represents the general posterior of the multifidelity inference procedure (14). Note that it is possible, with some abuse of notation, to represent the inference as an application of some nonlinear function,

$$Z^b = \mathcal{C}(X^b, \hat{U}^b, U^b), \tag{29}$$

with the function $\mathcal{C}$ defining an implicit assumption about the relationship between the four variates, such as the linear control variate assumption (8) or an optimal transport based assumption (27) if our variates are represented by ensembles. In the most general sense, $\mathcal{C}$ can represent some non-linear variance reduction technique that is informed by the distributions of the constituent variates (Nelson 1987).

The prior total variate is Z^b and the posterior total variate is Z^a, defined by the same function applied to its component variates:

$$Z^a = \mathcal{C}(X^a, \hat{U}^a, U^a). \tag{30}$$

The inference step from the prior total to the posterior total variates is a filtering step which explicitly combines information,

$$Z^a = \mathcal{F}(Z^b, \mathcal{H}(Z^b), Y), \tag{31}$$

with the function $\mathcal{F}$ standing in for some filter, such as the Kalman filter.

The principal variate can be propagated by some constituent filter,

$$X^a = \widetilde{\mathcal{F}}^Z_X(X^b, \mathcal{H}(X^b), Y), \tag{32}$$

which is dependent on Z^b, Z^a and the filter $\mathcal{F}$ that is implicitly applied between them. Similar formulations can be made for the other constituent variates.

Note that the goal of one step of a multifidelity filter is not to find the posterior total variate Z^a, but rather to find posteriors of its constituent variates, X^a, $\hat{U}^a$, and U^a. In fact, as the total variate is merely a synthetic construction, the multifidelity inference reduces to performing virtual inference on the total variate by manipulating the principal, control, and ancillary variates. In this way the explicit filtering of the total variate (31) is not performed, but only the constituent filtering problems (32) are explicitly solved.

While the general problem of finding the analysis principal variate X^a given only the analysis total variate Z^a is not well posed, the combined problem of finding the distributions of X^a, $\hat{U}^a$, and U^a may be posed in terms of a minimum cross entropy problem:

$$X^a, \hat{U}^a, U^a = \arg\min_{X^a, \hat{U}^a, U^a} H(X^b, X^a) + H(\hat{U}^b, \hat{U}^a) + H(U^b, U^a), \tag{33}$$

subject to the constraints,

$$Z^a = \mathcal{C}(X^a, \hat{U}^a, U^a), \quad Z^a = \mathcal{F}(Z^b, \mathcal{H}(Z^b), Y), \tag{34}$$

$$\mathbb{E}[\hat{U}^a] = \mathbb{E}[U^a], \qquad \hat{U}^a = \theta(X^a), \tag{35}$$

from which the constituent filters (32) are implicitly defined.

A powerful assumption that can be made is that the same control structure imposed on the prior is also imposed on the posterior. We call this the 'control structure consistency assumption'. One way in which this holds in the linear control variate approach is:

$$Z^b = X^b - \mathbf{S}^b(\hat{U}^b - U^b), \tag{36}$$

$$Z^a = X^a - \mathbf{S}^a(\hat{U}^a - U^a), \tag{37}$$

where the we impose the assumption that the (approximately) optimal prior and posterior gains are equivalent,

$$\mathbf{S}^b = \mathbf{S}^a, \tag{38}$$

meaning that we restrict all possible posterior constituent variates to ones that obey the same structure as their prior counterparts. One way in which this is achieved is by assuming a particular structure on the relationship between the principal and control variate (16) from Popov et al. (2020).

3.1 Multifidelity Kalman Filter

We now introduce the multifidelity Kalman filter (MFKF), fleshed out from Popov et al. (2020). As Gaussian random variables can be trivially combined through known formulas involving their means and covariances, the MFKF is not an algorithm that needs to exist for the purposes of practical implementation, but merely needs to exist to explain derivations of practical extensions thereof.

We restrict ourselves to a linear principal-control variate coupling (16),

$$\hat{U}_i = \theta(X_i) = \mathbf{\Theta} X_i, \tag{39}$$

with $\mathbf{\Theta}$ a projection operator from the n-dimensional space of the principal variate onto the r-dimensional space of the control variate. The corresponding interpolating

operator is denoted Φ (such that $\mathbf{\Theta}\,\mathbf{\Phi} = \mathbf{I}_r$). We decompose the principal variate into its control variate and residual variate components:

$$X_i = \mathbf{\Theta}\,\hat{U}_i + \Delta X_i. \tag{40}$$

Additionally, as is canonical, we restrict ourselves to the case of a linear observation operator $\mathbf{H}_i$.

For the rest of this chapter we assume that we seek to propagate the total variate,

$$Z_i = X_i - \mathbf{S}_i\,(\hat{U}_i - U_i), \tag{41}$$

through both a dynamical model (forecast step), and through the analysis step conditioned by observations.

We express the moments of the total variate in terms of the moments of the corresponding constituents:

$$\boldsymbol{\mu}_{Z_i} = \boldsymbol{\mu}_{X_i} - \mathbf{S}_i(\boldsymbol{\mu}_{\hat{U}_i} - \boldsymbol{\mu}_{U_i}), \tag{42}$$

$$\Sigma_{Z_i,Z_i} = \Sigma_{X_i,X_i} + \mathbf{S}_i\,\Sigma_{\hat{U}_i,\hat{U}_i}\mathbf{S}_i^T + \mathbf{S}_i\,\Sigma_{U_i,U_i}\mathbf{S}_i^T - \mathbf{S}_i\,\Sigma_{\hat{U}_i,X_i} - \Sigma_{X_i,\hat{U}_i}\mathbf{S}_i^T. \tag{43}$$

We are now ready to look at the MFKF. For the forecast step, assume that we have a linear fine fidelity model $\mathbf{M}_i^X$, and a linear coarse fidelity model $\mathbf{M}_i^U$. Assume that the error Ξ_i of the fine fidelity model is unbiased and is known to have covariance Σ_{Ξ_i,Ξ_i}. Assume additionally that the coarse fidelity model has no error in the coarse subspace in relation to the truth. This could be because the coarse fidelity model was built to capture this error through data driven closures.

Assume that we have the posterior information at the previous step $i-1$ about the principal, control, and ancillary variates, and that the relation between the principal and control variate (39) holds. We propagate the means as follows:

$$\boldsymbol{\mu}_{X_i^b} = \mathbf{M}_i^X\boldsymbol{\mu}_{X_{i-1}^a}, \tag{44}$$

$$\boldsymbol{\mu}_{\hat{U}_i^b} = \mathbf{M}_i^U\boldsymbol{\mu}_{\hat{U}_{i-1}^a}, \tag{45}$$

$$\boldsymbol{\mu}_{U_i^b} = \mathbf{M}_i^U\boldsymbol{\mu}_{U_{i-1}^a}, \tag{46}$$

with the covariances propagated as,

$$\Sigma_{X_i^b,X_i^b} = \mathbf{M}_i^X\,\Sigma_{X_{i-1}^a,X_{i-1}^a}\mathbf{M}_i^{X,T} + \Sigma_{\Xi_i,\Xi_i}, \tag{47}$$

$$\Sigma_{X_i^b,\hat{U}_i^b} = \mathbf{M}_i^X\,\Sigma_{X_{i-1}^a,\hat{U}_{i-1}^a}\mathbf{M}_i^{U,T}, \tag{48}$$

$$\Sigma_{\hat{U}_i^b,\hat{U}_i^b} = \mathbf{M}_i^U\,\Sigma_{\hat{U}_{i-1}^a,\hat{U}_{i-1}^a}\mathbf{M}_i^{U,T}, \tag{49}$$

$$\Sigma_{U_i^b,U_i^b} = \mathbf{M}_i^U\,\Sigma_{U_{i-1}^a,U_{i-1}^a}\mathbf{M}_i^{U,T}, \tag{50}$$

We note that unless the principal variate residual is not propagated in control space by the fine fidelity model,

$$\boldsymbol{\Theta}\mathbf{M}_i^X \Delta X_{i-1}^a = 0, \tag{51}$$

then the above propagation will violate (39). Therefore, as a useful heuristic, the propagation of the control variate moments can be replaced by the propagation of the projected principal variate moments in order for the relation (39) to hold more strongly at each step,

$$\boldsymbol{\mu}_{\hat{U}_i^b} = \mathbf{M}_i^U \boldsymbol{\Theta} \boldsymbol{\mu}_{X_{i-1}^a}, \tag{52}$$

$$\Sigma_{X_i^b, \hat{U}_i^b} = \mathbf{M}_i^X \Sigma_{X_{i-1}^a, X_{i-1}^a} \left(\mathbf{M}_i^U \boldsymbol{\Theta}\right)^T, \tag{53}$$

$$\Sigma_{\hat{U}_i^b, \hat{U}_i^b} = \mathbf{M}_i^U \boldsymbol{\Theta} \Sigma_{X_{i-1}^a, X_{i-1}^a} \left(\mathbf{M}_i^U \boldsymbol{\Theta}\right)^T. \tag{54}$$

This is especially useful if the models are non-linear, generalizing to the multifidelity extended Kalman filter, or in the case of the multifidelity ensemble Kalman filter later in the chapter in Sect. 3.2.

Lemma 1 *The fine fidelity model, coarse fidelity model, posterior optimal gain at step $i-1$, and prior optimal gain at time i are related as follows:*

$$\mathbf{M}_i^X \mathbf{S}_{i-1}^a = \mathbf{S}_i^b \mathbf{M}_i^U. \tag{55}$$

Proof By simple manipulation,

$$\mathbf{M}_i^X \mathbf{S}_{i-1}^a = \mathbf{M}_i^X \Sigma_{X_{i-1}^a, \hat{U}_{i-1}^a} \left(\Sigma_{\hat{U}_{i-1}^a, \hat{U}_{i-1}^a} + \Sigma_{U_{i-1}^a, U_{i-1}^a}\right)^{-1} \tag{56}$$

$$= \Sigma_{X_i^b, \hat{U}_{i-1}^a} \mathbf{M}_i^{U,T} \left[\mathbf{M}_i^U \left(\Sigma_{\hat{U}_{i-1}^a, \hat{U}_{i-1}^a} + \Sigma_{U_{i-1}^a, U_{i-1}^a}\right) \mathbf{M}_i^{U,T}\right]^{-1} \mathbf{M}_i^U \tag{57}$$

$$= \Sigma_{X_i^b, \hat{U}_i^b} \left(\Sigma_{\hat{U}_i^b, \hat{U}_i^b} + \Sigma_{U_i^b, U_i^b}\right)^{-1} \mathbf{M}_i^U = \mathbf{S}_i^b \mathbf{M}_i^U, \tag{58}$$

as required.

Theorem 3 *The MFKF forecast is the total variate forecast:*

$$\boldsymbol{\mu}_{Z_i^b} = \mathbf{M}_i^X \boldsymbol{\mu}_{Z_{i-1}^a}, \tag{59}$$

$$\Sigma_{Z_i^b, Z_i^b} = \mathbf{M}_i^X \Sigma_{Z_{i-1}^a, Z_{i-1}^a} \mathbf{M}_i^{X,T} + \Sigma_{\Xi_i, \Xi_i}. \tag{60}$$

Proof Using lemma 1, manipulate the formulation for the mean of Z in (42),

$$\begin{aligned}
\mathbf{M}_i^X \boldsymbol{\mu}_{Z_{i-1}^a} &= \mathbf{M}_i^X \boldsymbol{\mu}_{Z_{i-1}^a}, && (61)\\
&= \mathbf{M}_i^X \boldsymbol{\mu}_{X_{i-1}^a} - \mathbf{M}_i^X \mathbf{S}_{i-1}^a (\boldsymbol{\mu}_{\hat{U}_{i-1}^a} - \boldsymbol{\mu}_{U_{i-1}^a}), && (62)\\
&= \boldsymbol{\mu}_{X_i^b} - \mathbf{S}_i^b \mathbf{M}_i^U (\boldsymbol{\mu}_{\hat{U}_{i-1}^a} - \boldsymbol{\mu}_{U_{i-1}^a}), && (63)\\
&= \boldsymbol{\mu}_{X_i^b} - \mathbf{S}_i^b (\boldsymbol{\mu}_{\hat{U}_i^b} - \boldsymbol{\mu}_{U_i^b}) = \boldsymbol{\mu}_{Z_i^b}, && (64)
\end{aligned}$$

as required. A similar manipulation can be performed for the covariance.

In order to obtain an efficient implementation of the analysis step in the MFKF, we need to restrict the projection operator (39) to a class that has 'nice' properties. We assume that the joint variability of the principal variate in the orthogonal complement space and control variate is negligible,

$$\Sigma_{\Delta X, \hat{U}} \approx 0, \tag{65}$$

or alternatively that the projection operator $\boldsymbol{\Theta}$ captures the dominant linear modes of the variability in the dynamics of X. Common methods by which such operators can be obtained are POD and DMD, and variants thereof (Brunton and Kutz 2019).

Theorem 4 *If the first two moments of the control and ancillary variate are identical, and assumption* (65) *holds, then the optimal gain is,*

$$\mathbf{S} = \frac{1}{2}\boldsymbol{\Theta}. \tag{66}$$

Proof Observe by Theorem 2 and (65),

$$\mathbf{S} = \Sigma_{X,\hat{U}} \left(\Sigma_{\hat{U},\hat{U}} + \Sigma_{U,U}\right)^{-1} = \frac{1}{2}\boldsymbol{\Theta}\boldsymbol{\Sigma}_{\hat{U},\hat{U}} \Sigma_{\hat{U},\hat{U}}^{-1} = \frac{1}{2}\boldsymbol{\Theta}, \tag{67}$$

as required.

If we choose a projection operator for which (65) holds, then the optimal gain is constant and does not have to be estimated. Moreover this provides for a clear relationship between the projection operator $\boldsymbol{\Theta}$ and the optimal gain, such that $\boldsymbol{\Theta}\mathbf{S} = \frac{1}{2}\mathbf{I}_r$. For the rest of this section we assume that $\mathbf{S}$ is constant.

We next discuss the analysis step of the MFKF. Note first that the Kalman gain is the optimal gain when the principal variate is the prior information about the state of the dynamics, the control variate is that information cast into observation space, and the ancillary variate are the independent observations. Assume that the arbitrary variate W_i^b represents some prior information, we write the Kalman gain as a function of W_i^b,

$$\mathbf{K}_{W_i^b} = \Sigma_{W_i^b, \mathbf{H}_i W_i^b} \left(\Sigma_{\mathbf{H}_i W_i^b, \mathbf{H}_i W_i^b} + \Sigma_{Y_i, Y_i}\right)^{-1}. \tag{68}$$

The standard Kalman filter analysis step applied to the total variate, as described by (31):

$$Z_i^a(\mathbf{K}_{Z_i^b}) = Z_i^b - \mathbf{K}_{Z_i^b}(\mathbf{H}_i Z_i^b - Y_i), \tag{69}$$

can be decomposed into its constituent variates:

$$X^a - \mathbf{S}(\hat{U}^a - U^a) = X^b - \mathbf{S}(\hat{U}^b - U^b) - \mathbf{K}_{Z_i^b}\left(\mathbf{H}\left(X^b - \mathbf{S}(\hat{U}^b - U^b)\right) - Y\right). \tag{70}$$

Taking the 'natural' decomposition of this relation into components leads to, the constituent filters (32):

$$X^a = X^b - \mathbf{K}_{Z_i^b}\left(\mathbf{H}X^b - Y\right), \tag{71}$$

$$\hat{U}^a = \hat{U}^b - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}\Phi\hat{U}^b - Y\right), \tag{72}$$

$$U^a = U^b - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}\Phi U^b - Y\right), \tag{73}$$

which assumes that the control and ancillary variates do not carry any additional information from the orthogonal complement space of the principal variate.

The authors conjecture that the decomposition (71) approximately minimizes the cross entropy functional (33) out of all such decompositions, though there is no strong evidence for this claim as of yet.

The propagation of the total mean through its constituent variate means is:

$$\boldsymbol{\mu}_{X^a} = \boldsymbol{\mu}_{X^b} - \mathbf{K}_{Z_i^b}(\mathbf{H}\boldsymbol{\mu}_{X^b} - \mathbf{y}) \tag{74}$$

$$\boldsymbol{\mu}_{\hat{U}^a} = \boldsymbol{\mu}_{\hat{U}^b} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}\Phi\boldsymbol{\mu}_{\hat{U}^b} - \mathbf{y}\right), \tag{75}$$

$$\boldsymbol{\mu}_{U^a} = \boldsymbol{\mu}_{U^b} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}\Phi\boldsymbol{\mu}_{U^b} - \mathbf{y}\right). \tag{76}$$

The corresponding covariance update formulas are:

$$\Sigma_{X_i^a, X_i^a} = \left(\mathbf{I} - \mathbf{K}_{Z_i^b}\mathbf{H}\right)\Sigma_{X_i^b, X_i^b}\left(\mathbf{I} - \mathbf{K}_{Z_i^b}\mathbf{H}_i\right)^T + \mathbf{K}_{Z_i^b}\Sigma_{Y_i, Y_i}\mathbf{K}_{Z_i^b}^T, \tag{77}$$

$$\Sigma_{X_i^a, \hat{U}_i^a} = \left(\mathbf{I} - \mathbf{K}_{Z_i^b}\mathbf{H}_i\right)\Sigma_{X_i^b, \hat{U}_i^b}\left(\mathbf{I} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\mathbf{H}_i\Phi\right)^T, \tag{78}$$

$$+\mathbf{K}_{Z_i^b}\Sigma_{Y_i, Y_i}\left(\boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\right)^T, \tag{79}$$

$$\Sigma_{\hat{U}_i^a, \hat{U}_i^a} = \left(\mathbf{I} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\mathbf{H}_i\Phi\right)\Sigma_{\hat{U}_i^b, \hat{U}_i^b}\left(\mathbf{I} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\mathbf{H}_i\Phi\right)^T \tag{80}$$

$$+\boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\Sigma_{Y_i, Y_i}\left(\boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\right)^T, \tag{81}$$

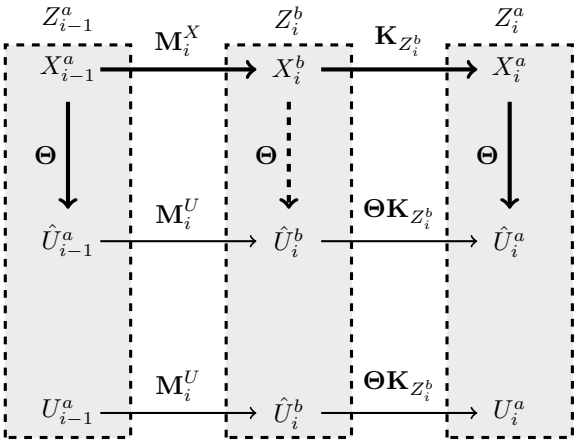

Fig. 1 A diagram of the MFKF computations, including the heuristic re-correlation of the principal and control variates (52), and the assumed relation between the forecast variates (16)

$$\Sigma_{U_i^a,U_i^a} = \left(\mathbf{I} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\mathbf{H}_i\Phi\right)\Sigma_{U_i^b,U_i^b}\left(\mathbf{I} - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\mathbf{H}_i\Phi\right)^T \tag{82}$$

$$+\boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\Sigma_{Y_i,Y_i}\left(\boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\right)^T. \tag{83}$$

The inner working of the MFKF is illustrated in Fig. 1.

Theorem 5 *Without proof, if the optimal gain interpolation projection step does not remove additional information from the control and ancillary variate (that is (65) is exact), then the 'natural' decomposition (71) is exact, thus the linear control variate combination of the mean is the total variate analysis mean,*

$$\boldsymbol{\mu}_{Z_i^a} = \boldsymbol{\mu}_{X_i^a} - \mathbf{S}(\boldsymbol{\mu}_{\hat{U}_i^a} - \boldsymbol{\mu}_{U_i^a}). \tag{84}$$

Similarly for the covariances.

Theorem 6 *Without proof, if* $\mathbf{S}$ *is the optimal gain (Theorems 2 and 4), then the simple relation between the covariances of the principal and total variates is,*

$$\Sigma_{Z_i,Z_i} = (\mathbf{I}_n - \mathbf{S}\boldsymbol{\Theta})\,\Sigma_{X_i,X_i}. \tag{85}$$

Thus we are able to obtain a covariance for the total variate by only knowing the covariance of the principal variate.

An alternate decomposition for which Theorem 5 is exact without qualification, that we will not be analyzing is:

$$X^a = X^b - \mathbf{K}_{Z_i^b}\left(\mathbf{H}Z^b - Y\right), \tag{86}$$

$$\hat{U}^a = \hat{U}^b - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}Z^b - Y\right), \tag{87}$$

$$U^a = U^b - \boldsymbol{\Theta}\mathbf{K}_{Z_i^b}\left(\mathbf{H}Z^b - Y\right), \tag{88}$$

These formulas, however, are difficult to implement using ensembles.

We show next how the variability of the total variate $Z_i^a(\mathbf{K}_{Z_i^b})$, the principal variate $X_i^a(\mathbf{K}_{Z_i^b})$, and the principal variate analyzed by itself $X_i^a(\mathbf{K}_{X_i^b})$ are related.

Theorem 7 *The covariances of $Z_i^a(\mathbf{K}_{Z_i^b})$, $X_i^a(\mathbf{K}_{Z_i^b})$, $X_i^a(\mathbf{K}_{X_i^b})$ are such that:*

$$\Sigma_{Z^a(\mathbf{K}_{Z_i^b}),Z^a(\mathbf{K}_{Z_i^b})} \leq \Sigma_{X^a(\mathbf{K}_{X_i^b}),X^a(\mathbf{K}_{X_i^b})} \leq \Sigma_{X^a(\mathbf{K}_{Z_i^b}),X^a(\mathbf{K}_{Z_i^b})}. \tag{89}$$

Proof By the optimality of the Kalman gain $\mathbf{K}_{Z_i^b}$ in Theorem 2,

$$\Sigma_{Z^a(\mathbf{K}_{Z_i^b}),Z^a(\mathbf{K}_{Z_i^b})} \leq \Sigma_{Z^a(\mathbf{K}_{X_i^b}),Z^a(\mathbf{K}_{X_i^b})}, \tag{90}$$

and by the optimality of the control variate relation $\mathbf{S}$ from Corollary 1,

$$\Sigma_{Z^a(\mathbf{K}_{X_i^b}),Z^a(\mathbf{K}_{X_i^b})} \leq \Sigma_{X^a(\mathbf{K}_{X_i^b}),X^a(\mathbf{K}_{X_i^b})}. \tag{91}$$

The second inequality similarly relies on the optimality of the Kalman gain $\mathbf{K}_{X_i^b}$.

Theorem 7 shows that the principal variate covariance is an upper bound on the covariance of the total variate.

Relations (89) in Theorem 7 are valid only when the means of the constituent variates are roughly equivalent. This is especially important in the 'extended' and 'ensemble' extensions to the MFKF. To achieve this, at each step we apply the following heuristic correction:

$$\boldsymbol{\mu}_{X_i^a} \leftarrow \boldsymbol{\mu}_{Z_i^a}, \quad \boldsymbol{\mu}_{\hat{U}_i^a} \leftarrow \boldsymbol{\Theta}\boldsymbol{\mu}_{Z_i^a}, \quad \boldsymbol{\mu}_{U_i^a} \leftarrow \boldsymbol{\Theta}\boldsymbol{\mu}_{Z_i^a}, \tag{92}$$

which additionally enforces the control variate relation (39), ensures that the principal and total variate means are equivalent, and that the control and ancillary variate means are equivalent.

3.2 Multifidelity Ensemble Kalman Filter

Following Popov et al. (2020), we present the multifidelity ensemble Kalman filter (MFEnKF).

Assume now that instead of manipulating the first two moments of our variates, we manipulate ensembles. Assume that we have N_X pairwise correlated samples of the principal and control variates $(\mathbf{X}^{(i)}, \hat{\mathbf{U}}^{(i)})$ represented by the ensembles $\mathbf{E}_X$ and $\mathbf{E}_{\hat{U}}$, respectively, and $N_U \geq N_X$ samples of the ancillary variate $\mathbf{U}^{(i)}$, represented by the ensemble $\mathbf{E}_U$. We wish to construct practical ensemble-based generalizations to the MFKF.

The forecast step, similar to the standard EnKF, and MFKF ((44) and (47)), propagates the ensemble members individually through their respective models,

$$\mathbf{X}_i^{b,(j)} = \mathcal{M}_i^X(\mathbf{X}_{i-1}^{a,(j)}) + \boldsymbol{\xi}_i^{(j)}, \quad 1 \leq j \leq N_X \tag{93}$$

$$\hat{\mathbf{U}}^{b,(j)} = \mathcal{M}_i^U(\hat{\mathbf{U}}_{i-1}^{a,(j)}), \quad 1 \leq j \leq N_X \tag{94}$$

$$\mathbf{U}^{b,(j)} = \mathcal{M}_i^U(\mathbf{U}_{i-1}^{a,(j)}), \quad 1 \leq j \leq N_U, \tag{95}$$

where each $\boldsymbol{\xi}_i^{(j)}$ is a random sample accounting for the fine fidelity model error. The coarse fidelity model bias is accounted for by the propagation of both the control variate ensemble and ancillary ensemble through the coarse fidelity model.

Assume that the sample means, anomalies, and covariances are readily available for the three constituent ensembles, from which it is possible to derive the empirical estimates of the first two moments of the total variate,

$$\widetilde{\boldsymbol{\mu}}_{Z_i} = \widetilde{\boldsymbol{\mu}}_{X_i} - \mathbf{S}_i(\widetilde{\boldsymbol{\mu}}_{\hat{U}_i} - \widetilde{\boldsymbol{\mu}}_{U_i}), \tag{96}$$

$$\widetilde{\Sigma}_{Z_i,Z_i} = \widetilde{\Sigma}_{X_i,X_i} + \mathbf{S}_i\widetilde{\Sigma}_{\hat{U}_i,\hat{U}_i}\mathbf{S}_i^T + \mathbf{S}_i\widetilde{\Sigma}_{U_i,U_i}\mathbf{S}_i^T - \mathbf{S}_i\widetilde{\Sigma}_{\hat{U}_i,X_i} - \widetilde{\Sigma}_{X_i,\hat{U}_i}\mathbf{S}_i^T, \tag{97}$$

where once again we assume that the optimal gain is constant (66) from Theorem 4.

Similar to the standard EnKF, it is not explicitly required to compute the full total background covariance, but merely the related cross-covariances with respect to the observation operator:

$$\widetilde{\Sigma}_{Z_i^b,\mathbf{H}_i Z_i^b} = \widetilde{\Sigma}_{Z_i^b,Z_i^b}\mathbf{H}_i^T, \tag{98}$$

$$\widetilde{\Sigma}_{\mathbf{H}_i Z_i^b,\mathbf{H}_i Z_i^b} = \mathbf{H}_i\widetilde{\Sigma}_{Z_i^b,Z_i^b}\mathbf{H}_i^T, \tag{99}$$

which can efficiently be computed by utilizing the observation ensemble anomalies. From this the sample Kalman gain can be computed.

Applying the MFKF formulas (71) to the MFEnKF statistics it is possible to gain access to the corresponding empirical Kalman gain,

$$\widetilde{\mathbf{K}}_{Z_i^b} = \widetilde{\Sigma}_{Z_i^b,\mathbf{H}_i Z_i^b}\left(\widetilde{\Sigma}_{\mathbf{H}_i Z_i^b,\mathbf{H}_i Z_i^b} + \Sigma_{Y_i,Y_i}\right)^{-1}, \tag{100}$$

and the corresponding analysis of the anomalies,

$$\mathbf{A}_{X_i^a} = \mathbf{A}_{X_i^b} - \widetilde{\mathbf{K}}_{Z_i^b}(\mathbf{A}_{\mathbf{H}_i X_i^b} - \mathbf{E}_{Y_i}^X), \tag{101}$$

$$\mathbf{A}_{\hat{U}_i^a} = \mathbf{A}_{\hat{U}_i^b} - \boldsymbol{\Theta}\widetilde{\mathbf{K}}_{Z_i^b}(\mathbf{A}_{\mathbf{H}_i \boldsymbol{\Theta} \hat{U}_i^b} - \mathbf{E}_{Y_i}^{\hat{U}}), \tag{102}$$

$$\mathbf{A}_{U_i^a} = \mathbf{A}_{U_i^b} - \boldsymbol{\Theta}\widetilde{\mathbf{K}}_{Z_i^b}(\mathbf{A}_{\mathbf{H}_i \boldsymbol{\Theta} U_i^b} - \mathbf{E}_{Y_i}^{U}), \tag{103}$$

where each $\mathbf{E}_{Y_i}$ is an ensemble of perturbed observations. In Popov et al. (2020) it was shown that there is no unique 'nice' solution to the problem of perturbed observations in the MFEnKF, thus we will leave this discussion aside in this chapter.

In order to get an ensemble of Z^a, we can look towards the ensemble transform Kalman filter, specifically at the 'left transform' variant (Sakov and Bertino 2011). Using known properties of the matrix shift lemma (Asch et al. 2016) and the linearity of the control variate relation (39) one can write the transformation of the ensemble of X^a into anomalies of Z^a given by the ETKF (23) as:

$$\mathbf{E}_{Z^a} = \widetilde{\boldsymbol{\mu}}_{Z^a} \mathbf{1}_{N_X}^T + (\mathbf{I}_n - \mathbf{S}\boldsymbol{\Theta})^{\frac{1}{2}} \mathbf{A}_{X^a}, \tag{104}$$

which can be implemented in any number of computationally efficient ways (Allen et al. 2000) beyond the scope of this chapter.

Note however that this methodology relies on the equality of the covariances of the control and ancillary variates, which especially in the ensemble case will be violated.

We now discuss the main advantage of the MFEnKF: utilizing the same amount of samples of the fine fidelity model as the standard EnKF, the MFEnKF provides a more accurate mean analysis.

Theorem 8 *Assume that we have access to the exact Kalman gains $K_{Z_i^b}$, $K_{X_i^b}$ from (68), of the theoretical Kalman filters. The variance of the empirical mean of the analysis total variate computed with the total variate Kalman gain is less that the variance in the empirical mean in the analysis principal variate computed with the principal variate Kalman gain,*

$$\Sigma_{\widetilde{\boldsymbol{\mu}}_{Z_i^a}(\mathbf{K}_{Z_i^b}), \widetilde{\boldsymbol{\mu}}_{Z_i^a}(\mathbf{K}_{Z_i^b})} \leq \Sigma_{\widetilde{\boldsymbol{\mu}}_{X_i^a}(\mathbf{K}_{X_i^b}), \widetilde{\boldsymbol{\mu}}_{X_i^a}(\mathbf{K}_{X_i^b})}. \tag{105}$$

Proof Assuming again that $N_U \geq N_X$, and by Theorem 7,

$$\begin{aligned}
\Sigma_{\widetilde{\boldsymbol{\mu}}_{Z_i^a(\mathbf{K}_{Z_i^b})},\widetilde{\boldsymbol{\mu}}_{Z_i^a(\mathbf{K}_{Z_i^b})}} &\leq N_X^{-1}\Sigma_{Z_i^a(\mathbf{K}_{Z_i^b}),Z_i^a(\mathbf{K}_{Z_i^b})} && (106)\\
&\quad +(N_U^{-1}-N_X^{-1})\mathbf{S}\Sigma_{U_i^a(\mathbf{K}_{Z_i^b}),U_i^a(\mathbf{K}_{Z_i^b})}\mathbf{S}^T, && (107)\\
&\leq N_X^{-1}\Sigma_{X_i^a(\mathbf{K}_{X_i^b}),X_i^a(\mathbf{K}_{X_i^b})}, && (108)\\
&= \Sigma_{\widetilde{\boldsymbol{\mu}}_{X_i^a(\mathbf{K}_{X_i^b})},\widetilde{\boldsymbol{\mu}}_{X_i^a(\mathbf{K}_{X_i^b})}}, && (109)
\end{aligned}$$

as required.

The perturbed observations MFEnKF is similar to a typical EnKF algorithm in the way in which inflation and B-localization can be applied. An important salient difference is that there is now an additional hyperparameter, namely the inflation factor for the ancillary ensemble α_U. Inflation for the principal and control ensembles α_X should be the same in order to keep them highly correlated. As optimal inflation is known to depend on the ensemble size (Popov and Sandu 2020), it should generally be the case that $\alpha_U < \alpha_X$.

3.3 Other 'Multi-x' Data Assimilation Algorithms

In this section we discuss other data assimilation algorithms in the 'multi-' family that do not, as-of-now, have rigorous multifidelity counterparts.

3.3.1 Particle Filters

In Giles (2008, 2015), Giles discusses 'multilevel' Monte-Carlo simulations. The case of projection and interpolation is ignored, and the optimal gain is explicitly set to be identity. The author examines the component variates as being coupled through their differences, which in a two-level control variate framework is equivalent to examining the variates:

$$\begin{cases} U \\ X - \hat{U} \end{cases}, \tag{110}$$

treating each as an independent source of information, with means:

$$\begin{cases} \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_X - \boldsymbol{\mu}_{\hat{U}} \end{cases}. \tag{111}$$

The implicit assumption is that the only important source of information is the mean-estimate, and that $\boldsymbol{\mu}_U$ carries negligible uncertainty. The intended use of such algorithms is in small-dimensional cases where large ensembles on coarse fidelity models can be created, thus this is not an unreasonable assumption.

In Gregory et al. (2016), Gregory and Cotter (2017), Gregory and co-authors propose 'multilevel' ensemble transform particle filters (ETPF). The authors again employ a linear control variate structure where the optimal gain is assumed to be the identity, and in which all variable operations are performed on the same empirical measures. The authors pay attention to the need for their principal and control variate ensemble to be related, but do not pay attention to the optimality of the couplings. Furthermore the authors utilize a coupling that is optimal for Gaussian random variables; an optimal coupling based on optimal transport could be utilized, while at the same time performing transformations between ensembles through optimal transport techniques.

It is of independent interest to develop more rigorous 'multifidelity' generalizations of such algorithms using the couplings outlined in this chapter.

3.3.2 Ensemble Kalman Filters

In Chernov et al. (2017), Hoel et al. (2016), the authors propose a 'multilevel' EnKF. The authors extend the empirical measures (110) to spatial relations. In a two-level framework the authors analyze the variables

$$\begin{cases} U \\ X - \Phi \hat{U} \end{cases}, \tag{112}$$

again treating each as an independent source of information, with the means:

$$\begin{cases} \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_X - \Phi \boldsymbol{\mu}_{\hat{U}} \end{cases}, \tag{113}$$

and the signed empirical measure covariance estimates:

$$\begin{cases} \Sigma_{U,U} \\ \Sigma_{X,X} - \Sigma_{\Phi\hat{U},\Phi\hat{U}} \end{cases}. \tag{114}$$

This covariance estimate is not guaranteed to be semi-positive definite. Additionally, no attention is paid to utilize an optimal gain linear control variate structure, and the enforcement the principal-control variate relation.

4 Multifidelity Observations

We now discuss an optimal way in which to combine observations from different sources at roughly similar physical locations of the same phenomenon. In operational literature this is commonly dubbed 'super-observations' (Cummings 2005; Oke et al. 2008), though such formulations are largely heuristic in that they take naive averages of interpolations of similar observations. The chief reason why observations are combined instead of used separately is to reduce the observation space dimension, making similar information represented in a denser format.

Assume that the true state is $\mathbf{X}^t$, and recall the multifidelity observation definition (6) where the observations Y^χ and Y^υ have the observation errors η^χ, and η^υ that are assumed to be unbiased and independent. We make the additional assumption that the fine fidelity and coarse fidelity observation operators are deterministically related by the coupling

$$\mathcal{H}^\upsilon(\mathbf{x}) = \theta(\mathcal{H}^\chi(\mathbf{x})), \tag{115}$$

similar to the state relation assumed in (16).

The truth in observation space is assumed to be the expected value of the observation for each fine and coarse observation. This can be alternatively reformulated as the truth in observation space is distributed according to a distribution with mean $\mathbf{Y}^\chi$ and $\mathbf{Y}^\upsilon$ for the fine and coarse observations respectively.

A canonical way of dealing with such a scenario is by 'stacking' the observations, and creating the observation operator

$$\mathcal{H}(\mathbf{X}) = \begin{bmatrix} \mathcal{H}^\chi(\mathbf{X}) \\ \mathcal{H}^\upsilon(\mathbf{X}) \end{bmatrix}. \tag{116}$$

We will not pursue this approach, as it increases the dimensionality of the observations without increasing the information content.

Under the linear control variate approach the total variate observation mean is defined to be:

$$\boldsymbol{\mu}_{Y^\zeta} = \boldsymbol{\mu}_{Y^\chi} - \mathbf{S}(\theta(\boldsymbol{\mu}_{Y^\chi}) - \boldsymbol{\mu}_{Y^\upsilon}), \tag{117}$$

where one implicitly assumes that $\mathbb{E}[\theta(Y^\chi)] = \mathbb{E}[Y^\upsilon]$. The optimal gain is,

$$\mathbf{S} = \Sigma_{Y^\chi,\theta(Y^\chi)}\left(\Sigma_{\theta(Y^\chi),\theta(Y^\chi)} + \Sigma_{Y^\upsilon,Y^\upsilon}\right)^{-1}, \tag{118}$$

with the new covariance of the total observation given by

$$\Sigma_{Y^\zeta,Y^\zeta} = \Sigma_{Y^\chi,Y^\chi} - \Sigma_{Y^\chi,\theta(Y^\chi)}\left(\Sigma_{\theta(Y^\chi),\theta(Y^\chi)} + \Sigma_{Y^\upsilon,Y^\upsilon}\right)^{-1}\Sigma_{\theta(Y^\chi),Y^\chi}. \tag{119}$$

Evaluation of this formula, however, requires knowledge of both $\Sigma_{Y^\chi,\theta(Y^\chi)}$ and $\Sigma_{\theta(Y^\chi),\theta(Y^\chi)}$, which might not be readily available.

An alternate approach is to utilize the importance sampling framework. Assume we have an ensemble of perturbed observations, $\mathbf{E}_{Y^\chi} = \left[\mathbf{Y}^{\chi,(1)}, \ldots, \mathbf{Y}^{\chi,(M)}\right]$, representing M independent samples from the assumed distribution of the fine fidelity observation π_{Y^χ}. Apply the importance sampling procedure to generate the weights:

$$[\mathbf{w}]_i \propto \pi_{Y^\upsilon}\left(\theta\left(\mathbf{Y}^{\chi,(i)}\right)\right). \tag{120}$$

The unbiased mean and covariance estimates of the total observation are given by

$$\widetilde{\boldsymbol{\mu}}_{Y^\zeta} = \mathbf{E}_{Y^\chi}\mathbf{w}, \quad \widetilde{\Sigma}_{Y^\zeta,Y^\zeta} = \frac{M}{M-1}\mathbf{E}_{Y^\chi}\left(\mathrm{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T\right)\mathbf{E}_{Y^\chi}^T. \tag{121}$$

Alternatively, an ensemble of equally weighted perturbed observations to be used with a perturbed observations EnKF can be derived by the optimal transport framework,

$$\mathbf{E}_{Y^\zeta} = \mathbf{E}_{Y^\chi}\mathbf{T}^*, \tag{122}$$

given by (27).

As many of these methods rely on empirical estimates of the total observation covariance matrix, methods such as localization can trivially be applied, especially since in most operational algorithms for physical systems the observation covariance is typically assumed to be diagonal.

5 Numerical Experiments

For the sake of completeness we provide a simple twin experiment on a simple dynamical system to test a two-fidelity MFEnKF.

For the fine fidelity model we use the 40-variable Lorenz '96 system (Lorenz 1996), posed as an ODE:

$$\mathbf{x}_i' = f_X(\mathbf{x}) = -\mathbf{x}_{i-1}(\mathbf{x}_{i-2} - \mathbf{x}_{i+1}) - \mathbf{x}_i + 8, \quad 1 \le i \le 40, \tag{123}$$

where $\mathbf{x}_0 := \mathbf{x}_{40}$, $\mathbf{x}_{-1} := \mathbf{x}_{39}$ and $\mathbf{x}_{41} := \mathbf{x}_1$.

We use the method of snapshots (Sirovich 1987) to construct linear projection and interpolation operators, Φ and $\boldsymbol{\Theta}$, utilizing 20000 snapshots over an expressive time interval of 1000 units.

For the coarse fidelity we consider a reduced order model built using a naive approach, where we evaluate the derivative in the full space and then project onto the reduced space:

$$\mathbf{u}' = f_U(\mathbf{u}) = \boldsymbol{\Theta} f_X(\Phi\mathbf{u}). \tag{124}$$

For the Lorenz '96 system this can be written equivalently as a multivariate quadratic equation.

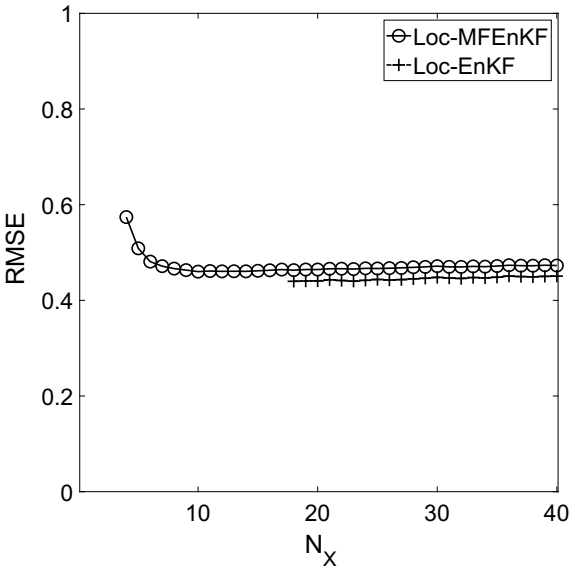

Fig. 2 Comparison of the analysis empirical mean RMSE of a localized perturbed observations MFEnKF with a localized perturbed observations EnKF, for various fine fidelity (full order) model ensemble sizes

The Lorenz '96 system is known to have a Kaplan-Yorke dimension of 27.1 (Popov and Sandu 2019). For this reason we take $r = 28$ reduced modes to describe the whole system (though this is only possible non-linearly). In the reduced model, this represents about 90% of the total energy of the system, as represented by the ratio of the captured eigenvalues to the total eigenvalues. In this context it is actually relatively difficult to build a reduced order model for the Lorenz '96 system.

We compare the algorithm to the standard perturbed observations ensemble Kalman filter. Both algorithms will use forecast anomaly inflation and Gaspi-Cohn covariance localization (Gaspari and Cohn 1999).

We observe every other variable every $\Delta t = 0.05$ time units, with a Gaussian error (2) of $\Sigma_{Y,Y} = \mathbf{I}_{20}$.

We perform localization and inflation as follows. For forecast anomaly inflation for the full system we will take $\alpha_X = 1.1$ and for the coarse system $\alpha_U = 1.00$ as the reduced order model is less stable than the full order model, thus not requiring inflation. To retain an undersampled ensemble for the ancillary variate, we choose an ensemble size of $N_U = 25$. The inner parameter of the localization function is selected to match that of a Gaussian kernel, and set the radius to be equal to 4 (Petrie and Dance 2010).

Figure 2 shows the relationship between the principal variate ensemble size and the spatio-temporal RMSE of the empirical analysis mean of the MFEnKF and the EnKF. As can be seen, the problem is comparatively difficult for the EnKF, as it requires at least 18 fine fidelity ensemble members for a stable behavior. The same RMSE can be achieved with less than 10 fine fidelity ensemble members in the MFEnKF

framework. Assuming the coarse fidelity model runs are significantly cheaper (not true in this trivial contrived example) then the MFEnKF is clearly superior.

We note that there is some loss of accuracy in the results, due in part to several assumptions that are violated. One is that the orthogonal complement space is uncorrelated with that of the full space (65). As we are capturing 90% of the energy of the system, the rest of the energy is not that negligible, and is no doubt highly correlated with the what is captured. Methods to diminish the influence of this error, would be needed for operational systems.

6 Discussion

Multifidelity data assimilation, and multifidelity inference in general, seek to leverage the availability of information about reality at multiple resolution levels. The field is still in its infancy, but the multifidelity methods are highly promising. This chapter provides a general philosophical and theoretical framework for the development of such methods. New multifidelity data assimilation approaches should utilize efficient coarse fidelity models to speed up high fidelity inference. The new methods should be grounded in sound statistical and probabilistic theory.

In this chapter we focus on the multifidelity stochastic EnKF. Variational multifidelity approaches have been developed in Stefanescu et al. (2015). Square root multifidelity Kalman filters, analogues to the perturbed observations MFEnKF, must be developed in the future. Particle filters that are appropriate for non-Gaussian probability densities, or even hybrid EnKF-PF systems where different variates are assimilated with different algorithms, might provide an avenue for development of multifidelity particle filtering. Multifidelity hybrid data assimilation, that combines multifidelity EnKF and multifidelity variational methods, are also a promising future venue. Finally, the construction of a hierarchy of coarser models to support data assimilation should be carefully investigated. For example, methods based on machine learning (e.g., as discussed in Moosavi et al. 2018a, b, or non-linear projections using autoencoders) are of considerable interest.

Acknowledgements The authors would like to acknowledge Traian Iliescu and Changhong Mou who have helped make some of the work underlying this possible. This work was supported by awards NSF CCF–1613905. NSF ACI–1709727, NSF CDS&E-MSS–1953113, and by the Computational Science Laboratory at Virginia Tech.

References

Acevedo W, de Wiljes J, Reich S (2017) Second-order accurate ensemble transform particle filters. SIAM J Sci Comput 39(5):A1834–A1850

Allen E, Baglama J, Boyd S (2000) Numerical approximation of the product of the square root of a matrix with a vector. Linear Algebra Appl 310(1–3):167–181

Asch M, Bocquet M, Nodet M (2016) Data assimilation: methods, algorithms, and applications. SIAM
Bishop C, Etherton B, Majumdar S (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. Mon Weather Rev 129:420–436
Brunton SL, Kutz JN (2019) Data-driven science and engineering: machine learning, dynamical systems, and control. Cambridge University Press
Chernov A, Hoel H, Law K, Nobile F, Tempone R (2017) Multilevel ensemble Kalman filtering for for spatio-temporal processes. MATHICSE technical report 22.2017, EPFL
Cummings JA (2005) Operational multivariate ocean data assimilation. Q J R Meteorol Soc 131(613):3583–3604
Gaspari G, Cohn S (1999) Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc 125:723–757
Giles MB (2008) Multilevel Monte Carlo path simulation. Oper Res 56(3):607–617
Giles MB (2015) Multilevel Monte Carlo methods. Acta Numer 24:259–328
Gregory A, Cotter CJ (2017) A seamless multilevel ensemble transform particle filter. SIAM J Sci Comput 39(6):A2684–A2701
Gregory A, Cotter CJ, Reich S (2016) Multilevel ensemble transform particle filtering. SIAM J Sci Comput 38(3):A1317–A1338
Hoel H, Law KJH, Tempone R (2016) Multilevel ensemble Kalman filtering. SIAM J Numer Anal **54**(3). https://doi.org/10.1137/15M100955X
Houtekamer P, Mitchell H (1998) Data assimilation using an ensemble Kalman filter technique. Mon Weather Rev 126:796–811
Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press
Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82:35–45
Lorenz EN (1996) Predictability: a problem partly solved. In: Proceedings of seminar on predictability, vol 1
Moosavi A, Stefanescu R, Sandu A (2018a) Efficient construction of local parametric reduced order models using machine learning techniques. Int J Numer Methods Eng 113(3):512–533
Moosavi A, Stefanescu R, Sandu A (2018b) Parametric domain decomposition for accurate reduced order models: applications of MP-LROM methodology. J Comput Appl Math 340:629–644
Nelson BL (1987) On control variate estimators. Comput Oper Res 14(3):219–225
Oke PR, Brassington GB, Griffin DA, Schiller A (2008) The Bluelink ocean data assimilation system (BODAS). Ocean Model 21(1–2):46–70
Peherstorfer B, Willcox K, Gunzburger M (2018) Survey of multifidelity methods in uncertainty propagation, inference, and optimization. SIAM Rev 60(3):550–591
Petrie RE, Dance S (2010) Ensemble-based data assimilation and the localisation problem. Weather 65(3):65–69
Popov AA, Mou C, Iliescu T, Sandu A (2020) A multifidelity ensemble Kalman filter with reduced order control variates. arXiv:2007.00793
Popov AA, Sandu A (2019) A Bayesian approach to multivariate adaptive localization in ensemble-based data assimilation with time-dependent extensions. Nonlinear Process Geophys 26(2):109–122
Popov AA, Sandu A (2020) An explicit probabilistic derivation of inflation in a scalar ensemble Kalman filter for finite step, finite ensemble convergence. arXiv:2003.13162
Reich S (2013) A nonparametric ensemble transform method for Bayesian inference. SIAM J Sci Comput 35(4):A2013–A2024
Reich S, Cotter C (2015) Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press
Rubinstein RY, Marcus R (1985) Efficiency of multivariate control variates in Monte Carlo simulation. Oper Res 33(3):661–677

Sakov P, Bertino L (2011) Relation between two common localisation methods for the EnKF. Comput Geosci 15(2):225–237

Sirovich L(1987) Turbulence and the dynamics of coherent structures. I. Coherent structures. Q Appl Math **45**(3), 561–571

Stefanescu R, Sandu A, Navon I (2015) POD/DEIM strategies for reduced data assimilation systems. J Comput Phys 295:569–595

Filtering with One-Step-Ahead Smoothing for Efficient Data Assimilation

Boujemaa Ait-El-Fquih and Ibrahim Hoteit

Abstract State-space systems arise in many applications in signal processing and data assimilation. In this context, the main goal is to estimate online the state of the system based on available observations, the so-called filtering problem. Standard filtering solutions are computed recursively as successive cycles of alternating time-update (forecast) and observation-update (analysis) steps. This path is however not the only recursive way to compute the filtering solution. In this context, new one-step-ahead smoothing (OSAS)-like filtering schemes have been introduced, reversing the order of the observation-update and the time-update steps. These involve two Bayesian-like update steps based on the same (present) observation: one for smoothing the previous state and one for analyzing the present one. These include new variants of Kalman filters (KF-OSAS), particle filters (PF-OSAS) and ensemble Kalman filters (EnKF-OSAS), depending on the size and the linear-Gaussian character of the underlying state-space system. While the standard KF and KF-OSAS provide the same (exact) estimator, the use of the same data twice in the estimation process generally leads to improved trade-off between estimation quality and computational burden for the PF-OSAS and EnKF-OSAS. This chapter offers a comprehensive presentation of the OSAS-like filtering algorithms, reviewing their derivations, detailing algorithmic and practical differences and similarities with their classical counterparts, and discussing their relevance for both small- and large-dimensional applications.

1 Introduction

The estimation of the state of a dynamical system based on incoming observations that are distributed in time has long been an active area of research (Hoteit et al. 2018; Chui and Chen 1999; Harvey 1989; Anderson and Moore 1979; Evensen

B. Ait-El-Fquih · I. Hoteit (✉)
King Abdullah University of Science and Technology (KAUST), PSE, Thuwal, Saudi Arabia
e-mail: ibrahim.hoteit@kaust.edu.sa

B. Ait-El-Fquih
e-mail: boujemaa.aitelfquih@kaust.edu.sa

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_3

2006). This is often formulated as the so-called state-space system composed by two stochastic equations: one describes the evolution of the state, called the dynamical model, and the other one maps the state into the observations' space through the so-called observational model. In the context of state-space systems, combining all available information from prior (initial) state and observations has proven to be the most viable approach to determine the best estimates of the state of an underlying physical system, a process that can be achieved through filtering or optimization approaches—often referred to as data assimilation (DA).

The theoretical framework of DA is now well-established, mostly based on the deterministic variational (optimization) theory or the statistical estimation (filtering) theory. Variational methods seek to fit the state model's trajectory to available observations by adjusting a well-chosen set of control parameters (Le Dimet and Talagrand 1986), whereas in the filtering methods the state of the system is estimated sequentially as the observations become available (Jazwinski 1970; Künsch 2001). More specifically, in the filtering framework, the estimation problem is typically split into successive cycles of alternating time-update (forecast) then observation-update (analysis) steps (Künsch 2001). The time-update step computes the forecast probability density function (pdf) of the state given past observations, by integrating the previous analysis pdf with the dynamical model. The forecast pdf is then updated in the observation-update step with the incoming observations, through Bayes' rule, to obtain the analysis pdf of the state given all observations up to the current time. The analysis pdf is conditioned on all available observations and as such, it provides any type of state estimates as for instance the posterior mean (PM), the optimum solution according to several criteria including the well-known mean-squared error (MSE) minimization (van Trees 1968; Sherman 1955).

In practice, however, analytic calculation of the forecast and analysis distributions, and associated estimates, is not straightforward unless the state-space system is linear and Gaussian, in which case the aforementioned generic forecast and analysis steps are reduced to the well-known Kalman filter (KF) (Kalman 1960; Jazwinski 1970; Anderson and Moore 1979; Ait-El-Fquih and Desbouvries 2006). The particle filter (PF)[1] is the most prominent among the vast toolbox of approximate numerical methods that have been proposed for nonlinear/non-Gaussian systems (Gordon et al. 1993; Doucet et al. 2001; Cappé et al. 2005). It is a sequential Monte Carlo (MC) algorithm that provides an approximation of the (continuous) forecast and analysis distributions by empirical (discrete) distributions of random samples, called particles. The theory behind PF is mathematically sound and its asymptotic (in the number of particles) convergence properties are well established (Doucet et al. 2001; Crisan and Doucet 2002). However, due to the finite number of particles that is used in practice, the weights of particles usually exhibit variances that increase exponentially in time, which may strongly limit the filter performances. Indeed, after few assimila-

[1] Without loss of generality, throughout the chapter PF refers to the particular *bootstrap* algorithm of Gordon et al. (1993), which is commonly used in geophysics problems. Sequential Importance Sampling (SIS) or Sampling Importance Resampling (SIR) algorithms will designate the other importance sampling (IS)-based algorithms.

tion cycles only, all but a few particles will have negligible weights, leading to the so-called weights' degeneracy phenomenon (Liu and Chen 1998; Doucet et al. 2001; Snyder et al. 2008; van Leeuwen 2009; Desbouvries et al. 2011). This happens in part because the incoming observations are not used to update the particles in the analysis step, but only their weights (Hoteit et al. 2008; van Leeuwen 2009; Hoteit et al. 2012). A standard solution to mitigate the degeneracy phenomenon is resampling, which basically consists of drawing "new" particles by duplicating those with large weights and abandoning those with low weights (Rubin 1988; Gordon et al. 1993; Liu and Chen 1998; Doucet et al. 2001; Lamberti et al. 2017). The PF with resampling has been proven to perform well for low-dimensional systems (Kivman 2003; Subramanian et al. 2012), but is still inefficient for large-dimensional systems due to the need of a prohibitive number of particles to sufficiently sample the state space (curse of dimensionality); with the number of particles needed scaling exponentially with the system dimension (Crisan and Doucet 2002; Snyder et al. 2008).

Despite the promising PF-based strategies that have recently been proposed to cope with the curse of dimensionality in large-dimensional systems (e.g., Spiller et al. 2008; Husz et al. 2011; Morzfelda et al. 2012; Ades and van Leeuwen 2013; Djuric and Bugallo 2013; Ait-El-Fquih and Hoteit 2015; Septier and Peters 2015; Ait-El-Fquih and Hoteit 2016), the ensemble Kalman filtering (EnKF) is still the most popular and most efficient framework for large-dimensional systems, e.g., in atmosphere, ocean, hydrology, etc. (Evensen 1994, 2006; Hoteit et al. 2015; Asch et al. 2016; Hoteit et al. 2018; Carrassi et al. 2018). EnKF is a MC implementation of the KF. It shares the same forecast step with the PF, but uses an analysis step that is derived from that of the KF based on the Gaussian assumption on the joint state-observation forecast pdf. In the EnKF analysis step, the analysis particles (called now "analysis members") can be sampled either by updating the forecast members directly with the KF correction step based on stochastically perturbed observations (referred to as the stochastic EnKF, e.g., Burgers et al. 1998; Houtekamer and Mitchell 1998; Evensen 2006), or through an update of the mean and a square-root form of the covariance of the forecast ensemble exactly as in the KF, without perturbing the observations (known as the deterministic EnKFs, e.g., Pham 2001; Bishop et al. 2001; Anderson 2001; Hoteit et al. 2002; Tippett et al. 2003; Hunt et al. 2007; Hoteit et al. 2015). Though the problem of weights' degeneracy is circumvented, these filters often suffer from the so-called filter inbreeding problem, in which the ensemble variance is increasingly under-estimated over time (Furrer and Bengtsson 2007; Hendricks Franssen and Kinzelbach 2008). This mostly happens when the filter is implemented without accounting for the various systematic errors in the system (mostly the dynamical model errors and those related to the Gaussian assumption), and sampling errors (Whitaker and Hamill 2002; Houtekamer and Mitchell 2005; Whitaker et al. 2008). Efficient auxiliary techniques have been introduced, most notably inflation (Anderson 2001; Ait-El-Fquih and Hoteit 2020) and localization (Houtekamer and Mitchell 1998) (see also Kivman 2003; Hoteit et al. 2008). With more studies demonstrating the efficiency of the EnKF in various applications, other EnKF-like extensions were introduced. These include the use of the less restrictive Gaussian mixture assumption (e.g., Hoteit et al. 2008; Stordal et al. 2011; Hoteit

et al. 2012; Frei and Künsch 2013; Liu et al. 2015); the development of ensemble Kalman smoothing algorithms (e.g., Evensen and van Leeuwen 2000; Dunne et al. 2007; Dreano et al. 2017); and the extension of the EnKF to the framework of state-parameters filtering problems (e.g., Moradkhani et al. 2005; Annan et al. 2005; Chen and Zhang 2006; Aksoy et al. 2006; Bellsky et al. 2014; Gharamti et al. 2014; Rasmussen et al. 2015; Ait-El-Fquih and Hoteit 2018).

Recent studies have investigated a reversed filtering path, i.e., "observation-update then time-update" to compute the analysis pdf recursively (Desbouvries and Ait-El-Fquih 2008; Desbouvries et al. 2011). This resulted in a new set of filtering algorithms that involve computing a one-step-ahead smoothing (OSAS) pdf between two successive analysis pdfs, within a fully consistent Bayesian framework. Starting from the analysis pdf at the previous time, the observation-update step aims at updating (constraining) this pdf with the current observation, using a (new) transformed observational model linking this observation with the previous state. This yields the so-called OSAS pdf of the previous state (i.e., given the history of observations, including the current one). The time-update step then integrates the resulting OSAS pdf with a (new) updated dynamical model linking the current state with the previous state and the current observation, which leads to the analysis pdf of the current state. The two "new" models are defined through a combination of the original dynamical and observational models. This means that the OSAS-like filtering algorithms involve the use of the original models twice (i.e., two integration and two update steps with the newest observation). At every assimilation cycle, this roughly doubles the overall computational cost of the standard filtering algorithms, but also exploits the available information twice in an attempt to enhance the filters performances (e.g., Desbouvries et al. 2011; Ait-El-Fquih et al. 2016; Raboudi et al. 2018).

Apart from the (exact) OSAS-like KF (KF-OSAS) (Desbouvries and Ait-El-Fquih 2008), several studies have suggested through numerical experiments that the recently introduced (approximate) OSAS-like PF (PF-OSAS) and EnKF (EnKF-OSAS) may provide a better trade-off between computational burden and estimates accuracy, mainly under challenging assimilation scenarios as for instance strong nonlinearities, sparse observational networks, large assimilation windows, etc. (e.g., Desbouvries et al. 2011; Ait-El-Fquih et al. 2016; Raboudi et al. 2018, 2019). This makes the OSAS formulations of the PF and EnKF promising approaches for deriving efficient schemes for large-dimensional nonlinear data assimilation problems (e.g., Desbouvries et al. 2011; Ait-El-Fquih et al. 2016; Raboudi et al. 2018, 2019, 2020). In such filters, a smoothing ensemble is first sampled, then forwarded in time with the new dynamical model, which also involves the newest observation. The smoothing ensemble is sampled in PF-OSAS according to weighting-resampling steps, and in EnKF-OSAS according to a Kalman-like update step.

This chapter aims at providing a general overview of the OSAS-like filtering methods, and discussing in particular their algorithmic and practical differences and similarities with their standard counterparts. It will also discuss the potential use of these filters with small- and large-dimensional problems. The remainder of this chapter is organized as follows. Section 2 recalls the Bayesian filtering problem and reviews its standard formulation. Section 3 introduces the generic formulation of

the OSAS-like filters, then design the associated "new" dynamical and observational models from the original ones. Section 4 discusses the implementation of the generic OSAS filtering scheme with small-dimensional models, including the KF-OSAS and PF-OSAS algorithms. Section 5 focuses on the large-dimensional systems, presenting the stochastic EnKF-OSAS, then discussing the key steps of its deterministic variants and also extending it to the more general state-parameters filtering framework. Concluding remarks are offered in Sect. 6.

2 Problem Formulation

Let $\mathbf{x} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{y} = \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{x}_N\}$, with $\mathbf{x}_n \in \mathbb{R}^{n_\mathbf{x}}$ and $\mathbf{y}_n \in \mathbb{R}^{n_\mathbf{y}}$, denote a discrete-time (unknown) state process and an observation process, respectively. In many DA applications, such as for instance geophysical fluid dynamics (Evensen 1994, 2006; Anderson 2009; Mandel et al. 2009; Yardim et al. 2011; Hoteit et al. 2018; Hoteit et al 2020), target tracking (Chui and Chen 1999; Bar-Shalom et al. 2001; Arulampalam et al. 2002; Yardim et al. 2011), navigation (Bar-Shalom et al. 2001; Durrant-Whyte and Bailey 2006; Gustafsson 2010), and wireless communications (Djuric et al. 2002; Ephraim and Merhav 2002; Haykin et al. 2004), these processes are related following a state-space system of the form:

$$\begin{cases} \mathbf{x}_n = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}) + \mathbf{u}_{n-1}, \\ \mathbf{y}_n = \mathbf{H}_n \mathbf{x}_n + \mathbf{v}_n; \end{cases} \tag{1}$$

$\mathbf{f}_{n-1}(.)$ being a (possibly nonlinear) dynamical operator[2] integrating the state of the system from time t_{n-1} to t_n, and $\mathbf{H}_n$ an observational operator at time t_n, assumed linear here for simplicity (the OSAS filtering schemes can be easily extended to the nonlinear case). The state noise process, $\mathbf{u} = \{\mathbf{u}_n\}_{n\in\mathbb{N}}$, and the observation noise process, $\mathbf{v} = \{\mathbf{v}_n\}_{n\in\mathbb{N}}$, are assumed to be independent (i.e., time-wise independent), jointly independent and independent of the initial state, $\mathbf{x}_0$. Also let $\mathbf{x}_0$, $\mathbf{u}_n$ and $\mathbf{v}_n$ be Gaussian with means $\hat{\mathbf{x}}_0$, $\mathbf{0}$ and $\mathbf{0}$, and covariances $\mathbf{P}_0$, $\mathbf{Q}_n$ and $\mathbf{R}_n$, respectively. Throughout this chapter, $\mathbf{x}_{0:n} = \{\mathbf{x}_0, \ldots, \mathbf{x}_n\}$, $\mathbf{y}_{0:n} = \{\mathbf{y}_0, \ldots, \mathbf{y}_n\}$, and $p(\boldsymbol{\xi})$ and $p(\boldsymbol{\xi}|\boldsymbol{\mu})$ denote the pdf of a random variable $\boldsymbol{\xi}$, and the conditional pdf of $\boldsymbol{\xi}$ given a realization of another random variable $\boldsymbol{\mu}$, respectively. Such independence properties yield,

[2] Without loss of generality, the term $\mathbf{g}(\mathbf{m})$ in any modeling equation $\mathbf{d} = \mathbf{g}(\mathbf{m}) + \textit{noise}$, will be called *operator*, whereas the whole equation will be called *model*, i.e., equations in (1) refer to as dynamical and observational *models*; ditto for those in (25).

$$p(\mathbf{x}_n|\mathbf{x}_{0:n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}), \tag{2}$$

$$p(\mathbf{y}_{0:n}|\mathbf{x}_{0:n}) = \prod_{k=0}^{n} p(\mathbf{y}_k|\mathbf{x}_{0:n}), \tag{3}$$

$$p(\mathbf{y}_k|\mathbf{x}_{0:n}) = p(\mathbf{y}_k|\mathbf{x}_k),\ k = 0, 1, \ldots, n, \tag{4}$$

which suggests that system (1) is a (continuous state) hidden Markov chain (HMC) of transition pdf $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ and likelihood $p(\mathbf{y}_n|\mathbf{x}_n)$ (Künsch 2001; Ait-El-Fquih and Desbouvries 2006). The aforementioned Gaussian assumptions, in turn, entail

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}) = \mathcal{N}_{\mathbf{x}_n}(\mathbf{f}_{n-1}(\mathbf{x}_{n-1}), \mathbf{Q}_{n-1}), \tag{5}$$

$$p(\mathbf{y}_n|\mathbf{x}_n) = \mathcal{N}_{\mathbf{y}_n}(\mathbf{H}_n\mathbf{x}_n, \mathbf{R}_n), \tag{6}$$

where $\mathcal{N}_{\mathbf{x}}(\mathbf{m}, \mathbf{C})$ stands for a Gaussian pdf of argument $\mathbf{x}$ and parameters $(\mathbf{m}, \mathbf{C})$. It is worth noticing from Eqs. (5)–(6) that the transition and likelihood laws are probabilistic formulations of the dynamical and observational models, respectively.

The Bayesian filtering problem consists of estimating the value taken by $\mathbf{x}_n$, at every time $t_n = t_0, \ldots, t_N$, from all available observations, $\mathbf{y}_{0:n}$. The posterior pdf, $p(\mathbf{x}_n|\mathbf{y}_{0:n})$, known as the analysis pdf, is a key quantity of such a problem, as it summarizes the entire information about $\mathbf{x}_n$ knowing $\mathbf{y}_{0:n}$, and allows one to compute any type of estimator depending on the choice of the optimization criteria. The PM estimator, which has the advantage to be the solution for a wide class of criteria, including the well-known MSE minimization, is the most popular (van Trees 1968; Sherman 1955). In a HMC, it is possible to calculate the analysis pdf in a *recursive* (efficient) way based on the transition and likelihood laws (Künsch 2001). Clearly, the recursivity is inherited from the equalities,

$$p(\mathbf{x}_n|\mathbf{x}_{0:n-1}, \mathbf{y}_{0:n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}), \tag{7}$$

$$p(\mathbf{y}_n|\mathbf{x}_{0:n}, \mathbf{y}_{0:n-1}) = p(\mathbf{y}_n|\mathbf{x}_n), \tag{8}$$

which, indeed, follow from (2)–(4) (Ait-El-Fquih and Desbouvries 2006). More precisely, the path from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ to $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ can be achieved with a succession of a Markovian step (or integration) where the transition pdf is used to obtain the forecast pdf,

$$p(\mathbf{x}_n|\mathbf{y}_{0:n-1}) \overset{(7)}{=} \int p(\mathbf{x}_n|\mathbf{x}_{n-1})p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})d\mathbf{x}_{n-1}, \tag{9}$$

and a Bayesian step (or correction) in which the likelihood is combined with the forecast pdf using Bayes' rule,

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) \propto p(\mathbf{y}_n, \mathbf{x}_n|\mathbf{y}_{0:n-1}), \tag{10}$$

$$\overset{(8)}{\propto} p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{y}_{0:n-1}). \tag{11}$$

In the linear-Gaussian systems, (9)–(11) can be computed exactly by KF techniques (Kalman 1960; Anderson and Moore 1979; Kailath et al. 2000). In a general system, however, calculating $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ analytically is most often not possible. Many approximate techniques such as the sequential Monte Carlo (SMC) methods, most notably SIRs/PF (e.g., Gordon et al. 1993; Doucet et al. 2001) and EnKFs (e.g., Evensen 1994, 2006; Hoteit et al. 2015), which aim at providing MC approximations of $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ have been developed. In this chapter we do not aim to modify the standard ("time-update then observation-update"-like) KF, PF or EnKF algorithms. We focus instead on the alternate estimation path "observation-update then time-update" to revisit new filtering algorithms, which involve the computation of an OSAS pdf between two successive analysis pdfs.

3 One-Step-Ahead Smoothing (OSAS) Formulation of Bayesian Filtering

The classical "time-update then observation-update" path (9)–(11) is not the only one to compute the analysis pdf $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$. One indeed may reverse the order of the time- and observation-update steps, without requiring any further assumptions (Desbouvries et al. 2011).

3.1 The Generic Algorithm

3.1.1 The Observation-Update Step

Constraining $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ by the current observation, $\mathbf{y}_n$, is performed using Bayes' rule, which results in the one-step-ahead smoothing pdf $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ (Desbouvries et al. 2011):

$$p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n}) \propto p(\mathbf{y}_n|\mathbf{x}_{n-1})p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1}). \tag{12}$$

Equation (12) holds under the condition $p(\mathbf{y}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n-1}) = p(\mathbf{y}_n|\mathbf{x}_{n-1})$, which is already fulfilled from (7)–(8). Unlike the classical Bayesian step (11), which makes use of the likelihood $p(\mathbf{y}_n|\mathbf{x}_n)$, defined in (6), $p(\mathbf{y}_n|\mathbf{x}_{n-1})$ is not known and needs to be calculated before applying the Bayesian step (12). This can be achieved following an integration based on the original transition and likelihood laws (Desbouvries et al. 2011):

$$p(\mathbf{y}_n|\mathbf{x}_{n-1}) = \int p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{x}_{n-1})d\mathbf{x}_n. \tag{13}$$

3.1.2 The Time-Update Step

Given the set of observations $\mathbf{y}_{0:n}$, this step leads to move from the law of $\mathbf{x}_{n-1}$ (i.e., from the resulting smoothing pdf $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$) to the law of $\mathbf{x}_n$ (i.e., to the analysis pdf of interest $p(\mathbf{x}_n|\mathbf{y}_{0:n})$), hence the term "time-update". This can be done through the integration (Desbouvries et al. 2011),

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) = \int p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)\, p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n}) d\mathbf{x}_{n-1}. \tag{14}$$

Equation (14) holds under the condition $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)$, which is already fulfilled from (7)–(8). Noticing that (14) shares a similar structure with the standard integration step (9), with (14) coinciding with (9) "given the observation $\mathbf{y}_n$", step (14) will henceforth be referred to as the conditional integration step (driven by the conditional transition density, $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)$). Unlike the (original) transition density in (9), the conditional transition density is not known and its calculation is required before applying the conditional integration step (14). This can be achieved using a Bayesian mechanism based on the original transition and likelihood laws (Desbouvries et al. 2011):

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n) \propto p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathbf{x}_{n-1}). \tag{15}$$

In summary, the OSAS-like filtering approach has a similar structure as the classical filtering one, with each assimilation cycle involving a Bayesian step and an integration step. The key difference lies in the fact that the likelihood and transition densities that govern these steps are known in the classical approach, but are unknown in the OSAS formulation, a reason for which further integration- and Bayesian-like steps arise to calculate them.

3.2 State-Space Transform

Substituting (5)–(6) in (13) and (15), this results in "new" transition and likelihood laws that are also Gaussian with,

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n) = \mathcal{N}_{\mathbf{x}_n}(\tilde{\mathbf{f}}_{n-1}(\mathbf{x}_{n-1}, \mathbf{y}_n), \tilde{\mathbf{Q}}_{n-1}), \tag{16}$$

$$p(\mathbf{y}_n|\mathbf{x}_{n-1}) = \mathcal{N}_{\mathbf{y}_n}(\tilde{\mathbf{h}}_{n-1}(\mathbf{x}_{n-1}), \tilde{\mathbf{R}}_n), \tag{17}$$

where,

$$\tilde{\mathbf{f}}_{n-1}(\mathbf{x}_{n-1}, \mathbf{y}_n) = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}) + \mathbf{G}_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1})), \tag{18}$$

$$\tilde{\mathbf{h}}_n(\mathbf{x}_{n-1}) = \mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1}), \tag{19}$$

$$\tilde{\mathbf{R}}_n = \mathbf{H}_n\mathbf{Q}_{n-1}\mathbf{H}_n^T + \mathbf{R}_n, \tag{20}$$

$$\mathbf{G}_n = \mathbf{Q}_{n-1}\mathbf{H}_n^T\tilde{\mathbf{R}}_n^{-1}, \tag{21}$$

$$\tilde{\mathbf{Q}}_{n-1} = \mathbf{Q}_{n-1} - \mathbf{G}_n\mathbf{H}_n\mathbf{Q}_{n-1}. \tag{22}$$

As will be further discussed hereafter, (16) and (17) are probabilistic representations of "new" dynamical and observational models.

3.2.1 The New Dynamical Model

As can be seen from (16), (18), updating the Gaussian prior pdf (5) with the Gaussian likelihood (6) following the Bayesian update (15), amounts to update the moments of this prior with the observation $\mathbf{y}_n$ and the original observational model following a Kalman-like update with a gain $\mathbf{G}_n$ as in (21). One can then easily show that the resulting conditional (posterior) transition pdf (16) is a probabilistic formulation of the "new" dynamical model,

$$\mathbf{x}_n = \tilde{\mathbf{f}}_{n-1}(\mathbf{x}_{n-1}, \mathbf{y}_n) + \tilde{\mathbf{u}}_{n-1}; \quad \tilde{\mathbf{u}}_{n-1} \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{Q}}_{n-1}), \tag{23}$$

which, as can be seen from (18), is, somehow, a Kalman-like update of the original dynamical model based on the observation $\mathbf{y}_n$ and the original observational model. The correction terms, $\mathbf{G}_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1}))$ in (18) and $\mathbf{G}_n\mathbf{H}_n\mathbf{Q}_{n-1}$ in (22), make the new model computationally more expensive than the original one. However, because of its fully consistent Bayesian derivation, the new model should be "more informative".

3.2.2 The New Observational Model

As can be seen from[3] (13) and (17), (19), "averaging" the likelihood $p(\mathbf{y}_n|\mathbf{x}_n)$ w.r.t. $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ to obtain $p(\mathbf{y}_n|\mathbf{x}_{n-1})$ amounts to "fix" the variable $\mathbf{x}_n$ in the observational model, replacing it by the mean of $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ (which turns $\mathbf{H}_n\mathbf{x}_n$ in (6) into $\mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1})$ in (17), (19)), and to inflate the likelihood covariance additively by a factor $\mathbf{H}_n\mathbf{Q}_{n-1}\mathbf{H}_n^T$ (which turns $\mathbf{R}_n$ in (6) into $\tilde{\mathbf{R}}_n$ in (17)). This inflation of the covariance comes from the fact that $p(\mathbf{y}_n|\mathbf{x}_n)$ is nothing but $p(\mathbf{y}_n|\mathbf{x}_{n-1})$, constrained with "more information" that is contained in $\mathbf{x}_n$, i.e.,

[3] Notice that (13) is simply an "averaging" of $p(\mathbf{y}_n|\mathbf{x}_n)$ w.r.t. $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ to obtain $p(\mathbf{y}_n|\mathbf{x}_{n-1})$, i.e., $p(\mathbf{y}_n|\mathbf{x}_{n-1}) \overset{(13)}{=} \mathbb{E}_{p(\mathbf{x}_n|\mathbf{x}_{n-1})}[p(\mathbf{y}_n|\mathbf{x}_n)]$.

$$\underbrace{p(\mathbf{y}_n|\mathbf{x}_n)}_{\text{posterior}} = p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{x}_{n-1}),$$
$$\propto \underbrace{p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)}_{\text{likelihood}} \times \underbrace{p(\mathbf{y}_n|\mathbf{x}_{n-1})}_{\text{prior}}.$$

On the other hand, one can also show that the new likelihood (17) is a probabilistic representation of a new observational model,

$$\mathbf{y}_n = \tilde{\mathbf{h}}_{n-1}(\mathbf{x}_{n-1}) + \tilde{\mathbf{v}}_n; \quad \tilde{\mathbf{v}}_n \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}_n), \tag{24}$$

which, unlike the *linear* original model, maps *non-linearly* the previous state to the current observation. However, since $\tilde{\mathbf{R}}_n$ has been already calculated to build the new dynamical model, no additional computations are needed compared to the original observational model.

3.2.3 The New State-Space System

Grouping together the new models leads to a new state-space system,[4]

$$\begin{cases} \mathbf{y}_n = \quad \tilde{\mathbf{h}}_n(\mathbf{x}_{n-1}) + \tilde{\mathbf{v}}_n, \\ \mathbf{x}_n = \tilde{\mathbf{f}}_{n-1}(\mathbf{x}_{n-1}, \mathbf{y}_n) + \tilde{\mathbf{u}}_{n-1}, \end{cases} \tag{25}$$

where the associated noise processes, $\tilde{\mathbf{v}} = \{\tilde{\mathbf{v}}_n\}_{n\in\mathbb{N}}$ and $\mathbf{u} = \{\tilde{\mathbf{u}}_n\}_{n\in\mathbb{N}}$, are Gaussian, independent, jointly independent and independent of the initial state, $\mathbf{x}_0$. Defining $\tilde{\mathbf{v}}_n$ and $\tilde{\mathbf{u}}_{n-1}$ as,

$$\tilde{\mathbf{v}}_n = \mathbf{H}_n\mathbf{u}_{n-1} + \mathbf{v}_n, \tag{26}$$
$$\tilde{\mathbf{u}}_{n-1} = [\mathbb{I}_{n_\mathbf{x}} - \mathbf{G}_n\mathbf{H}_n]\mathbf{u}_{n-1} - \mathbf{G}_n\mathbf{v}_n, \tag{27}$$

one can easily show that these Gaussian and independence properties follow from those of the "original" noises, $\mathbf{u}$ and $\mathbf{v}$ (see Sect. 2). More specifically:

- *Gaussianity.* $\mathbf{u}_{n-1}$ and $\mathbf{v}_n$ are independent and Gaussian variables with zero means and covariances $\mathbf{Q}_{n-1}$ and $\mathbf{R}_n$, respectively, implies that $\tilde{\mathbf{v}}_n$ and $\tilde{\mathbf{u}}_{n-1}$ are Gaussian with zero means and covariances $\tilde{\mathbf{R}}_n$ (20) and $\tilde{\mathbf{Q}}_{n-1}$ (22), respectively.
- *Independence of the noises.* $\mathbf{u}$ and $\mathbf{v}$ are independent and jointly independent processes, implies that $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{u}}$ are independent, i.e., $\mathbb{E}\left[\tilde{\mathbf{v}}_n\tilde{\mathbf{v}}_\ell^T\right] = \mathbf{0}$ and $\mathbb{E}\left[\tilde{\mathbf{u}}_n\tilde{\mathbf{u}}_\ell^T\right] = \mathbf{0}$ for any $n, \ell \in \mathbb{N}$ and $n \neq \ell$.
- *Joint independence between the noises.* $\mathbf{u}$ and $\mathbf{v}$ are independent and jointly independent processes, implies that $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{u}}$ are jointly independent, i.e., $\mathbb{E}\left[\tilde{\mathbf{v}}_n\tilde{\mathbf{u}}_\ell^T\right] = \mathbf{0}$

[4] Recall that the OSAS-like filter reverts the order of the time-update and observation-update steps (i.e., step (12) then (14)), involving the use of the new likelihood before the new transition density. Thus, for the sake of consistency, we revert the conventional order of the models in the new system.

for any $n, \ell \in \mathbb{N}$. This include the particular case in which the noises $\tilde{\mathbf{v}}_n$ and $\tilde{\mathbf{u}}_{n-1}$ in (25) are independent. More specifically, the independence between $\mathbf{u}_{n-1}$ and $\mathbf{v}_n$ implies that $\mathbb{E}\left[\tilde{\mathbf{v}}_n \tilde{\mathbf{u}}_{n-1}^T\right] = \mathbf{0}$, which means that $\tilde{\mathbf{v}}_n$ and $\tilde{\mathbf{u}}_{n-1}$ are independent.

- *Independence between the initial state and the noises.* $\mathbf{u}$ and $\mathbf{v}$ are independent of $\mathbf{x}_0$, implies that $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{u}}$ are independent of $\mathbf{x}_0$, i.e., $\mathbb{E}\left[\tilde{\mathbf{v}}_n (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T\right] = \mathbf{0}$ and $\mathbb{E}\left[\tilde{\mathbf{u}}_n (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T\right] = \mathbf{0}$ for any $n \in \mathbb{N}$.

The new state-space system (25), and equivalently the associated conditional transition and likelihood laws (16)–(17), establish the context in which implementations of the observation-update step (12) and the time-update step (14) of the generic OSAS-like filter are performed. This is presented in details in the next Sections, including KF-OSAS and PF-OSAS for DA into small-dimensional systems, and (stochastic) EnKF-OSAS and its deterministic variants as well as its extension to state-parameters filtering framework for DA into large-dimensional systems.

4 OSAS-Like Filtering for Small-Dimensional Systems

As shown in Desbouvries and Ait-El-Fquih (2008), Desbouvries et al. (2011), KF-OSAS is an exact implementation of the generic OSAS filter (12)–(15) in the particular case of linear-Gaussian systems,[5] and PF-OSAS is an approximate implementation in the general case of nonlinear systems.

4.1 The OSAS-Based Kalman Filter (KF-OSAS)

Considering linear dynamics for the state process, i.e., $\mathbf{f}_{n-1}(\mathbf{x}_{n-1}) = \mathbf{F}_{n-1}\mathbf{x}_{n-1}$, this makes the transformed system (25) linear-Gaussian, a case in which the forecast, analysis and smoothing pdfs of interest are Gaussian and are thus entirely described by their first two moments (Desbouvries and Ait-El-Fquih 2008). Let hereafter for any time t_n, $\hat{\mathbf{x}}_n^f$, $\hat{\mathbf{x}}_n^a$ and $\hat{\mathbf{x}}_n^s$ denote the (true) mean respectively of the forecast pdf, $p(\mathbf{x}_n|\mathbf{y}_{0:n-1})$, the analysis pdf, $p(\mathbf{x}_n|\mathbf{y}_{0:n})$, and the smoothing pdf, $p(\mathbf{x}_n|\mathbf{y}_{0:n+1})$. The associated covariances are further denoted by $\mathbf{P}_n^f$, $\mathbf{P}_n^a$ and $\mathbf{P}_n^s$, respectively.

4.1.1 The Observation-Update Step (Smoothing)

The Bayesian step (12) that calculates $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ based on the likelihood (17), reduces to a Kalman-like correction that calculates $(\hat{\mathbf{x}}_{n-1}^s, \mathbf{P}_{n-1}^s)$

[5] Notice that another OSAS-like KF has been introduced in Ait-El-Fquih and Hoteit (2015), but is left out in this chapter. Its difference with the KF-OSAS lies in the fact that (i) it does not follow from the generic filter (12)–(15), but from splitting the state vector using the variational Bayesian approach, and (ii) it only computes an approximation of the analysis and smoothing pdfs.

from $(\hat{\mathbf{x}}^a_{n-1}, \mathbf{P}^a_{n-1})$ based on the observational model (24). Denoting $\tilde{\mathbf{H}}_n = \mathbf{H}_n \mathbf{F}_{n-1}$ one has (Desbouvries and Ait-El-Fquih 2008; Desbouvries et al. 2011):

$$\boldsymbol{\upsilon}_n = \mathbf{y}_n - \tilde{\mathbf{H}}_n \hat{\mathbf{x}}^a_{n-1}, \tag{28}$$

$$\mathbf{V}_n = \tilde{\mathbf{H}}_n \mathbf{P}^a_{n-1} \tilde{\mathbf{H}}_n^T + \tilde{\mathbf{R}}_n, \tag{29}$$

$$\mathbf{K}^s_{n-1} = \mathbf{P}^a_{n-1} \tilde{\mathbf{H}}_n^T \mathbf{V}_n^{-1}, \tag{30}$$

$$\hat{\mathbf{x}}^s_{n-1} = \hat{\mathbf{x}}^a_{n-1} + \mathbf{K}^s_{n-1} \boldsymbol{\upsilon}_n, \tag{31}$$

$$\mathbf{P}^s_{n-1} = \mathbf{P}^a_{n-1} - \mathbf{K}^s_{n-1} \tilde{\mathbf{H}}_n \mathbf{P}^a_{n-1}. \tag{32}$$

KF-OSAS and KF are algorithmically different. However, their observation-update steps share a common structure, and both are driven by the same so-called innovation process $\boldsymbol{\upsilon}_n$ (28) and associated covariance $\mathbf{V}_n$ (29). This can be shown by inserting the forecast equations of the standard KF in (28) and (29), which leads to (Anderson and Moore 1979):

$$\boldsymbol{\upsilon}_n = \mathbf{y}_n - \mathbf{H}_n \hat{\mathbf{x}}^f_n, \tag{33}$$

$$\mathbf{V}_n = \mathbf{H}_n \mathbf{P}^f_n \mathbf{H}_n^T + \mathbf{R}_n, \tag{34}$$

the innovation parameters of the standard KF. The common (Kalman-type) structure of the observation-update steps of the two filters stems from the fact that these steps originate from:

$$p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n}) \overset{\text{smoothing(12)}}{=} \frac{p(\mathbf{x}_{n-1}, \mathbf{y}_n|\mathbf{y}_{0:n-1})}{p(\mathbf{y}_n|\mathbf{y}_{0:n-1})}, \tag{35}$$

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) \overset{\text{standard KF}}{=} \frac{p(\mathbf{x}_n, \mathbf{y}_n|\mathbf{y}_{0:n-1})}{p(\mathbf{y}_n|\mathbf{y}_{0:n-1})}, \tag{36}$$

which share a common Bayesian structure. Furthermore, the common evidence (denominator) in (35) and (36),

$$p(\mathbf{y}_n|\mathbf{y}_{0:n-1}) \propto \mathrm{e}^{-\frac{1}{2}\boldsymbol{\upsilon}_n^T \mathbf{V}_n^{-1} \boldsymbol{\upsilon}_n}, \tag{37}$$

explains why these steps use the same innovation parameters $(\boldsymbol{\upsilon}_n, \mathbf{V}_n)$.

4.1.2 The Time-Update Step (Analysis)

The integration step (14) that calculates $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ based on the conditional transition density (16) reduces to a Kalman-like integration step that computes $(\hat{\mathbf{x}}^a_n, \mathbf{P}^a_n)$ from $(\hat{\mathbf{x}}^s_{n-1}, \mathbf{P}^s_{n-1})$ based on the dynamical model (23). One has (Desbouvries et al. 2011):

$$\hat{\mathbf{x}}_n^a = \mathbf{F}_{n-1}\hat{\mathbf{x}}_{n-1}^s + \underbrace{\mathbf{G}_n(\mathbf{y}_n - \tilde{\mathbf{H}}_n\hat{\mathbf{x}}_{n-1}^s)}_{\boldsymbol{\upsilon}_n}, \tag{38}$$

$$= \tilde{\mathbf{F}}_{n-1}\hat{\mathbf{x}}_{n-1}^s + \mathbf{G}_n\mathbf{y}_n, \tag{39}$$

$$\mathbf{P}_n^a = \tilde{\mathbf{F}}_{n-1}\mathbf{P}_{n-1}^s\tilde{\mathbf{F}}_{n-1}^T + \tilde{\mathbf{Q}}_{n-1}, \tag{40}$$

where $\tilde{\mathbf{F}}_{n-1} = \mathbf{F}_{n-1} - \mathbf{G}_n\tilde{\mathbf{H}}_n$.

As can be seen, the time-update steps of KF and KF-OSAS have a common structure, which consists in a linear propagation of the means and a quadratic propagation of the covariances, based on the original dynamical model (for KF) and the new dynamical model (for KF-OSAS). The aforementioned Kalman-like update in the new dynamical model shows up in the KF-OSAS analysis mean (i.e., term $\boldsymbol{\upsilon}_n$ in (38)). On the other hand, as discussed in Desbouvries et al. (2011), such a term can be expressed as, $\boldsymbol{\upsilon}_n = \tilde{\mathbf{Q}}_{n-1}\mathbf{H}_n^T\mathbf{V}_n^{-1}\boldsymbol{\upsilon}_n$, suggesting that the analysis mean (38), in turn, involves the innovation parameters $(\boldsymbol{\upsilon}_n, \mathbf{V}_n)$, i.e.,

$$\hat{\mathbf{x}}_n^a = \mathbf{F}_{n-1}\hat{\mathbf{x}}_{n-1}^s + \tilde{\mathbf{Q}}_{n-1}\mathbf{H}_n^T\mathbf{V}_n^{-1}\boldsymbol{\upsilon}_n, \tag{41}$$

just as the KF-OSAS smoothing mean (31) and the standard KF analysis mean.

4.2 The OSAS-Based Particle Filter (PF-OSAS)

Consider now a nonlinear dynamical model, $\mathbf{f}_{n-1}(\mathbf{x}_{n-1})$. In such a case, the posterior pdfs are no longer Gaussian and KFs can no longer be applied. PF-OSAS is a MC implementation of the generic algorithm (12), (14) in the transformed system (25), or equivalently the new transition and likelihood laws (16)–(17) (Desbouvries et al. 2011). Let $\overline{\mathbf{u}}$ and $\mathbf{P}_{\mathbf{u}}$ respectively denote the ensemble mean and covariance for an ensemble $\{\mathbf{u}^m\}_m$. Starting from a set of analysis particles $\{\mathbf{x}_{n-1}^{a,m}\}_{m=1}^M$ independently and identically distributed (iid)[6] of $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$, the PF-OSAS computes a sample $\{\mathbf{x}_n^{a,m}\}_{m=1}^M$ of $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ as described below.

4.2.1 The Observation-Update Step (Smoothing)

The Bayesian step (12) that calculates $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ based on the likelihood (17), consists of a weighting step with this likelihood then a resampling step, resulting in a smoothing ensemble $\{\mathbf{x}_{n-1}^{s,m}\}_{m=1}^M$ (of $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$).

[6] Without abuse of language, we use iid throughout the chapter even though this is true only *asymptotically* (in M). Indeed, with finite M, the particles are identically distributed (id) from the associated density, but not independent.

- *Weighting (W) step*: The normalized weight of each particle, $\mathbf{x}_{n-1}^{a,m}$, is computed based on the likelihood (17) as,[7]

$$w_n^{s,m} \propto \underbrace{p(\mathbf{y}_n|\mathbf{x}_{n-1}^{a,m})}_{\mathcal{N}_{\mathbf{y}_n}(\mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{a,m}),\tilde{\mathbf{R}}_n)} \quad ; \quad \text{with} \quad \sum_{m=1}^{M} w_n^{s,m} = 1. \tag{42}$$

- *Resampling (R) step*: The resulting weighted ensemble is next (re)sampled with replacement to obtain the smoothing ensemble, i.e.,

$$\mathbf{x}_{n-1}^{s,m} \sim \sum_{\ell=1}^{M} w_n^{s,\ell}\,\delta(\mathbf{x}_{n-1} - \mathbf{x}_{n-1}^{a,\ell}), \tag{43}$$

where $\delta(. - \mathbf{x}_{n-1}^{a,\ell})$ stands for the Dirac mass at point $\mathbf{x}_{n-1}^{a,\ell}$.

4.2.2 The Time-Update Step (Analysis)

The propagation step (14) that calculates $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ based on the conditional transition density (16), reduces to a sampling, according to this density, of an iid analysis ensemble $\{\mathbf{x}_n^{a,m}\}_{m=1}^M$, given the smoothing one, $\{\mathbf{x}_{n-1}^{s,m}\}_{m=1}^M$. This amounts to integrate $\{\mathbf{x}_{n-1}^{s,m}\}_{m=1}^M$ with the new dynamical operator in (23) as,

$$\mathbf{x}_n^{a,m} = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{s,m}) + \mathbf{G}_n(\mathbf{y}_n - \mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{s,m})) + \tilde{\mathbf{u}}_{n-1}^m, \tag{44}$$

with $\tilde{\mathbf{u}}_{n-1}^m$ being a sample of $\mathcal{N}(\mathbf{0}, \tilde{\mathbf{Q}}_{n-1})$. Once the M particles (44) are sampled, ensemble-based approximations of the analysis PM estimate $\hat{\mathbf{x}}_n^a$ (38) and associated covariance $\mathbf{P}_n^a$ (40) are then obtained as $\overline{\mathbf{x}}_n^a$ and $\mathbf{P}_{\mathbf{x}_n}^a$, respectively.

Remark 1 According to the expression (27) of $\tilde{\mathbf{u}}_{n-1}$, sampling (44) can be implemented differently, in two steps:

- *Pseudo-forecast step*: The smoothing ensemble is first integrated forward in time with the original dynamical operator, which results in a pseudo-forecast ensemble:

$$\boldsymbol{\xi}_n^m = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{s,m}) + \mathbf{u}_{n-1}^m; \quad \mathbf{u}_{n-1}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1}). \tag{45}$$

- *Correction step*: The resulting ensemble is then updated based on perturbed observations, which yields the analysis ensemble:

$$\mathbf{x}_n^{a,m} = \boldsymbol{\xi}_n^m + \mathbf{G}_n(\mathbf{y}_n - \tilde{\mathbf{y}}_n^{f,m}); \tag{46}$$

$$\tilde{\mathbf{y}}_n^{f,m} = \mathbf{H}_n\boldsymbol{\xi}_n^m + \mathbf{v}_n^m, \quad \text{and} \quad \mathbf{v}_n^m \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_n). \tag{47}$$

[7] The term $p(\mathbf{y}_n|\mathbf{x}_{n-1}^{a,m})$ in (42) is not a probability density function, but the value of this function at point $\mathbf{y}_n$ (i.e., the observed data).

As will be seen in the next section, such pseudo-forecast and correction steps are actually nothing but those of EnKF-OSAS analysis step in the particular case of linear-Gaussian systems.

Linking this with the KF-OSAS analysis step in the linear case, it follows that the equations of the mean $\hat{\mathbf{x}}_n^a$ (38) and the members $\mathbf{x}_n^{a,m}$ (46) suggest similar forms. Indeed, both share the same correction mechanism with a gain $\mathbf{G}_n$, with the former updating the mean $\hat{\boldsymbol{\xi}}_n = \mathbf{F}_{n-1}\hat{\mathbf{x}}_{n-1}^s$ based on $\mathbf{y}_n$ and the latter updating the members $\boldsymbol{\xi}_n^m$ based on perturbed $\mathbf{y}_n, \mathbf{y}_n^m = \mathbf{y}_n - \mathbf{v}_n^m$, using the same $\mathbf{G}_n$. Furthermore, averaging (46) yields a similar correction mechanism operating on the empirical means, i.e.,

$$\overline{\mathbf{x}}_n^a = \overline{\boldsymbol{\xi}}_n + \mathbf{G}_n(\mathbf{y}_n - \overline{\tilde{\mathbf{y}}}_n^f), \tag{48}$$

and which is an ensemble-like variant of (38).

4.2.3 Similarities and Differences with Other Standard Filtering Algorithms

- PF-OSAS is a natural implementation (using classical hierarchical and Rubin sampling tools, e.g., Desbouvries et al. 2011) of the "observation-update then time-update" path (12)–(15). The observation-update turns into a weighting-resampling (W, R) step and the time-update turns into a sampling (S) step, thereby leading to a (W, R, S) scheme. In contrast, the (bootstrap) PF, which originates from the "time-update then observation-update" generic path (9)–(11), suggests a (S, W, R) scheme, in which sampling is done before (W, R)-like update. As is well known, the sampling in the PF is performed according to the (prior) transition pdf, $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{a,m})$, and the particles' weights are given by the original likelihood, i.e., $w_n^{a,m} \propto p(\mathbf{y}_n|\mathbf{x}_n^{f,m})$. These laws have the advantage to be available and easy to sample (in particular $p(\mathbf{x}_n|\mathbf{x}_{n-1})$), which explains the widespread popularity of the PF. Nevertheless, in this algorithm, the particles are sampled (in S step) blindly with regard to the observation, $\mathbf{y}_n$, which is indeed exploited solely in the computation of the weights (in step W). In PF-OSAS, however, the observation is exploited explicitly both in W and S steps, which should lead to higher accuracy (as has been numerically shown, e.g. in Desbouvries et al. 2011). Concretely, the information from $\mathbf{y}_n$ that is exploited in W step spreads through R step then the pseudo-forecast step (45), yielding a constrained background, $\{\boldsymbol{\xi}_n^m\}_m$, which, in turn, is linearly updated in S step with $\mathbf{y}_n$, as in (46), to obtain the analysis particles of interest, $\{\mathbf{x}_n^{a,m}\}_m$. As a matter of fact, unlike the PF, the analysis particles are sampled in an optimal way in PF-OSAS, in the sense that they lead to weights with minimum variance conditionally on the observations, $\mathbf{y}_{0:n}$, and past samples, $\mathbf{x}_{0:n-1}^{a,m}$. More precisely, for each trajectory m, the conditional (posterior) transition sampling density, $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{a,m}, \mathbf{y}_n)$ (which is equal to $p(\mathbf{x}_n|\mathbf{x}_{0:n-1}^{a,m}, \mathbf{y}_{0:n})$), leads to weights, w_n^{s,i_m}, with i_m satisfying $\mathbf{x}_{n-1}^{s,i_m} = \mathbf{x}_{n-1}^{a,m}$ (R step), that are deterministic given $(\mathbf{x}_{0:n-1}^{a,m}, \mathbf{y}_{0:n})$ and thus with zero variance.

- PF-OSAS turns out to be a particular case of the well-known auxiliary particle filter (APF) (Pitt and Shephard 1999; Desbouvries et al. 2011). To show that, let us first briefly recall how does APF sample the analysis pdf $p(\mathbf{x}_n|\mathbf{y}_{0:n})$. Replacing in (14) $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ by its expression (12), then $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ by its MC approximation $\sum_{m=1}^{M} 1/M\delta(\mathbf{x}_{n-1}-\mathbf{x}_{n-1}^{a,m})$, one obtains a continuous finite mixture pdf approximating $p(\mathbf{x}_n|\mathbf{y}_{0:n})$:

$$p^c(\mathbf{x}_n|\mathbf{y}_{0:n}) = \sum_{m=1}^{M} w_n^{s,m}\, p(\mathbf{x}_n|\mathbf{x}_{n-1}^{a,m}, \mathbf{y}_n). \tag{49}$$

  If direct sampling (of particles $\mathbf{x}_n^{a,m}$) from p^c is not straightforward, one could approximate p^c by a mixture pdf,

$$q(\mathbf{x}_n) = \sum_{m=1}^{M} \lambda_n^m\, \tilde{q}(\mathbf{x}_n|\mathbf{x}_{n-1}^{a,m}, \mathbf{y}_{0:n}), \tag{50}$$

  sample from q (the so-called first-stage weights λ_n^m and pdf $\tilde{q}$ are degrees of freedom used for designing importance density q), and use importance sampling (IS). Since the target density p^c in (49) and the importance density q in (50) are both mixture densities, the (so-called second-stage) weights computed by this IS method would be the ratio of a sum of M terms over another sum of M terms. APF circumvents this drawback by using data augmentation, which consists in sampling (jointly) both from q and from the index of the mixture in q, and taking the marginal, according to the hierarchical sampling mechanism (e.g., Ait-El-Fquih and Hoteit 2016), i.e., $i_m \sim p(m|\mathbf{y}_{0:n})$, $\mathbf{x}_n^{i_m} \sim q(\mathbf{x}_n|m=i_m)$, then take $\mathbf{x}_n^{i_m}$ as sample from $q(\mathbf{x}_n)$. Based on (50), sampling of i_m and $\mathbf{x}_n^{i_m}$ read $i_m \sim \{m, \lambda_n^m\}_{m=1}^M$ and $\mathbf{x}_n^{i_m} \sim \tilde{q}(\mathbf{x}_n|\mathbf{x}_{n-1}^{a,i_m}, \mathbf{y}_{0:n})$. Now, to link this with PF-OSAS, one can see that this algorithm is a particular case of APF when $q = p^c$ (i.e., $\lambda_n^m = w_n^{s,m}$ and $\tilde{q}(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n})$).
- Looking at the filtering problem in a batch way, one can see that, by construction, each trajectory $\mathbf{z}_{0:n-1}^m$, with $\mathbf{z}_{n-2:n-1}^m = (\mathbf{x}_{n-2}^{s,m}, \mathbf{x}_{n-1}^{a,m})$, is sampled from $p(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1})$. This suggests that PF-OSAS can be interpreted as a batch IS algorithm for which the (joint) target pdf plays the role of importance density, i.e., $q(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1}) = p(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1})$. However, as stated above, the sequential version of this algorithm (i.e., PF-OSAS) involves a scheme (W, R, S), which differs from that of existing sequential IS (SIS, or SIR) algorithms. The key reason of such a difference is that SIR algorithms are founded based on the (sufficient) condition

$$q(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = q(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n}) q(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1}), \tag{51}$$

  whereas, by nature, PF-OSAS is a direct MC implementation of the recursion (remember that $q = p$ in PF-OSAS),

$$q(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n}) \times \underbrace{\left[\frac{p(\mathbf{y}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n-1})}{p(\mathbf{y}_n|\mathbf{y}_{0:n-1})} q(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1})\right]}_{p(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n})}, \quad (52)$$

which involves an extra term, $p(\mathbf{y}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n-1})/p(\mathbf{y}_n|\mathbf{y}_{0:n-1})$, compared to (51). More precisely, PF-OSAS is a reordering of the SIR algorithm with optimal conditional importance density (CID), $q(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_{0:n})$, and with systematic resampling (Doucet et al. 2000; Ades and van Leeuwen 2013). However, these two algorithms are not simply related through a shift in time. Indeed, SIR with CID (SIR-CID) involves a (S, W, R) scheme, or equivalently[8] (W, S, R), suggesting that some (not necessarily all) particles $\mathbf{z}_n^m$ that are sampled in S step with $p(\mathbf{x}_n|\mathbf{z}_{n-1}^m, \mathbf{y}_{0:n})$ can be taken as *approximate* analysis particles, after that W and R steps are achieved (i.e., $\{\mathbf{x}_n^{a,m}\}_{m=1}^M \approx \{\mathbf{z}_n^{i_m}\}_{i_m=1}^M$, with i_m being the indexes sampled in R step). This is clearly different from the (W, R, S)-like PF-OSAS algorithm, in which all particles $\mathbf{z}_n^m$ are *exact* analysis samples (i.e., $\mathbf{x}_n^{a,m} = \mathbf{z}_n^m$, for any m). Furthermore, each old particle $\mathbf{z}_{n-1}^m$ is taken uniformly into account in S step of SIR-CID, whereas only those with high weights do contribute to the updated trajectory in PF-OSAS.

5 OSAS-Like Filtering for Large-Dimensional Systems

In large-dimensional systems, KF-OSAS becomes impractical because of the prohibitive computational cost, $\mathcal{O}(n_{\mathbf{x}}^3)$, required for the calculations of the error covariance matrices. PF-OSAS also becomes impractical as it would require a prohibitive number of particles to overcome the curse of dimensionality. EnKF-OSAS can tackle these issues with reasonable-size ensembles, as it avoids explicit computations of the covariance matrices and mitigates for the curse of dimensionality through an efficient Gaussian sampling of the state-space.

5.1 The OSAS-Based Ensemble Kalman Filter (EnKF-OSAS)

EnKF-OSAS is a Gaussian-based MC implementation of the generic algorithm (12), (14) in the transformed system (25) (Ait-El-Fquih et al. 2016; Raboudi et al. 2018). Similarly to the standard EnKF, its derivation is founded on the assumption that $p(\mathbf{x}_{n-1}, \mathbf{x}_n, \mathbf{y}_n|\mathbf{y}_{0:n-1})$ is Gaussian.

[8] Steps W and S commute in SIR-CID since the weights do not depend on the new particles $\mathbf{z}_n^m$.

5.1.1 The Observation-Update Step (Smoothing)

The Bayesian step (12) that calculates $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$ based on the likelihood (17), reduces to a Kalman-like correction that computes $\{\mathbf{x}_{n-1}^{s,m}\}_{m=1}^{M}$ from $\{\mathbf{x}_{n-1}^{a,m}\}_{m=1}^{M}$ based on the observational model (24) as,

$$\mathbf{x}_{n-1}^{s,m} = \mathbf{x}_{n-1}^{a,m} + \overline{\mathbf{K}}_{n-1}^{s}(\mathbf{y}_n - \mathbf{y}_n^{f,m}), \tag{53}$$

with[9]

$$\mathbf{y}_n^{f,m} = \tilde{\mathbf{h}}_{n-1}(\mathbf{x}_{n-1}^{a,m}) + \tilde{\mathbf{v}}_n^m, \tag{54}$$

and $\tilde{\mathbf{v}}_n^m$ are samples from $\mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}_n)$. Samples $\mathbf{y}_n^{f,m}$ represent the forecast members in the observation space ($\sim p(\mathbf{y}_n|\mathbf{y}_{0:n-1})$), which, based on (19)–(20) and (26), can be also computed from the state forecast members, $\mathbf{x}_n^{f,m}$, in a similar way as the standard EnKF, i.e., (54) is equivalent to set,

$$\mathbf{x}_n^{f,m} = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{a,m}) + \mathbf{u}_{n-1}^m, \tag{55}$$

$$\mathbf{y}_n^{f,m} = \mathbf{H}_n\mathbf{x}_n^{f,m} + \mathbf{v}_n^m, \tag{56}$$

where $\mathbf{u}_{n-1}^m$ and $\mathbf{v}_n^m$ are samples respectively from $\mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1})$ and $\mathcal{N}(\mathbf{0}, \mathbf{R}_n)$. The matrix $\overline{\mathbf{K}}_{n-1}^{s}$ is an ensemble-based approximation of the Kalman gain (30), which can be written as (Ait-El-Fquih and Hoteit 2020),

$$\mathbf{K}_{n-1}^{s} = \mathrm{cov}[\mathbf{x}_{n-1}, \mathbf{y}_n|\mathbf{y}_{0:n-1}] \times \mathrm{cov}[\mathbf{y}_n|\mathbf{y}_{0:n-1}]^{-1}, \tag{57}$$

$$= \mathrm{cov}[\mathbf{x}_{n-1}, \mathbf{x}_n|\mathbf{y}_{0:n-1}]\mathbf{H}_n^T \times \left(\mathbf{H}_n\mathrm{cov}[\mathbf{x}_n|\mathbf{y}_{0:n-1}]\mathbf{H}_n^T + \mathrm{cov}[\mathbf{v}_n]\right)^{-1}, \tag{58}$$

$$= \mathrm{cov}[\mathbf{x}_{n-1}, \mathbf{r}_n|\mathbf{y}_{0:n-1}]\mathbf{H}_n^T \times \left(\mathbf{H}_n\mathrm{cov}[\mathbf{r}_n|\mathbf{y}_{0:n-1}]\mathbf{H}_n^T + \mathrm{cov}[\tilde{\mathbf{v}}_n]\right)^{-1}, \tag{59}$$

with $\mathbf{r}_n = \mathbf{f}_{n-1}(\mathbf{x}_{n-1})$, and $\mathrm{cov}[\boldsymbol{\xi}]$ and $\mathrm{cov}[\boldsymbol{\xi}, \boldsymbol{\mu}]$ respectively denote the covariance of $\boldsymbol{\xi}$ and the cross-covariance between $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$. More precisely, expressions (58) and (59) coincide with (30) using the forms (34) and (29), respectively, of the innovation matrix, $\mathbf{V}_n$. Denoting by $\mathbf{P}_{\boldsymbol{\xi},\boldsymbol{\mu}}$ the sample cross-covariance between ensembles $\{\boldsymbol{\xi}^m\}_{m=1}^{M}$ and $\{\boldsymbol{\mu}^m\}_{m=1}^{M}$, $\overline{\mathbf{K}}_{n-1}^{s}$ can be written as

$$\overline{\mathbf{K}}_{n-1}^{s} = \mathbf{P}_{\mathbf{x}_{n-1}^a, \mathbf{y}_n^f}\mathbf{P}_{\mathbf{y}_n^f}^{-1}, \tag{60}$$

if (57) is used,

[9] It follows from (54) that $\mathbf{y}_n - \mathbf{y}_n^{f,m} = \mathbf{y}_n - \tilde{\mathbf{v}}_n^m - \tilde{\mathbf{h}}_{n-1}(\mathbf{x}_{n-1}^{a,m})$, suggesting that the Kalman correction in (53) is based on observations perturbed with $-\tilde{\mathbf{v}}_n^m \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}_n)$. Similarly to the classical EnKF, perturbing the observations for each ensemble member enables matching with the error statistics in the KF-OSAS smoothing step when the state-space system is linear-Gaussian.

$$\overline{\mathbf{K}}^s_{n-1} = \mathbf{P}_{\mathbf{x}^a_{n-1},\mathbf{x}^f_n}\mathbf{H}_n^T \times \left(\mathbf{H}_n\mathbf{P}_{\mathbf{x}^f_n}\mathbf{H}_n^T + \mathbf{R}_n\right)^{-1}, \tag{61}$$

if (58) is used, or as

$$\overline{\mathbf{K}}^s_{n-1} = \mathbf{P}_{\mathbf{x}^a_{n-1},\mathbf{r}_n}\mathbf{H}_n^T \times \left(\mathbf{H}_n\mathbf{P}_{\mathbf{r}_n}\mathbf{H}_n^T + \tilde{\mathbf{R}}_n\right)^{-1}, \tag{62}$$

with $\mathbf{r}^m_n = \mathbf{f}_{n-1}(\mathbf{x}^{a,m}_{n-1})$, if (59) is used instead. The expressions (61) and (62) are generally more suitable in practice than (60) because they involve less sampling errors, especially when dealing with small ensembles. This is because $\mathrm{cov}[\mathbf{x}_n, \mathbf{v}_n|\mathbf{y}_{0:n-1}]$ and $\mathrm{cov}[\mathbf{v}_n]$ are set in (61) and (62) to their exact values, $\mathbf{0}$ and $\mathbf{R}_n$, respectively, while (60) rather uses ensemble-based estimates of these quantities. In turn, (62) should be, to some extent, more suitable than (61), as the former (inherently) uses the true value of $\mathrm{cov}[\mathbf{u}_{n-1}]$, $\mathbf{Q}_{n-1}$, whereas the latter uses an ensemble-based approximation of it through the samples $\mathbf{u}^m_{n-1}$, that are used to generate the forecast members $\mathbf{x}^{f,m}_n$ (see (55): $\mathbf{x}^{f,m}_n = \mathbf{r}^m_n + \mathbf{u}^m_{n-1}$).

In large-dimensional systems, the explicit computation of the $n_\mathbf{x} \times n_\mathbf{x}$ ensemble-based covariance matrices $\mathbf{P}_{\mathbf{x}^a_{n-1},\mathbf{x}^f_n}$ and $\mathbf{P}_{\mathbf{x}^f_n}$ in (61), or $\mathbf{P}_{\mathbf{x}^a_{n-1},\mathbf{r}_n}$ and $\mathbf{P}_{\mathbf{r}_n}$ in (62), is not possible. One could instead follow the classical EnKF and use the $n_\mathbf{x} \times M$ (square-root) ensemble perturbation matrices of these (cross-)covariances, to respectively compute (61) and (62) (e.g., Hoteit et al. 2018).

Once the M members (53) are sampled, an ensemble-based approximation of the smoothing PM estimate $\hat{\mathbf{x}}^s_{n-1}$ (31) can then be computed as $\overline{\mathbf{x}}^s_{n-1}$; the associated error covariance $\mathbf{P}^s_{n-1}$ could be also estimated, if the computational cost allows it.

Remark 2 Similarities arise between the expressions of the mean $\hat{\mathbf{x}}^s_{n-1}$ (31) and members $\mathbf{x}^{s,m}_{n-1}$ (53). Indeed, both share the same Kalman-like mechanism, with the former updating the mean $\hat{\mathbf{x}}^a_{n-1}$ based on $\mathbf{y}_n$ and using the gain $\mathbf{K}^s_{n-1}$, whereas the latter updating the members $\mathbf{x}^{a,m}_{n-1}$ based on perturbations of $\mathbf{y}_n$, $\mathbf{y}^m_n = \mathbf{y}_n - \tilde{\mathbf{v}}^m_n$, and using an ensemble approximation of $\mathbf{K}^s_{n-1}$, $\overline{\mathbf{K}}^s_{n-1}$. As such, similarly to $\hat{\mathbf{x}}^s_{n-1}$, the updates $\mathbf{x}^{s,m}_{n-1}$ are nudged via the innovation members, $\boldsymbol{\upsilon}^m_n = \mathbf{y}^m_n - \mathbf{H}_n\mathbf{r}^m_n$, and associated covariance[10] $\mathbf{P}_{\boldsymbol{\upsilon}_n}$, i.e., for $m = 1, \cdots, M$,

$$\mathbf{x}^{s,m}_{n-1} = \mathbf{x}^{a,m}_{n-1} + \overline{\mathbf{K}}^s_{n-1}\boldsymbol{\upsilon}^m_n. \tag{63}$$

Furthermore, averaging (63) leads to a similar process operating on the empirical means (i.e., involving $\bar{\cdot}$ instead of $\cdot^m$), which is an ensemble-like implementation of (31). Finally, one may notice that the same reasoning holds for the analysis steps of the standard KF and EnKFs; in particular,

[10] Note that members $\boldsymbol{\upsilon}^m_n$ correspond to (28) and their covariance, $\mathbf{P}_{\boldsymbol{\upsilon}_n}$, is an ensemble-based approximation of the form (29) of the innovation matrix $\mathbf{V}_n$. One could also perturb the observations by $\mathbf{v}^m_n$ instead of $\tilde{\mathbf{v}}^m_n$ (i.e., $\mathbf{y}^m_n$ become $\mathbf{y}^m_n = \mathbf{y}_n - \mathbf{v}^m_n$); in this case the members $\boldsymbol{\upsilon}^m_n$ become $\boldsymbol{\upsilon}^m_n = \mathbf{y}^m_n - \mathbf{H}_n\mathbf{x}^{f,m}_n$ (which correspond to the form (33) of the innovation) and their covariance, $\mathbf{P}_{\boldsymbol{\upsilon}_n}$, corresponds to the form (34) of $\mathbf{V}_n$.

$$\mathbf{x}_n^{a,m} \overset{\text{EnKF}}{=} \mathbf{x}_n^{f,m} + \underbrace{\mathbf{P}_{\mathbf{x}_n^f}\mathbf{H}_n^T \times \mathbf{P}_{\boldsymbol{\upsilon}_n}^{-1}}_{\overline{\mathbf{K}}_n^a} \times \boldsymbol{\upsilon}_n^m, \tag{64}$$

$$\overline{\mathbf{x}}_n^a \overset{\text{EnKF}}{=} \overline{\mathbf{x}}_n^f + \overline{\mathbf{K}}_n^a \times \overline{\boldsymbol{\upsilon}}_n, \tag{65}$$

$$\hat{\mathbf{x}}_n^a \overset{\text{KF}}{=} \hat{\mathbf{x}}_n^f + \underbrace{\mathbf{P}_n^f\mathbf{H}_n^T \times \mathbf{V}_n^{-1}}_{\mathbf{K}_n^a} \times \boldsymbol{\upsilon}_n. \tag{66}$$

5.1.2 The Time-Update Step (Analysis)

As discussed in Remark 1, the integration step (14) that calculates $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n})$ based on the conditional transition density $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{y}_n)$, given in (16), can be reduced into a two-step sampling (pseudo-forecast and correction steps (45)–(47)) of an analysis ensemble $\{\mathbf{x}_n^{a,m}\}_{m=1}^M$ from the smoothing one, $\{\mathbf{x}_{n-1}^{s,m}\}_{m=1}^M$. This amounts to sample the members $\mathbf{x}_n^{a,m}$ from $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_n)$ as in (44), which can be also directly obtained via Property 1 in Appendix. However, unlike in Remark 1, the state dimension $n_\mathbf{x}$ is assumed here to be large, making the computation cost of the gain $\mathbf{G}_n$ (21) prohibitive. One way to avoid this problem is to seek a different sampling strategy from $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_n)$ to avoid the explicit computation of $\mathbf{G}_n$. Noticing that $\boldsymbol{\xi}_n^m$ (45) and $\tilde{\mathbf{y}}_n^{f,m}$ (47) are respectively samples of $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1})$ and $p(\mathbf{y}_n|\mathbf{x}_{n-1}^{s,m}) = p(\mathbf{y}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1})$, one can use Property 2 in Appendix to show that

$$\mathbf{x}_n^{a,m} = \boldsymbol{\xi}_n^m + \underbrace{\text{cov}[\mathbf{x}_n, \mathbf{y}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1}]\,\text{cov}[\mathbf{y}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1}]^{-1}}_{\mathbf{L}_n} \left(\mathbf{y}_n - \tilde{\mathbf{y}}_n^{f,m}\right), \tag{67}$$

are samples of $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n})$, and then Property 1 to show that these are also samples of $p(\mathbf{x}_n|\mathbf{y}_{0:n})$, i.e., analysis members. In practice, for each member m, the (cross-)covariances in (67) can be estimated based on ensembles, $\{\boldsymbol{\xi}_n^{m,j}\}_{j=1}^J$ and $\{\tilde{\mathbf{y}}_n^{f,m,j}\}_{j=1}^J$, respectively sampled from $p(\mathbf{x}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1})$ and $p(\mathbf{y}_n|\mathbf{x}_{n-1}^{s,m}, \mathbf{y}_{0:n-1})$ as,

$$\boldsymbol{\xi}_n^{m,j} = \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{s,m}) + \mathbf{u}_{n-1}^j, \tag{68}$$

$$\begin{aligned}\tilde{\mathbf{y}}_n^{f,m,j} &= \mathbf{H}_n\boldsymbol{\xi}_n^{m,j} + \mathbf{v}_n^j,\\ &= \mathbf{H}_n\mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{s,m}) + \underbrace{\mathbf{H}_n\mathbf{u}_{n-1}^j + \mathbf{v}_n^j}_{\tilde{\mathbf{v}}_n^j},\end{aligned} \tag{69}$$

with $\mathbf{u}_{n-1}^j \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1})$ and $\mathbf{v}_n^j \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_n)$. The gain $\mathbf{L}_n$ can be then approximated by,

$$\overline{\mathbf{L}}_n = \overline{\mathbf{P}}_{\boldsymbol{\xi}_n, \tilde{\mathbf{y}}_n^f}\overline{\mathbf{P}}_{\tilde{\mathbf{y}}_n^f}^{-1} = \overline{\mathbf{P}}_{\mathbf{u}_{n-1}, \tilde{\mathbf{v}}_n}\overline{\mathbf{P}}_{\tilde{\mathbf{v}}_n}^{-1}, \tag{70}$$

where the sample (cross-)covariances $\overline{\mathbf{P}}_{\boldsymbol{\xi}_n,\tilde{\mathbf{y}}_n^f}$ and $\overline{\mathbf{P}}_{\tilde{\mathbf{y}}_n^f}$, both independent of m, are computed from $\{\boldsymbol{\xi}_n^{m,j}\}_{j=1}^J$ and $\{\tilde{\mathbf{y}}_n^{f,m,j}\}_{j=1}^J$, while $\overline{\mathbf{P}}_{\mathbf{u}_{n-1},\tilde{\mathbf{v}}_n}$ and $\overline{\mathbf{P}}_{\tilde{\mathbf{v}}_n}$ are computed from $\{\mathbf{u}_{n-1}^j\}_{j=1}^J$ and $\{\tilde{\mathbf{v}}_n^j\}_{j=1}^J$. As such, $\mathbf{L}_n$ is equal to the KF-OSAS gain $\mathbf{G}_n$, as can be shown from (70) when J goes to infinity. Therefore, the discussion about the similarities with the KF-OSAS in Sect. 4.2.2 remains valid here.

After sampling the M members $\mathbf{x}_n^{a,m}$, one can compute ensemble-based estimates of the analysis PM estimate $\hat{\mathbf{x}}_n^a$ (38) and eventually its associated error covariance $\mathbf{P}_n^a$.

Remark 3 (Ait-El-Fquih et al. 2016; Raboudi et al. 2018) approximated $\mathbf{L}_n$ by $\tilde{\mathbf{L}}_n = \mathbf{P}_{\boldsymbol{\xi}_n,\tilde{\mathbf{y}}_n^f}\mathbf{P}_{\tilde{\mathbf{y}}_n^f}^{-1}$, based on the ensembles $\{\boldsymbol{\xi}_n^m\}_{m=1}^M$ (45) and $\{\tilde{\mathbf{y}}_n^{f,m}\}_{m=1}^M$ (47), instead of (70). The analysis step then becomes,

$$\mathbf{x}_n^{a,m} = \boldsymbol{\xi}_n^m + \tilde{\mathbf{L}}_n\left(\mathbf{y}_n - \tilde{\mathbf{y}}_n^{f,m}\right); \quad m = 1, \cdots, M. \tag{71}$$

While in the linear-Gaussian systems the gain $\tilde{\mathbf{L}}_n$ coincides with $\mathbf{L}_n$, in the limit of large ensembles, its derivation in the nonlinear systems considers the members $\mathbf{x}_{n-1}^{s,m}$ as samples from $p(\mathbf{x}_{n-1}|\mathbf{y}_{0:n-1})$, which holds under the assumption[11] that $\mathbf{x}_{n-1}$ and $\mathbf{y}_n$ are independent conditionally on $\mathbf{y}_{0:n-1}$. Despite its empirical nature in the nonlinear systems, this analysis step may have some practical advantages in realistic large-scale DA problems. In such applications, the state noise covariance, $\mathbf{Q}_{n-1}$, is usually not known and often ignored. If so, the correction term in (67) vanishes (as $\mathbf{L}_n = \mathbf{0}$), and therefore this second update step will no longer be needed. One may recover (71) (Ait-El-Fquih et al. 2016; Raboudi et al. 2018) by taking $\mathbf{Q}_{n-1}$ as a fraction of the ensemble covariance matrix of $\{\boldsymbol{\xi}_n^m\}_m$, $\mathbf{P}_{\boldsymbol{\xi}_n}$, which could be considered as some sort of parameterizing the state noise with an inflation term (Pham et al. 1998). This form of the second update step was indeed shown to improve upon the performances of the standard EnKFs especially in challenging settings (e.g. strong nonlinearities, poorly known state noises, small ensembles, sparse observations, etc.). It may further provide a good approximation in the case of a non-additive state noise that is accounted for through sampled noise in the dynamical operator. Moreover, and as argued by Pham et al. (1998), even in the situation of a perfect dynamical operator (free of state noise), it is always beneficial to err on the safe side and add some state noise, which could partly account for various approximations in the filtering process.

5.1.3 Discussion

In the context of large-dimensional applications, the EnKF-OSAS suggests an efficient Gaussian-based sampling scheme of the analysis pdf of interest, as well as the forecast and OSAS ones. Given the analysis ensemble, $\{\mathbf{x}_{n-1}^{a,m}\}_m$, the smoothing step starts by computing the forecast ensemble, $\{\mathbf{x}_n^{f,m}\}_m$, as in (55), then updating

[11] This assumption was forgotten in Ait-El-Fquih et al. (2016), Raboudi et al. (2018).

it based on the current observation, $\mathbf{y}_n$, following a Kalman-like correction (53), to obtain the smoothing ensemble, $\{\mathbf{x}_{n-1}^{s,m}\}_m$. The analysis step then takes place, starting from an integration of the resulting smoothing members in the (original) dynamical operator as in (45) to obtain a pseudo-forecast ensemble, $\{\boldsymbol{\xi}_n^m\}_m$; the latter is then updated based on the same observation following a Kalman-like correction as in (71), which results in the analysis ensemble of interest, $\{\mathbf{x}_n^{a,m}\}_m$. Thus, compared to the standard EnKF, the EnKF-OSAS not only introduces another Kalman-like update of the state, but also involves a (new) smoothing step that constraints the ensemble sampling with the future observation. This should provide an improved background ensemble, $\{\boldsymbol{\xi}_n^m\}_m$, in the analysis step, as it includes more information in the estimation process that may help mitigating for the suboptimal character of the EnKF, usually implemented with limited ensembles and crude noise statistics (see e.g. (Raboudi et al. 2018) which investigated the benefit of the improved background, $\{\boldsymbol{\xi}_n^m\}_m$, with extensive numerical experiments).

In terms of computational complexity, the EnKF-OSAS is roughly twice more expensive than the EnKF, as it involves two "forecast" steps with the (original) dynamical model (to compute $\{\mathbf{x}_n^{f,m}\}_m$ and $\{\boldsymbol{\xi}_n^m\}_m$), and two Kalman-like update steps (to compute $\{\mathbf{x}_{n-1}^{s,m}\}_m$ and $\{\mathbf{x}_n^{a,m}\}_m$). Nevertheless, results of numerical experiments have suggested that exploiting the information in the observation more efficiently, enhances the ensemble filter's performance even in situations in which both schemes are implemented at comparable computational costs (Raboudi et al. 2018).

Deterministic OSAS-like ensemble filters have been also recently introduced. The goal is to build efficient schemes that combine the benefits of the OSAS-like filtering approach in exploiting the data twice to obtain improved backgrounds, with those of the deterministic EnKF-formulation in avoiding the undersampling of observational errors, that typically arise in the stochastic ensemble formulation when dealing with small ensembles (Bishop et al. 2001; Anderson 2001; Tippett et al. 2003; Hoteit et al. 2015). In particular, two deterministic schemes have been derived, a singular interpolated ensemble Kalman-like filter (SEIK-OSAS) (Raboudi et al. 2018), and an ensemble transform Kalman-like filter (ETKF-OSAS) (Raboudi et al. 2019), generating smoothing and analysis ensembles deterministically, following moments matching procedures. These algorithms, in which the observations are not perturbed, were derived based on square-root formulations of the KF-OSAS, under the assumption that the pseudo-forecast error $(\mathbf{x}_n - \hat{\boldsymbol{\xi}}_n)$ is uncorrelated with the observation noise $\mathbf{v}_n$, and following exactly the same reasoning with which their standard counterparts were derived from the standard KF.

5.2 *State-Parameters Estimation with OSAS-Based Ensemble Filtering*

Dynamical models are often functions of poorly known parameters, $\boldsymbol{\theta}$, and are therefore expressed here as, $\mathbf{f}_{n-1}(\mathbf{x}_{n-1}, \boldsymbol{\theta})$. Estimating the values of these parameters along

with the state should provide improved filtering results, and is therefore of great interest in many applications. Several studies argued that concatenating the state and the parameters as, $\boldsymbol{\vartheta}_n = [\mathbf{x}_n^T, \boldsymbol{\theta}^T]^T$, then applying the EnKF on the augmented system $(\boldsymbol{\vartheta}, \mathbf{y})$, the so-called joint-EnKF, can lead to inconsistency between the estimated state and parameters that could degrade the filter performance (e.g., Moradkhani et al. 2005; Chen and Zhang 2006; Wen and Chen 2007). One classical technique that has been proposed to tackle this issue is the so-called dual EnKF, which successively updates the parameters and the state using two interactive (sub-)EnKFs, one acting on the parameters and the other on the state conditioned on the solution of the former (Moradkhani et al. 2005). In other words, the state filter uses in its forecast step the output of the parameters' filter, $\boldsymbol{\theta}_n^{a,m}$ ($\sim p(\boldsymbol{\theta}|\mathbf{y}_{0:n})$), and not $\boldsymbol{\theta}_{n-1}^{a,m}$ as does the joint EnKF, i.e.,

$$\tilde{\mathbf{x}}_n^{f,m} \overset{\text{dual EnKF}}{=} \mathbf{f}_{n-1}(\mathbf{x}_{n-1}^{a,m}, \boldsymbol{\theta}_n^{a,m}) + \mathbf{u}_{n-1}^m; \quad \mathbf{u}_{n-1}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1}). \tag{72}$$

This suggests that updating $\tilde{\mathbf{x}}_n^{f,m}$ in the analysis step, following a Kalman-like update, to obtain $\mathbf{x}_n^{a,m}$, leads to a two fold exploitation of $\mathbf{y}_n$. Such a dual update was indeed proven to provide more accurate state and parameters' estimates than the joint EnKF (Gharamti et al. 2014). However, the dual EnKF has been introduced as a heuristic scheme and is not consistent with the Bayesian filtering framework (Hendricks Franssen and Kinzelbach 2008). The so-called joint EnKF-OSAS of Gharamti et al. (2015) has thus been proposed in an attempt to built a Bayesian consistent dual-like filter, and was derived in the context of the augmented system $(\boldsymbol{\vartheta}, \mathbf{y})$, under the assumption,

$$p(\mathbf{x}_n|\boldsymbol{\vartheta}_{n-1}, \mathbf{y}_n) = p(\mathbf{x}_n|\boldsymbol{\vartheta}_{n-1}), \tag{73}$$

the goal of which is to avoid the Kalman-like update in the time-update (analysis) step. As such, the joint EnKF-OSAS corresponds to the above EnKF-OSAS algorithm (Sect. 5.1), when replacing $\mathbf{x}$ by $\boldsymbol{\vartheta}$ and ignoring the second Kalman-like update, in which case $\mathbf{x}_n^{a,m} = \boldsymbol{\xi}_n^m$.

The joint EnKF-OSAS shares the same parameters' filter with the standard dual EnKF, but has a different state filter involving, among others, an additional Kalman-like update of the state (i.e., smoothing of $\mathbf{x}_{n-1}$ based on $\mathbf{y}_n$). More precisely, the dual EnKF integrates first the members $(\mathbf{x}_{n-1}^{a,m}, \boldsymbol{\theta}_n^{a,m})$ with the dynamical operator (as in (72)) before updating the resulting members $\tilde{\mathbf{x}}_n^{f,m}$ with the observation $\mathbf{y}_n$, whereas the joint EnKF-OSAS updates first the members $\mathbf{x}_{n-1}^{a,m}$, which provides the smoothing samples $\mathbf{x}_{n-1}^{s,m}$, before integrating $(\mathbf{x}_{n-1}^{s,m}, \boldsymbol{\theta}_n^{a,m})$ with the dynamical operator. Motivated by the good results provided by the joint EnKF-OSAS, a more general algorithm, a dual EnKF-OSAS, has subsequently been introduced in Ait-El-Fquih et al. (2016) (see also Khaki et al. 2020) by relaxing the assumption (73) and assuming instead

$$p(\boldsymbol{\vartheta}_{n-1}|\mathbf{y}_{0:n}) = p(\boldsymbol{\vartheta}_{n-1}|\mathbf{y}_{0:n-1}) \tag{74}$$

in the derivation step of the Kalman gain when the system is nonlinear (see Remark 3); it actually coincides with the EnKF-OSAS (in Remark 3) when $\mathbf{x}$ is replaced by $\boldsymbol{\vartheta}$. In the dual EnKF-OSAS, accounting for the conditional dependence between $\mathbf{x}_n$ and $\mathbf{y}_n$ given $\boldsymbol{\vartheta}_{n-1}$ (by relaxing (73)), introduces one more Kalman-like update of the state (in the analysis step), even under (74), which should generally enhance the estimation accuracy compared to the joint EnKF-OSAS.

6 Summary

This chapter reviewed the one-step-ahead smoothing (OSAS) formulation of the Bayesian filtering problem both for small- and large-dimensional state-space systems. After describing the standard filtering algorithms, which involve a "time-update (forecast) then observation-update (analysis)" path at each assimilation cycle, we presented an overview of alternative OSAS-like filters, which are based on the reversed "observation-update then time-update" path. The new formulation involves two Bayesian-like update steps based on the same (current) observation: one for smoothing the previous state and one for analyzing the current one. The OSAS-like filters, including new variants of Kalman filters (KF-OSAS), particle filters (PF-OSAS) and ensemble Kalman filters (EnKF-OSAS), depending on the size and the linear-Gaussian character of the underlying state-space system, were then derived and reviewed. While the standard KF and KF-OSAS provide the same (exact) estimator, exploiting the data twice in the estimation process generally leads to a better trade-off between estimation quality and computational burden for the PF-OSAS and EnKF-OSAS, compared to their standard counterparts. Deterministic variants of EnKF-OSAS and extensions to the more general state-parameters' filtering problem were also discussed.

Future work will investigate the unsupervised framework in which one or some hyper-(statistical) parameters of the system, as for instance the observational or dynamical noise statistics, are not well known. Another interesting direction is to explore the use of the OSAS-based ensemble strategy in the context of Gaussian mixture filtering, which was suggested to be suitable for strongly nonlinear large-dimensional models.

Appendix

Property 1 (Hierarchical sampling Ait-El-Fquih et al. 2016) Assuming that one can sample from $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2|\mathbf{x}_1)$, then a sample, $\mathbf{x}_2^*$, from $p(\mathbf{x}_2)$ can be drawn as follows:

1. $\mathbf{x}_1^* \sim p(\mathbf{x}_1)$;
2. $\mathbf{x}_2^* \sim p(\mathbf{x}_2|\mathbf{x}_1^*)$.

Property 2 (Conditional sampling Ait-El-Fquih et al. 2016) Consider a Gaussian pdf, $p(\mathbf{x}, \mathbf{y})$, with $\mathbf{P}_{xy}$ and $\mathbf{P}_y$ denoting the cross-covariance of $\mathbf{x}$ and $\mathbf{y}$ and the covariance of $\mathbf{y}$, respectively. Then a sample, $\mathbf{x}^*$, from $p(\mathbf{x}|\mathbf{y})$, can be drawn as follows:

1. $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \sim p(\mathbf{x}, \mathbf{y})$;
2. $\mathbf{x}^* = \tilde{\mathbf{x}} + \mathbf{P}_{xy}\mathbf{P}_y^{-1}[\mathbf{y} - \tilde{\mathbf{y}}]$.

References

Ades M, van Leeuwen PJ (2013) An exploration of the equivalent weights particle filter. Q J R Meteorol Soc 139:820–840

Ait-El-Fquih B, Desbouvries F (2006) Kalman filtering in triplet Markov chains. IEEE Trans Signal Process 54(8):2957–2963

Ait-El-Fquih B, Hoteit I (2015) Fast Kalman-like filtering in large-dimensional linear and Gaussian state-space models. IEEE Trans Signal Process 63(21):5853–5867

Ait-El-Fquih B, Hoteit I (2016) A variational Bayesian multiple particle filtering scheme for large-dimensional systems. IEEE Trans Signal Process 64(20):5409–5422

Ait-El-Fquih B, Hoteit I (2018) An efficient state-parameter filtering scheme combining ensemble Kalman and particle filters. Mon Weather Rev 146:871–887

Ait-El-Fquih B, Hoteit I (2020) A particle filter-based adaptive inflation scheme for the ensemble Kalman filter. Q J R Meteorol Soc 146:922–937

Ait-El-Fquih B, Gharamti ME, Hoteit I (2016) A Bayesian consistent dual ensemble Kalman filter for state-parameter estimation in subsurface hydrology. Hydrol Earth Syst Sci 20:3289–3307

Ait-El-Fquih B, Hoteit I (2015) An efficient multiple particle filter based on the variational Bayesian approach. In: Proceedings of the IEEE international ISSPIT symposium

Aksoy A, Zhang F, Nielsen-Gammon JW (2006) Ensemble-based simultaneous state and parameter estimation with MM5. Geophys Res Lett 33:L12801

Anderson JL (2001) An ensemble adjustment Kalman filter for data assimilation. Mon Weather Rev 129:2884–2903

Anderson JL (2009) Ensemble Kalman filters for large geophysical applications. IEEE Control Syst Mag 29(3):66–82

Anderson BDO, Moore JB (1979) Optimal filtering. Prentice Hall, Englewood Cliffs, New Jersey

Annan JD, Lunt DJ, Hargreaves JC, Valdes PJ (2005) Parameter estimation in an atmospheric GCM using the ensemble Kalman filter. Nonlinear Process Geophys 12:363–371

Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50(2):174–188

Asch M, Bocquet M, Nodet M (2016) Data assimilation: methods, algorithms, and applications. SIAM

Bar-Shalom Y, Li X, Kirubarajan T (2001) Estimation with applications to tracking and navigation. Wiley, New York

Bellsky T, Berwald J, Mitchell L (2014) Nonglobal parameter estimation using local ensemble Kalman filtering. Mon Weather Rev 142:2150–2164

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman Filter. Part I: Theoretical aspects. Mon Weather Rev 129:420–436

Burgers G, van Leeuwen PJ, Evensen G (1998) Analysis scheme in the ensemble Kalman filter. Mon Weather Rev 126:1719–1724

Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer

Carrassi A, Bocquet M, Bertino L, Evensen G (2018) Data assimilation in the geosciences: an overview of methods, issues, and perspectives. In: Wiley interdisciplinary reviews: climate change, p e535

Chen Y, Zhang D (2006) Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv Water Resour 29:1107–1122

Chui CK, Chen G (1999) Kalman filtering with real-time applications. Springer, Berlin, DE

Crisan D, Doucet A (2002) A survey on convergence results on particle filtering methods for practitioners. IEEE Trans Signal Process 50(3):736–746

Desbouvries F, Petetin Y, Ait-El-Fquih B (2011) Direct, prediction- and smoothing-based Kalman and particle filter algorithms. Signal Process 91(8):2064–2077

Desbouvries F, Ait-El-Fquih B (2008) Direct, prediction-based and smoothing-based particle filter algorithms. In: Proceedings of the international conference on IASC

Djuric PM, Bugallo MF (2013) Particle filtering for high-dimensional systems. In: Proceedings of the IEEE international workshop on CAMSAP

Djuric PM, Zhang J, Ghirmai T, Huang Y, Kotecha JH (2002) Applications of particle filtering to communications: a review. In: Proceedings of the European conference on EUSIPCO

Doucet A, Godsill SJ, Andrieu C (2000) On sequential monte Carlo sampling methods for Bayesian filtering. Stat Comput 10:197–208

Doucet A, de Freitas N, Gordon N (eds) (2001) Sequential Monte Carlo methods in practice. Statistics for engineering and information science. Springer, New York

Dreano D, Tandeo P, Pulido M, Ait-El-Fquih B, Chonavel T, Hoteit I (2017) Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation-maximization algorithm. Q J R Meteorol Soc 143:1877–1885

Dunne SC, Entekhabi D, Njoku EG (2007) Impact of multiresolution active and passive microwave measurements on soil moisture estimation using the ensemble Kalman smoother. IEEE Trans Geosci Remote Sens 45(4):1016–1028

Durrant-Whyte H, Bailey T (2006) Simultaneous localisation and mapping (SLAM): Part I: The essential algorithms. IEEE Robot Autom Mag 13:99–110

Ephraim Y, Merhav N (2002) Hidden Markov processes. IEEE Trans Inf Theory 48(6):1518–1569

Evensen G (1994) Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99(C5):143–162

Evensen G (2006) Data assimilation: the ensemble Kalman filter. Springer, New York

Evensen G, van Leeuwen PJ (2000) An ensemble Kalman smoother for nonlinear dynamics. Mon Weather Rev 128(6):1852–1867

Frei M, Künsch H (2013) Bridging the ensemble Kalman and particle filters. Biometrika 100:781–800

Furrer R, Bengtsson T (2007) Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J Multivar Anal 98(2):227–255

Gharamti ME, Valstar J, Hoteit I (2014) An adaptive hybrid EnKF-OI scheme for efficient state-parameter estimation of reactive contaminant transport models. Adv Water Resour 71:1–15

Gharamti ME, Kadoura A, Valstar J, Sun S, Hoteit I (2014) Constraining a compositional flow model with flow-chemical data using an ensemble-based Kalman filter. Water Resour Res 50:2444–2467

Gharamti ME, Ait-El-Fquih B, Hoteit I (2015) An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. J Hydrol 527:442–457

Gordon NJ, Salmond DJ, Smith AFM (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc F 140:107–113

Gustafsson F (2010) Particle filter theory and practice with positioning applications. IEEE Trans Aerosp Electron Syst Mag 25(7):53–82

Harvey AC (1989) Forecasting, structural time series models and the Kalman filter. Cambridge University Press

Haykin S, Huber K, Chen Z (2004) Bayesian sequential state estimation for MIMO wireless communications. Proc IEEE 92(3):439–454

Hendricks Franssen HJ, Kinzelbach W (2008) Real-time groundwater flow modeling with the ensemble Kalman filter: joint estimation of states and parameters and the filter inbreeding problem. Water Resour Res 44(9)

Hoteit I et al (2020) Towards an end-to-end analysis and prediction system for weather, climate, and marine applications in the red sea. Bull Am Meteorol Soc. https://doi.org/10.1175/BAMS-D-19-0005.1

Hoteit I, Pham D-T, Blum J (2002) A simplified reduced order Kalman filtering and application to altimetric data assimilation in tropical pacific. J Mar Syst 36(1–2):101–127

Hoteit I, Pham D-T, Triantafyllou G, Korres G (2008) A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. Mon Weather Rev 136(1):317–334

Hoteit I, Luo X, Pham D-T (2012) Particle Kalman filtering: a nonlinear Bayesian framework for ensemble Kalman filters. Mon Weather Rev 140:528–542

Hoteit I, Pham D-T, Gharamti ME, Luo X (2015) Mitigating observation perturbation sampling errors in the stochastic EnKF. Mon Weather Rev 143:2918–2936

Hoteit I, Luo X, Bocquet M, Köhl A, Ait-El-Fquih B (2018) New frontiers in operational oceanography. In: Data assimilation in oceanography: current status and new directions. GODAE Ocean View, pp 465–512

Houtekamer PL, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. Mon Weather Rev 126:796–811

Houtekamer PL, Mitchell HL (2005) Ensemble Kalman filtering. Q J R Meteorol Soc 131:3269–3289

Hunt BR, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126

Husz ZL, Wallace AM, Green PR (2011) Tracking with a hierarchical partitioned particle filter and movement modelling. IEEE Trans Syst Man Cybern B Cybern 41(6):1571–1584

Jazwinski AH (1970) Stochastic processes and filtering theory. Mathematics in science and engineering, vol 64. Academic Press, San Diego

Kailath T, Sayed AH, Hassibi B (2000) Linear estimation. Prentice Hall information and system sciences series. Prentice Hall, Upper Saddle River, NJ

Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng Ser D 82(1):35–45

Khaki M, Ait-El-Fquih B, Hoteit I (2020) Calibrating land hydrological models and enhancing their forecasting skills using an ensemble Kalman filter with one-step-ahead smoothing. J Hydrol 584

Kivman GA (2003) Sequential parameter estimation for stochastic systems. Nonlinear Process Geophys 10(3):253–259

Künsch H (2001) State space and hidden Markov models. In: Barndorff-Nielsen OE, Cox DR, Klüppelberg C (eds) Complex stochastic systems. CRC Press, pp 109–173

Lamberti R, Petetin Y, Desbouvries F, Septier F (2017) Independent resampling sequential Monte Carlo algorithms. IEEE Trans Signal Process 65(20):5318–5333

Le Dimet F-X, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus 38A:97–110

Liu JS, Chen R (1998) Sequential Monte Carlo methods for dynamic systems. J Am Stat Assoc 93(443):1032–1044

Liu B, Ait-El-Fquih B, Hoteit I (2015) Efficient Kernel-based ensemble Gaussian mixture filtering. Mon Weather Rev 144:781–800

Mandel J, Beezley JD, Coen JL, Minjeong K (2009) Data assimilation for wildland fires. IEEE Control Syst Mag 29(3):47–65

Moradkhani H, Sorooshian S, Gupta HV, Houser PR (2005) Dual state-parameter estimation of hydrological models using ensemble Kalman filter. Adv Water Resour 28(2):135–147

Morzfelda M, Tub X, Atkins E, Chorina AJ (2012) A random map implementation of implicit filters. J Comput Phys 231:2049–2066

Pham D-T (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon Weather Rev 129:1194–1207
Pham D-T, Verron J, Rouband MC (1998) Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography. J Mar Syst 16(3–4):323–340
Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filter. J Am Stat Assoc 94(446):550–599
Raboudi NF, Ait-El-Fquih B, Subramanian AC, Hoteit I (2020) Enhancing ensemble data assimilation into one-way-coupled models with one-step-ahead smoothing. Quart J R Meteorol Soc. https://doi.org/10.1002/qj.3916
Raboudi NF, Ait-El-Fquih B, Hoteit I (2018) Ensemble Kalman filtering with one-step-ahead smoothing. Mon Weather Rev 146:561–581
Raboudi NF, Ait-El-Fquih B, Dawson C, Hoteit I (2019) Combining hybrid and one-step-ahead smoothing for efficient/short-range storm surge forecasting with an ensemble Kalman filter. Mon Weather Rev 147:3283–3300
Rasmussen J, Madsen H, Jensen KH, Refsgaard JC (2015) Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: evaluating the effect of ensemble size and localization on filter performance. Hydrol Earth Syst Sci 19:2999–3013
Rubin DB (1988) Using the SIR algorithm to simulate posterior distributions. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Bayesian statistics, vol 3. Oxford University Press, pp 395–402
Septier F, Peters GW (2015) An overview of recent advances in Monte-Carlo methods for Bayesian filtering in high-dimensional spaces. In: Peters GW, Matsui T (eds) Theoretical aspects of spatial-temporal modeling. Briefs - JSS research series in statistics. Springer
Sherman S (1955) A theorem on convex sets with applications. Ann Math Stat 26:763–767
Snyder C, Bengtsson T, Bickel P, Anderson J (2008) Obstacles to high-dimensional particle filtering. Mon Weather Rev 136(12):4629–4640
Spiller ET, Budhiraja A, Ide K, Jones CK (2008) Modified particle filter methods for assimilating Lagrangian data into a point-vortex model. Phys D 237(10):1498–1506
Stordal AS, Karlsen HA, Naevdal G, Skaug HJ, Valles B (2011) Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. Comput Geosci 15(2):293–305
Subramanian A, Hoteit I, Cornuelle B, Song H (2012) Linear vs. nonlinear filtering with scale selective corrections for balanced dynamics in a simple atmospheric model. J Atmos Sci 69:3405–3419
Tippett M, Anderson J, Bishop C, Hamill T, Whitaker J (2003) Ensemble square root filters. Mon Weather Rev 131(7):1485–1490
van Leeuwen PJ (2009) Particle filtering in geophysical systems. Mon Weather Rev 137(12):4089–4114
van Trees HL (1968) Detection, estimation, and modulation theory: Part I. Wiley, New York
Wen XH, Chen WH (2007) Real-time reservoir updating using ensemble Kalman filter: the confirming approach. Soc Pet Eng 11:431–442
Whitaker JS, Hamill TS (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 130:1913–1924
Whitaker JS, Hamill HT, Wei X, Song Y, Toth Z (2008) Ensemble data assimilation with the NCEP global forecast system. Mon Weather Rev 136:463–481
Yardim C, Michalopoulou Z-H, Gerstoft P (2011) An overview of sequential Bayesian filtering in ocean acoustics. IEEE J Oceanic Eng 36(1):71–89

Sparsity-Based Kalman Filters for Data Assimilation

Wei Kang and Liang Xu

Abstract Several variations of the Kalman filter, such as the extended Kalman filter (EKF) and the unscented Kalman filter (UKF), are widely used in science and engineering applications. However, traditional UKFs or EKFs cannot assimilate big data sets associated with models that have high dimensions, such as those in operational numerical weather prediction. In this chapter, we introduce two sparsity-based Kalman filters, namely the sparse-UKF and the progressive-EKF. The filters are designed specifically for problems with high dimensions. Different from ensemble Kalman filters (EnKFs) in which the error covariance is approximated using a set of dense ensemble vectors, the algorithms developed in this chapter are based on the sparse matrix approximation of error covariance. The new algorithms enjoy several advantages. The error covariance has full rank without being limited within a subspace generated by a set of ensembles. In addition to the estimated states, the algorithms provide updated error covariance in every assimilation cycle. Taking the advantage of sparsity, the required memory size and computational load can be significantly reduced.

1 Introduction

For dynamical systems, data assimilation is a process that integrates observational data with a numerical model for the purpose of estimating the system's state. Data assimilation is essential to numerical weather prediction (NWP). The estimate of the state value is used as the initial condition for weather forecast. If the dimension is relatively low and the data set is small, various linear and nonlinear estimators can

This work was supported in part by U.S. Naval Research Laboratory—Monterey, CA.

W. Kang (✉)
Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, USA
e-mail: wkang@nps.edu

L. Xu
Naval Research Laboratory, Monterey, CA, USA
e-mail: liang.xu@nrlmry.navy.mil

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_4

be found in the literature that have optimal or suboptimal performances. However, to assimilate big data sets with models that have high dimensions, such as those in operational NWP systems with tens of millions of variables, achieving reliable state estimation and error probability distributions is a challenging problem that have been studied for decades with a huge literature.

There are two categories of methods widely used in NWP, namely the variational method and the ensemble Kalman filter (EnKF) (Xu et al. 2005; Houtekamer and Zhang 2016). The former is based on a weighted least-square optimization, such as the four dimensional variational data assimilation (4D-Var) in a fixed time window or the three dimensional version (3D-Var) that excludes the time variable. The EnKF algorithm is based on the Kalman filter except that the error covariance is approximated using a set of ensembles. 4D-Var methods are used in operational NWP systems by many meteorological centers. While it serves as an effective method of data assimilation, 4D-Var algorithms have difficulty to explicitly track the evolution of error covariance within its estimation process due to high computational costs and input/output (I/O) loads required by the process of high dimensional matrices. EnKF, on the other hand, updates information about the error covariance in the form of ensembles. However, it is common in practical applications that the number of vectors in an ensemble is significantly smaller than the number of state variables. As a result, the rank deficiency of error covariance tends to deteriorate the integrity of the estimation process unless remedies to the algorithm, such as localization and covariance inflation, are applied.

Different types of Kalman filters have been developed and widely used in science and engineering applications, such as the EnKF, the extended Kalman filter (EKF) and the unscented Kalman filter (UKF). In this chapter, we introduce two sparsity-based Kalman filters, namely the sparse-UKF and the progressive-EKF. The goal of the work is to explore innovative ideas that take the advantage of the sparsity structure of matrices so that analysis and error covariance can be updated effectively and efficiently without the drawback of rank deficiency. The filters are developed specifically for problems with high dimensions. Different from EnKFs in which the error covariance is represented by a set of dense vectors in an ensemble, the new algorithms in this chapter are based on a sparse but full rank matrix as an approximation of the error covariance. This is made possible because of two assumptions: (a) the error covariance is approximately a sparse matrix; (b) the system model is component based, i.e. the state vectors are divided into components that can be computed independently in parallel. In Sect. 2, analysis is provided to justify that assumption (a) is expected to hold for a large family of system models. Assumption (b) is about the numerical method used for the system model, a topic that is not addressed in this chapter. In Sect. 3, the sparse-UKF is introduced. Its performance is exemplified using a Lorenz-96 model. In Sect. 4, the progressive-EKF is introduced and exemplified using the same Loren-96 model as in Sect. 3. In all examples, the new estimation methods developed in this chapter are compared to an EnKF as well as a traditional UKF with a full rank dense covariance.

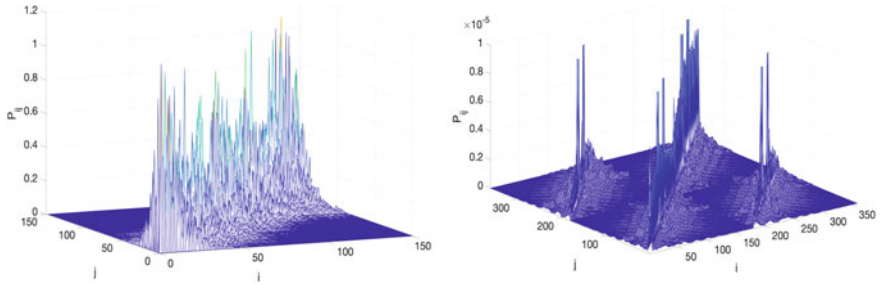

Fig. 1 Examples of Kalman filter error covariance

2 The Sparsity of Error Covariance

Shown in Fig. 1 are two error covariance matrices of Kalman filters, one for a linear system of ordinary differential equations and the other for a discretized shallow water equation. The dimensions of the state spaces are $n = 150$ and $n = 350$, respectively. The x- and y-axes represent row and column indices, i and j; the z-axis represents the absolute value of the error covariance, $|P_{ij}|$. Both matrices are approximately sparse, i.e., the majority of entries are relatively small. The matrices have peak value around one or multiple diagonals only. Approximating the covariance using a sparse matrix by setting all small entries to zeros, one can significantly reduce the computational cost, I/O loads and the amount of memory usage.

The approximate sparsity shown in Fig. 1 is not unusual. In fact, a theorem in Kang and Xu (2021) indicates that this type of sparse covariance is expected for a family of dynamic systems. In the following, a matrix A is said to be banded with bandwidth $s \geq 0$ if $A_{ij} = 0$ whenever $|i - j| > s$. If $s = 0$, then the matrix is diagonal. We say that a symmetric matrix P is less than or equal to another symmetric matrix G, or $P \leq G$, if $G - P$ is positive semidefinite. Consider the following system of ordinary differential equations (ODEs)

$$\begin{aligned} \dot{x}(t) &= Ax(t) + \eta(t), \quad x, \eta \in \mathbb{R}^n, \\ y(t) &= Hx(t) + \delta(t), \quad y, \delta \in \mathbb{R}^m \end{aligned} \tag{1}$$

where x is the state variable, $\dot{x}$ represents its time derivative, y is the observation variable, η and δ are zero-mean Gaussian white noise with covariances Q and R, respectively.

Theorem (Kang and Xu 2021) *Suppose that A and Q in (1) are banded. Let $P(t)$ be the error covariance of the Kalman filter that estimates $x(t)$. Then*

$$0 \leq P(t) \leq e^{At} P(0) e^{A^T t} + G^C(t) \tag{2}$$

where $G^C(t)$ is a symmetric matrix. Its entries have an upper bound

$$|(G^C(t))_{ij}| \leq \frac{\bar{G}\alpha^{(|i-j|+\beta)/\gamma}}{((|i-j|+\beta)/\gamma)^{(|i-j|+\beta)/\gamma}}$$

for some constants $\bar{G}$, α, β, and γ. The upper bound decreases at a rate greater than exponential as $|i-j| \to \infty$.

This result implies that $|(G^C(t))_{ij}|$ is almost zero when $|i-j|$ is large. Although $|i-j|$ is bounded by $n-1$, the result is applicable if n is significantly larger than the bandwidth of the banded matrices A and Q. In this case, G^C is approximately a sparse matrix. For the term of initial error covariance, $e^{At}P(0)e^{A^T t}$, it can be proved that this matrix is also approximately sparse if $P(0)$ is banded with a bandwidth significantly smaller than n. If the ODE in (1) is the discretization of partial differential equations, which is the case in many NWP problems, then A is banded provided that local discretization algorithms are used. Inspired by this theorem, we assume in the following sections that the covariance of Kalman filters can be approximated by a sparse matrix. We present two new algorithms of data assimilation in which the matrix of error covariance is computationally tractable.

3 Sparse-UKF

Consider a dynamical system model in which the state variable is $x(t)$, where $t = 1, 2, 3, \ldots$ represents time steps. The value of observation at $t = k$ is denoted by $y(k)$. The system model is defined as follows,

$$\begin{aligned} x(k) &= \mathcal{M}(x(k-1)) + \eta_{k-1}, \quad && x(k), \eta_{k-1} \in \mathbb{R}^n, \\ y(k) &= \mathcal{H}(x(k)) + \delta_k, && y_k, \delta_k \in \mathbb{R}^m, \end{aligned} \tag{3}$$

where η_{k-1} is a random variable representing the model error. Its covariance is Q. The observational error, δ_k, has a covariance R. In data assimilation, the goal is to estimate the value of $x(k)$ given the observations $y(1), y(2), \ldots, y(k)$ and the model (3). If (3) is linear and if all random variables are Gaussian, then the Kalman filter is an optimal state estimator. For nonlinear systems with non-Gaussian randomness, various types of Kalman filters exist in the literature with successful applications in science and engineering. If a system has a very high dimension, the conventional form of Kalman filter based on a dense error covariance is not applicable. In this section, we introduce an algorithm that is a variation of UKF for problems with approximately sparse matrices of covariance.

3.1 Sparse Matrix Algebra

In a sparse matrix, most entries are zeros. For some dense matrices in which most entries are relatively small, we approximate them using sparse matrices. In this chapter, we use an underbar to represent an operator that maps a vector or matrix to a sparse one. For instance, given a vector $x \in \mathbb{R}^n$. Let N_{sp} be an integer representing the size of the sparsity and

$$\mathcal{I} = \{i_1, i_2, \ldots, i_{N_{sp}}\}$$

be an index consisting of a sequence of integers. Then the underbar operator maps x to a vector $\underline{x} \in \mathbb{R}^{N_{sp}}$ in which

$$\underline{x}_k = x_{i_k} \text{ for } k = 1, 2, \ldots, N_{sp}.$$

It is equivalently to say that $\underline{x}$ is obtained from x by removing all x_i if $i \notin \mathcal{I}$. Usually, the removed entries are either zeros or relatively small (in absolute value). We would like to emphasize that, although $\underline{x}$ is formally a vector in $\mathbb{R}^{N_{sp}}$, it is treated as a vector in $\mathbb{R}^n$ by setting the ith entry zero if $i \notin \mathcal{I}$. This vector in $\mathbb{R}^n$ is, in general, different from the original vector, x, if the latter is a dense vector that is only approximately sparse.

Similarly, we can define the underbar operator for matrices. Given $P \in \mathbb{R}^{n\times n}$. Its columns may have different numbers of nonzero, or relatively large, entries. The largest such number is denoted by N_{sp}. Then $\underline{P}$ is a set of vectors associated with index sets

$$\underline{P} = \{\underline{P}_1, \underline{P}_2, \ldots, \underline{P}_n\}, \quad \text{sparsity index set } \mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\} \tag{4}$$

where $\underline{P}_i$ associated with $\mathcal{I}_i$ is the sparse vector approximation of the ith column of P. In all algebraic derivations, $\underline{P}$ is treated as a matrix in $\mathbb{R}^{n\times n}$ in which all entries are zeros except for those included in $\mathcal{I}$.

In sparsity-based algorithms, a full model evaluation is not always necessary. Using a *component-based* model can significantly reduce the computational load. In the notation, a component-based model has three inputs: state variable (either dense or sparse), its index, and the index of the output state. More specifically,

$$\underline{x}(k) = \mathcal{M}(\underline{x}(k-1); \mathcal{I}_1; \mathcal{I}_2), \tag{5}$$

where $\mathcal{I}_1$ is the index set of the sparse vector $\underline{x}(k-1)$ and $\mathcal{I}_2$ is the index set of $\underline{x}(k)$. The model evaluates only the entries with indices in $\mathcal{I}_2$, setting all other entries as zeros. The indices in $\mathcal{I}_2$ represent those entries in $x(k)$ that are most sensitive to the variation of the entries in $x(k-1)$ with indices in $\mathcal{I}_1$. For instance, a discretization of PDE using finite difference results in a model, $\mathcal{M}$, such that each entry in $x(k)$ is sensitive only to the variation of its adjacent entries in $x(k-1)$. If the input vector, $x(k-1)$, is dense, we omit $\mathcal{I}_1$ in the notation, i.e.

Table 1 Notations

Notation	Definition	Notation	Definition
x	State variable	y	Observation variable
$\mathcal{M}$	Model function	$\mathcal{H}$	Observation operator
n	State space dimension	$t = 1, 2, \ldots$	(Discrete) time variable
x_i^σ	ith σ-point		
x_i^b	ith background state	y_i^b	Output of observation operator $\mathcal{H}(x_i^b)$
$\bar{x}^b$	Average of x_i^b	$\bar{y}^b$	Average of y_i^b
P^b	Background error covariance		
x^a	Analysis—state vector	P^a	Analysis–error covariance

$$\underline{x}(k) = \mathcal{M}(x(k-1); \mathcal{I}), \tag{6}$$

where $\mathcal{I}$ is the same as $\mathcal{I}_2$ in (5).

Additions and multiplications of vectors/matrices, in both dense and sparse formats, are carried out in their original spaces, $\mathbb{R}^n$ or $\mathbb{R}^{n\times n}$. For instance,

$$\underline{P}\,\underline{x} \text{ or } \underline{P}x$$

are both evaluated using matrix multiplication in $\mathbb{R}^n$ in the dense format. If one needs the value at only a sparse set of locations, it is denoted by another underbar mapping

$$\underline{(\underline{P}x)}$$

Similarly, the summation of a dense vector and a vector in sparse format makes sense. For instance,

$$\underline{x} + x$$

is a vector in $\mathbb{R}^n$ in which all entries of x is unchanged except that those with indices in $\mathcal{I}$ are doubled. A new operation, called merging, between a sparse vector and a dense vector is defined as follows,

$$z = \underline{x} \triangleright w, \begin{cases} i\text{th component of } z = i\text{th component of } x, & \text{if } i \in \mathcal{I}. \\ i\text{th component of } z = i\text{th component of } w, & \text{if } i \notin \mathcal{I}. \end{cases} \tag{7}$$

A summary of notations is listed in the following Table 1.

3.2 UKF

The unscented Kalman filter has been increasingly popular in engineering applications since its introduction about twenty years ago (Julier et al. 2000; Julier and Uhlmann 2004). In a UKF, the error covariance is propagated with the dynamics using a set of vectors, or σ-points denoted by x^σ. Their definition is given in (8)–(9). The σ-points are computed at each time step using a square root of the error covariance. In most UKF applications, σ-points are computed using either Cholesky factorization or matrix diagonalization. In the notation, a variable with a superscript 'a', such as x^a, represents the *analysis* value of the variable, i.e., the updated value based on observations. A variable with a superscript 'b', such as y^b, represents the background, i.e., the propagated value of analysis using the system model. The algorithm is summarized as follows. At $t = k-1$, suppose we have the analysis and error covariance as well as its square root

$$\begin{aligned} &x^a(k-1),\ P^a(k-1),\\ &X^a(k-1) = \sqrt{(n+\kappa)P^a(k-1)}, \end{aligned} \tag{8}$$

where κ is a scaling factor for the fine tuning of the higher order moments of the approximation error (Julier et al. 2000). How to tune the value of κ for a sparsity-based UKF is an open problem that needs further study. In this chapter, $\kappa = 0$ is used in all examples. A set of σ-points is generated as follows,

$$\begin{aligned} &x_0^\sigma(k-1) = x^a(k-1),\\ &x_i^\sigma(k-1) = x^a(k-1) + X_i^a(k-1),\ 1 \le i \le n,\\ &x_i^\sigma(k-1) = x^a(k-1) - X_i^a(k-1),\ n+1 \le i \le 2n. \end{aligned} \tag{9}$$

where $X_i^a(k-1)$ is the ith column vector of $X^a(k-1)$. The next step is to propagate the σ-points, which represent the background at $t = k$. For simplicity of notations, the time variable 'k' in the kth time-step is omitted.

$$\begin{aligned} &x_i^b = \mathcal{M}(x_i^\sigma(k-1)), \quad y_i^b = \mathcal{H}(x_i^b), \qquad 0 \le i \le 2n,\\ &\bar{x}^b = \sum_{i=0}^{2n} w_i x_i^b, \qquad \bar{y}^b = \sum_{i=0}^{2n} w_i y_i^b, \end{aligned} \tag{10}$$

where the weights are defined as follows

$$w_0 = \frac{\kappa}{n+\kappa},\ w_i = \frac{1}{2(n+\kappa)}, \tag{11}$$

for $i = 1, 2, \ldots, 2n$. Define the variations

$$X_i^b = x_i^b - \bar{x}^b,\ Y_i^b = y_i^b - \bar{y}^b. \tag{12}$$

The background covariances are

$$\begin{aligned} P^b &= \sum_{i=0}^{2n} w_i X_i^b (X_i^b)^T + Q, \\ P_{xy} &= \sum_{i=0}^{2n} w_i X_i^b (Y_i^b)^T, \\ P_{yy} &= \sum_{i=0}^{2n} w_i Y_i^b (Y_i^b)^T + R. \end{aligned} \tag{13}$$

The Kalman gain, K, satisfies the following equation,

$$K P_{yy} = P_{xy}. \tag{14}$$

The analysis is updated as follows

$$\begin{aligned} x^a &= \bar{x}^b + K(y_o - \bar{y}^b), \\ P^a &= P^b - K(P_{xy})^T, \end{aligned} \tag{15}$$

where y_o is the observation at $t = k$. This completes one iteration of the filter. For the next step, $t = k + 1$, go back to (8) replacing the analysis by the updated value of x^a and P^a.

3.3 *Sparse-UKF*

The square root factorization of a matrix is not unique. For large and sparse matrices, various algorithms and their implementations on different computing platforms have been studied for many years. The literature can be traced back to the early days of electronic computers (Davis et al. 2016). In the case of Cholesky factorization, the square root of a sparse matrix is still sparse, although the computation may require larger amounts of processor memory than the original matrix (Davis 2006; Rozin and Toledo 2005).

A dense error covariance is intractable in computation for global models used in NWP. In the following approach, we assume that P and $\sqrt{P}$ are approximately sparse. In the algorithm, they are replaced by their sparse approximations, $\underline{P}$ and $\underline{(\sqrt{P})}$. Their sparsity index sets are denoted by $\mathcal{I}$ and $\mathcal{I}^\sigma$, respectively. When propagating the σ-points using a component-based model, only a sparse subset of the elements is computed. The indices of the subset form an index set, $\mathcal{I}^b$. How to determine the index sets for sparse vectors and matrices is discussed later in this section

Algorithm I (sparse-UKF)
Given the initial analysis,

$$x^a(k-1),\ \underline{P^a}(k-1). \tag{16}$$

Step 1. σ-points and forecast

$$\underline{X^a}(k-1) = \underline{\sqrt{(n+\kappa)P^a(k-1)}},\ \text{sparsity index set}\,\mathcal{I}^\sigma \tag{17}$$

For $i = 0$,

$$x_0^b = \mathcal{M}(x^a(k-1)), \quad y_0^b = \mathcal{H}(x_0^b). \tag{18}$$

For $i = 1, 2, 3, \ldots, 2n$,

$$\begin{aligned} x_i^\sigma(k-1) &= x^a(k-1) + \underline{X_i^a}(k-1), && 1 \le i \le n, \\ x_i^\sigma(k-1) &= x^a(k-1) - \underline{X_i^a}(k-1), && n+1 \le i \le 2n. \\ \underline{x_i^b} &= \mathcal{M}(x_i^\sigma(k-1); \mathcal{I}_i^b), \qquad y_i^b = \mathcal{H}(\underline{x_i^b} \triangleright x_0^b), && 1 \le i \le 2n. \end{aligned} \tag{19}$$

Step 2. Background covariances

$$\bar{x}^b = w_0 x_0^b + \sum_{i=1}^{2n} w_i(\underline{x_i^b} \triangleright x_0^b),\ \bar{y}^b = \sum_{i=0}^{2n} w_i y_i^b \tag{20}$$

$$\begin{aligned} \underline{P^b} &= w_0 \underline{(x_0^b - \bar{x}^b)(x_0^b - \bar{x}^b)^T} \\ &\quad + \sum_{i=1}^{2n} w_i \underline{(\underline{x_i^b} \triangleright x_0^b - \bar{x}^b)(\underline{x_i^b} \triangleright x_0^b - \bar{x}^b)^T} + \underline{Q}, \quad \text{sparsity index set}\,\mathcal{I}, \\ P_{xy} &= w_0(x_0^b - \bar{x}^b)(y_i^b - \bar{y}^b)^T + \sum_{i=1}^{2n} w_i(\underline{x_i^b} \triangleright x_0^b - \bar{x}^b)(y_i^b - \bar{y}^b)^T, \\ P_{yy} &= \sum_{i=0}^{2n} w_i(y_i^b - \bar{y}^b)(y_i^b - \bar{y}^b)^T + R. \end{aligned} \tag{21}$$

Step 3. Kalman gain and analysis

$$\begin{aligned} K P_{yy} &= P_{xy}, \\ x^a &= \bar{x}^b + K(y_o - \bar{y}^b), \\ \underline{P^a} &= \underline{P^b} - \underline{K(P_{xy})^T} + \gamma I, \quad \text{sparsity index set}\,\mathcal{I}. \end{aligned} \tag{22}$$

The constant term γI in (22) is a diagonal matrix. The value of γ is selected so that $\underline{P^a}$ is positive definite, which is guaranteed if γ is larger than the smallest negative eigenvalue of

$$\underline{P^b} - \underline{K(P_{xy})^T}. \tag{23}$$

If this matrix is positive definite, then $\gamma = 0$. In the case that a fixed lower bound of eigenvalues is unknown, the value of γ can be adaptively changed in every cycle depending on the smallest negative eigenvalue of (23). Numerical algorithms of finding the smallest eigenvalue for high dimensional matrices is needed for the determination of γ. Studies about this problem is out the scope of this paper. A survey on this topic can be found in Davidson (1989).

How to determine the index sets for sparse vectors/matrices? This is a problem for which we do not have a complete answer. The selection of $\mathcal{I}$ for $\underline{P}^a$ is a trade-off between the computational cost and the approximation accuracy. If $|\mathcal{I}|$ is small (highly sparse), $\underline{P}^a$ may not be a good approximation of P^a because too many nonzero entries are set to zero; if $|\mathcal{I}|$ is large, it increases the computational cost. The sparsity index set, $\mathcal{I}^\sigma$, of $(\sqrt{\underline{P}^a})$ is determined by the sparsity of $\sqrt{\underline{P}^a}$. If the square root is the Cholesky factorization, this matrix is already sparse. The number of nonzero entries in $\sqrt{\underline{P}^a}$ is larger than the size of $\mathcal{I}$ (Rozin and Toledo 2005). However, one may use a smaller set as $\mathcal{I}^\sigma$ to speed up the computation. Once again, this is a trade-off between computational cost and accuracy. In (19), $\mathcal{I}_i^b$ is the index set of $\underline{x}_i^b$, which is the propagation of the ith σ-point. The indices in $\mathcal{I}_i^b$ represent those entries in x_i^b that are most sensitive to the variation of the entries of $x^\sigma(k-1)$ with indices in $\mathcal{I}_i^\sigma$. Or equivalently, $\underline{x}_i^b$ contains those entries of x_i^b that have relatively large change when the value of $\underline{X}_i^a$ is changed. In general, these sparsity index sets are different from each other. However, as a means of reducing computational loads, we may use one index set for all three, $\mathcal{I}_i$, $\mathcal{I}_i^\sigma$ and $\mathcal{I}_i^b$. This idea is tested in the next section on a Lorenz-96 model.

In the sparse-UKF, the assumption is that P^a can be approximated by a sparse matrix $\underline{P}^a$. Although the σ-points in the algorithm play a similar role as that of ensembles in EnKF, using sparse-UKF one can avoid the problem of rank deficiency. For systems with very high dimensions, the number of ensemble members used in an EnKF is much smaller than the dimension. As shown in Fig. 2 (left plot), the narrow and tall matrix of ensemble vectors makes EnKF fundamentally a rank deficient approach. In contrast, the block diagonal matrix $\underline{P}^a$ shown in Fig. 2 (middle plot) as a sparse approximation of P^a has full rank.

The computational load required by (19) in Step 1 is extremely high if full state vectors are computed. Thanks to the sparsity, we only need to compute the entries with indices in $\mathcal{I}^b$. For a sparse-UKF to be successful for high dimensional problems, it is critical to have component-based numerical models so that only the entries with indices in $\mathcal{I}^b$ are computed; and most entries of the state vector are not evaluated at all. It is also important to point out that individual terms for $i = 1, 2, \ldots, 2n$ in (19), (20) and (21) can be computed independent of each other, making the computation perfectly parallel. Because matrices of covariance are symmetric, the memory size and I/O usage for the computation of covariance can be significantly reduced. For instance, the number of nonzero entries in the upper half of $\underline{P}^a$ is less than or equal to

$$n\left(\frac{N_{sp}-1}{2}+1\right) \tag{24}$$

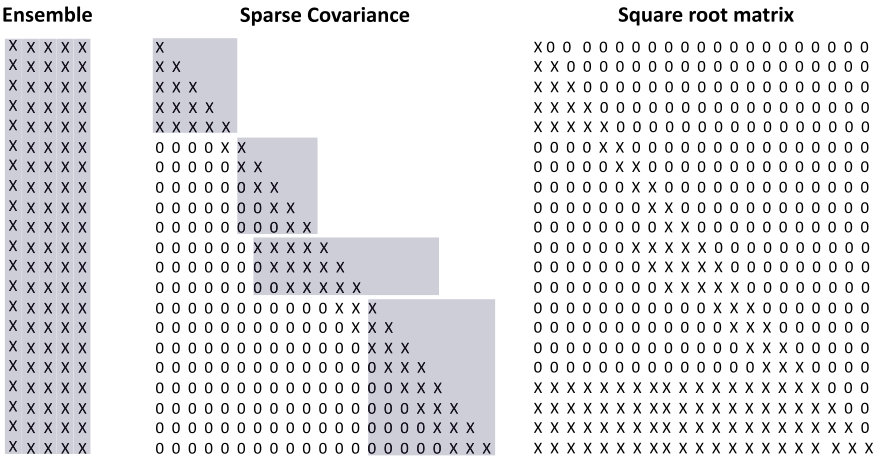

Fig. 2 Patterns of ensemble vectors and sparse error covariances

If N_{sp} is an integer close to the ensemble size of an EnKF, then (24) is smaller than the number of entries in the ensemble matrix, which is dense and nonsymmetric. Shown in Fig. 2 (middle and right plots) are the sparsity patterns of $\underline{P^a}$ and $\sqrt{\underline{P^a}}$ that we find in some examples. Note that the number of nonlinear entries in each column may vary. An advantage of the sparse-UKF is the capability of easily assigning different sparsity to different columns in $\underline{P^a}$ by using the index sets $\mathcal{I}_i$, $\mathcal{I}_i^\sigma$ and $\mathcal{I}_i^b$.

3.4 Lorenz-96 Model

In this section, we use a Lorenz-96 model that was first introduced in Lorenz (1996) to test the performance of the sparse-UKF. Consider

$$\begin{aligned} \frac{dx_i}{dt} &= (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \quad i = 1, 2, \ldots, n, \\ x_{n+1} &= x_1, \\ n &= 40, \\ \Delta t &= 0.025, \\ F &= 8. \end{aligned} \tag{25}$$

The system has chaotic trajectories as shown in Fig. 3, a plot of $x_1(t), x_2(t), x_3(t)$. The simulations are conducted based on a 4th-order Runge-Kutta discretization. The trajectories are used as the ground truth. The sparsity pattern for $\underline{P^a}$ and $\sqrt{\underline{P^a}}$ are assumed to be centered along the diagonal line with a fix length. The total number of nonzero entries in each column is N_{sp}. We would like to point out that the sparse

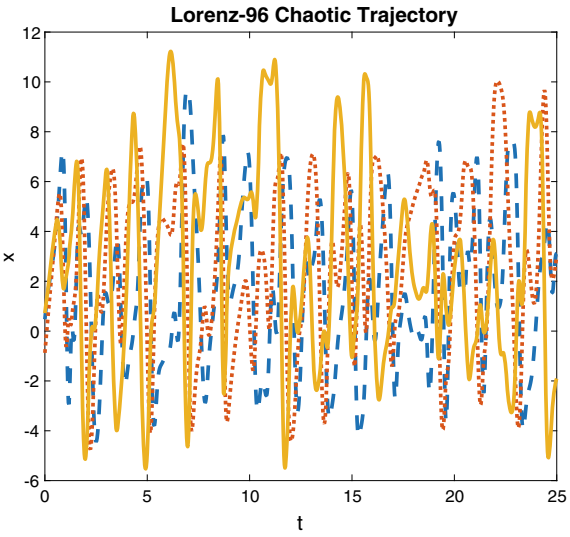

Fig. 3 A chaotic trajectory of the Lorenz-96 model, x_1(solid), x_2(dash), x_3(dot)

matrices are approximations of the true error covariance and its square root. The true sparsity pattern of $\sqrt{\underline{P^a}}$ is, in fact, different from that of $\underline{P^a}$. In the approximation, however, we ignore the difference and use the same sparsity pattern for both. This idea of simplifying index sets works fine for the Lorenz-96 model. A systematic way of choosing the sparsity pattern for $\sqrt{\underline{P^a}}$ based on given $\underline{P^a}$ is an open problem that needs further study.

The numerical experimentation is based on $N = 1000$ uniformly distributed random initial states in $[-1\ 1]$. The time step size is $\Delta t = 0.025$. The total number of time steps for each simulation is $N_t = 4000$. The number of observations at any given time is $m = 20$, i.e. every other state variable is measured,

$$y(k) = \left[x_1(k)\ x_3(k)\ x_5(k) \cdots x_{39}(k) \right]^T . \tag{26}$$

The observational error has the Gaussian distribution. Its covariance is $R = I$, the identity matrix. The initial background error covariance is $P^b(0) = 0.2I$. The estimation error is defined by the following RMSE

$$RMSE = \sqrt{\frac{1}{n(N_t+1)} \sum_{k=0}^{N_t} ||x^a(k) - x^{truth}(k)||_2^2}. \tag{27}$$

For comparison, an EnKF is also applied to the same data set. The localization radius is $\rho = 4$ and the inflation factor is $\sqrt{1.08}$. A full scale UKF based on dense error covariance is applied as the best estimator in the study. The number of nonzero entry evaluations in the computation of $\underline{P^a}$ and $\sqrt{\underline{P^a}}$, an indicator of computational load, is

Table 2 Summary of simulation results

Filter	Size	Nonzero Entries	Entries EVAL	Error Median	Error Mean	Error STD
EnKF	$N_{ens} = 10$	400 in ensemble	400	0.3462	1.0741	1.0652
S-UKF	$N_{sp} = 7$	160 in $\underline{P}^a$	600	0.3061	0.3067	0.0071
S-UKF	$N_{sp} = 11$	240 in $\underline{P}^a$	920	0.2691	0.2691	0.0048
UKF	Full covariance	820 in P^a	3200	0.2358	0.2360	0.0039

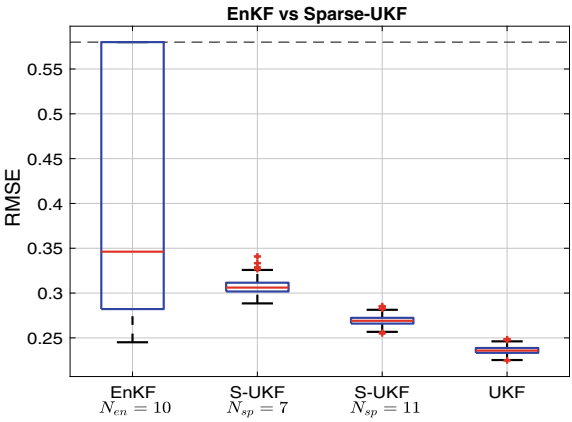

Fig. 4 Boxplot of RMSE

$$2nN_{sp} + n \tag{28}$$

Reducing the number of entries being evaluated, such as using a smaller set of σ-points, and testing the impact of Cholesky factorization on the efficiency of UKF are ongoing research topics not addressed in this chapter.

Shown in Table 2, the EnKF has $N_{ens} = 10$ ensemble vectors with a total of 400 nonzero entries. In comparison, the sparse-UKFs with $N_{sp} = 7$ and 11 have much smaller numbers of nonzero entries to be evaluated and stored in memory. A smaller number is desirable because it implies reduce I/O load and the amount of memory usage. In terms of computational load, the number of entry evaluations for the sparse-UKFs are higher. This is mainly due to the propagation of the $2n$ σ-points. Studies show that reducing the number of σ-points to n is possible. However, its impact on the estimation accuracy has to be studied case by case, which is beyond the scope of this chapter. In the columns under error median and mean, the numbers show that both sparse-UKFs achieve more accurate estimation than EnKF. The most significant advantage of sparse-UKFs are the small variation of estimation error. In Table 2, the error standard deviation of the sparse-UKFs are 0.0071 and 0.0048, which is in sharpe contrast to 1.0652 of EnKF. The error of EnKF has large variation due to the

method's high dependency on the subspace in which the ensembles are selected. This problem does not exist for the sparse-UKF because $\underline{P}^a$ has full rank. The variation of errors is clearly shown in the boxplot in Fig. 4. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points the algorithm considers to be not outliers, and the outliers are plotted individually. For comparison, the error boxplot of the full size UKF is included in Fig. 4.

4 Progressive-EKF

In a sparse-UKF, the σ-points are computed by taking a square root of the error covariance, such as the Cholesky factorization. In this section, we propose a progressive algorithm of approximating error covariance without taking square roots.

4.1 Basic Ideas

The main assumption for this algorithm is the following progressive relationship

$$M_{k-1}P^a(k-1)M_{k-1}^T = P^a(k-1) + \Delta P^b, \tag{29}$$

where ΔP^b is assumed to be small. In (29), M_{k-1} is the Jacobian of $\mathcal{M}$ at $x^a(k-1)$. Similarly, the Jacobian of $\mathcal{H}$ is H_k. To estimate ΔP^b, assume

$$M_{k-1} = I + \Delta M_{k-1}. \tag{30}$$

where we assume that ΔM_{k-1} is small. If the system model is based on the discretization of a differential equation with a small time step size, then

$$\mathcal{M}(x(k-1)) = x(k-1) + O(\Delta t^\alpha), \quad \alpha > 1. \tag{31}$$

The Jacobian of $O_{k-1}(\Delta t^\alpha)$ in space variables is expected to have small value if Δt is small, which makes (30) a reasonable assumption. Then we have

$$\begin{aligned} &M_{k-1}P^a(k-1)M_{k-1}^T \\ &= (I + \Delta M_{k-1})P^a(k-1)(I + \Delta M_{k-1}^T) \\ &= P^a(k-1) + \Delta M_{k-1}P^a(k-1) + \left(\Delta M_{k-1}P^a(k-1)\right)^T \\ &\quad + \Delta M_{k-1}P^a(k-1)\Delta M_{k-1}^T \\ &\approx P^a(k-1) + \Delta M_{k-1}P^a(k-1) + \left(\Delta M_{k-1}P^a(k-1)\right)^T. \end{aligned} \tag{32}$$

This is in consistent with (29). It can be computed using a tangent linear model. Or it can be approximated using the dynamical model

$$\begin{aligned} & M_{k-1} P^a(k-1) M_{k-1}^T \\ & = (I + \Delta M_{k-1}) P^a(k-1)(I + \Delta M_{k-1}^T) \\ & \approx \left(\mathcal{M}(x(k-1) + \delta P^a(k-1)) - \mathcal{M}(x(k-1))\right)/\delta \\ & + \left(\mathcal{M}(x(k-1) + \delta P^a(k-1)) - \mathcal{M}(x(k-1))\right)^T/\delta - P^a. \end{aligned} \tag{33}$$

where $\delta > 0$ is the step size of a finite difference approximation of $\Delta M_{k-1} P^a$. Its value should be determined depending on the numerical model and its linearization. In (33), a vector and matrix summation is a new matrix resulting from adding the vector to every column in the matrix. Applying an operator to a matrix is to apply the operator to every column in the matrix.

4.2 Progressive-EKF

The column vectors in the matrices in (32) and (33) are sparse. However, the number of column vectors equals n, which can be as high as $10^6 - 10^7$ for some atmospheric models. Applying a full model to all the vectors is impractical because of the high computational and I/O loads. Similar to the idea that we used in sparse-UKF, we approximate the error covariance using a given sparsity, i.e., only a small portion of the entries in each column vector is evaluated. Evaluating the entire state vector is unnecessary. This is the reason we need a component-based model. Then the algorithm of progressive-EKF is summarized as follows.

Algorithm II (progressive-EKF)
Given the initial analysis at $t = k - 1$,

$$x^a(k-1) \text{ and } \underline{P}^a(k-1). \tag{34}$$

Step 1. Forecast

$$\begin{aligned} x^b &= \mathcal{M}(x^a(k-1)), \\ y^b &= \mathcal{H}(x^b). \end{aligned} \tag{35}$$

Step 2. Background error covariance

$$\begin{aligned} \underline{P}^b = & \underline{\left(\mathcal{M}\left(x^a(k-1) + \delta \underline{P}^a(k-1), \mathcal{I}\right) - x^b\right)/\delta} \\ & + \underline{\left(\mathcal{M}\left(x^a(k-1) + \delta \underline{P}^a(k-1), \mathcal{I}\right) - x^b\right)^T}/\delta - \underline{P}^a + Q. \end{aligned} \tag{36}$$

Step 3. Kalman gain and analysis

$$\begin{aligned} K &= \underline{P}^b H_k^T (H_k \underline{P}^b H_k^T + R)^{-1}, \\ x^a &= x^b + K(y_o - y^b), \\ \underline{P}^a &= (I - K H_k)\underline{P}^b. \end{aligned} \tag{37}$$

Different from the sparse-UKF, this algorithm avoids the computation of matrix square roots. However, the algorithm requires that ΔP^b in (29) can be approximated effectively. From (31), the method is expected to work better for a small time step-size. If Δt is large, ΔM_{k-1} in (30) may not be small enough. A remedy is to use a refined step-size in an inner-loop computation. More specifically, the discrete model is a discretization of a continuous-time model. The discrete time moment $k-1$ corresponds to the continuous time moment $(k-1)\Delta t$. We refine the step size by dividing the time interval into n_p subintervals. In our examples, we choose $n_p = 2$. The refined time steps are

$$(k-1)\Delta t, (k-1)\Delta t + \frac{\Delta t}{n_p}, \ldots, (k-1)\Delta t + s\frac{\Delta t}{n_p}, \ldots, k\Delta t, \qquad 0 \le s \le n_p \tag{38}$$

For the inner loop, one can compute a sequence of backgrounds, $\tilde{x}^b(s)$.

$$\begin{aligned} t_s &= (k-1)\Delta t + s\frac{\Delta t}{n_p}, \\ \tilde{x}^b(s) &= \tilde{\mathcal{M}}_{t_s}(x^a(k-1)),\ s = 1, 2, \ldots, n_p. \end{aligned} \tag{39}$$

where $\tilde{\mathcal{M}}_{t_s}$ represents the refined model function in the time interval from $t = (k-1)\Delta t$ to $t = t_s$. In Step 2, repeat (36) n_p times along the sequence of background states, $\tilde{x}^b(s)$, without adding Q until the last round. This refined Step 2 increases the computational load, while improving the accuracy of the progressive estimation.

4.3 Examples

In the following, we apply the progressive-EKF to the Lorenz-96 model using the same parameters given in (25). The analysis is based on the simulation data using $N = 1000$ random initial states, in which the value of each state variable is uniformly distributed in $[-1, 1]$. The error covariance is approximated using sparsity matrices with $N_{sp} = 7, 11, 17$. For $N_{sp} = 11$, we tested the idea of refining step-size using $n_p = 1$ and $n_p = 2$. The results are summarized in Table 3. The boxplots of error variation are shown in Fig. 5. Comparing to EnKF, the error variations of the progressive-EKFs are significantly smaller. If $N_{sp} = 7$, which is smaller than the ensemble size $N_{ens} = 10$, the median value of estimation error is larger than that of the EnKF. The median error for $N_{sp} = 11$ is comparable to that of the EnKF. If a refined step-size in (39) is applied, for instance $n_p = 2$, the median estimation error is further reduced. Comparing to the performance of the sparse-UKF in Table 2, the

Table 3 Summary of simulation results

Filter	Size	Nonzero Entries	Entries EVAL	Error Median	Error Mean	Error STD
EnKF	$N_{ens} = 10$	400 in ensemble	400	0.3462	1.0741	1.0652
P-EKF	$N_{sp} = 7$ $N_p = 1$	160 in $\underline{P}^a$	320	0.3845	0.3846	0.0055
P-EKF	$N_{sp} = 11$ $N_p = 1$	240 in $\underline{P}^a$	480	0.3455	0.3458	0.0050
P-EKF	$N_{sp} = 11$ $N_p = 2$	240 in $\underline{P}^a$	480×2	0.3041	0.3041	0.0044
P-EKF	$N_{sp} = 17$ $N_p = 3$	360 in $\underline{P}^a$	720×2	0.2872	0.2873	0.0046

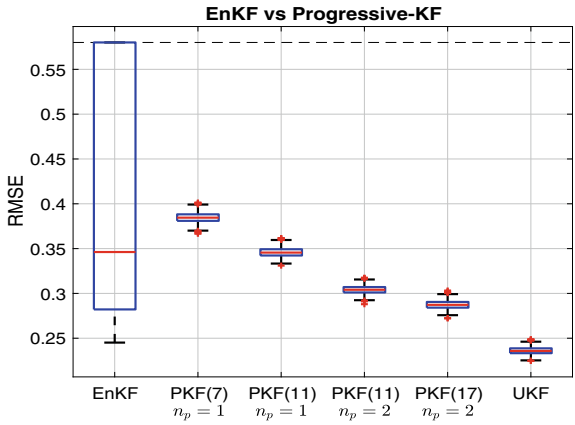

Fig. 5 Boxplot of RMSE. For Prograssive-KF, N_{sp} = 7, 11, and 17

error variations are similar. However, the estimation error of the sparse-UKF has a smaller median in all cases. For example, to achieve a similar performance as the sparse-UKF when $N_{sp} = 11$, one has to use a larger sparsity index $N_{sp} = 17$ for the progressive-EKF.

5 Conclusions

Two Kalman type filters, sparse-UKF and progressive-EKF, based on sparse error covariances are introduced. They are tested using the Lorenz-96 model with 40 state variables and chaotic trajectories. Both algorithms share the same basic idea: the error covariance is approximated using a sparse matrix. Thanks to the sparsity, the required memory size is significantly reduced. The symmetry of the error covariance

can potentially reduce the I/O load. The analysis error covariance can be updated as a sparse matrix in each cycle using a deterministic process, either a square root matrix or a progressive algorithm. The updated sparse matrix is then used as the background error covariance for the next cycle. Relative to EnKFs, the main advantage of the proposed methods is that the estimation process do not need an ensemble; and the error covariance has a full rank. The algorithms do not suffer issues of rank deficiency as in EnKFs. As a result, the variation of analysis error is constantly small in all examples. Techniques of localization and covariance inflation are unnecessary. Relative to 4D-Var methods, the proposed algorithms are highly parallel in computation. They provide not only the estimate of states but also the analysis error covariance. For the purpose of scalability, we suggest that the proposed methods are applied with component-based numerical models. From the examples, the sparse-UKF has better accuracy than the progressive-EKF. On the other hand, the progressive-EKF is a simple algorithm that avoids taking square roots of large matrices, provided that the progressive approximation of error covariance is adequately accurate. The limited number of examples in this chapter is not enough for drawing a comprehensive comparison between the two filters. More numerical experimentations and further study of the methods using different types of system models are main topics of our future work.

References

Davidson ER (1989) Super-matrix methods. Comput Phys Commun 53(1–3):49–60

Davis TA (2006) Direct methods for sparse linear systems. SIAM

Davis T, Rajamanickam S, Sid-Lakhdar WM (2016) A survey of direct methods for sparse linear systems. Acta Numer 25:383–566

Houtekamer PL, Zhang F (2016) Review of the ensemble Kalman filter for atmospheric data assimilation. Mon Weather Rev 144:4489–4532

Julier S, Uhlmann J, Durrant-Whyte HF (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans Autom Control 45(3):477–482

Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc IEEE 92(3):401–422

Kang W, Xu L (2021) Some quantitative characteristics of error covariance for Kalman filters. Tellus A: Dyn Meteorol Ocenogr 73(1):1–19

Lorenz E (1996) Predictability—a problem partly solved. In: Seminar on predictability, vol I. ECMWF

Rozin E, Toledo S (2005) Locality of reference in sparse Cholesky factorization methods. Electron Trans Numer Anal 21:81–106

Xu L, Rosmond R, Daley R (2005) Development of NAVDAS-AR: formulation and initial tests of the linear problem. Tellus 57A:546–559

Perturbations by the Ensemble Transform

Kazuo Saito, Le Duc, Takumi Matsunobu, and Takuya Kurihana

Abstract In the ensemble data assimilation, the background error covariance is estimated from perturbations of the ensemble forecast, while characteristics of the ensemble forecast strongly depend on how the initial ensemble is generated. The ensemble transform is a popular perturbation method that widely used as an ensemble perturbation generator, however, linear combinations of different perturbations in the ensemble transform (off-diagonal components of the transform matrix) may harm the global balance of the meteorological field. In this paper, we discuss this issue and show the structure of initial perturbations. Results of forecast experiments using the local ensemble transform Kalman filter (LETKF) for a simplified global model and a regional NWP model are shown. The spin-up issue in a cloud resolving model is shown with the comparison to an alternative method (diagonal LETKF).

K. Saito (✉)
Atmosphere and Ocean Research Institute, University of Tokyo, Chiba, Japan
e-mail: k_saito@aori.u-tokyo.ac.jp

K. Saito · L. Duc
Meteorological Research Institute, Tsukuba, Japan

Japan Meteorological Business Support Center, Tokyo, Japan

L. Duc
e-mail: leduc@mri-jma.go.jp

T. Matsunobu · T. Kurihana
Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan
e-mail: tkmi@ccs.tsukuba.ac.jp

T. Kurihana
e-mail: tkurihana@uchicago.edu

T. Matsunobu
Meteorologisches Institut, Ludwig-Maximilians-Universität, Munich, Germany

T. Kurihana
Department of Computer Science, University of Chicago, Chicago, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_5

1 Introduction

Ensemble data assimilation (DA) methods (e.g., LETKF,[1] EnKF,[2] EnVAR[3]) are becoming widely used as the analysis methods for research and operational NWP. Ensemble DA methods have an advantage against the variational (VAR) methods in terms of the development cost where the adjoint models are not necessary. In ensemble DA, magnitudes of forecast error are locally estimated from the ensemble forecast, reflecting the flow-dependent uncertainty of the first guess. Their performances, however, are still arguable and likely have a room for further improvement.

It is well known that characteristics of the ensemble forecast error strongly depend on how the initial ensemble is generated. In the current most ensemble DA methods including LETKF and EnVAR, the ensemble transform (ET; Bishop et al. 2001; Wang et al. 2004), the eigen-decomposition of the analysis error covariance matrix in ensemble space, is widely used for generating the initial ensemble perturbation. ET corresponds to the principal component analysis and initial ensemble perturbations are given by a linear combination of forecast perturbations obtained in the former DA cycle. On the other hand, another perturbation scheme, the breeding growing mode (BGM) method scales the bred vectors to a certain magnitude, where perturbations are composed by the diagonal components only in terms of the transform matrix (Bowler et al. 2008). The ET method has an advantage where the magnitudes of perturbations (initial ensemble spreads) reflect the magnitudes of the analysis error. Another advantage is that the orthogonalization of perturbation vectors is not required. In case of BGM, the bred vectors tend to converge to the first Lyapunov vector through the breeding cycles (Toth and Kalnay 1997).

On the other hand, it is known that the growth of the LETKF perturbations is slower than other perturbation methods. Figure 1 shows time evolution of ensemble spreads of 500 hPa height field (Z) and meridional wind (V) by different perturbation methods in the same regional EPS (Saito et al. 2011). During the 36-h simulation period, the global singular vector (GSV) method showed the largest growth rate of the ensemble spreads. Growths of ensemble spreads in LETKF (LET) and downscale of global ensemble (WEP) were rather sluggish among the five perturbation methods. Other perturbation methods based on the mesoscale model SV (MSV), mesoscale BGM (MBD) show similar tendency, but the growth of LETKF was relatively sluggish.

Figure 2 indicates root mean square errors (RMSEs) of the ensemble mean forecasts at FT = 24 for height (Z), meridional winds (V), temperature (T) and relative humidity (RH) at 500 hPa level and the Receiver Operating Characteristics (ROC) Area Skill Scores for 6-h precipitation. Again, performances of the forecast from LETKF perturbations were not necessarily best compared with other perturbation methods. The tendency of forecasts from the LETKF perturbations (LET) is similar

[1] Local ensemble transform Kalman filter.

[2] Ensemble Kalman filter.

[3] Ensemble-based variational method.

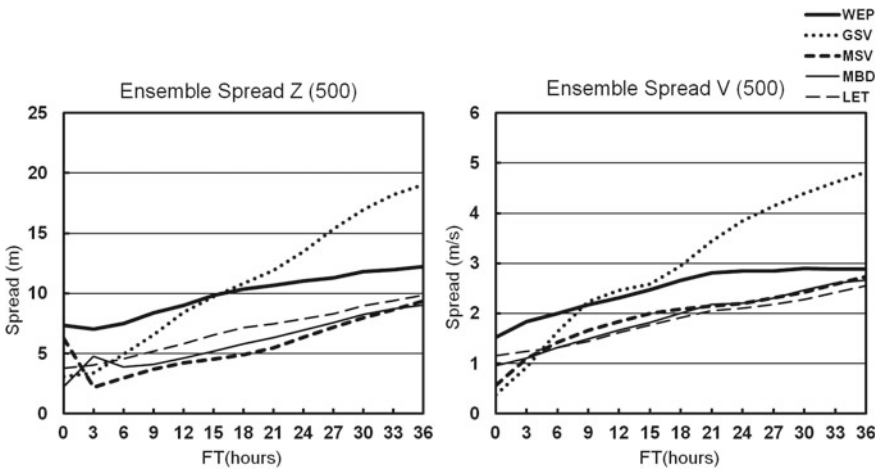

Fig. 1 Growth of ensemble spread by different perturbation methods in a regional EPS for the Word Weather Research Programme Beijing 2008 Olympics Research and Development Project. WEP; downscale of global ensemble, GSV; Global SV, MSV; mesoscale model SV, MBD; mesoscale model BGM, LET; mesoscale model LETKF. After Saito et al. (2011)

to that of mesoscale breeding (MBD) but the performance of LET is slightly worse than MBD in several cases.

Figure 3 compares time evolutions of central pressures of cyclone Nargis predicted by regional EPSs. In downscale ensemble from global BGM perturbations (Fig. 3a; Saito et al. 2010), about a half of perturbed members predicted the TC stronger than the control run (non-perturbed member), while in the LETKF ensemble (Fig. 3b), TC intensities in the most perturbed members were weaker than the control run (LETKF analysis).

Reich et al. (2011) pointed-out that those properties of ensemble transform perturbations in LETKF induce an initial imbalance in analysis ensemble members. The Relaxation-to-prior-perturbations (RTPP; Zhang et al. 2004a, b) method corresponds to increasing the diagonal components in ET. Ota et al. (2019) added singular vectors to perturbations of LETKF in the operational global EPS at JMA to inflate perturbations in the ensemble forecast.

Required conditions of initial perturbations in ensemble DA may be listed as follows:

- Zero-mean at initial (mean-preserving property) and no biases in forecast.
- Forecast perturbations express the forecast error.
- Magnitude of perturbations reflects the analysis error.
- Good orthogonality to assure that the perturbations compose the basis to span the space of a large freedom.
- Accuracy of the ensemble mean is good.
- Performance of the forecast by each perturbed member is not poor, so that each forecast can be regarded as a possible scenario of the future states.

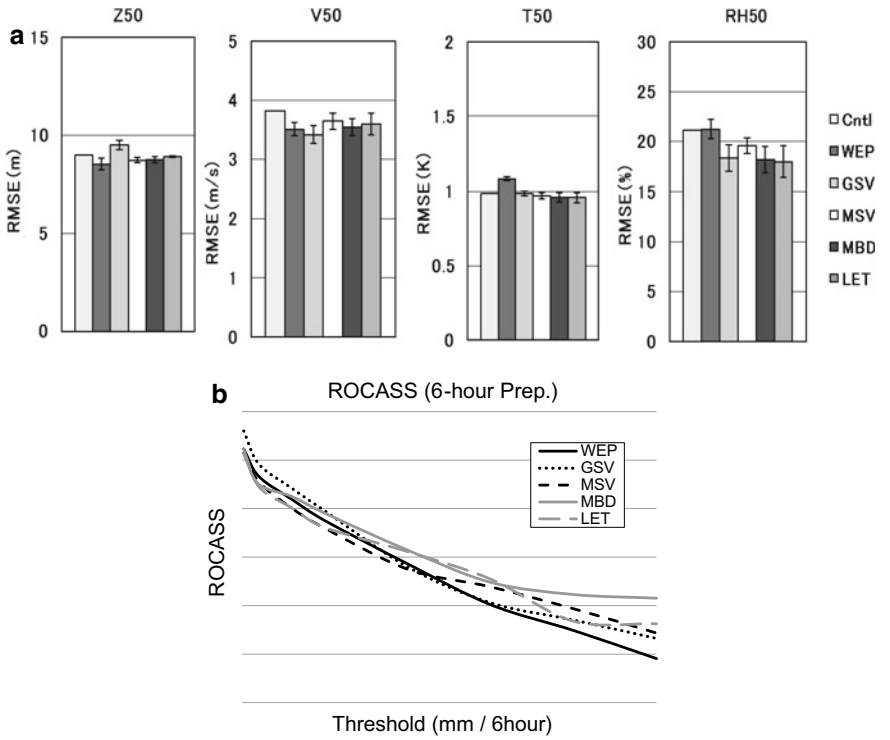

Fig. 2 **a** RMSE of ensemble mean (FT = 24) with various perturbation methods for 500 hPa height field (Z50), meridional wind (V50), temperature (T50) and relative humidity (RH50). **b** ROC Area Skill Scores for 6-h accumulated precipitation. After Saito et al. (2011)

- Good growth rates of perturbations.

As for orthogonality, Table 1 shows orthogonality of perturbation vectors (similarity index) of BGM (upper) and LETKF (lower) perturbations with and without lateral boundary perturbations (LBPs) for the case of a regional ensemble prediction system (EPS) (Saito et al. 2012). Here, p1–p5 are positive perturbations and m1–m5 are negative perturbations. In this table, upper triangular matrix components indicate the case without LBPs in the breeding cycles, whereas lower triangular matrix components indicate the case with LBPs in breeding cycles. In BGM (Table 1a), about 30% of bred vectors (12 of 40) were similar to each other in case without LBPs, and the number of such similar vector pairs decreased to 10% (4 of 40) with LBPs. In LETKF (Table 1b), orthogonality between the initial perturbations was generally good even without LBPs in the ensemble Kalman filter data assimilation cycles.

Recently, Duc et al. (2020a) presented the mathematical proof of derivation of ET matrix in the sense of the Frobenius norm. They showed that the positive symmetric square root of the analysis error covariance is the closest matrix to the identity I, and

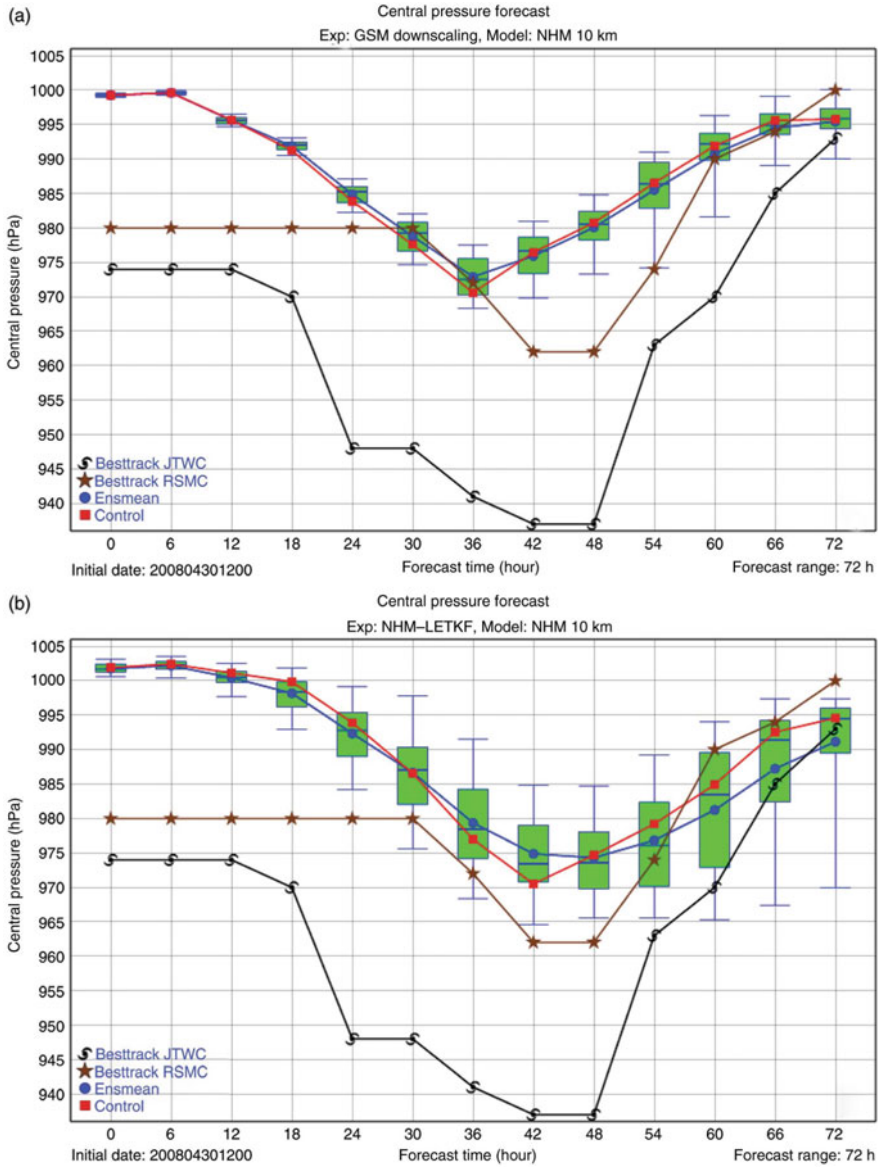

Fig. 3 Time evolution of central pressures of cyclone Nargis predicted by regional EPSs. **a** Downscale ensemble from global EPS perturbations. **b** Same as in **a** but for ensemble forecast from NHM-LETKF perturbations. After Duc et al. (2015)

Table 1 Orthogonality of perturbation vectors (similarity index) of BGM (upper) and LETKF (lower) with and without LBPs. After Saito et al. (2012)

(a)

	p1	p2	p3	p4	p5	m1	m2	m3	m4	m5
p1	**1.00**	0.25	**0.50**	0.18	0.21	**−1.00**	−0.25	**−0.49**	−0.18	−0.21
p2	0.09	**1.00**	0.04	0.02	**0.62**	−0.25	**−0.99**	−0.03	−0.01	**−0.61**
p3	0.39	0.28	**1.00**	0.65	−0.05	**−0.49**	−0.03	**−0.99**	**−0.64**	0.06
p4	0.21	0.03	0.20	**1.00**	−0.18	−0.18	0.00	**−0.64**	**−0.98**	0.19
p5	−0.07	**0.53**	0.00	0.25	**1.00**	−0.20	**−0.61**	0.06	0.19	**−0.99**
m1	**−0.99**	−0.08	−0.37	−0.21	0.07	**1.00**	0.25	**0.50**	0.19	0.21
m2	−0.08	**−0.98**	−0.25	−0.0 2	**−0.52**	0.08	**1.00**	0.03	0.00	**0.62**
m3	−0.37	−0.26	**−0.97**	−0.19	0.01	0.37	0.27	**1.00**	**0.65**	−0.06
m4	−0.20	−0.02	−0.18	**−0.99**	−0.24	0.21	0.02	0.19	**1.00**	−0.19
m5	0.07	**−0.52**	0.02	−0.24	**−0.99**	−0.07	**0.53**	−0.01	0.24	**1.00**

(b)

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	**1.00**	0.07	−0.23	−0.18	0.01	0.09	−0.03	**−0.46**	−0.30	−0.19
M2	−0.10	**1.00**	−0.11	−0.07	0.01	0.12	−0.31	−0.22	−0.33	−0.36
M3	0.19	−0.51	**1.00**	−0.06	**−0.40**	−0.34	−0.12	−0.03	0.27	0.03
M4	−0.18	−0.25	0.01	**1.00**	−0.22	−0.23	−0.02	−0.11	0.03	−0.15
M5	−0.10	0.32	−0.37	−0.26	**1.00**	0.06	−0.16	−0.01	−0.25	−0.02
M6	**−0.56**	0.12	−0.19	−0.02	0.06	**1.00**	−0.11	−0.07	−0.36	−0.18
M7	−0.01	−0.07	−0.07	−0.11	−0.17	−0.09	**1.00**	−0.14	0.10	−0.07
M8	0.03	−0.14	0.07	−0.10	−0.17	−0.17	−0.31	**1.00**	0.00	0.11
M9	−0.02	−0.27	0.09	−0.07	−0.15	−0.20	−0.20	−0.17	**1.00**	−0.02
M10	−0.23	−0.07	−0.26	−0.03	−0.14	0.07	0.00	−0.22	0.04	**1.00**

discussed the diagonally predominant property. In this study, we show detailed structures of the LETKF perturbations in NWP models and test results on the potential application of the diagonal ET in LETKF. In Sect. 2, we first refer to mathematical treatment of ET and refer to the diagonally predominant property discussed in Duc et al. (2020a). Section 3 discusses characteristics of ET perturbations in a simple global model and regional NWP model. Section 4 shows spin-up aspect of ET perturbations with deep convection in a cloud resolving model and cycle DA test. Verifications of precipitation forecasts using the fraction skill score are also shown. Summary and concluding remarks are given in Sect. 5.

2 Ensemble Perturbations and Ensemble Transform

In the LETKF, analysis perturbations are obtained by applying a linear transformation on background perturbations. In this section, we remind mathematical treatment of ET and decomposition of ETM according to Duc et al. (2020a). Denoting analysis and background perturbations by $n \times k$ matrices $\mathbf{X}^{\mathrm{a}}$ and $\mathbf{X}^{\mathrm{b}}$, respectively, ET is represented by a right-multiplication of $\mathbf{X}^{\mathrm{b}}$ by a $k \times k$ matrix $\mathbf{T}$:

$$\mathbf{X}^{\mathrm{a}} = \mathbf{X}^{\mathrm{b}}\mathbf{T}, \tag{1}$$

where n is the size of the state vectors, and k is the ensemble size. Here each column of $\mathbf{X}^{\mathrm{b}}$ represents the difference $\mathbf{x}^{\mathrm{b}} - \overline{\mathbf{x}^{\mathrm{b}}}$ between each forecast member $\mathbf{x}^{\mathrm{b}}$ and the ensemble mean $\overline{\mathbf{x}^{\mathrm{b}}}$ calculated from all forecast members. A similar definition is applied for the columns of $\mathbf{X}^{\mathrm{a}}$.

The matrix $\mathbf{T}$ is ETM, and it is formulated so that the analysis error covariance $\mathbf{P}^{\mathrm{a}} = \mathbf{X}^{\mathrm{a}}(\mathbf{X}^{\mathrm{a}})^{\mathrm{T}}/(k-1)$ obeys the Kalman filter equation:

$$\mathbf{P}^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^{\mathrm{b}} = \mathbf{P}^{\mathrm{b}} - \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\mathbf{H}\mathbf{P}^{\mathrm{b}}, \tag{2}$$

where $\mathbf{R}$ is the observation error covariance, $\mathbf{H}$ is the observation operator, and $\mathbf{P}^{\mathrm{b}} = \mathbf{X}^{\mathrm{b}}(\mathbf{X}^{\mathrm{b}})^{\mathrm{T}}/(k-1)$ is the background error covariance. Using background perturbations in observation space $\mathbf{Y}^{\mathrm{b}} = \mathbf{H}\mathbf{X}^{\mathrm{b}}$, $\mathbf{P}^{\mathrm{a}}$ can be rewritten as

$$\mathbf{P}^{\mathrm{a}} = \mathbf{X}^{\mathrm{b}}\left(\mathbf{I} + \left(\mathbf{Y}^{\mathrm{b}}\right)^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}^{\mathrm{b}}/(k-1)\right)^{-1}\left(\mathbf{X}^{\mathrm{b}}\right)^{\mathrm{T}}\Big/(k-1). \tag{3}$$

Suppose that the eigen-decomposition of $\mathbf{Y}^{\mathrm{bT}}\mathbf{R}^{-1}\mathbf{Y}^{\mathrm{b}}/(k-1)$ is given by $\mathbf{C}\boldsymbol{\Gamma}\mathbf{C}^{\mathrm{T}}$, where the orthogonal matrix $\mathbf{C}$ contains the eigenvectors in its columns, and the diagonal matrix $\boldsymbol{\Gamma}$ contains the eigenvalues γ_{i} on its diagonal. Since $\mathbf{C}$ is orthogonal, we have $\mathbf{I} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$, $\tilde{\mathbf{P}}^{a} = (\mathbf{I} + (\mathbf{Y}^{\mathrm{b}})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}^{\mathrm{b}}/(k-1))^{-1}$ is simplified as

$$\tilde{\mathbf{P}}^{a} = \mathbf{C}(\mathbf{I} + \Gamma)^{-1}\mathbf{C}^{\mathrm{T}}. \tag{4}$$

Wang et al. (2004) pointed out that the positive symmetric square root $\mathbf{T}^{\mathrm{s}} = \mathbf{C}(\mathbf{I} + \boldsymbol{\Gamma})^{-1/2}\mathbf{C}^{\mathrm{T}}$ is the ETM which possesses the mean-preserving property.

Duc et al. (2020a) presented the mathematical proof of derivation of $\mathbf{T}^{\mathrm{s}}$, showing that for all ETMs $\mathbf{S}$ the squared distance in Frobenius norm attains its minimum when $\mathbf{S}$ is the positive symmetric square root of $\tilde{\mathbf{P}}^{a}$

$$\left\|\mathbf{T}^{\mathrm{s}} - \mathbf{I}\right\|_{\mathrm{F}}^{2} = \sum (\lambda_{\mathrm{i}} - 1)^{2}, \tag{5}$$

where λ_i are the eigenvalues of $\tilde{\mathbf{P}}^a$. They extended this proof to another minimization problem for the squared distance $\|\mathbf{S} - \alpha\mathbf{I}\|_F^2$, by introducing a new variable α. They found that $\mathbf{S}$ resembles a scalar multiple of $\mathbf{I}$ most when the scalar multiple of $\mathbf{I}$ is $\bar{\lambda}$. This result suggests that we can decompose $\mathbf{T}^s$ into the sum of a diagonal matrix $\mathbf{D}$ and a perturbation matrix $\mathbf{P}$

$$\mathbf{T}^s = \mathbf{D} + \mathbf{P} = \bar{\lambda}\mathbf{I} + \mathbf{P}, \tag{6}$$

where the Frobenius norm of D is $\sqrt{k}\bar{\lambda}$, and the typical magnitudes for the entries of of $\mathbf{P}$ are estimated by $\sigma_\lambda/\sqrt{k}$ (σ_λ is the standard deviation of λ).

Upper panel of Fig. 4 shows entries of three 50 × 50 $\mathbf{T}^s$ matrices computed at three different grid points obtained from an LETKF experiment with real observations using 50 ensemble members (Duc et al. 2015). Diagonal terms are dominant over

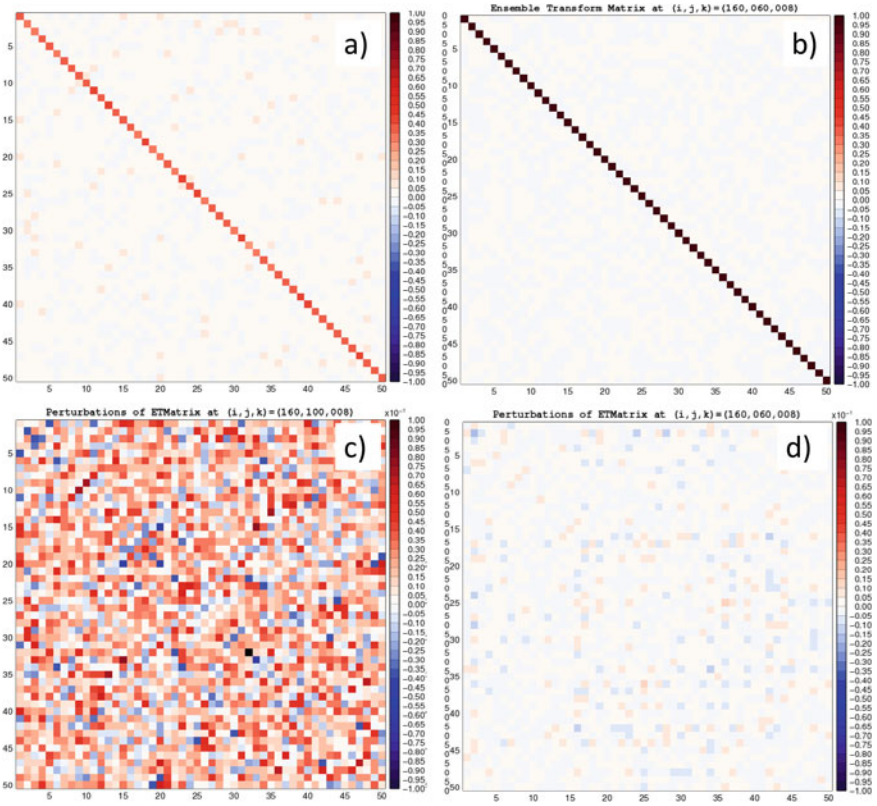

Fig. 4 Entries of 50 x 50 ET matrix at different grid points obtained from an LETKF experiment in the region **a** with dense observations and **b** far from observations, **c** and **d** the same as in **a** and **b**, respectively, but for the perturbation matrix **P**. Note that the numbers at color bars are multiplied by 10. Editing from Duc et al. (2020a)

non-diagonal terms. In Fig. 4b, the average value of the diagonal terms is about 0.99 at a grid point in an observations sparse area. This value in Fig. 4a is 0.39, at a grid point in the region with dense observations. The typical magnitudes for the entries of **P** in Fig. 4c, d are as 0.0235 and 0.0031, respectively.

Figure 5 shows an example of, background, perturbation, and analysis of surface (mean sea level) pressures in an LETKF analysis for a case of typhoon. Here, upper panels show result of normal LETKF which employs full matrix of $\mathbf{T}^s$, while lower panels are result when $\mathbf{T}^s$ is replaced by **D** ('diagonal LETKF'). Perturbations by full-T LETKF (Fig. 5b) are noisier than those of diagonal LETKF (Fig. 5e) due to the mixture of off-diagonal perturbations, and two pressure minimums are seen in the resultant analysis (Fig. 5c). In case of diagonal LETKF, the perturbation is a simple dipole pattern which adjusts the position of TC center southwest ward.

After the next section, we show detailed structures of the LETKF perturbations in NWP models comparing with those from diagonal LETKF.

3 Perturbations in LETKF in NWP Models

3.1 Cases of SPEEDY-LETKF

In order to check the problem of ET, a test of LETKF using the "simplified parameterization primitive-equation dynamics" global model (SPEEDY; Molteni 2003) was conducted. The SPEEDY model is an atmospheric global circulation model with a T30L7 resolution (represented by $96 \times 48 \times 7$ grid points), and the model variables are zonal and meridional winds (u, v), temperature T, specific humidity q, and surface pressure ps. These variables are also used as control variables in assimilation with LETKF (SPEEDY-LETKF; Miyoshi 2005; Yokota et al. 2016). Adaptive inflation (Miyoshi 2011) is used and the number of ensemble members is 20. Different horizontal localization scales with 500, 750, 1000, and 2000 km were tested with 0.1 (ln p) vertical localization.

Pseudo sonde observation data with errors followed by a Gaussian distribution set for observation system simulation experiment (OSSE) are assimilated with 40 days spin-up from January 1 1982. Figure 6a, b shows the analysis field of 500 hPa (level 4) geopotential height (Z) and wind speed of zonal wind (U) for 9 February 1982. Figure 6c shows perturbation of horizontal winds at level 2 (835 hPa) of member 01 for the case of horizontal localization of 750 km. Figure 6d, e indicate perturbation vectors reproduced from diagonal and off-diagonal components of the ET matrix. By the diagonal predominant property, the perturbation produced by the diagonal components is similar to the original perturbation, while the one produced by the off-diagonal part appears noisy. Inconsistency of off-diagonal perturbations is also clear vertically. Figure 6f–h show corresponding horizontal wind perturbations at 685 hPa (level 3). Even 685 hPa is just adjacent level above level 2, the horizontal

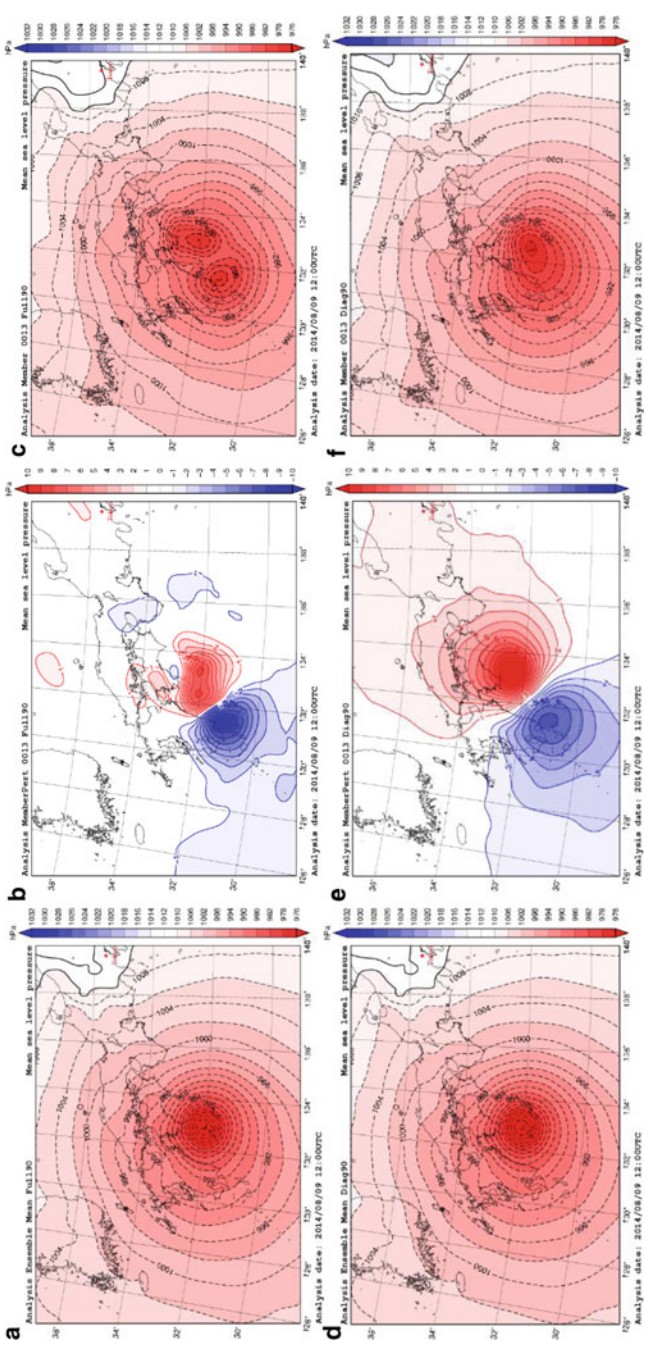

Fig. 5 Example of horizontal perturbations of surface (mean sea level) pressures in LETKF for **a** first guess, **b** perturbation, and **c** analysis. **d**, **e** and **f** are the same as in **a**, **b** and **c**, respectively, but for the diagonal LETKF

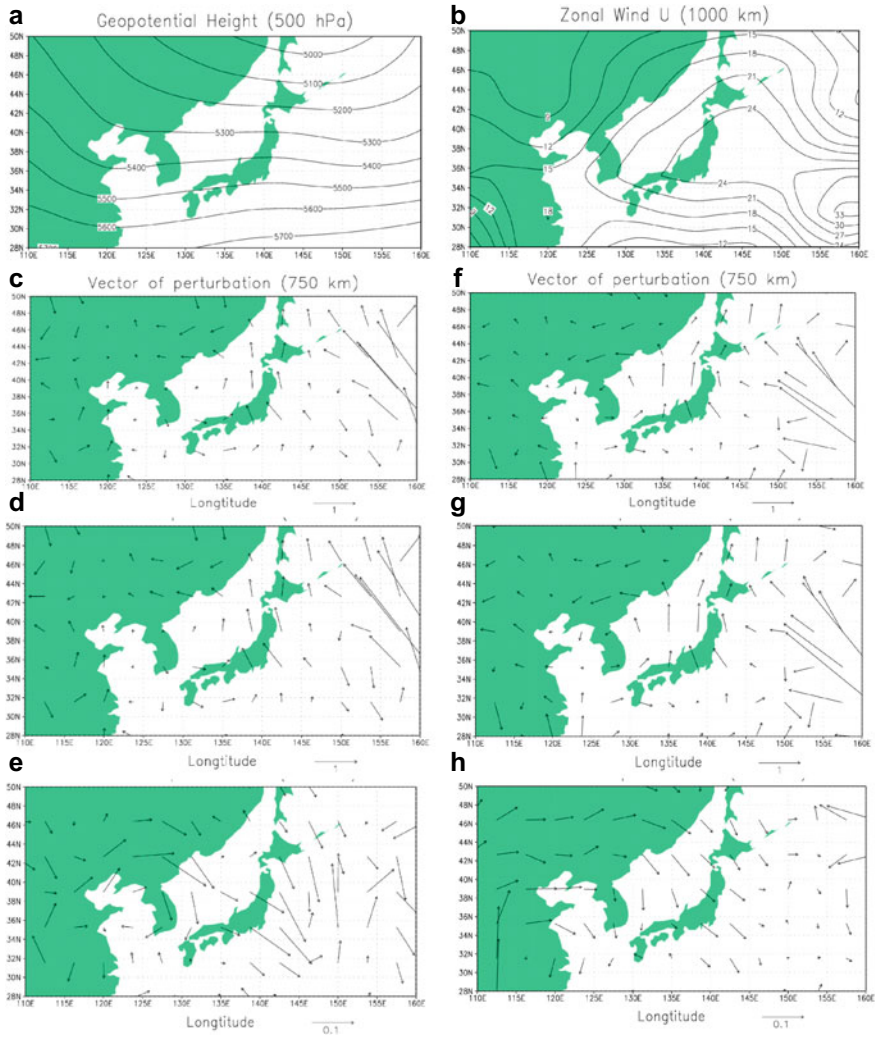

Fig. 6 Analysis fields and perturbation vectors by speedy LETKF for 9 February 1982 for **a** height field at 500 hPa, **b** zonal wind speed, and **c** perturbation of horizontal winds at level 2 (835 hPa) of member 01, **d** and **e** are the same as in **c** but for diagonal and off-diagonal components in the ensemble transform, respectively, **f**, **g** and **h** are the same as in **c**, **d** and **e**, respectively, but for level 3 (685 hPa). Note that the vector scale is enlarged 10 times

pattern of the off-diagonal perturbation (Fig. 6h) shows a quite different appearance from that at level 3 (Fig. 6e).

Figure 7a depicts power spectra of LETKF perturbation vectors for meridional winds at level 3 (685 hPa) for member 01 with various horizontal localization scales. Spectral powers decrease with wave numbers. Power spectra of diagonal components

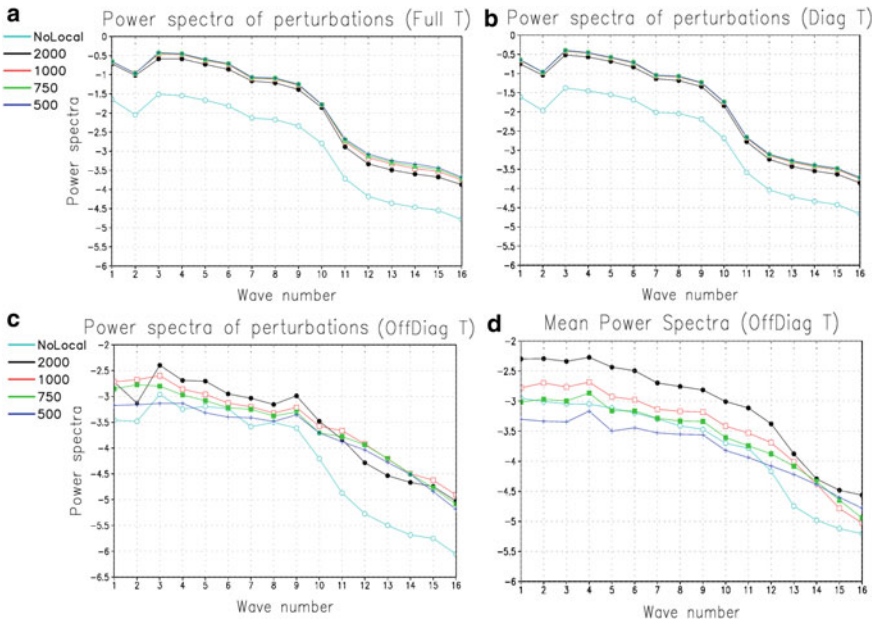

Fig. 7 Power spectra of perturbation vectors for meridional winds at level 3 (685 hPa) for member 01 with various horizonal localization scales for **a** full matrix, **b** diagonal components, and **c** off-diagonal components in the ensemble transform, **d** is the same as in **c** but for mean power spectra for all 20 members

(Fig. 7b) is similar to those in Fig. 7a, while power spectra of off-diagonal part (Fig. 7c) differ depending on the horizontal localization scales. When a smaller localization scale (e.g., 500 km) was applied, powers in small wave numbers relatively decrease and powers in large wave number relatively increase, which means that perturbations by off-diagonal components are noisier in smaller localization scales. Figure 7d shows mean power spectra for all 20 members for off-diagonal components. The same tendency is seen in the mean power spectral.

Figure 8 indicates root mean square errors (RMSEs) and ensemble spreads of horizontal wind by the 48-h extended ensemble forecast from LETKF analyses. In this experiment, RMSEs of the ensemble mean were slightly larger than the control run. RMSE of each perturbed member is much larger than that of the control run. RMSEs of the control run, the ensemble mean and most perturbed members increase at the first 6 h and then slightly decrease at FT = 12. After that RMSEs gradually increase with time. The ensemble spreads decrease at the first 6 h and then gradually increase after FT=12. These behaviors of RMSEs and ensemble spreads suggest a spin-up problem in the forecasts from ET perturbations.

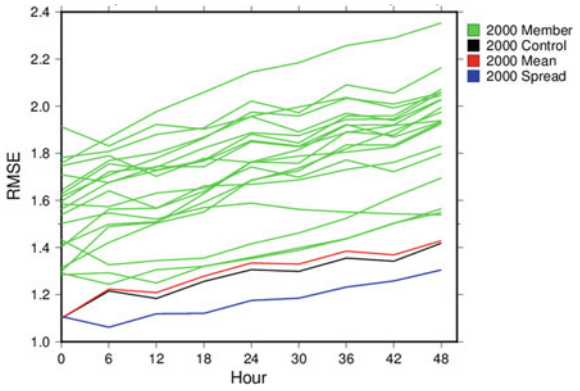

Fig. 8 RMSEs of control run (black), ensemble mean (red) and perturbed members (green) and ensemble spread (blue) for horizontal winds (U) at level 3 (685 hPa) for the case of horizontal localization scale of 2000 km

3.2 Case of NHM-LETKF

A similar tendency as described in the former subsection was confirmed by an experiment using a regional NWP model. The Japan Meteorological Agency (JMA) nonhydrostatic model (Saito et al. 2006, 2007; Saito 2012) and its LETKF version (Miyoshi and Aranami 2006; Kunii 2014; Yokota et al. 2018) were used. In this subsection, first we show the result with a horizontal resolution of 15 km. The horizontal domain size of the JMA-NHM is 4,080 km $\times$ 3,300 km and the number of vertical levels is 50. The 3-ice bulk cloud microphysics scheme which predicts six water species and the Kain-Fritsch (K-F) convection parameterization scheme were simultaneously employed as the precipitation process.

Specification of NHM-LETKF is 15 km 50 layers with 20 members. Real observations similar to the JMA operational mesoscale DA system except for satellite radiance and precipitation analysis were assimilated. NHM-LETKF system was started from Aug 1st, 2016 with 6 hourly DA cycles, using the JMA operational global EPS as the very initial and lateral boundary perturbations. Relaxation to prior spread is employed as the inflation method. Different horizontal localization scales with 100, 200, and 400 km were tested with 0.1 (ln p) vertical localization.

Height and temperature fields at 500 hPa at 0000 UTC 5 August 2016 are shown in Fig. 9. Figure 10a, b show horizontal wind fields at z = 5721 m at the same time analyzed by NHM-LETKF (CNTL and member 01, respectively). The analysis perturbation of member 01 is shown in Fig. 10c. Figure 10d, e indicate the diagonal ($\mathbf{D} = \overline{\lambda}\,\mathrm{I}$) and perturbation ($\mathbf{P}$) components of analysis perturbations. As similar to the SPEEDY-LETKF case, perturbations by the $\mathbf{D}$ components (Fig. 10d) represent most part of the total analysis perturbations (Fig. 10c), and the residual part, the $\mathbf{P}$ components (Fig. 10e) are noisy. Figure 10f shows the $\mathbf{P}$ part perturbations at z = 6665 m. Again, vertical inconsistency of $\mathbf{P}$ part is evident.

Figure 10g, h show the $\mathbf{P}$ perturbations by experiments with different horizontal localization scales. With a small localization scale (Fig. 10g), $\mathbf{P}$ perturbations are unevenly distributed mainly only around land area suggesting that the diagonal

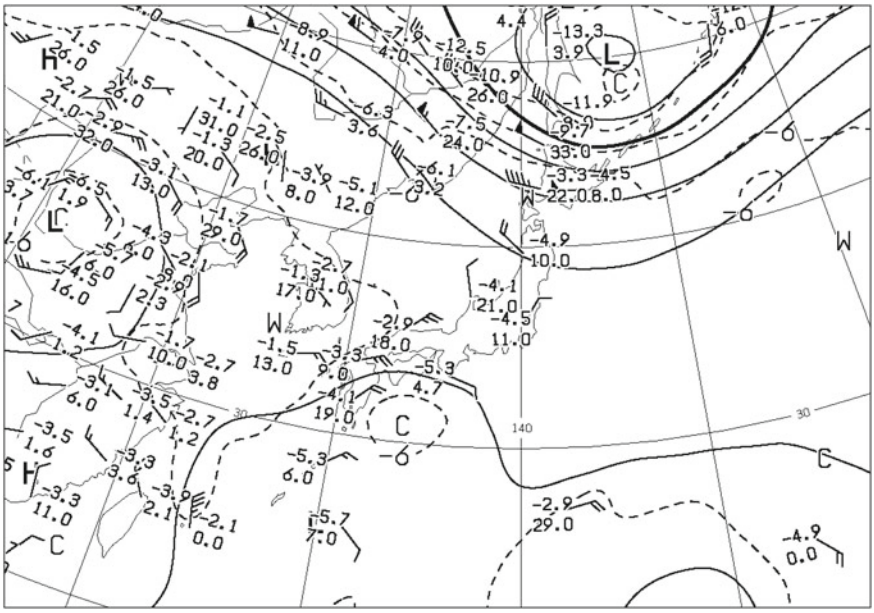

Fig. 9 Height and temperature field at 500 hPa at 0000 UTC, 5 Aug 2016

components are dominant over the sea, in the region far from observations (see Fig. 4b, d). When a larger localization scale is applied, **P** part increases according to the increase of assimilated data (Fig. 10h).

A similar property of ensemble perturbations shown above is also seen in the vertical motion, but in 15 km NHM-LETKF, vertical motion is generally small because deep convection is suppressed by the cumulus parametrization scheme (not shown). Vertical inconsistency of off-diagonal perturbations is likely more problematic when penetrative deep convection exists in the analysis field.

4 Perturbations by Ensemble Transform in a Cloud Resolving Model

4.1 2 km NHM-LETKF

In this section, we focus more on the detailed structures of the ET perturbations and check spin-up of cloud resolving models after LETKF analysis. High-resolution (2 km) version of NHM-LETKF is used, and its domain size is 240 × 240 km. Specification of NHM-LETKF is 50 layers with 20 members. The system is nested with the 15 km NHM-LETKF described in the former section. To permit deep convection,

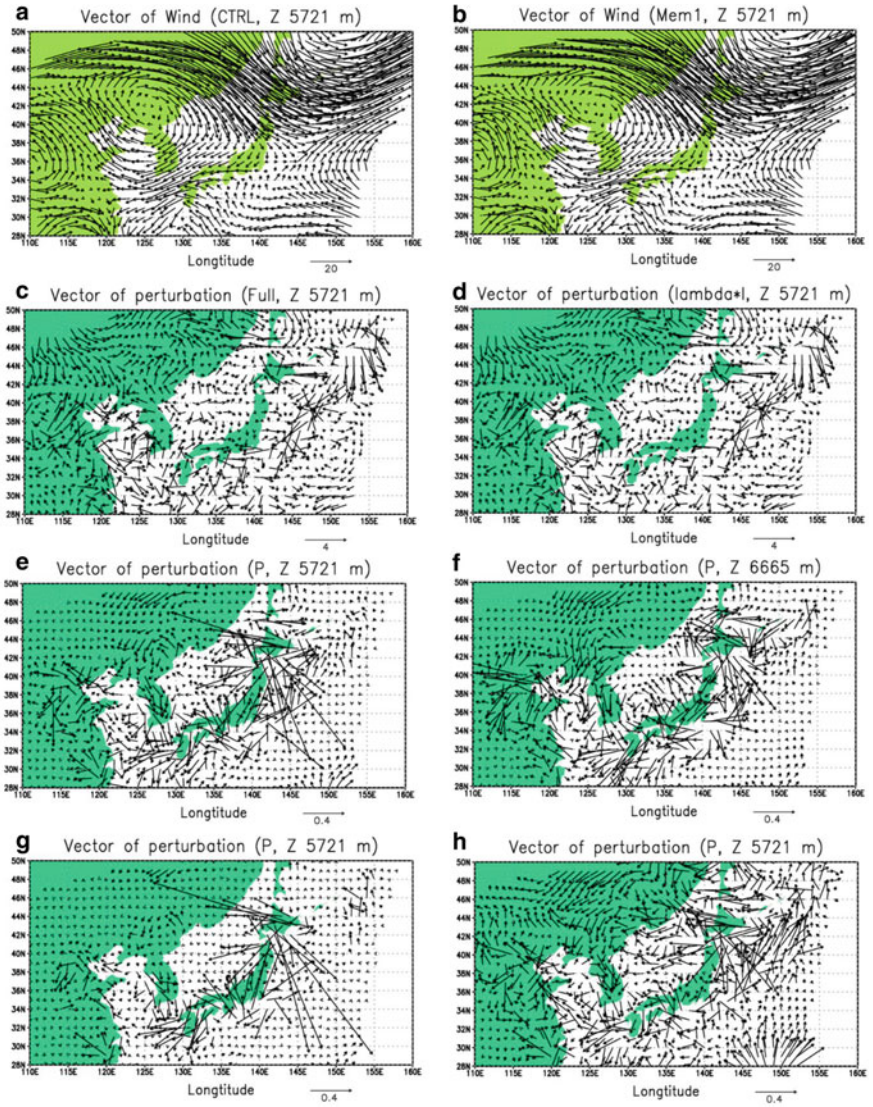

Fig. 10 **a** Horizontal wind field at z = 5721 m on 5 Aug 2016 by NHM-LETKF analysis with the horizontal localization scale of 200 km. **b** Same as in **a** but for member 01. **c** Analysis perturbation of member 01. **d** Diagonal ($\mathbf{D} = \overline{\lambda}\ \mathrm{I}$) components of analysis perturbation in Eq. (6). **e** Same as in **d** but for **P** components. **f** Same as in **e** but for z = 6665 m. **g** Same as in **e** but the horizontal localization scale is 100 km. **h** Same as in **e** but the horizontal localization scale is 400 km

the K-F cumulus parameterization scheme was switched off. Real observations same as in the 15 km NHM-LETKF were assimilated. The 2 km LETKF was started from 1200 UTC 2 August, 2016 with 3 hourly DA cycles, targeting on analysis at 1800 UTC 2 August, using 15 km NHM-LETKF results as the very initial and lateral boundary perturbations. Relaxation to prior spread or multiplicative inflation is used as in the 15 km NHM-LETKF. Horizontal localization scale was fixed to 200 km, while vertically, 0.2, 0.1 and 0.05 (ln p) vertical localizations were tested.

Figure 11a–c show observed horizontal surface winds and three-hour accumulated precipitation from 1500 to 2100 UTC 2 August 2016. After 1800 UTC, an area of intense local rain is seen around Tokyo. Corresponding forecasts by 15 km NHM-LETKF and 2 km NHM-LETKF are indicated in Fig. 11d–i, respectively. 15 km NHM-LETKF generally well reproduced the southeasterly surface wind field but tended to miss the intense rains. The 2 km NHM-LEKF reproduces localized intense rains around Tokyo for some extent at 1800 and 2100 UTC.

Figure 12 indicates horizontal wind field at 24th level at 1500 UTC 2 Aug 2016 by 2 km NHM-LETKF. At this height (z = 5276 m), southwesterly winds prevailed suggesting warm and humid air advection over the Kanto area. Analysis perturbations (Fig. 12c) and its diagonal (**D**) components (Fig. 2d) are very similar. Perturbations by **P** components (Fig. 12e) are much smaller than **D** components and noisy.

Figure 13 indicates vertical wind field at the same level. The vertical motion in the control analysis (Fig. 13a) is weak because of the ensemble mean. In the analysis field of member 01 (Fig. 13b), spots of intense updrafts are seen in the southwest of Tokyo, which corresponds to deep convection bred in the guess forecast. Several weak spots of up/down drafts are reflection of the off-diagonal components. These randomly distributed spots of up/down drafts are distinct in the perturbations by **P** components (Fig. 13e).

Figure 14a–c show the vertical cross-section of vertical winds of member 14 by 2 km NHM-LETKF along a north–south plane from 34.8 N to 35.8 N through a convective cell in different vertical localization scales. Here, we show the result of member 14 as an example which showed intense rains near Tokyo in both full-T LETKF and diagonal LETKF (see Fig. 16). Intense upward motions are analyzed corresponding to deep convection at the guess field. Seemingly, these analyses are similar each other, but perturbations by **P** components (Fig. 14d–f) are very different depending on the vertical localization scale. In case of a small vertical localization scale (Fig. 14d; 0.05 ln(p)), the patterns are noisy while in case of a larger localization (Fig. 14f; 0.2 ln (p)), the perturbation patterns have longer structure in vertical. The small-scale perturbations vanish rapidly when the model time integration started. At FT = 2 min, the patterns in three figures become somewhat similar (Fig. 14g–i).

Figure 15a shows time evolutions of the maximum horizontal divergence at 19th level (Z = 3321 m) in full-T LETKF (blue) and in diagonal LETKF (red) in member 14 by 2 km NHM-LETKF. Maximum values of horizontal divergence in full-T experiment tend to be larger than those in diagonal LETKF, suggesting the start-up stirring in the full-T LETKF. Figure 15b shows time evolutions of the maximum updraft for entire model domains. The maximum updrafts in full-T LETKF tend to be larger than that of diagonal LETKF at the first 35 min.

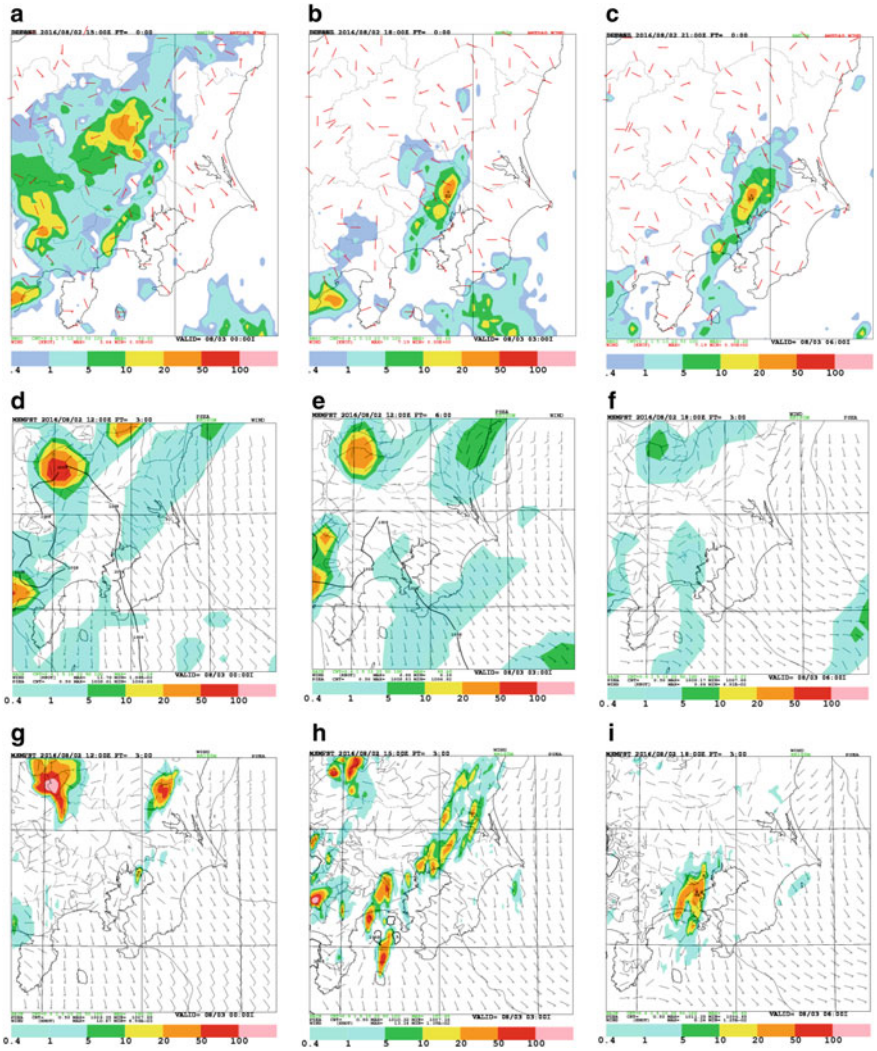

Fig. 11 Observed horizontal surface winds and three-hour accumulated precipitation at **a** 1500 UTC, **b** 1800 UTC, and **c** 2100 UTC on 2 August 2016. **d**, **e** and **f** are the same as in **a**, **b**, and **c**, respectively, but for forecast by the 15 km NHM-LETKF, **g**, **h** and **i** are the same as in **d**, **e** and **f**, respectively, but for the 2 km NHM-LETKF

4.2 Cycle Experiments and Verification

Next, we conducted cycle DA experiments with full-T and diagonal LETKF. Ensemble forecast with 2 km started at 1800 UTC 1 August using downscale of the 15 km NHM-LETKF ensemble as the initial conditions. Then, analyses of the cycle DA were conducted eight times from 2100 UTC 1 August to 1800 UTC 2

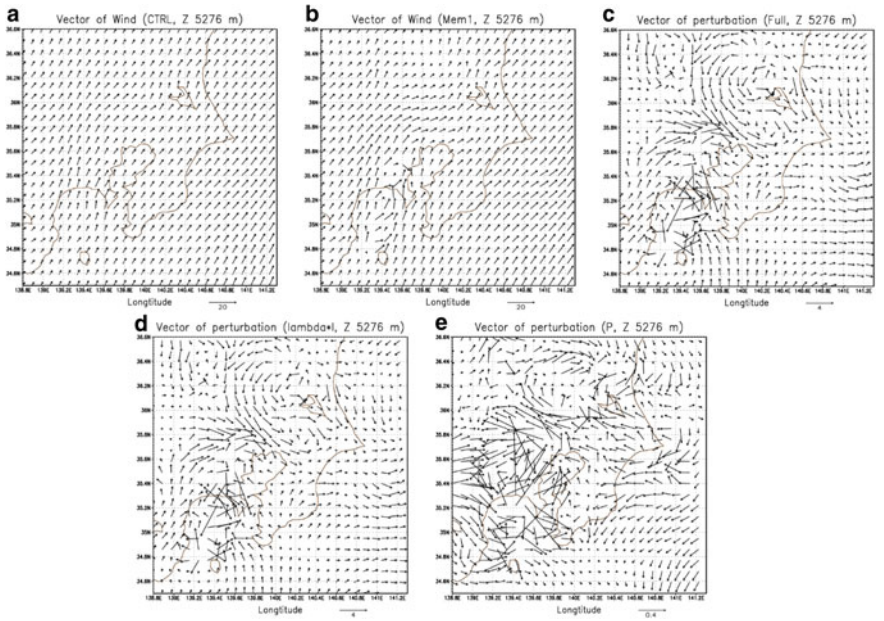

Fig. 12 Horizontal wind field at 24th level (z = 5276 m) at 1500 UTC 2 Aug 2016 by 2 km NHM-LETKF. **a** Control analysis. **b** Analysis of member 01. **c** Analysis perturbation of member 01. **d** Diagonal ($\mathbf{D} = \overline{\lambda}\ \mathrm{I}$) components of analysis perturbation in Eq. (6). **e** Same as in **d** but for **P** components

August. Other specifications of JMA-NHM and LETKF are the same as in the former subsection.

Figure 16 shows precipitation forecast by the two LETKFs (member 14). Figure 16a, c show three-hour precipitation by 2 km full-T LETKF at 1800 UTC and 2100 UTC, respectively, and Fig. 16b is one-hour precipitation from 1500 to 1600 UTC to show the initial spin-up after the analysis at 1500 UTC. Corresponding forecasts by the diagonal LETKF are shown in the lower panels (Fig. 16d–f). Comparing with the observed three-hour precipitation on the day (Fig. 11b, c), the diagonal LETKF better reproduced the rainfall around Tokyo. The large difference between the two LETKFs is in the precipitation amount at spin-up period. In the forecast from full-T LETKF (Fig. 16b) rainfall amount in the initial 1-h is smaller than diagonal LETKF (Fig. 16e).

Figure 17a, b show composite of time evolutions of precipitation intensity for every time step at nine forecast cycles (colored) and their average (bold black) by the full-T and diagonal LETKFs. Here, time evolutions of precipitation intensity in member 14 for every forecast cycle are superimposed and the average is plotted by a black line. In case of full-T LETKF (Fig. 17a), precipitation intensities change rapidly within the first 10 min just after the model start-up with a small bump and then precipitation tends to decrease after that in many DA cycles. Similar tendency

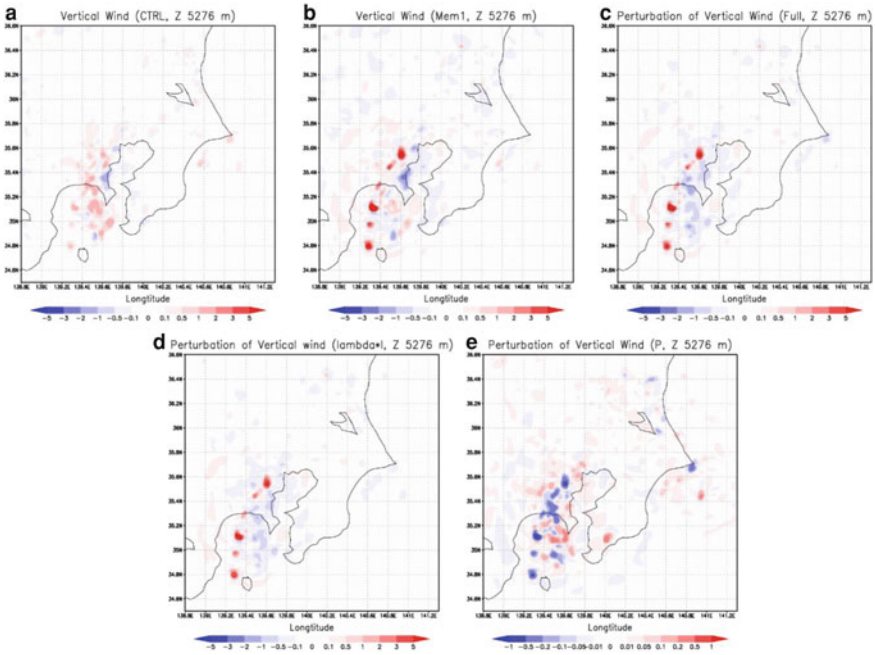

Fig. 13 Same as in Fig. 12 but for vertical wind field at 24th level (z = 5276 m)

is also seen in some members in the diagonal LETKF (Fig. 17b), but generally the change of precipitation at the model spin-up is small and the average precipitation intensity (MEAN) is more constant than that in full-T LETKF. As for period 1500 UTC to 1800 UTC (brown), difference of rainfall amount in the first one-hour shown in Fig. 16 is evident. Difference in the average precipitation (black) is not large but the decrease of rainfall in the model spin-up in full-T LETKF is observable. Figure 17c, d show the verification of precipitation forecast against the radar-rain gauge precipitation analysis of JMA. Here intensity-scale diagrams for fraction skill scores (FSS; Ebert 2008; Duc et al. 2013) for three-hour accumulated precipitations in the forecast period are indicated. Difference of the performances of the two LETKFs is evident especially for weak to moderate rains on 0.1–10 mm/h (Fig. 17e).

Figure 18 is precipitation forecast at 1800 and 2100 UTC by the two LETKFs for control run. Different from the forecast in member 14 shown in Fig. 16, the difference between the full-T LETKF and the diagonal LETKF is not large. Precipitation amount at the first one hour by the diagonal LETKF (Fig. 18e) for 1800 to 1900 UTC is still slightly larger than that by the full-T LETKF (Fig. 18b). Figure 19a–c show composites of time evolution of precipitation intensity for every time step at nine

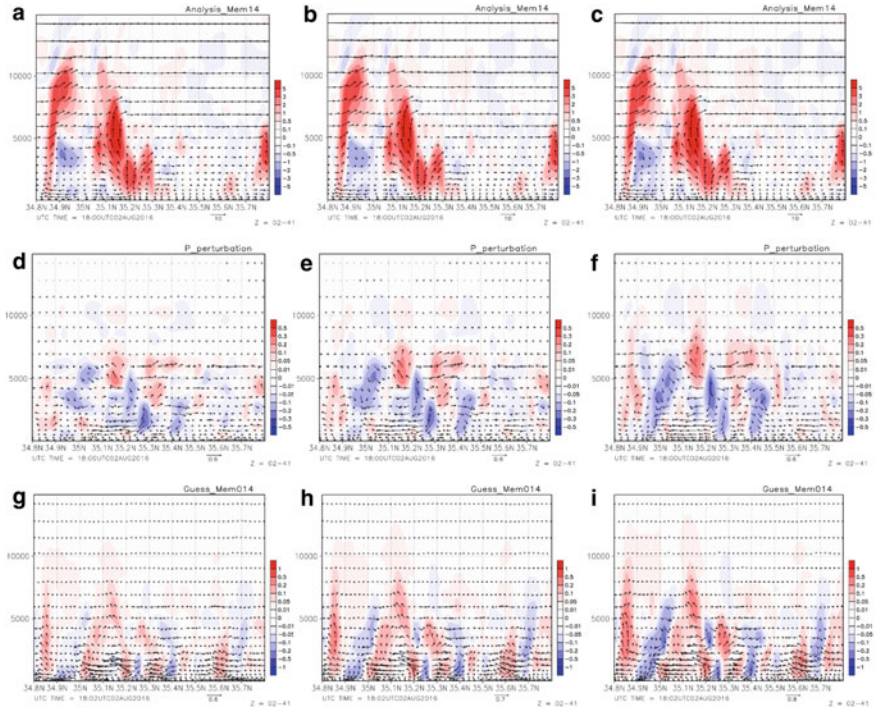

Fig. 14 Vertical (north–south) cross-section of vertical winds of member 14 by 2 km NHM-LETKF: analysis at 1500 UTC 2 Aug 2016 for the case of **a** vertical localization 0.05 ln(p), **b** vertical localization 0.1 ln(p), and **c** vertical localization 0.2 ln(p), **d**, **e** and **f** are the same as in **a**, **b** and **c**, respectively, but for perturbations by **P** components, **g**, **h** and **i** are the same as in **d**, **e** and **f**, respectively, but for the forecast at 2 min

DA cycles and verification scores in the two LETKFs for control forecast. The large difference in precipitation intensity from Fig. 17 is the decrease of precipitation intensity after the spin-up. This depression of precipitation comes from the lack of the intense updraft in the initial condition (Fig. 13a), where the ensemble mean is used in the first guess. Fukui et al. (2018) pointed out lack of intense rains from LETKF analysis using ensemble mean for first guess. They attributed this property to the lack of high wave number spectra in the ensemble mean, and used the deterministic forecast as the background (first guess) field to increase rainfall in their regional reanalysis. The depression of initial rains is seen both in full-T and diagonal LETKFs, because even the diagonal LETKF, the ensemble mean was used as the first guess in our experiment. Figure 19c, d show the FSS verification results for control run. Differences of FSSs in the two LETKFs are also small but diagonal LETKF (Fig. 19d) was slightly better than the full-T LETKF (Fig. 19c) for weak to moderate rains (Fig. 19e).

Recently, Duc et al. (2020b) proposed a unifying theory for covariance inflation in the ensemble Kaman filter, and discussed an inflation function which transforms

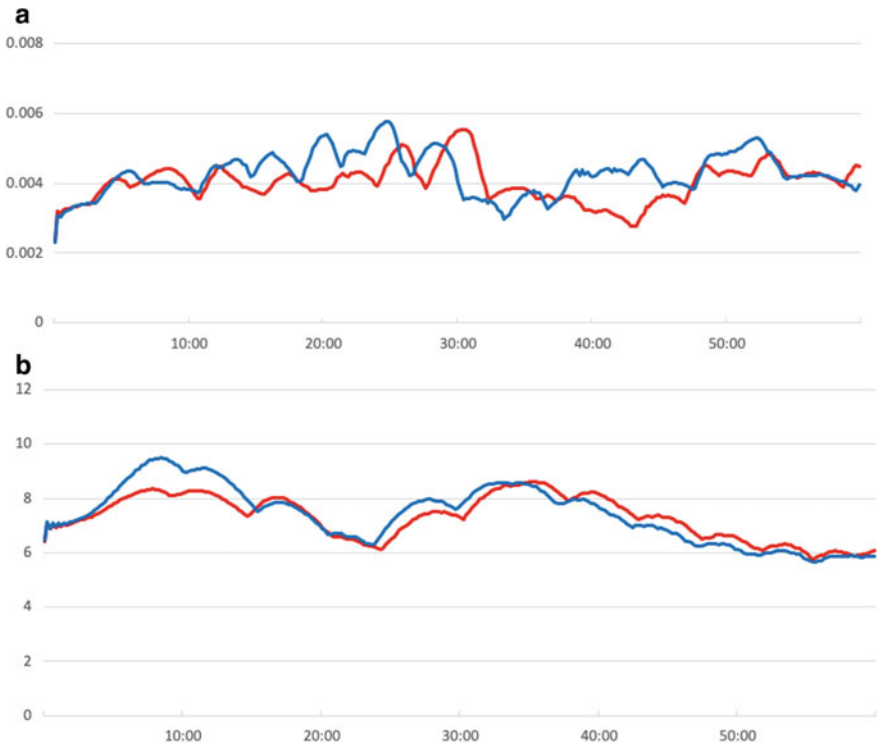

Fig. 15 Time evolution of **a** the maximum horizontal divergence (in s^{-1}) at 19th level, and **b** the maximum updraft (in ms^{-1}) for entire model domains, from the full-T LETKF (blue) and the diagonal LETKF (red) in member 14 by the 2 km NHM-LETKF. Horizontal axes indicate the forecast time (in min)

background perturbations to analysis perturbations. They compared several covariance inflation methods as the response function which reduces factors of background singular values in analysis, and pointed out that RTPP is a mixture of the diagonal ETKF and RTPS and puts more inflation on large-scale processes, while RTPS puts more inflation on small-scale processes. Diagonal ETKF is the constant inflation function and has a similar inflation property to RTPS in terms of the response function. In our experiments, we applied diagonal LETKF with RTPS but this inflation property of diagonal LETKF likely influenced the results. Additional experiments with multiplicative inflation and without inflation were conduct with full-T and diagonal ETM. In this case, even no inflation did not yield filter divergence because the computational domain is small and the meteorological field is restricted by the boundary condition.

Figure 20 shows time evolutions of ensemble spreads of meridional winds in eight DA cycles from 2100 UTC 1 August to 2100 UTC 2 August in the 2 km NHM-LETKF by six experiments using different inflation methods (RTPS, multiplicative, and no inflation) with diagonal and full-T ETM. Ensemble spreads decrease at analyses

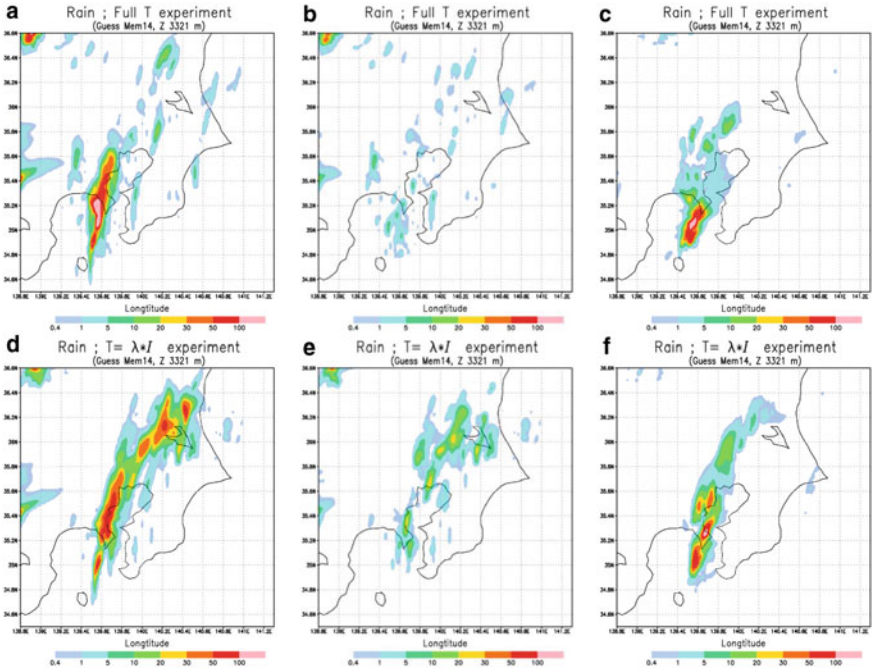

Fig. 16 **a** Three-hour accumulated precipitation at 1800 UTC, **b** one-hour precipitation for 1500 to 1600 UTC, and **c** three-hour accumulated precipitation at 2100 UTC on 2 August 2016 by member 14 in the 2 km LETKF for the cycle experiment. **d**, **e** and **f** are the same as in **a**, **b** and **c**, respectively, but for the 2 km diagonal LETKF

and in the most DA cycles, they increase with time except RTPS after 0900 UTC 2 August. Ensemble spreads of diagonal LETKF are always larger than full-T LEKF because of covariance inflation property in the diagonal LETKF. Verifications of QPF performance were checked but there were no large differences in diagonal and full-T LETKF for multiplicative inflation (figure not shown).

5 Summary and Concluding Remarks

We checked analysis perturbations in LETKF of SPEEDY and JMA-NHM and confirmed that most part of the perturbations are from diagonal part of ETM, due to the diagonally predominant property discussed by Duc et al. (2020a). Off-diagonal part is spatially noisy both in horizontally and vertically, depending on the localization scales. Power spectrum of the off-diagonal perturbation indicated that the power of higher wave numbers were relatively higher in small horizontal localization scales. This tendency suggests that the off-diagonal perturbations do not hold

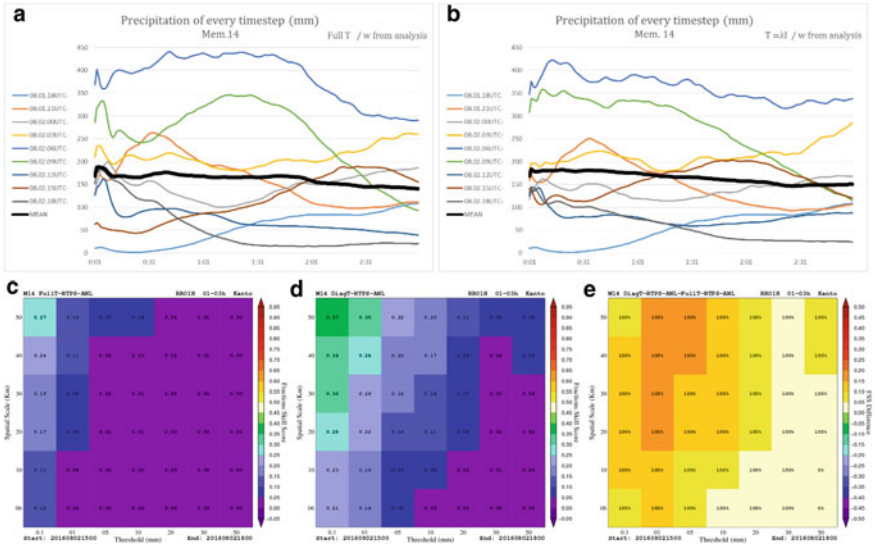

Fig. 17 **a** Composite of time evolution of precipitation intensity for every time step at nine DA cycles (colored) and the average (bold black) in the 2 km LETKF. Horizontal axis indicates the forecast time from analysis (in hour:min). **b** Same as in **a** except for the diagonal LETKF. **c** Intensity-scale diagram on fraction skill scores for three-hour accumulated precipitations in the DA cycle period. Horizontal axis is thresholds of rain intensity and vertical axis indicates spatial scales which allow positional lags. **d** Same as in **d** except for the diagonal LETKF. **e** Difference of **c** and **d**

the global structures of the meteorological field and may be harmful for the model spin-up and the succeeding forecast.

Vertical structures of ensemble perturbations in a cloud resolving model (2 km resolution JMA-NHM) and model spin-up were checked. Small scale noises in the initial perturbation rapidly vanish after the model start, but the horizontal divergence and vertical updraft of full-T LETKF tended to be larger than the diagonal LETKF until 30 to 40 min after the model start. Verification of precipitation shows that rainfalls at the very first one hour in DA cycles of full-T LETKF are smaller than diagonal LETKF and the FSS of diagonal LETKF was better than full-T LETKF for a specific perturbed member. This difference was much smaller in the control run.

As pointed out by Duc et al. (2020b), diagonal LETKF has inflation effect. Thus, evaluation of the diagonal LETKF likely should be done in a situation of well-tuned inflation, or more sophisticated experimental setting. Diagonal LETKF may be useful at least as a perturbation generator for extended runs because it removes the global inconsistency in the meteorological field, but for DA cycles, orthogonal property must be secured by other methods (e.g., lateral boundary perturbations in a regional model). Side effect of the full-T ETM may be reduced in a system of the large number of ensemble member, because we can apply a longer localization scale in such the system. Further studies are needed to evaluate ET perturbations.

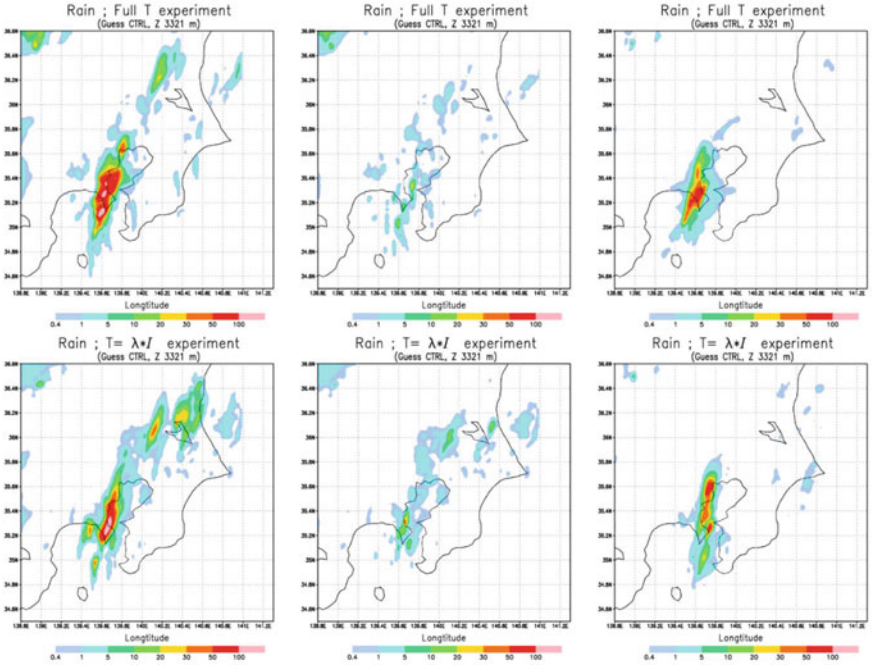

Fig. 18 Same as in Fig. 16 except for control run

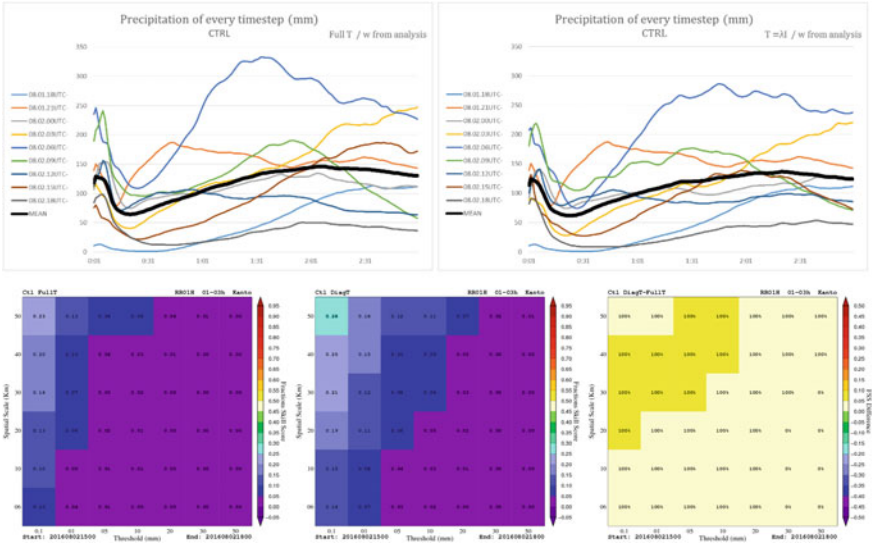

Fig. 19 Same as in Fig. 17 except for control run

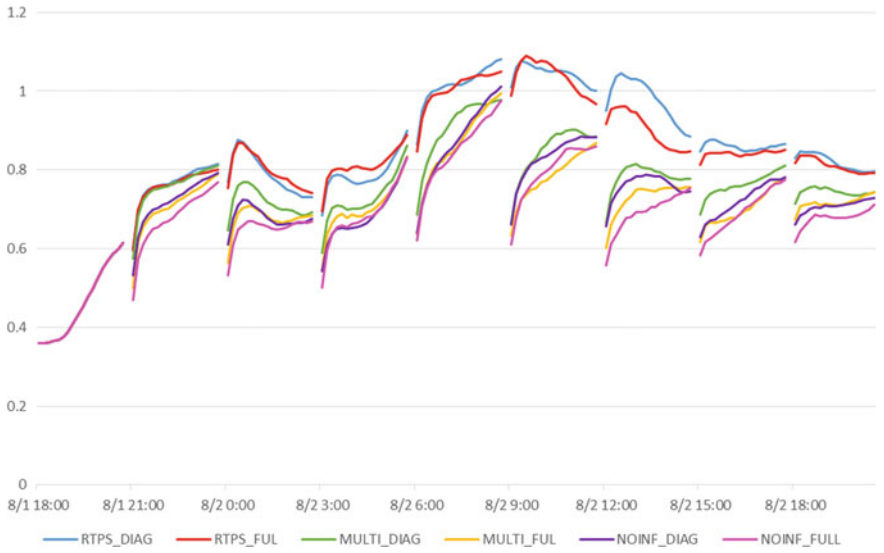

Fig. 20 Time evolution of ensemble spreads of meridional winds in eight DA cycles from 2100 UTC 1 August to 2100 UTC 2 August in the 2 km NHM-LETKF. 'RTPS', 'MULTI', and 'NOINF' before underbars indicate inflation method using RTPS, multiplicative, and no inflation, respectively. 'DIAG' and 'FUL' after underbars mean the diagonal LETKF and the full-T LETKF, respectively

Acknowledgements This study was partly supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) as "Program for Promoting Researches on the Supercomputer Fugaku (hp200128, hp210166)" and Grant-in-Aid for Scientific Research (B) (16H04054) from Japan Society for the Promotion of Science (JSPS). The authors thank Takemasa Miyoshi of RIKEN, Takuya Kawabata of Meteorological Research Institute, Sho Yokota and Masaru Kunii of the Japan Meteorological Agency for their help and comments on data assimilation experiments. We appreciate careful check by an anonymous reviewer which improved the maturity of the manuscript.

References

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter. Part 1: theoretical aspects. Mon Weather Rev 129:420–436

Bowler NE, Arribas A, Mylne KR, Robertson KB, Beare SE (2008) The MOGREPS short-range ensemble prediction system. Q J R Meteorol Soc 134:703–722

Duc L, Saito K, Seko H (2013) Spatial-temporal fractions verification for high resolution ensemble forecasts. Tellus 65:18171. https://doi.org/10.3402/tellusa.v65i0.18171

Duc L, Kuroda T, Saito K, Fujita T (2015) Ensemble Kalman Filter data assimilation and storm surge experiments of tropical cyclone Nargis. Tellus A 67:25941. https://doi.org/10.3402/tellusa.v67.25941

Duc L, Saito K, Hotta D (2020a) An explanation for the diagonally predominant property of the positive symmetric ensemble transform matrix. J Meteorol Soc Jpn 98:445–462

Duc L, Saito K, Hotta D (2020b) Analysis and design of covariance inflation methods using inflation functions. Part 1: theoretical framework. Q J R Meterol Soc (in press)
Ebert E (2008) Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol Appl 15:51–64
Fukui S, Iwasaki T, Saito K, Seko H, Kunii M (2018) A feasibility study of the high-resolution regional reanalysis over Japan assimilating only conventional observations as an alternative to the dynamical downscaling. J Meteorol Soc Jpn 96:565–585
Kunii M (2014) 2014: Mesoscale data assimilation for a local severe rainfall event with the NHM–LETKF system. Weather Forecast 29:1093–1105
Miyoshi T (2005) Ensemble Kalman filter experiments with a primitive-equation global model. PhD dissertation, University of Maryland, College Park, 197 pp
Miyoshi T (2011) The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. Mon Weather Rev 139:1519–1535
Miyoshi T, Aranami K (2006) Applying a four-dimensional local ensemble transform Kalman filter (4D-LETKF) to the JMA Nonhydrostatic Model (NHM). SOLA 2:128–131
Molteni F (2003) Atmospheric simulations using a GCM with simplified physical parametrizations. I: model climatology and variability in multi-decadal experiments. Clim Dyn 20(2–3):175–191
Ota Y, Ikegami M, Yamaguchi H (2019) Upgrade of initial perturbations made using the Local Ensemble Transform Kalman Filter in JMA's Global EPS. CAS/JSC WGNE Res Act Atmos Ocean Model. http://bluebook.meteoinfo.ru/uploads/2019/docs/06_Ota_Yoichiro_Ensemble_forecast,_Ensemble_Kalman_Filter.pdf
Reich H, Rhodin A, Schraff C (2011) LETKF for the nonhydrostatic regional model COSMO-DE. COSMO Newsl 11:27–31
Saito K (2012) The Japan Meteorological Agency nonhydrostatic model and its applications to operation and research. In: Atmospheric model applications. InTech, pp 85–110.https://doi.org/10.5772/35368
Saito K, Fujita T, Yamada Y, Ishida J, Kumagai Y, Aranami K, Ohmori S, Nagasawa R, Kumagai S, Muroi C, Kato T, Eito H, Yamazaki Y (2006) The operational JMA nonhydrostatic mesoscale model. Mon Weather Rev 134:1266–1298
Saito K, Ishida J, Aranami K, Hara T, Segawa T, Narita M, Honda Y (2007) Nonhydrostatic atmospheric models and operational development at JMA. J Meteorol Soc Jpn 85B:271–304
Saito K, Kuroda T, Kunii M, Kohno N (2010) Numerical simulations of Myanmar cyclone Nargis and the associated storm surge. Part 2: ensemble prediction. J Meteorol Soc Jpn 88:547–570
Saito K, Hara M, Kunii M, Seko H, Yamaguchi M (2011) Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the Meteorological Research Institute for the WWRP Beijing 2008 Olympics Research and Development Project (B08RDP). Tellus 63A:445–467
Saito K, Seko H, Kunii M, Miyoshi T (2012) Effect of lateral boundary perturbations on the breeding method and the local ensemble transform Kalman filter for mesoscale ensemble prediction. Tellus 64:11594. https://doi.org/10.3402/tellusa.v64i0.11594
Toth Z, Kalnay E (1997) Ensemble forecasting at NCEP and the breeding method. Mon Weather Rev 125:3297–3319
Wang X, Bishop CH, Julier SJ (2004) Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? Mon Weather Rev 132:1590–1605
Yokota S, Kunii M, Aonashi K, Origuchi S (2016) Comparison between four-dimensional LETKF and ensemble-based variational data assimilation with observation localization. SOLA 12:80–85
Yokota S, Niino H, Seko H, Kunii M, Yamauchi H (2018) Important factors for tornadogenesis as revealed by high-resolution ensemble forecasts of the Tsukuba supercell tornado of 6 May 2012 in Japan. Mon Weather Rev 146:1109–1132
Zhang F, Snyder C, Juanzhen S (2004a) Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. Mon Weather Rev 132:1238–1253. https://doi.org/10.1175/1520-0493(2004)132<1238:ioieao>2.0.CO;2

Zhang F, Snyder C, Juanzhen S (2004b) Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. Mon Weather Rev 132:1238–1253. https://doi.org/10.1175/1520-0493(2004)132%3c1238:ioieao%3e2.0.CO;2

Stochastic Representations for Model Uncertainty in the Ensemble Data Assimilation System

Sujeong Lim and Seon Ki Park

Abstract The ensemble data assimilation system is beneficial to express flow-dependent model errors. Furthermore, the effectiveness of this system depends on the accuracy of the flow-dependent background error covariance. However, the background error covariance is often underestimated due to limited ensemble size, sampling errors and model errors, which causes a filter divergence problem—the analysis state diverges from the nature stage ignoring the observation influence. As one of the remedies to solve this problem, the stochastic representations address the model-related uncertainties by perturbing the model tendency or parameters using a random forcing to replenish the insufficient model errors. In this study, we implemented a stochastic perturbation hybrid tendencies (SPHT) scheme, which perturbs both physical tendency and dynamical tendency using the random forcing, and assessed its impact on the spread of ensemble forecast and ensemble mean error.

1 Introduction

Ensemble data assimilation (EnsDA) finds the best initial conditions of the numerical weather prediction (NWP) model using model forecasts and their error covariance as well as observations Evensen (1994). In particular, it describes the flow-dependent forecast error covariance through an ensemble of the model forecasts. Therefore, it contains uncertainties in both the initial conditions and the model. Model uncertainty representations can be distinguished from actual model errors: the former samples model perturbations from some distribution while the latter presents only one realization per model and forecast Leutbecher et al. (2017). In this study, we focus on the model uncertainty in the EnsDA system using the stochastic representations that simulate the errors of model tendencies from random components.

S. Lim · S. K. Park (✉)
Department of Climate and Energy Systems Engineering, Ewha Womans University, Seoul 03760, Republic of Korea
e-mail: spark@ewha.ac.kr

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_6

In EnsDA, the model uncertainty is used in the ensemble background error covariance (BEC) through the 6-h forecasts. However, it is found to be underdispersive due to the limited ensemble size, sampling error, and imperfect model parametrization, resulting in over-confidence in ensemble forecasts Buizza et al. (2005). This problem is usually covered by covariance inflation, e.g., multiplicative inflation Anderson and Anderson (1999), additive inflation Mitchell and Houtekamer (2000), combined multiplicative and additive inflation Whitaker and Hamill (2012), relaxation to the prior Zhang et al. (2004), multischeme ensembles Meng and Zhang (2007), and so on.

In general, the ensemble BEC, composed of an ensemble spread ($\sigma(x)$), i.e., the standard deviation to the ensemble mean ($\bar{x}$), should reflect the ensemble mean error ($e(\bar{x})$) because the ensemble spread distribution determines the analysis status (see Fig. 1). Here, the model error is expressed by the distance between the ensemble mean and the observation while the ensemble spread is represented by the pre-described ensemble distribution. The optimal ensemble spread is expected to have a spread similar to the ensemble mean (Fig. 1a), i.e.,

$$\sigma(x) \approx e(\bar{x}); \tag{1}$$

then, the analysis includes reliable information from the model and observations. The underdispersive ensemble members show a smaller ensemble spread compared to the model error, i.e.,

$$\sigma(x) \ll e(\bar{x}), \tag{2}$$

where the analysis ignores the observation and trusts the model more due to small ensemble BEC (Fig. 1b). The overdispersive ensemble members show the larger ensemble spread compared to the model error, i.e.,

$$\sigma(x) \gg e(\bar{x}). \tag{3}$$

In this case, the analysis ignores the model errors and relies more on the observation due to the larger ensemble BEC (Fig. 1c).

To remedy the general problem of the underdispersive model error, various stochastic schemes can be used, e.g., Buizza et al. (1999), Shutts (2005), Palmer et al. (2009). It is based on the fact that the NWP models represent the physical process with simplifications and approximations due to incomplete knowledge and computational costs. The European Centre for Medium-Range Weather Forecasts (ECMWF) developed the original version of the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme—called the Buizza-Miller-Palmer (BMP) scheme Buizza et al. (1999)—for the first time and introduced the stochastic representation of model uncertainty that perturbs the total parametrized physics tendencies using the random forcing. After major revisions on random patterns and distribution of perturbations in 2009 Palmer et al. (2009), the BMP scheme has evolved into the SPPT scheme. Since

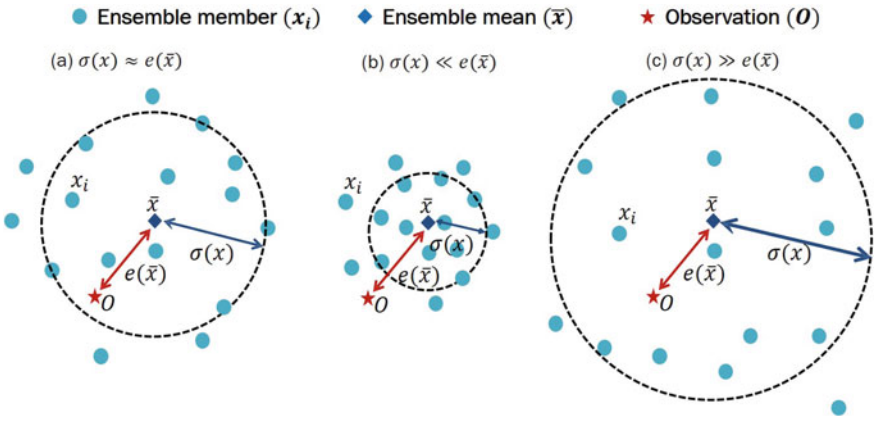

Fig. 1 Schematic diagram of the ensemble spread, $\sigma(x)$, and the ensemble mean error, $e(\bar{x})$: **a** optimal ensemble spread, **b** underdispersive ensemble spread, and **c** overdispersive ensemble spread

then, the SPPT scheme has been employed by many operational NWP centers, e.g., Environment Canada (EC), Japan Meteorological Agency (JMA), the United States National Centers for Environmental Prediction (NCEP), and the United Kingdom Met Office (UKMO), for their global prediction systems Charron et al. (2010), Leutbecher et al. (2017). It has also been implemented in regional prediction systems, e.g., the Application of Research to Operations at Mesoscale convection-permitting model (AROME) of Meteo-France Bouttier et al. (2012) and the Weather Research and Forecasting (WRF) model Romine et al. (2014), Berner et al. (2015).

Afterward, the Stochastic Kinetic Energy Backscatter (SKEB) scheme was introduced to represent the model uncertainties for scale interactions what is absent in a truncated numerical model by randomly perturbing the stream function and potential temperature tendencies Shutts (2005). The SKEB has also been used for global ensembles in many operational center, e.g., ECMWF, EC, and UKMO Charron et al. (2010), Tennant et al. (2011), Sanchez et al. (2016) as well as regional ensembles (e.g., WRF) Berner et al. (2011), Berner et al. (2015). Recently, a stochastic convective backscatter scheme has been introduced Shutts (2015).

Alternately, the Stochastically Perturbed Dynamical Tendencies (SPDT) scheme, which perturbs the total dynamical tendencies using the random forcing, was introduced: the scheme proved to be effective in global ensemble forecasting Koo and Hong (2014), e.g., in the Global/Regional Integrated Model system (GRIMs) Hong et al. (2013). By combining the SPDT and SPPT schemes, Lim et al. Lim et al. (2020) devised the stochastic perturbation hybrid tendencies (SPHT) scheme to improve the underestimated BEC in the EnsDA system of the Korean Integrated Model (KIM) Hong et al. (2018)—a global model developed at the Korea Institute of Atmospheric Prediction System (KIAPS).

In this study, we introduce the SPHT scheme in the EnsDA system as a covariance inflation method in order to solve the underestimated ensemble BEC by taking into account that model variables are integrated by changes in physical and dynamic tendencies every time. The methodology is described in Sect. 2, and the experimental designs and results are in Sects. 3 and 4, respectively. Section 5 provides the summary and suggests the direction to further development.

2 Methodology

2.1 Local Ensemble Transform Kalman Filter (LETKF)

We employ the EnsDA system of KIAPS, which is a four-dimensional local ensemble transform Kalman filter (4D-LETKF). The analysis is obtained by assimilating the available observations within a local region Hunt et al. (2007), Shin et al. (2016), Shin et al. (2018). This LETKF also provides ensemble perturbations to the hybrid four-dimensional ensemble-variational (H4DEV) system, another data assimilation system operated by KIAPS. In this study, we use the LETKF results just to distinguish the changes of ensemble BEC.

The control variables are zonal wind, meridional wind, potential temperature, mixing ratio, and surface pressure. The KIM Package for Observation Processing (KPOP) provides quality-controlled real observations to the data assimilation system Kang et al. (2018), including the sonde, surface, aircraft, Global Positioning System-Radio Occultation (GPS-RO), Infrared Atmospheric Sounding Interferometer (IASI), Advanced Microwave Sounding Unit-A (AMSU-A), Cross-track Infrared Sounder (CrIS), Microwave Humidity Sounder (MHS), Advanced Technology Microwave Sounder (ATMS), Atmospheric Motion Vectors (AMVs), and tropical cyclone initialization.

In resolving the filter divergence problems in LETKF, three approaches used to be applied in terms of ensemble size, localization, and inflation method, which are specified below for this study:

1. *Ensemble size*: Increasing the ensemble size is commonly limited due to computational costs. At the early stage of developing LETKF, we used an ensemble size of 30 members, which is now increased to 50 members and is used in this study.
2. *Localization*: We implemented both horizontal and vertical localizations. The horizontal localization is expressed by a Gaussian-like piecewise fifth-order rational function Gaspari and Cohn (1999), Miyoshi (2011) varying from 660 to 1800 km depending on vertical levels Kleist and Ide (2015). The vertical localization varies depending on the observational types (e.g., conventional versus satellite data). For conventional data, it is defined by a Gaussian-like rational function, represented by $2\sqrt{10/3} \cdot \sigma_v$ where σ_v is chosen to be $0.2 \ln p$ for wind and surface pressure and $0.1 \ln p$ for mass variables. For the satellite radiance data, the vertical weight-

ing function is defined by the gradient of transmittance of the measured radiance Thépaut (2003).

3. *Inflation method*: We used two inflation methods in this study. The additive inflation adds the perturbations randomly sampled from the bias-corrected lagged forecast differences to each ensemble member after the analysis step Whitaker et al. (2008). The relaxation to prior spread (RTPS) relaxes the ensemble standard deviation of analysis back to the background Whitaker and Hamill (2012). However, the LETKF still requires additional inflation method to make a sufficient ensemble BEC: we will cover it through the stochastic representation method in this study.

2.2 Numerical Weather Prediction (NWP) Model

We employ the KIM, a global NWP model developed at KIAPS Hong et al. (2018), which has recently been operationally implemented in the Korea Meteorological Administration (KMA). It is a non-hydrostatic model on a cubed sphere with state-of-the-art physics parametrization packages—including radiation, gravity wave drag, vertical mixing, convection, cloud physics, and so on.

Since our concern is a BEC in the data assimilation process, we only deal with the 6-h forecast (prior) results. The ensemble size is 50 members and the horizontal resolution for the ensemble forecast is 50 km. The initial conditions of the ensemble forecast have been generated by the lagged forecast difference samples, which are used to generate the static BEC in H4DEV Kwon et al. (2018).

2.3 Stochastic Perturbation Hybrid Tendencies (SPHT) Scheme

In this study, we introduce a stochastic perturbation hybrid tendencies (SPHT) scheme that perturbs the dynamic tendency $\left(\frac{\partial \mathbf{x}}{\partial t}\right)_{dyn}$ and the physical tendency $\left(\frac{\partial \mathbf{x}}{\partial t}\right)_{phy}$ of the model variables $\mathbf{x}$ at each time step n using the multiplicative random forcing (r):

$$\begin{aligned} \mathbf{x}^{n*} &= \mathbf{x}^{n} + (1+\mu r)\left(\frac{\partial \mathbf{x}^{n}}{\partial t}\right)_{dyn} \Delta t \\ \mathbf{x}^{n+1} &= \mathbf{x}^{n*} + (1+\mu r)\left(\frac{\partial \mathbf{x}^{n*}}{\partial t}\right)_{phy} \Delta t \end{aligned} \tag{4}$$

where $\mu \in \{0, 1\}$ represents the vertical tapering function ($e^{\eta-1}$) in the generalized vertical coordinate η. The model variable $\mathbf{x}$ consists of temperature and humidity mixing ratio only. Note that in the KIM, physics and dynamics are coupled by time-splitting method; thus, this approach differs from the method of perturbing total model tendency by simply adding up perturbations of two tendencies, i.e.,

$$(1 + \mu r)\left[\left(\frac{\partial \mathbf{x}}{\partial t}\right)_{dyn} + \left(\frac{\partial \mathbf{x}}{\partial t}\right)_{phy}\right].$$

Here, r is a 2-dimensional value following the Gaussian distributed zero-mean perturbation considering spatial and temporal correlation. Occasionally, μ is applied to perturbations for the upper model levels to avoid the instability issue. The amplitude is determined by the standard-deviation (σ), and the length and time scales are based on the decorrelation lengths (L) and times (t), respectively.

The SPPT assumes that the model errors from the parametrized physical tendency are proportional to the total physical tendency Buizza et al. (1999), Palmer et al. (2009) while the SPDT assumes that the model errors from the dynamic tendency concern with the computational representations of the underlying partial differential equations Koo and Hong (2014). Since both methods deal with the model tendency, we devised a hybrid stochastic scheme (i.e., SPHT) by combining the two perturbation tendencies based on Eq. (4). The SPHT scheme is applied to the ensemble forecasting in LETKF to obtain an ensemble BEC.

3 Experimental Designs

To identify how the SPHT scheme increases the ensemble spread, we designed two experiments: CTRL (representing the *control* run) is without the SPHT scheme and STOC (representing the *stochastic* run) uses the SPHT scheme to perturb the model variables (e.g., temperature and specific humidity). To avoid instability due to excessive inflation, we suppressed perturbation of wind variables. To test the effectiveness of the inflation method, the warm cycle is started from 1200 UTC 22 June 2018 and ended on 1200 UTC 7 July 2018.

The random forcing in both SPPT and SPDT is described in Fig. 2. The SPPT (Fig. 2a) has horizontal correlation scales of 500 km (mesoscale), decorrelation times of 6 h, and standard deviations of 1.0. On the other hand, the SPDT (Fig. 2b) has horizontal correlation scales of 500 km, decorrelation times of 3 h, and a standard deviation of 0.5. The SPDT especially used a tapering function that decreased exponentially with height (e.g., in the upper level) to prevent instability. It is designed to generate a smaller random forcing to ensure stability because the dynamic tendency variability is sensitive to r.

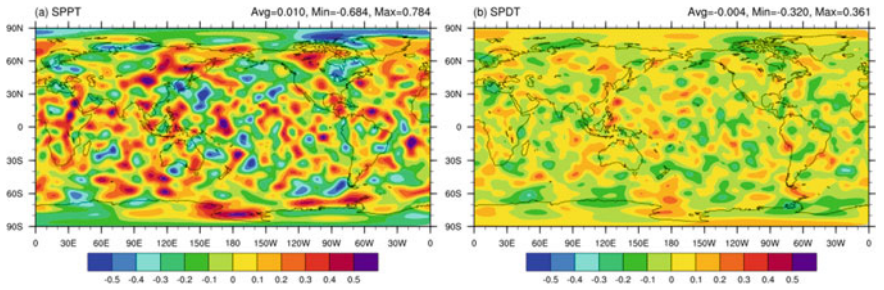

Fig. 2 Random patterns of (**a**) SPPT and (**b**) SPDT at model 90-th level, represented as an ensemble mean of the 6-h forecast initiated from 1200 UTC 6 June 2018

4 Results

We have evaluated the SPHT scheme through the root-mean-square difference (RMSD) and ensemble spread. The RMSD represents the model error while the ensemble spread explains the model uncertainty. Here, we assume that the true state is the ECMWF Integrated Forecast System (IFS) analysis, which is well known for high forecast skill. To identify the vertical distribution of ensemble mean spread and ensemble mean error globally, we analyzed the zonal mean during the experiment periods, excluding the spin-up. The STOC, including both SPPT and SPDT, increases the ensemble spread below 700 hPa in the troposphere and above 10 hPa in the stratosphere (Fig. 3).

Increase in ensemble spread is evident because the model variables are perturbed by the random forcing. Therefore, it is essential to check if the increase in ensemble spread induces reduction in ensemble mean error: if the ensemble mean error increased, the ensemble spread increase is ineffectual. The augmented ensemble spread reduces the ensemble mean RMSD as well, mostly in the tropical troposphere for wind and in the low to mid-troposphere for temperature and specific humidity (Fig. 4).

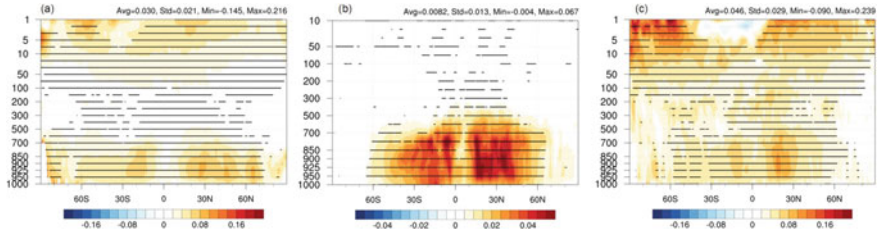

Fig. 3 The difference of zonal mean ensemble spread (STOC − CTRL) for the prior for (**a**) temperature (in K), (**b**) specific humidity (in g kg^{-1}), and (c) zonal wind (in m s^{-1}), averaged for the period of 1800 UTC 25 June – 1800 UTC 7 July 2018. Black dots indicate 95% statistical significance based on two-tail *t*-test

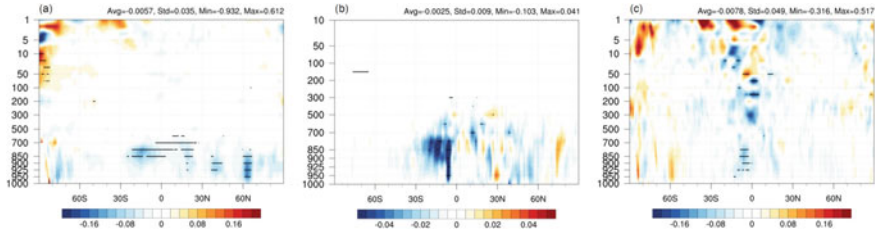

Fig. 4 Same as in Fig. 3 but for the difference of zonal mean RMSD against IFS analysis

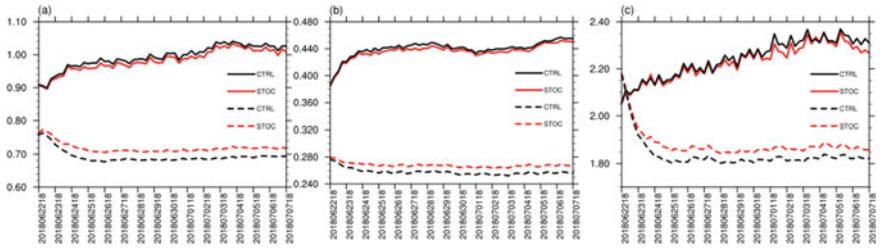

Fig. 5 Time series of the globally-averaged ensemble mean spread (dotted line) and the ensemble mean RMSD (solid line) in the prior for STOC (red) and CTRL (black) for (**a**) temperature (in K), (**b**) specific humidity (in g kg^{-1}), and (**c**) zonal wind (in m s^{-1})

We also have assessed the performance of the SPHT scheme, which is applied to the KIM global model, by evaluating the ensemble quality via the globally-averaged RMSD and ensemble spread. The global performance at each prior (the 6-h background) is evaluated with regard to the vertically-averaged RMSD from 1000 hPa to 10 hPa over the globe (see Fig. 5). As shown in Fig. 5, the ensemble spread augmentation obviously brought on the RMSD reduction during the whole experiment period. For temperature, the ensemble mean RMSD decreased by 1% when the ensemble mean spread increased by 3.7%; for specific humidity, the former decreased by 0.65% when the latter increased by 2.0%.

5 Summary

We implemented the stochastic perturbation hybrid tendencies (SPHT) scheme perturbing both the physical tendency and the dynamic tendency in a global numerical weather prediction model—the Korean Integrated Model (KIM)—which has recently been operational in the Korea Meteorological Administration. The SPHT scheme inflates the insufficient ensemble background error covariance coupled with the local ensemble transform Kalman filter system: it leads to an increase in ensemble spread as well as a decrease in the ensemble mean errors, thus improving the ensemble background error covariance and ensemble prediction.

The stochastic schemes can be used in various fields as the demand for ensemble systems increases. Recently, Ollinaho et al. Ollinaho et al. (2017) developed the stochastically perturbed parametrizations (SPP) scheme to perturb the parameters and variables in physical parametrizations. For example, many physical processes, including turbulent diffusion, sub-grid orography, convection, cloud, large-scale precipitation, and radiation, used to be perturbed to cover the model uncertainty in the European Centre for Medium-Range Weather Forecasts (ECMWF). As demonstrated, we can extend the use of the stochastic perturbation schemes on demand to overcome underestimation of model uncertainty.

Acknowledgements This work has been carried out as part of the R&D project on the development of global numerical weather prediction systems of the Korea Institute of Atmospheric Prediction System funded by the Korea Meteorological Administration. S. Lim is partly supported by the Ewha Womans University scholarship of 2018. S. K. Park is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1A6A1A08025520).

References

Anderson JL, Anderson SL (1999) A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon Weather Rev 127:2741–2758

Berner J, Ha SY, Hacker JP, Fournier A, Snyder C (2011) Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Mon Weather Rev 139:1972–1995

Berner J, Fossell KR, Ha SY, Hacker JP, Snyder C (2015) Increasing the skill of probabilistic forecasts: understanding performance improvements from model-error representations. Mon Weather Rev 143:1295–1320

Bouttier F, Vié B, Nuissier O, Raynaud L (2012) Impact of stochastic physics in a convection-permitting ensemble. Mon Weather Rev 140:3706–3721

Buizza R, Miller M, Palmer TN (1999) Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. J R Meteorol Soc 125:2887–2908

Buizza R, Houtekamer PL, Toth Z, Pellerin G, Wei M, Zhu Y (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Mon Weather Rev 133:1076–1097

Charron M, Pellerin G, Spacek L, Houtekamer PL, Gagnon N, Mitchell HL, Michelin L (2010) Toward random sampling of model error in the Canadian ensemble prediction system. Mon Weather Rev 138:1877–1901

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99:10143–10162

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J R Meteor Soc 125:723–757

Hong SY, Park H, Cheong HB, Kim JEE, Koo MS, Jang J, Ham S, Hwang SO, Park BK, Chang EC, Li H (2013) The global/regional integrated model system (GRIMs). Asia-Pacific J Atmos Sci 49:219–243

Hong S-Y, Kwon YC, Kim T-H, Kim J-EE, Choi S-J, Kwon I-H, Kim J, Lee E-H, Park R-S, Kim D-I (2018) The Korean Integrated Model (KIM) system for global weather forecasting. Asia-Pac J Atmos Sci 54:267–292

Hunt B, Kostelich E, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126

Kang J-H, Chun HW, Lee S, Song H-J, Ha J-H, Kwon I-H, Han H-J, Jeong H, Kwon H-N (2018) Development of an observation processing package for data assimilation in KIAPS. Asia-Pac J Atmos Sci 54:303–318

Kleist DT, Ide K (2015) An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. Mon Weather Rev 143:433–451

Koo M-S, Hong SY (2014) Stochastic representation of dynamic model tendency: formulation and preliminary results. Asia-Pac J Atmos Sci 50:497–506

Kwon I-H, Song H-J, Ha J-H, Chun H-W, Kang J-H, Lee S, Lim S, Jo Y, Han H-J, Jeong H, Kwon H-N, Shin S, Kim T-H (2018) Development of operational hybrid data assimilation system at KIAPS. Asia-Pac J Atmos Sci 54:319–335

Leutbecher M, Lock S-J, Ollinaho P, Lang ST, Balsamo G, Bechtold P, Bonavita M, Christensen HM, Diamantakis M, Dutra E, English S, Fisher M, Forbes RM, Goddard J, Haiden T, Hogan RJ, Juricke S, Lawrence H, MacLeod D, Magnusson L, Malardel S, Massart S, Sandu I, Smolarkiewicz PK, Subramanian A, Vitart F, Wedi N, Weisheimer A (2017) Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. Quart J Roy Meteor Soc 143:2315–2339

Lim S, Koo M-S, Kwon I-H, Park SK (2020) Model error representation using the stochastically perturbed hybrid physical-dynamical tendencies in ensemble data assimilation system. Appl Sci 2020 (in press)

Meng Z, Zhang F (2007) Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: imperfect model experiments. Mon Weather Rev 135:1403–1423

Mitchell HL, Houtekamer PL (2000) An adaptive ensemble Kalman filter. Mon Weather Rev 128:416–433

Miyoshi T (2011) The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. Mon Weather Rev 139:1519–1535

Ollinaho P, Lock SJ, Leutbecher M, Bechtold P, Beljaars A, Bozzo A, Forbes RM, Haiden T, Hogan RJ, Sandu I (2017) Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. Q J R Meteor Soc 143:408–422

Palmer TN, Buizza R, D-R F, Jung T, Leutbecher M, Shutts GJ, Steinheimer M, Weisheimer A, (2009) Stochastic parametrization and model uncertainty. ECMWF Tech Memo, vol 598, p 42

Romine GS, Schwartz CS, Berner J, Fossell KR, Snyder C, Anderson JL, Weisman ML (2014) Representing forecast error in a convection-permitting ensemble system. Mon Weather Rev 142:4519–4541

Sanchez C, Williams KD, Collins M (2016) Improved stochastic physics schemes for global weather and climate models. Q J R Meteor Soc 142:147–159

Shin S, Kang J-S, Jo Y (2016) The local ensemble transform Kalman filter (LETKF) with a global NWP model on the cubed sphere. Pure Appl Geophys 173:2555–2570

Shin S, Kang J-H, Chun H-W, Lee S, Sung K, Cho K, Jo Y, Kim J-E, Kwon I-H, Lim S, Kang J-S (2018) Real data assimilation using the local ensemble transform Kalman Filter (LETKF) system for a global non-hydrostatic NWP model on the cubed-sphere. Asia-Pacific J Atmos Sci 54:351–360

Shutts G (2005) A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q J R Meteor Soc 131:3079–3102

Shutts G (2015) A stochastic convective backscatter scheme for use in ensemble prediction systems. Q J R Meteor Soc 141:2602–2616

Tennant WJ, Shutts GJ, Arribas A, Thompson SA (2011) Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. Mon Weather Rev 139:1190–1206

Thépaut JN (2003) Satellite data assimilation in numerical weather prediction: an overview. Meteorological Training Course Lecture Series, ECMWF, Reading https://www.ecmwf.int/node/12657

Whitaker JS, Hamill TM (2012) Evaluating methods to account for system errors in ensemble data assimilation. Mon Weather Rev 140:3078–3089

Whitaker JS, Hamill TM, Wei X, Song Y, Toth Z (2008) Ensemble data assimilation with the NCEP global forecast system. Mon Weather Rev 136:463–482

Zhang F, Snyder C, Sun J (2004) Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. Mon Weather Rev 132:1238–1253

Second-Order Methods in Variational Data Assimilation

François-Xavier Le Dimet and Victor Shutyaev

Abstract The chapter presents the second-order methods in variational data assimilation. The algorithms to compute the Hessian of the cost function are discussed, the second-order adjoint method among them. General sensitivity analysis for the optimality system is presented. Using the Hessian, the sensitivity of the optimal solution and its functionals is studied with respect to observations and uncertainties in model parameters. Numerical examples for joint state and parameter estimation for a sea thermodynamics model are presented.

1 Introduction

The methods of data assimilation (DA) have become an important tool for analysis of complex physical phenomena in various fields of science and technology. These methods allow us to combine mathematical models, data from observations and a priori information.

Currently, there is an increasing interest in computational technologies that combine the flows of real data and hydrodynamic forecasts using mathematical models. This is especially true for 4D technologies - the combination of the flows of observational data and forecasts in a certain spatio-temporal domain. These methods have received the greatest applications in meteorology and oceanography, where observations are assimilated into numerical models. Geophysical flows are governed by equations derived from fluid dynamics: a set of nonlinear partial differential equations of the first order with respect to time. Formally, it is a Cauchy problem, and an initial condition is necessary to integrate these equations, to carry out a prediction. The purpose of assimilation procedures is to construct or refine the initial and bound-

F.-X. Le Dimet
LJK, Université Grenoble Alpes, 700 Avenue Centrale, 38401 Domaine Universitaire de Saint-Martin-d'Hères, Grenoble, France

V. Shutyaev (✉)
Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences, 119333 Gubkina 8, Moscow, Russia
e-mail: victor.shutyaev@mail.ru

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_7

ary conditions (or other model parameters) to improve the accuracy of a prediction model Le Dimet and Talagrand (1986), Asch et al. (2016), Fletcher (2017), Carrassi et al. (2018).

At present, two main approaches are well known for the assimilation of observational data in models of geophysical hydrodynamics and oceanography. The first is the Statistical approach which is based on the methods of probability theory and mathematical statistics. Historically, its rigorous justification and limits of applicability were given by Markov (1900) and Kolmogorov (1946). From a methodological point of view, this method gave rise to the methods of optimal interpolation, the Kalman filter methods and their subsequent modifications, widely used in various fields of science and technology. This approach is used to estimate unknown quantities from measurement data, taking into account the random nature of measurement errors.

The second approach is based on the methods of calculus of variations, optimal control (see, e.g., Lions (1968), Pontryagin et al. (1964)) and the theory of adjoint equations (see Marchuk (1995)). Compared to the statistical method, the variational method has greater versatility. It allows, on a unified methodological basis, to solve the problems of initializing hydrophysical fields, assessing the sensitivity of a model solution, identifying model parameters, etc. The variational approach can be applied by assimilating information of various types and measuring systems. In this case this approach is reffered to as variational data assimilation (VDA) Le Dimet and Talagrand (1986), Asch et al. (2016), Fletcher (2017), Carrassi et al. (2018). The main idea of the method is to minimize some functional that describes the deviation of the model solution from the observational data, and the minimum of this functional is sought on the model trajectories, in other words, in the subspace of model solutions.

Basically, as seen as a problem of optimal control, VDA is an optimization problem and as such we need to exhibit a necessary optimality condition derived from the evaluation of the gradient of the cost function, which should be zero at the optimum. Information on the gradient of the cost function (first-order information) is used to construct the optimality system (OS). To this aim and for the numerical solution of the optimization problem, the representation of the gradient through adjoint equations (first-order adjoint problem) is often used Le Dimet and Talagrand (1986), Marchuk (1995). In the case of discontinuous processes in the physics (rain, deep convection, etc.) the cost function is no longer differentiable and the formal application of the adjoint operator will evaluate a sub-gradient.

To study the variational data assimilation problem (as an optimal control problem) and to develop efficient algorithms for its numerical solution, second-order information is needed. This is information about the Hessian of the cost function. A necessary and sufficient optimality condition is to get the Hessian positive definite at the optimum; therefore, a second-order analysis must be carried out. Often, to construct the Hessian, it is necessary to differentiate the optimality system. In this case, a second-order adjoint problem arises Le Dimet et al. (2002). The investigation of the second-order adjoint equations and the Hessian of the cost functional plays an important role in the study of the solvability of the variational assimilation problem, the construction of algorithms for its numerical solution based on the modification

of Newton type methods, the identification of model parameters, and the study of the sensitivity of the optimal solution and its functionals. These issues are the subject of this chapter.

2 Variational Data Assimilation

Variational methods were introduced in meteorology in 1958 by Sasaki (1958). These methods consider the equations governing the flow as constraints and the problem is closed by using a variational principle, e.g. the minimization of the discrepancy between the model and the observations. Using Optimal Control Techniques (Lions (1968)) was proposed by Le Dimet (1982), Le Dimet and Talagrand (1986), Talagrand and Courtier (1987), Penenko and Obraztsov (1976), Marchuk et al. (1978).

Consider the mathematical model of a physical process that is described by the nonlinear evolution problem

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \end{cases} \tag{1}$$

where the initial state u is supposed to be from a Hilbert space X, the unknown function $\varphi = \varphi(t)$ belongs to $Y = L_2(0, T; X)$ with the norm $\|\varphi\|_Y = (\varphi, \varphi)_Y^{1/2} = (\int_0^T \|\varphi(t)\|_X^2 dt)^{1/2}$, F is a nonlinear operator mapping Y into Y, $f \in Y$. We suppose that for given $u \in X$, $f \in Y$ there exists a unique solution $\varphi \in Y$ to (1) with $\frac{\partial \varphi}{\partial t} \in Y$.

Often, the the initial state u is supposed to be unknown, and one would like to find it using the information from observations. Let us introduce the cost function as a functional on X in the form

$$J(u) = \frac{1}{2}(V_1(u - u_b), u - u_b)_X + \frac{1}{2}(V_2(C\varphi - \varphi_{obs}), C\varphi - \varphi_{obs})_{Y_{obs}}, \tag{2}$$

where $u_b \in X$ is a prior (background) function, $\varphi_{obs} \in Y_{obs}$ is a prescribed function (observational data), Y_{obs} is a Hilbert space (observation space), $C : Y \to Y_{obs}$ is a linear bounded operator (observation operator), $V_1 : X \to X$ and $V_2 : Y_{obs} \to Y_{obs}$ are symmetric positive definite bounded operators. Usually, V_1, V_2 are chosen as inverse covarianvce operators of background and observation errors, respectively, Asch et al. (2016), Carrassi et al. (2018).

Let us consider the following data assimilation problem with the aim to find the initial value u: for given $f \in Y$, $\varphi_{obs} \in Y_{obs}$, $u_b \in X$, find $u \in X$ and $\varphi \in Y$ such that they satisfy (1), and on the set of solutions to (1), the functional $J(u)$ takes the minimum value, i.e.

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \\ J(u) = \inf\limits_{w \in X} J(w). \end{cases} \tag{3}$$

This is a so-called hind-cast (initialization) variational DA problem, a typical DA problem often considered in numerical weather prediction and oceanographic applications Le Dimet and Talagrand (1986), Asch et al. (2016), Fletcher (2017), Carrassi et al. (2018). We suppose that the solution of (3) exists. To derive the optimality system, we assume the solution φ and the operator $F(\varphi)$ in (1)–(2) are regular enough, and for $u, w \in X$ introduce the directional (Gâteaux) derivative with respect to u in the direction w (Gâteaux differential):

$$dJ(u, w) = \lim_{\tau \to 0} \frac{J(u + \tau w) - J(u)}{\tau} = \frac{d}{d\tau} J(u + \tau w)\Big|_{\tau=0}.$$

If $dJ(u, w)$ is linear with respect to w, then it may be represented as follows:

$$dJ(u, w) = J'(u)w,$$

where $J'(u)$ is the gradient of J with respect to u. From (1)–(2) we get

$$dJ(u, w) = (V_1(u - u_b), w)_X + (C^* V_2(C\varphi - \varphi_{obs}), \tilde{\phi})_Y, \tag{4}$$

where $\tilde{\phi}$ is the solution to the tangent linear problem:

$$\begin{cases} \dfrac{\partial \tilde{\phi}}{\partial t} = F'_\varphi(\varphi)\tilde{\phi}, \quad t \in (0, T), \\ \phi\big|_{t=0} = w. \end{cases} \tag{5}$$

Here $F'_\varphi(\varphi) : Y \to Y$ is the Fréchet derivative of F Marchuk et al. (1996) with respect to φ, and C^* is the adjoint operator to C defined by $(C\varphi, \psi)_{Y_{obs}} = (\varphi, C^*\psi)_Y$, $\varphi \in Y$, $\psi \in Y_{obs}$.

Let us introduce the adjoint operator $(F'_\varphi(\varphi))^* : Y \to Y$ and consider the adjoint problem:

$$\begin{cases} \dfrac{\partial \varphi^*}{\partial t} + (F'_\varphi(\varphi))^* \varphi^* = C^* V_3(C\varphi - \varphi_{obs}), \quad t \in (0, T) \\ \varphi^*\big|_{t=T} = 0. \end{cases} \tag{6}$$

The problem (6) is adjoint with respect to the linearized (tangent linear) problem (5), therefore, it is also linear in φ^*, however, it is still nonlinear in φ.

In the below consideration, we assume that the direct and adjoint linear problems of the form

$$\begin{cases} \dfrac{\partial \phi}{\partial t} - F'_\varphi(\varphi)\phi = p, \quad t \in (0, T) \\ \phi\big|_{t=0} = q, \end{cases}$$

$$\begin{cases} -\dfrac{\partial \phi^*}{\partial t} - (F'_\varphi(\varphi))^* \phi^* = g, \quad t \in (0, T) \\ \phi^*\big|_{t=T} = 0 \end{cases}$$

with $p, g \in Y, q \in X$ have the unique solutions $\phi, \phi^* \in Y$ and $\dfrac{\partial \phi}{\partial t}, \dfrac{\partial \phi^*}{\partial t} \in Y$. From (4)–(6) we get

$$dJ(u, w) = (V_1(u - u_b), w)_X - (\varphi^*\big|_{t=0}, w)_X. \tag{7}$$

The relation (7) exibits the linear dependence of $dJ(u, w)$ with respect to w. Thus, $dJ(u, w) = J'(u)w$, and the gradient of J with respect to u is defined by

$$J'(u) = V_1(u - u_b) - \varphi^*\big|_{t=0}.$$

The necessary optimality condition Lions (1968) is $J'(u) = 0$. From (3)–(7) we obtain the Optimality System :

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi) + f, \quad t \in (0, T), \\ \varphi\big|_{t=0} = u, \end{cases} \tag{8}$$

$$\begin{cases} \dfrac{\partial \varphi^*}{\partial t} + (F'_\varphi(\varphi))^* \varphi^* = C^* V_3(C\varphi - \varphi_{obs}), \quad t \in (0, T) \\ \varphi^*\big|_{t=T} = 0, \end{cases} \tag{9}$$

$$V_1(u - u_b) - \varphi^*\big|_{t=0} = 0. \tag{10}$$

It worth to point out that there is no approximation in the derivation of the optimality system and the only assumption is the differentiability of the operator of the model. Some authors consider, at this level, a so-called "linear tangent approximation", it is fully unnecessary.

We suppose that the system (8)–(10) has a unique solution $\varphi, \varphi^* \in Y, u \in X$. The system (8)–(10) may be considered as a generalized model of the form $\mathcal{A}(U) = 0$ with the state variable $U = (\varphi, \varphi^*, u)$, and it contains the information on the observation data $\varphi_{obs} \in Y_{obs}$. The optimality system plays a fundamental role in studying the solvability of the original data assimilation problem, searching efficient algorithms for its solution, and studying the sensitivity of the optimal solution with respect to observations.

3 Computing the Hessian

Consider the Hessian $\mathcal{H}(u)$ of the functional (2); it depends on $u \in X$ (which may be the exact solution, the optimal solution, or some arbitrary function $u \in X$). For a fixed $u \in X$ the Hessian $\mathcal{H}(u)$ is defined by the successive solutions of the below-

formulated problems. First we find φ and φ^* by solving the direct and adjoint problems (like in the optimality system):

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \end{cases} \tag{11}$$

$$\begin{cases} -\dfrac{\partial \varphi^*}{\partial t} - (F'(\varphi))^* \varphi^* = -C^* V_2 (C\varphi - \varphi_{obs}), \quad t \in (0, T) \\ \varphi^*\big|_{t=T} = 0. \end{cases} \tag{12}$$

Note that here u is not necessarily the optimal solution from the optimality system (8)–(10); it is just some fixed function at which we would like to compute the Hessian. (Hence, in general, the functions φ and φ^* do not satisfy the optimality system). Note also that (11)–(12) are usual two steps when we compute the gradient of the functional $J(u)$ (at the point u) using the adjoint problem. If for a fixed u the functions φ, φ^* are computed from (11)–(12), the gradient of J with respect to u is defined by

$$J'(u) = V_1(u - u_b) - \varphi^*\big|_{t=0}. \tag{13}$$

To find the Hessian we should differentiate (11)–(13) with respect to u. Then, the action of the Hessian $\mathcal{H}(u)$ on the function $v \in X$ is defined by the successive solutions of the following problems:

$$\begin{cases} \dfrac{\partial \psi}{\partial t} - F'(\varphi)\psi = 0, \ t \in (0, T), \\ \psi|_{t=0} = v, \end{cases} \tag{14}$$

$$\begin{cases} -\dfrac{\partial \psi^*}{\partial t} - (F'(\varphi))^* \psi^* = (F''(\varphi)\psi)^* \varphi^* - C^* V_2 C \psi, \quad t \in (0, T) \\ \psi^*\big|_{t=T} = 0, \end{cases} \tag{15}$$

$$\mathcal{H}(u)v = V_1 v - \psi^*|_{t=0}. \tag{16}$$

Here φ and φ^* are involved, being taken from (11)–(12). The problem (15) is a so-called *second-order adjoint* problem Le Dimet et al. (2002). It involves the second derivative $F''(\varphi)$ of the model operator $F(\varphi)$ and depends on the solution φ^* of the first-order adjoint problem (12).

If u is the optimal solution, then φ and φ^* are exactly the functions from the optimality system (8)–(10).

Formulas (11)–(16) may be used to compute the Hessian of the original cost functional. To solve the second-order adjoint problem (15), no additional software is needed to be developed. To this aim, one can use the existing code for the first-order adjoint problem (12), taking into account the new right-hand side involving the term $(F''(\varphi)\psi)^*\varphi^*$. An alternative method to compute the Hessian $\mathcal{H}$ is the method of finite differences described in Gill et al. (1981). However, this method is not sufficiently accurate due to truncations used in a local Taylor expansion and

is expensive for practical implementation. The sensitivity matrix method Thacker (1989) is computationally efficient if the dimension of the observation vector is much smaller than the dimension of the state vector, and so is feasible mainly for the 3D-VAR applications. It requires full storage of the resulting matrix. The above-formulated second-order adjoint method allows the action $\mathcal{H}v$ to be computed, thus does not require full storage of $\mathcal{H}$.

In the finite-dimensional space, $\mathcal{H}(u)$ is a matrix. To obtain the first column of this matrix, one can choose v in (14)–(16) to be the first basis vector $v = (1, 0, \ldots, 0)$. To obtain the second column of this matrix, one can choose v in (14)–(16) to be the second basis vector $v = (0, 1, 0, \ldots, 0)$, and so on.

In the linear case, the solution is unique if the Hessian is positive definite. In this case, the necessary optimality condition given by the optimality system is also a sufficient condition. From a general point of view the information given by the Hessian is important for theoretical, numerical and practical issues. For operational models it is impossible to compute the Hessian itself, as it is a square matrix with around 10^{18} terms, nevertheless the most important information can be extracted from the spectrum of the Hessian which can be estimated without an explicit determination of this matrix. This information is of importance for estimating the condition number of the Hessian for preparing an efficient preconditioning.

The above-obtained system with the second order adjoint is used to compute the product of the Hessian by any vector. Of course, if we consider all the vectors of the canonical base, then it will be possible to get the complete Hessian.

The determination of this product permits to access some information. So, by using Lanczos type methods and deflation, it is possible to compute the eigenvectors and eigenvalues of the Hessian. Also, to solve the variational data assimilation problem, second-order optimization methods of Newton-type are used for equations of the form:

$$J'(u) = 0.$$

The iterations are

$$u_{n+1} = u_n - \mathcal{H}^{-1}(u_n)\, J'(u_n),$$

where $\mathcal{H}$ is the Hessian of J, or its approximation. At each iteration a linear system should be solved. This is done by carrying out some iterations of a conjugate gradient methods which require computing the Hessian-vector product. To construct an approximation of the inverse Hessian, the quasi-Newton BFGS algorithm may be used Polak (1997). This algorithm generates $\mathcal{H}^{-1}$ in the course of a minimization process.

In some applications (such as sensitivity analysis) one needs to solve the system of equations in the form $\mathcal{H}(u)v = p$. In this case, computing the Hessian-vector product by (11)–(16) may be efficient for using iterative algorithms. The following directions to construct a specialized solver for the equation $\mathcal{H}(u)v = p$ could be considered: the use of a multi-grid strategy; the use of reduced order models (Proper Orthogonal Decomposition) or local approximations (splines, wavelets); decomposition of the

spatial domain by the 'region of influence' principle, hence decomposition of a global DA problem into a set of local open boundary DA problems.

The inverse Hessian or its approximations may be used also to estimate the optimal solution error covariances Gejadze et al. (2008, 2011, 2013), Shutyaev et al. (2012). Assuming the so-called tangent linear hypothesis (TLH), the covariance is often approximated by the inverse Hessian of the objective function. In practice, the same approximation could be valid even though the TLH is clearly violated. However, often we deal with such a highly nonlinear dynamics that the inverse Hessian approach is no longer valid. In this case a new method for computing the covariance matrix named the 'effective inverse Hessian' method can be used Shutyaev et al. (2012). This method yields a significant improvement of the covariance estimate as compared to the inverse Hessian. The method is potentially feasible for large-scale applications because it can be used in the multiprocessor environment and operates in terms of the Hessian-vector products. The software blocks needed for its implementation are the standard blocks of any existing 4D-Var system. The results given by the method are consistent with the assumption on a 'close-to-normal' nature of the optimal solution error. This should be expected taking into account the consistency and asymptotic normality of the estimator and the fact that the observation window in variational DA is usually quite large.

4 Parameter Estimation

We should mention the importance of the parameter estimation problem itself. A precise determination of the initial condition is very important in view of forecasting, however the use of variational data assimilation is not limited to operational forecasting. In many domains (e.g. hydrology) the uncertainty in the parameters is more crucial that the uncertainty in the initial condition (e.g. White et al. (2003)). In some problems the quantity of interest can be represented directly by the estimated parameters as controls. For example, in Agoshkov et al. (2015) the sea surface heat flux is estimated in order to understand its spatial and temporal variability. The problems of parameter estimation are common inverse problems considered in geophysics and in engineering applications (see Alifanov et al. (1996), Sun (1994), Zhu and Navon (1999), Storch et al. (2007)). Last years an interest is rising to the parameter estimation using 4D-Var (Bocquet (2012), Schirber et al. (2013), Smith et al. (2013), Yuepeng et al. (2018), Agoshkov and Sheloput (2017)).

We consider a dynamic formulation of variational data assimilation problem for parameter estimation in a continuous form. Of course, the initial condition function may be also considered as a parameter, however, in our dynamic formulation we have two equations for the model: one equation for describing an evolution of the model operator (involving model parameters such as right-hand sides, coefficients, boundary conditions etc.), and another equation is considered as an initial condition.

Let the model be governed by the evolution problem of the form (1):

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi, \lambda) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \end{cases} \tag{17}$$

where F is a nonlinear operator mapping $Y \times Y_p$ into Y, Y_p is a Hilbert space (space of control parameters, or control space). Suppose that for given $u \in X$, $f \in Y$ and $\lambda \in Y_p$ there exists a unique solution $\varphi \in Y$ to (17) with $\dfrac{\partial \varphi}{\partial t} \in Y$. The function λ is an unknown model parameter.

Let us introduce the cost function

$$J(\lambda) = \frac{1}{2}(V_1(\lambda - \lambda_b), \lambda - \lambda_b)_{Y_p} + \frac{1}{2}(V_2(C\varphi - \varphi_{obs}), C\varphi - \varphi_{obs})_{Y_{obs}}, \tag{18}$$

where $\lambda_b \in Y_p$ is a prior (background) function, $\varphi_{obs} \in Y_{obs}$ is a prescribed function (observational data), Y_{obs} is a Hilbert space (observation space), $C : Y \to Y_{obs}$ is a linear bounded observation operator, $V_1 : Y_p \to Y_p$ and $V_2 : Y_{obs} \to Y_{obs}$ are symmetric positive definite bounded operators.

Let us consider the following data assimilation problem with the aim to estimate the parameter λ: for given $u \in X$, $f \in Y$, find $\lambda \in Y_p$ and $\varphi \in Y$ such that they satisfy (17), and on the set of solutions to (17), the functional $J(\lambda)$ takes the minimum value, i.e.

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi, \lambda) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \\ J(\lambda) = \inf\limits_{v \in Y_p} J(v). \end{cases} \tag{19}$$

We suppose that the solution of (19) exists. Let us note that the solvability of the parameter estimation problems (or identifiability) has been addressed, e.g., in Chavent (1983), Navon (1998). To derive the optimality system, we assume the solution φ and the operator $F(\varphi, \lambda)$ in (17)–(18) are regular enough, and for $v \in Y_p$ find the gradient of the functional J with respect to λ:

$$\begin{aligned} J'(\lambda)v &= (V_1(\lambda - \lambda_b), v)_{Y_p} + (V_2(C\varphi - \varphi_{obs}), C\phi)_{Y_{obs}} \\ &= (V_1(\lambda - \lambda_b), v)_{Y_p} + (C^* V_2(C\varphi - \varphi_{obs}), \phi)_Y, \end{aligned} \tag{20}$$

where ϕ is the solution to the problem:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = F'_\varphi(\varphi, \lambda)\phi + F'_\lambda(\varphi, \lambda)v, \\ \phi\big|_{t=0} = 0. \end{cases} \tag{21}$$

Here $F'_\varphi(\varphi, \lambda) : Y \to Y$, $F'_\lambda(\varphi, \lambda) : Y_p \to Y$ are the Fréchet derivatives of F Marchuk et al. (1996) with respect to φ and λ, correspondingly, and C^* is the adjoint operator to C defined by $(C\varphi, \psi)_{Y_{obs}} = (\varphi, C^*\psi)_Y$, $\varphi \in Y$, $\psi \in Y_{obs}$.

Let us consider the adjoint operator $(F'_\varphi(\varphi,\lambda))^* : Y \to Y$ and introduce the adjoint problem:

$$\begin{cases} \dfrac{\partial \varphi^*}{\partial t} + (F'_\varphi(\varphi,\lambda))^*\varphi^* = C^*V_2(C\varphi - \varphi_{obs}), \\ \varphi^*\big|_{t=T} = 0. \end{cases} \tag{22}$$

Then (20) with (21) and (22) gives

$$J'(\lambda)v = (V_1(\lambda - \lambda_b), v)_{Y_p} - (\varphi^*, F'_\lambda(\varphi,\lambda)v)_Y =$$

$$(V_1(\lambda - \lambda_b), v)_{Y_p} - ((F'_\lambda(\varphi,\lambda))^*\varphi^*, v)_{Y_p},$$

where $(F'_\lambda(\varphi,\lambda))^* : Y \to Y_p$ is the adjoint operator to $F'_\lambda(\varphi,\lambda)$. Therefore, the gradient of J is defined by

$$J'(\lambda) = V_1(\lambda - \lambda_b) - (F'_\lambda(\varphi,\lambda))^*\varphi^*. \tag{23}$$

From (20)–(23) we get the optimality system (the necessary optimality conditions, Lions (1968)):

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi,\lambda) + f, \quad t \in (0,T), \\ \varphi\big|_{t=0} = u, \end{cases} \tag{24}$$

$$\begin{cases} \dfrac{\partial \varphi^*}{\partial t} + (F'_\varphi(\varphi,\lambda))^*\varphi^* = C^*V_2(C\varphi - \varphi_{obs}), \\ \varphi^*\big|_{t=T} = 0, \end{cases} \tag{25}$$

$$V_1(\lambda - \lambda_b) - (F'_\lambda(\varphi,\lambda))^*\varphi^* = 0. \tag{26}$$

We assume that the system (24)–(26) has a unique solution. The system (24)–(26) may be considered as a generalized model $\mathcal{A}(U) = 0$ with the state variable $U = (\varphi, \varphi^*, \lambda)$, and it contains information about observations.

If the observation operator C is nonlinear, i.e. $C\varphi = C(\varphi)$, then the right-hand side of the adjoint equation (25) contains $(C'_\varphi)^*$ instead of C^* and all the analysis presented below is similar.

To compute the Hessian $\mathcal{H}(\lambda)$ of the cost function (18) one should differentiate (24)–(25) and (23) with respect to λ, following Sect. 3. Then, the action of the Hessian $\mathcal{H}(\lambda)$ on a function $w \in Y_p$ is defined by the successive solutions of the following problems:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} - F'_\varphi(\varphi,\lambda)\phi = F'_\lambda(\varphi,\lambda)w, \quad t \in (0,T) \\ \phi\big|_{t=0} = 0, \end{cases} \tag{27}$$

$$\begin{cases} -\dfrac{\partial \phi^*}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^*\phi^* - (F''_{\varphi\varphi}(\varphi,\lambda)\phi)^*\varphi^* = (F''_{\lambda\varphi}(\varphi,\lambda)w)^*\varphi^* - C^*V_2C\phi, \\ \phi^*\big|_{t=T} = 0, \end{cases} \tag{28}$$

$$\mathcal{H}(\lambda)w = V_1 w - (F''_{\varphi\lambda}(\varphi,\lambda)\phi)^*\varphi^* - (F''_{\lambda\lambda}(\varphi,\lambda)w)^*\varphi^* - (F'_{\lambda}(\varphi,\lambda))^*\phi^*. \tag{29}$$

The definition of the Hessian $\mathcal{H}(\lambda)$ by (27)–(29) involves the second-order derivatives of the model operator F with respect to φ and λ.

Numerical examples for computing the Hessian for the parameter estimation problems are presented in Gejadze et al. (2010).

5 Sensitivity Analysis

In the environmental sciences the mathematical models contain parameters which cannot be estimated precisely, because they are used to parametrize some subgrid processes and therefore can not be physically measured. Thus, it is important to be able to estimate the impact of uncertainties on the outputs of the model after assimilation. The optimal solution depends on the parameters, which may contain uncertainties, and for the forecasts it is very important to study the sensitivity of the optimal solution and its functionals with respect to these parameters Marchuk (1995), Cacuci (1981), Dontchev (1983), Griesse and Vexler (2007).

The necessary optimality condition is related to the gradient of the original cost function, thus to study the sensitivity of the optimal solution, one should differentiate the optimality system with respect to imprecisely known parameters. In this case, we come to the second-order adjoint problem Le Dimet et al. (2002). The first studies of sensitivity of the response functions after assimilation with the use of second-order adjoint were done by Le Dimet et al. (1997) for variational data assimilation problem aimed at restoration of initial condition, where sensitivity with respect to model parameters was considered. The equations of the forecast sensitivity to observations in a four-dimensional (4D-Var) data assimilation were derived by Daescu (2008). Based on these results, a practical computational approach was given by Cioaca et al. (2013) to quantify the effect of observations in 4D-Var data assimilation.

Sensitivity of the optimal solution is related to its statistical properties (see Gejadze et al. (2008, 2011, 2013), Shutyaev et al. (2012)). General sensitivity analysis in variational data assimilation with respect to observations for a nonlinear dynamic model was given in Shutyaev et al. (2017)–Shutyaev et al. (2018) to control the initial-value function and the model parameters.

Consider the mathematical model of a physical process that is described by the evolution problem of the form (17):

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi,\lambda), \quad t \in (0,T) \\ \varphi\big|_{t=0} = u. \end{cases} \tag{30}$$

Suppose that for given $u \in X$ and $\lambda \in Y_p$ there exists a unique solution $\varphi \in Y$ to (30).

We introduce the functional

$$J(u) = \frac{1}{2}(V_1(u - u_0), u - u_0)_X + \frac{1}{2}(V_2(C\varphi - \varphi_{obs}), C\varphi - \varphi_{obs})_{Y_{obs}}, \tag{31}$$

where $u_0 \in X$ is a prior initial-value function (background state), $\varphi_{obs} \in Y_{obs}$ is a prescribed function (observational data), Y_{obs} is a Hilbert space (observation space), $C : Y \to Y_{obs}$ is a linear bounded operator, $V_1 : X \to X$ and $V_2 : Y_{obs} \to Y_{obs}$ are symmetric positive definite operators.

Consider the variational data assimilation problem with the aim to identify the initial condition: for given $\lambda \in Y_p$ find $u \in X$ and $\varphi \in Y$ such that they satisfy (30), and on the set of solutions to (30), the functional $J(u)$ takes the minimum value, i.e.

$$\begin{cases} \frac{\partial \varphi}{\partial t} = F(\varphi, \lambda), \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \\ J(u) = \inf\limits_{v} J(v). \end{cases} \tag{32}$$

The necessary optimality condition reduces the problem (32) to the optimality system:

$$\begin{cases} \frac{\partial \varphi}{\partial t} = F(\varphi, \lambda), \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \end{cases} \tag{33}$$

$$\begin{cases} -\frac{\partial \varphi^*}{\partial t} - (F'_\varphi(\varphi, \lambda))^* \varphi^* = -C^* V_2(C\varphi - \varphi_{obs}), \quad t \in (0, T) \\ \varphi^*\big|_{t=T} = 0, \end{cases} \tag{34}$$

$$V_1(u - u_0) - \varphi^*\big|_{t=0} = 0 \tag{35}$$

with the unknowns φ, φ^*, u, where $(F'_\varphi(\varphi, \lambda))^*$ is the adjoint to the Frechet derivative of F with respect to φ.

We assume that the system (33)–(35) has a unique solution. The system (33)–(35) may be considered as a generalized model $\mathcal{F}(U, \lambda) = 0$ with the state variable $U = (\varphi, \varphi^*, u)$, and it contains all the available information. All the components of U depend on the parameters $\lambda \in Y_p$, which may contain uncertainties. An important issue is to study the sensitivity of this generalized model with respect to the parameters.

Let us introduce a response function $G(\varphi, u, \lambda)$, which is supposed to be a real-valued function and can be considered as a functional on $Y \times X \times Y_p$. We are interested in the sensitivity of G with respect to λ, with φ and u obtained from the optimality system (33)–(35). As is known Marchuk (1995), Cacuci (1981), Dontchev (1983), sensitivity is defined by the gradient of G with respect to λ, which is a functional derivative:

$$\frac{dG}{d\lambda} = \frac{\partial G}{\partial \varphi}\frac{\partial \varphi}{\partial \lambda} + \frac{\partial G}{\partial u}\frac{\partial u}{\partial \lambda} + \frac{\partial G}{\partial \lambda}. \tag{36}$$

If $\delta\lambda$ is a perturbation on λ, we get from the optimality system:

$$\begin{cases} \dfrac{\partial \delta\varphi}{\partial t} = F'_{\varphi}(\varphi,\lambda)\delta\varphi + F'_{\lambda}(\varphi,\lambda)\delta\lambda, \quad t \in (0,T) \\ \delta\varphi\big|_{t=0} = \delta u, \end{cases} \tag{37}$$

$$\begin{cases} -\dfrac{\partial \delta\varphi^*}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^*\delta\varphi^* - (F''_{\varphi\varphi}(\varphi,\lambda)\delta\varphi + F''_{\varphi\lambda}(\varphi,\lambda)\delta\lambda)^*\varphi^* = -C^*V_2C\delta\varphi, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \delta\varphi^*\big|_{t=T} = 0, \end{cases} \tag{38}$$

$$V_1\delta u - \delta\varphi^*\big|_{t=0} = 0, \tag{39}$$

and

$$\left(\frac{dG}{d\lambda}, \delta\lambda\right)_{Y_p} = \left(\frac{\partial G}{\partial \varphi}, \delta\varphi\right)_{Y} + \left(\frac{\partial G}{\partial u}, \delta u\right)_{X} + \left(\frac{\partial G}{\partial \lambda}, \delta\lambda\right)_{Y_p}, \tag{40}$$

where $\delta\varphi$, $\delta\varphi^*$ and δu are the Gâteaux derivatives of φ, φ^* and u in the direction $\delta\lambda$ (for example, $\delta\varphi = \frac{\partial\varphi}{\partial\lambda}\delta\lambda$).

To compute the gradient $\nabla_\lambda G(\varphi, u, \lambda)$, let us introduce three adjoint variables $P_1 \in Y$, $P_2 \in Y$ and $P_3 \in X$. By taking the inner product of (37) by P_1, (38) by P_2 and of (39) by P_3 and adding them, we obtain:

$$\left(\delta\varphi, -\frac{\partial P_1}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^*P_1 - (F''_{\varphi\varphi}(\varphi,\lambda)P_2)^*\varphi^* + C^*V_2CP_2\right)_Y + \left(\delta\varphi\big|_{t=T}, P_1\big|_{t=T}\right)_X +$$

$$+\left(\delta\varphi^*, \frac{\partial P_2}{\partial t} - F'_{\varphi}(\varphi,\lambda)P_2\right)_Y + \left(\delta\varphi^*\big|_{t=0}, P_2\big|_{t=0} - P_3\right)_X +$$

$$+\left(\delta u, -P_1\big|_{t=0} + V_1P_3\right)_X + \left(\delta\lambda, -(F'_{\lambda}(\varphi,\lambda))^*P_1 - (F''_{\varphi\lambda}(\varphi,\lambda)P_2)^*\varphi^*\right)_{Y_p} = 0. \tag{41}$$

Here we put

$$-\frac{\partial P_1}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^*P_1 - (F''_{\varphi\varphi}(\varphi,\lambda)P_2)^*\varphi^* + C^*V_2CP_2 = \frac{\partial G}{\partial \varphi},$$

and

$$-P_1\big|_{t=0} + V_1P_3 = \frac{\partial G}{\partial u}, \quad P_1\big|_{t=T} = 0, \quad \frac{\partial P_2}{\partial t} - F'_{\varphi}(\varphi,\lambda)P_2 = 0, \quad P_2\big|_{t=0} - P_3 = 0.$$

Hence, we can exclude the variable P_3 by

$$P_3 = P_2\big|_{t=0}$$

and obtain the initial condition for P_2 in the form:

$$V_1 P_2\big|_{t=0} = \frac{\partial G}{\partial u} + P_1\big|_{t=0}.$$

Thus, if P_1, P_2 are the solutions of the following system of equations

$$\begin{cases} -\dfrac{\partial P_1}{\partial t} - (F'_\varphi(\varphi,\lambda))^* P_1 - (F''_{\varphi\varphi}(\varphi,\lambda)P_2)^*\varphi^* + C^* V_2 C P_2 = \dfrac{\partial G}{\partial \varphi}, \quad t \in (0,T) \\ P_1\big|_{t=T} = 0, \end{cases} \tag{42}$$

$$\begin{cases} \dfrac{\partial P_2}{\partial t} - F'_\varphi(\varphi,\lambda)P_2 = 0, \quad t \in (0,T) \\ V_1 P_2\big|_{t=0} = \dfrac{\partial G}{\partial u} + P_1\big|_{t=0}, \end{cases} \tag{43}$$

then from (41) we get

$$\left(\frac{\partial G}{\partial \varphi}, \delta\varphi\right)_Y + \left(\frac{\partial G}{\partial u}, \delta u\right)_X = \left(\delta\lambda, (F'_\lambda(\varphi,\lambda))^* P_1 + (F''_{\varphi\lambda}(\varphi,\lambda)P_2)^*\varphi^*\right)_{Y_p},$$

and the gradient of G is given by

$$\frac{dG}{d\lambda} = (F'_\lambda(\varphi,\lambda))^* P_1 + (F''_{\varphi\lambda}(\varphi,\lambda)P_2)^*\varphi^* + \frac{\partial G}{\partial \lambda}. \tag{44}$$

We get a coupled system of two differential equations (42) and (43) of the first order with respect to time. One equation has a final condition (backward problem) while the other has an initial condition (forward problem) depending on the initial value for the first equation: it is a non-standard problem.

Let us represent the non-standard problem (42)–(43) in an equivalent form:

$$\begin{cases} -\dfrac{\partial P_1}{\partial t} - (F'_\varphi(\varphi,\lambda))^* P_1 - (F''_{\varphi\varphi}(\varphi,\lambda)P_2)^*\varphi^* + C^* V_2 C P_2 = \dfrac{\partial G}{\partial \varphi}, \quad t \in (0,T) \\ P_1\big|_{t=T} = 0, \end{cases} \tag{45}$$

$$\begin{cases} \dfrac{\partial P_2}{\partial t} - F'_\varphi(\varphi,\lambda)P_2 = 0, \quad t \in (0,T) \\ P_2\big|_{t=0} = v, \end{cases} \tag{46}$$

$$V_1 v - P_1\big|_{t=0} = \frac{\partial G}{\partial u}. \tag{47}$$

Here we have three unknowns: $v \in X$, $P_1, P_2 \in Y$. Let us write (45)–(47) in the form of an operator equation for v. We define the operator $\mathcal{H}$ by the successive solution of the following problems:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} - F'_{\varphi}(\varphi, \lambda)\phi = 0, \quad t \in (0, T) \\ \\ \phi\big|_{t=0} = w, \end{cases} \tag{48}$$

$$\begin{cases} -\dfrac{\partial \phi^*}{\partial t} - (F'_{\varphi}(\varphi, \lambda))^*\phi^* - (F''_{\varphi\varphi}(\varphi, \lambda)\phi)^*\varphi^* = -C^*V_2C\phi, \quad t \in (0, T) \\ \\ \phi^*\big|_{t=T} = 0, \end{cases} \tag{49}$$

$$\mathcal{H}w = V_1 w - \phi^*\big|_{t=0}. \tag{50}$$

Then (45)–(47) is equivalent to the following equation in X:

$$\mathcal{H}v = \mathcal{F} \tag{51}$$

with the right-hand side $\mathcal{F}$ defined by

$$\mathcal{F} = \frac{\partial G}{\partial u} + \tilde{\phi}^*\big|_{t=0}, \tag{52}$$

where $\tilde{\phi}^*$ is the solution to the adjoint problem:

$$\begin{cases} -\dfrac{\partial \tilde{\phi}^*}{\partial t} - (F'_{\varphi}(\varphi, \lambda))^*\tilde{\phi}^* = \dfrac{\partial G}{\partial \varphi}, \quad t \in (0, T) \\ \\ \tilde{\phi}^*\big|_{t=T} = 0. \end{cases} \tag{53}$$

It is easily seen that the operator $\mathcal{H}$ defined by (48)–(50) is the Hessian of the original functional J considered on the optimal solution u of the problem (33)–(35): $J''(u) = \mathcal{H}$. Under the assumption that $\mathcal{H}$ is positive definite, the operator equation (51) is correctly and everywhere solvable in X, i.e. for every $\mathcal{F}$ there exists a unique solution $v \in X$ and

$$\|v\|_X \le c\|\mathcal{H}\|_X, \quad c = const > 0.$$

Therefore, under the assumption that $J''(u)$ is positive definite on the optimal solution, the non-standard problem (42)–(43) has a unique solution $P_1, P_2 \in Y$.

Based on the above consideration, we can formulate the following algorithm to solve the non-standard problem:

(1) For $\frac{\partial G}{\partial u} \in X$, $\frac{\partial G}{\partial \varphi} \in Y$ solve the adjoint problem

$$\begin{cases} -\dfrac{\partial \tilde{\phi}^*}{\partial t} - (F'_\varphi(\varphi,\lambda))^* \tilde{\phi}^* = \dfrac{\partial G}{\partial \varphi}, \quad t \in (0,T) \\ \tilde{\phi}^*\big|_{t=T} = 0 \end{cases} \tag{54}$$

and put

$$\mathcal{F} = \frac{\partial G}{\partial u} + \tilde{\phi}^*\big|_{t=0}.$$

(2) Find v by solving

$$\mathcal{H}v = \mathcal{F}$$

with the Hessian of the original functional J defined by (48)–(50).

(3) Solve successively the direct and adjoint problems

$$\begin{cases} \dfrac{\partial P_2}{\partial t} - F'_\varphi(\varphi,\lambda) P_2 = 0, \quad t \in (0,T) \\ P_2\big|_{t=0} = v, \end{cases} \tag{55}$$

$$\begin{cases} -\dfrac{\partial \tilde{P}_1}{\partial t} - (F'_\varphi(\varphi,\lambda))^* \tilde{P}_1 - (F''_{\varphi\varphi}(\varphi,\lambda) P_2)^* \varphi^* + C^* V_2 C P_2 = 0, \quad t \in (0,T) \\ \tilde{P}_1\big|_{t=T} = 0, \end{cases} \tag{56}$$

and put

$$P_1 = \tilde{P}_1 + \tilde{\phi}^*.$$

Thus, we obtain $P_1, P_2 \in Y$ as the solutions to the non-standard problem (42)–(43), which determine the sensitivity of the response function with respect to imprecisely known parameters according to (44).

6 Sensitivity with Respect to Observations

In geophysical applications a usual request is the estimation of the sensitivity with respect to observations Langland and Baker (2004), Daescu and Langland (2013), Kalnay et al. (2012), Godinez and Daescu (2009). What will be the impact of an uncertainty on the prediction? It is clear that observations are not directly used in the forward model, they are involved only as a forcing term in the adjoint model. Therefore to apply the general formalism of sensitivity analysis we should apply it not to the model itself but to the optimality system, i.e. the model plus the adjoint model. A very simple example with a scalar ordinary differential equation is given in Le Dimet et al. (2002) showing that the only model is not sufficient to carry out

sensitivity analysis in the presence of data. Differentiating the optimality system will introduce second order derivatives.

Consider the mathematical model governed by the nonlinear evolution problem of the form (17):

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi, \lambda) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u. \end{cases} \tag{57}$$

We suppose that for given $u \in X$, $f \in Y$ and $\lambda \in Y_p$ there exists a unique solution $\varphi \in Y$ to (57) with $\dfrac{\partial \varphi}{\partial t} \in Y$. The function λ is an unknown model parameter, and we suppose that the initial state u is also unknown, so we will consider joint parameter and state estimation problem.

Let us introduce the cost function as a functional on $X \times Y_p$ in the form

$$J(u, \lambda) = \frac{1}{2}\|V_1^{1/2}(u - u_b)\|_X + \frac{1}{2}\|V_2^{1/2}(\lambda - \lambda_b)\|_{Y_p} + \frac{1}{2}\|V_3^{1/2}(C\varphi - \varphi_{obs})\|_{Y_{obs}}, \tag{58}$$

where $u \in X$, $\lambda_b \in Y_p$ are prior (background) functions, $\varphi_{obs} \in Y_{obs}$ is a prescribed function (observational data), Y_{obs} is a Hilbert space (observation space), $C : Y \to Y_{obs}$ is a linear bounded operator (observation operator), $V_1 : X \to X$, $V_2 : Y_p \to Y_p$ and $V_3 : Y_{obs} \to Y_{obs}$ are symmetric positive definite bounded operators.

Let us consider the following data assimilation problem with the aim to find the initial value u and the parameter λ: for given $f \in Y$, $\varphi_{obs} \in Y_{obs}$, find $u \in X$, $\lambda \in Y_p$ and $\varphi \in Y$ such that they satisfy (57), and on the set of solutions to (57), the functional $J(u, \lambda)$ takes the minimum value, i.e.

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi, \lambda) + f, \quad t \in (0, T) \\ \varphi\big|_{t=0} = u, \\ J(u, \lambda) = \inf\limits_{w \in X, v \in Y_p} J(w, v). \end{cases} \tag{59}$$

We suppose that the solution of (59) exists. The necessary optimality condition reduces (59) to the optimality system:

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = F(\varphi, \lambda) + f, \quad t \in (0, T), \\ \varphi\big|_{t=0} = u, \end{cases} \tag{60}$$

$$\begin{cases} \dfrac{\partial \varphi^*}{\partial t} + (F'_\varphi(\varphi, \lambda))^* \varphi^* = C^* V_3 (C\varphi - \varphi_{obs}), \quad t \in (0, T) \\ \varphi^*\big|_{t=T} = 0, \end{cases} \tag{61}$$

$$V_1(u - u_b) - \varphi^*\big|_{t=0} = 0, \tag{62}$$

$$V_2(\lambda - \lambda_b) - (F'_\lambda(\varphi, \lambda))^* \varphi^* = 0. \tag{63}$$

Here $F'_\varphi(\varphi,\lambda): Y \to Y,\ F'_\lambda(\varphi,\lambda): Y_p \to Y$ are the Fréchet derivatives of F with respect to φ and λ, correspondingly, and C^* is the adjoint operator to C defined by $(C\varphi,\psi)_{Y_{obs}} = (\varphi, C^*\psi)_Y,\ \varphi \in Y, \psi \in Y_{obs}$.

Supposing that the system (60)–(63) has a unique solution $\varphi, \varphi^* \in Y, u \in X, \lambda \in Y_p$, we will study the sensitivity of functionals of the optimal solution with respect to the observation data φ_{obs}.

We introduce a response function $G(\varphi, u, \lambda)$, which is supposed to be a real-valued function and can be considered as a functional on $Z = Y \times X \times Y_p$. We are interested in the sensitivity of G with respect to φ_{obs}, with φ, u and λ obtained from the optimality system (60)–(63). By definition, the sensitivity is defined by the gradient of G with respect to φ_{obs}:

$$\frac{dG}{d\varphi_{obs}} = \frac{\partial G}{\partial \varphi}\frac{\partial \varphi}{\partial \varphi_{obs}} + \frac{\partial G}{\partial \lambda}\frac{\partial \lambda}{\partial \varphi_{obs}} + \frac{\partial G}{\partial u}\frac{\partial u}{\partial \varphi_{obs}}, \tag{64}$$

where $\frac{\partial G}{\partial \varphi}: Z \to Y$, $\frac{\partial G}{\partial \lambda}: Z \to Y_p$, $\frac{\partial G}{\partial u}: Z \to X$, and $\frac{\partial \varphi}{\partial \varphi_{obs}}$, $\frac{\partial \lambda}{\partial \varphi_{obs}}$, $\frac{\partial u}{\partial \varphi_{obs}}$ are the Gâteaux derivatives of φ, λ, u with respect to φ_{obs}.

Let $\delta\varphi_{obs}$ be a perturbation on φ_{obs}, then we obtain from the optimality system (60)–(63):

$$\begin{cases} \dfrac{\partial \delta\varphi}{\partial t} = F'_\varphi(\varphi,\lambda)\delta\varphi + F'_\lambda(\varphi,\lambda)\delta\lambda, \quad t \in (0,T) \\ \delta\varphi\big|_{t=0} = \delta u, \end{cases} \tag{65}$$

$$\begin{cases} -\dfrac{\partial \delta\varphi^*}{\partial t} - (F'_\varphi(\varphi,\lambda))^*\delta\varphi^* - (F''_{\varphi\varphi}(\varphi,\lambda)\delta\varphi)^*\varphi^* = (F''_{\varphi\lambda}(\varphi,\lambda)\delta\lambda)^*\varphi^* \\ \qquad\qquad -C^*V_3(C\delta\varphi - \delta\varphi_{obs}), \\ \qquad\qquad \delta\varphi^*\big|_{t=T} = 0, \end{cases} \tag{66}$$

$$V_1\delta u - \delta\varphi^*\big|_{t=0} = 0, \tag{67}$$

$$V_2\delta\lambda - (F''_{\lambda\varphi}(\varphi,\lambda)\delta\varphi)^*\varphi^* - (F''_{\lambda\lambda}(\varphi,\lambda)\delta\lambda)^*\varphi^* - (F'_\lambda(\varphi,\lambda))^*\delta\varphi^* = 0, \tag{68}$$

and

$$\left(\frac{dG}{d\varphi_{obs}}, \delta\varphi_{obs}\right)_{Y_{obs}} = \left(\frac{\partial G}{\partial \varphi}, \delta\varphi\right)_Y + \left(\frac{\partial G}{\partial \lambda}, \delta\lambda\right)_{Y_p} + \left(\frac{\partial G}{\partial u}, \delta u\right)_X, \tag{69}$$

where $\delta\varphi$, $\delta\varphi^*$, $\delta\lambda$, δu are the solutions of (65)–(68).

Following the methodology presented in Sect. 5, we obtain the gradient of G through solutions of a non-standard problem.

Let $P_1, P_2 \in Y$, $P_3 \in Y_p$, $P_4 \in X$ be the solutions of the following system of equations

$$\begin{cases} -\dfrac{\partial P_1}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^* P_1 - (F''_{\varphi\varphi}(\varphi,\lambda)P_2)^*\varphi^* = (F''_{\lambda\varphi}(\varphi,\lambda)P_3)^*\varphi^* - C^*V_3CP_2 \\ \qquad\qquad + \dfrac{\partial G}{\partial \varphi}, \\ P_1\big|_{t=T} = 0, \end{cases} \tag{70}$$

$$\begin{cases} \dfrac{\partial P_2}{\partial t} - F'_{\varphi}(\varphi,\lambda)P_2 - F'_{\lambda}(\varphi,\lambda)P_3 = 0, \quad t \in (0,T) \\ P_2\big|_{t=0} - P_4 = 0, \end{cases} \tag{71}$$

$$V_1P_4 - P_1\big|_{t=0} = \frac{\partial G}{\partial u}, \tag{72}$$

$$V_2P_3 - (F''_{\varphi\lambda}(\varphi,\lambda)P_2)^*\varphi^* - (F''_{\lambda\lambda}(\varphi,\lambda)P_3)^*\varphi^* - (F'_{\lambda}(\varphi,\lambda))^*P_1 = \frac{\partial G}{\partial \lambda}, \tag{73}$$

where $\varphi, \varphi^* \in Y$, $u \in X$, $\lambda \in Y_p$ are the solution of the optimality system (60)–(63). Then the gradient of G with respect to φ_{obs} is given by

$$\frac{dG}{d\varphi_{obs}} = V_3CP_2. \tag{74}$$

We obtain a coupled system of two differential equations (70) and (71) of the first order with respect to time, with additional conditions (72)–(73). To study this non-standard problem (70)–(73) with mutually dependent initial conditions for P_1, P_2, we reduce it to a single operator equation involving the Hessian of the original cost function $J(u,\lambda)$.

The Hessian $\mathcal{H} : X \times Y_p \to X \times Y_p$ acts on $U = (w,v)^T \in X \times Y_p$ and is defined by the successive solution of the following problems:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} - F'_{\varphi}(\varphi,\lambda)\phi = F'_{\lambda}(\varphi,\lambda)v, \quad t \in (0,T) \\ \phi\big|_{t=0} = w, \end{cases} \tag{75}$$

$$\begin{cases} -\dfrac{\partial \phi^*}{\partial t} - (F'_{\varphi}(\varphi,\lambda))^*\phi^* - (F''_{\varphi\varphi}(\varphi,\lambda)\phi)^*\varphi^* = (F''_{\lambda\varphi}(\varphi,\lambda)w)^*\varphi^* - C^*V_3C\phi, \\ \phi^*\big|_{t=T} = 0, \end{cases} \tag{76}$$

$$\mathcal{H}U = \left(V_1w - \phi^*\big|_{t=0},\ V_2v - (F''_{\varphi\lambda}(\varphi,\lambda)\phi)^*\varphi^* - (F''_{\lambda\lambda}(\varphi,\lambda)w)^*\varphi^* - (F'_{\lambda}(\varphi,\lambda))^*\phi^* \right)^T, \tag{77}$$

where λ, u, φ and φ^* are the solutions of the optimality system (60)–(63). It is easily seen that (70)–(73) is equivalent to the following equation in $X \times Y_p$:

$$\mathcal{H}U = \mathcal{F} \tag{78}$$

with some $\mathcal{F} \in X \times Y_p$.

Under the assumption that $\mathcal{H}$ is positive definite, the operator equation (78) is correctly and everywhere solvable in $X \times Y_p$, i.e. for every $\mathcal{F}$ there exists a unique solution $U \in X \times Y_p$ and the estimate is valid:

$$\|U\|_{X\times Y_p} \le c\|\mathcal{F}\|_{X\times Y_p}, \quad c = const > 0.$$

Therefore, under the assumption that $J''(u, \lambda)$ is positive definite on the optimal solution, the non-standard problem (70)–(73) has a unique solution P_1, $P_2 \in Y$, $P_3 \in Y_p$, $P_4 \in X$.

Based on (70)–(74),we can formulate the following algorithm to compute the gradient of the response function G:

(1) For $\frac{\partial G}{\partial \lambda} \in Y_p$, $\frac{\partial G}{\partial \varphi} \in Y$, $\frac{\partial G}{\partial u} \in X$ solve the adjoint problem

$$\begin{cases} -\frac{\partial \tilde{\phi}^*}{\partial t} - (F'_\varphi(\varphi, \lambda))^* \tilde{\phi}^* = \frac{\partial G}{\partial \varphi}, \quad t \in (0, T) \\ \tilde{\phi}^*\big|_{t=T} = 0 \end{cases} \tag{79}$$

and put

$$\mathcal{F} = \left(\frac{\partial G}{\partial u} + \tilde{\phi}^*\big|_{t=0}, \ \frac{\partial G}{\partial \lambda} + (F'_\lambda(\varphi, \lambda))^* \tilde{\phi}^* \right)^T.$$

(2) Find $U = (w, v)^T$ by solving

$$\mathcal{H}U = \mathcal{F}$$

with the Hessian of the original functional J defined by (75)–(77).

(3) Solve the direct problem

$$\begin{cases} \frac{\partial P_2}{\partial t} - F'_\varphi(\varphi, \lambda) P_2 = F'_\lambda(\varphi, \lambda) v, \quad t \in (0, T) \\ P_2\big|_{t=0} = w. \end{cases} \tag{80}$$

(4) Compute the gradient of the response function as

$$\frac{dG}{d\varphi_{obs}} = V_3 C P_2. \tag{81}$$

The last formula allows us to estimate the sensitivity of the response functions related to the optimal solution after assimilation, with respect to observation data.

7 Application for a Sea Thermodynamics Model

We consider the sea thermodynamics problem in the form Marchuk et al. (1987):

$$
\begin{gathered}
T_t + (\bar{U}, \mathrm{Grad})T - \mathrm{Div}(\hat{a}_T \cdot \mathrm{Grad}\, T) = f_T \ \text{ in } \ D \times (t_0, t_1), \\
T = T_0 \text{ for } t = t_0 \text{ in } D, \\
-\nu_T \frac{\partial T}{\partial z} = Q \text{ on } \Gamma_S \times (t_0, t_1), \quad \frac{\partial T}{\partial n} = 0 \text{ on } \Gamma_{w,c} \times (t_0, t_1), \\
\bar{U}_n^{(-)} T + \frac{\partial T}{\partial n} = Q_T \text{ on } \Gamma_{w,op} \times (t_0, t_1), \\
\frac{\partial T}{\partial n} = 0 \text{ on } \Gamma_H \times (t_0, t_1),
\end{gathered}
\tag{82}
$$

where $T = T(x, y, z, t)$ is an unknown temperature function, $t \in (t_0, t_1)$, $(x, y, z) \in D = \Omega \times (0, H)$, $\Omega \subset R^2$, $H = H(x, y)$ is the function of the bottom releif, $Q = Q(x, y, t)$ is the total heat flux, $\bar{U} = (u, v, w)$, $\widehat{a}_T = \mathrm{diag}((a_T)_{ii})$, $(a_T)_{11} = (a_T)_{22} = \mu_T$, $(a_T)_{33} = \nu_T$, $f_T = f_T(x, y, z, t)$ are given functions. The boundary of the domain $\Gamma \equiv \partial D$ is represented as a union of four disjoint parts $\Gamma_S, \Gamma_{w,op}, \Gamma_{w,c}$, Γ_H, where $\Gamma_S = \Omega$ (the unperturbed sea surface), $\Gamma_{w,op}$ is the liquid (open) part of vertical lateral boundary, $\Gamma_{w,c}$ is the solid part of the vertical lateral boundary, Γ_H is the sea bottom, $\bar{U}_n^{(-)} = (|\bar{U}_n| - \bar{U}_n)/2$, and $\bar{U}_n$ is the normal component of $\bar{U}$. The other notations and a detailed description of the problem statement can be found in Agoshkov et al. (2008).

Problem (82) can be written in the form of an operator equation:

$$
\begin{gathered}
T_t + LT = \mathcal{F} + BQ, \quad t \in (t_0, t_1), \\
T = T_0, \qquad t = t_0,
\end{gathered}
\tag{83}
$$

where the equality is understood in the weak sense, namely,

$$
(T_t, \widehat{T}) + (LT, \widehat{T}) = \mathcal{F}(\widehat{T}) + (BQ, \widehat{T}) \ \ \forall \widehat{T} \in W_2^1(D), \tag{84}
$$

in this case $L, \mathcal{F}, B$ are defined by the following relations:

$$
(LT, \widehat{T}) \equiv \int_D (-T\mathrm{Div}(\bar{U}\widehat{T}))dD + \int_{\Gamma_{w,op}} \bar{U}_n^{(+)} T\widehat{T} d\Gamma + \int_D \widehat{a}_T \mathrm{Grad}(T) \cdot \mathrm{Grad}(\widehat{T}) dD,
$$

$$
\mathcal{F}(\widehat{T}) = \int_{\Gamma_{w,op}} Q_T \widehat{T} d\Gamma + \int_D f_T \widehat{T} dD, (T_t, \widehat{T}) = \int_D T_t \widehat{T} dD, (BQ, \widehat{T}) = \int_\Omega Q\widehat{T}\big|_{z=0} d\Omega,
$$

and the functions $\widehat{a}_T$, Q_T, f_T, Q are such that equality (84) makes sense. The properties of the operator L were studied in Agoshkov et al. (2008).

Problem (82) is linear in T, Q, however, written in the form (83), it is a particular case of the original problem (57), and all the reasoning and the methodology presented in Sect. 6 are easily transferred to the case of problem (83), understood in a weak sense (83).

We consider the data assimilation problem for the sea surface temperature (see Agoshkov et al. (2008)). Suppose that the functions $Q \in L_2(\Omega \times (t_0, t_1))$ and $T_0 \in L_2(D)$ are unknown in problem (82). Let also $T_{obs}(x, y, t) \in L_2(\Omega \times (t_0, t_1))$ be the function on Ω obtained for $t \in (t_0, t_1)$ by processing the observation data, and this function in its physical sense is an approximation to the surface temperature function on Ω, i.e. to $T\big|_{z=0}$. We admit the case when T_{obs} is defined only on some subset of $\Omega \times (t_0, t_1)$ and denote the indicator (characteristic) function of this set by m_0. For definiteness sake, we assume that T_{obs} is zero outside this subset.

Consider the data assimilation problem for the surface temperature in the following form: find T_0 and Q such that

$$\begin{cases} T_t + LT = \mathcal{F} + BQ \ \text{ in } \ D \times (t_0, t_1), \\ \qquad T = T_0, \quad t = t_0 \\ J(T_0, Q) = \inf\limits_{w,v} J(w, v), \end{cases} \tag{85}$$

where

$$J(T_0, Q) = \frac{\alpha}{2} \int\limits_{t_0}^{t_1} \int\limits_{\Omega} |Q - Q^{(0)}|^2 d\Omega dt + \frac{\beta}{2} \int\limits_{D} |T_0 - T^{(0)}|^2 dD +$$

$$+ \frac{1}{2} \int\limits_{t_0}^{t_1} \int\limits_{\Omega} m_0 |T\big|_{z=0} - T_{obs}|^2 d\Omega dt, \tag{86}$$

and $Q^{(0)} = Q^{(0)}(x, y, t)$, $T^{(0)} = T^{(0)}(x, y, z)$ are given functions, $\alpha, \beta = const > 0$.

For $\alpha, \beta > 0$ this variational data assimilation problem has a unique solution. The existence of the optimal solution follows from the classic results of the theory of optimal control problems Lions (1968).

The optimality system determining the solution of the formulated variational data assimilation problem according to the necessary condition $\text{grad} J = 0$ has the form:

$$\begin{gathered} T_t + LT = \mathcal{F} + BQ \quad \text{in} \quad D \times (t_0, t_1), \\ T = T_0, \qquad t = t_0, \end{gathered} \tag{87}$$

$$-(T^*)_t + L^*T^* = Bm_0(T_{\text{obs}} - T) \text{ in } D \times (t_0, t_1),$$
$$T^* = 0, \quad t = t_1, \tag{88}$$

$$\alpha(Q - Q^{(0)}) - T^* = 0 \quad \text{on } \Omega \times (t_0, t_1), \tag{89}$$

$$\beta(T_0 - T^{(0)}) - T^*\big|_{t=0} = 0 \quad \text{in } D, \tag{90}$$

where L^* is the operator adjoint to L.

Here the boundary-value function Q plays the role of λ from Sect. 6, $\varphi = T$, the operator F has the form $F(T, Q) = -LT + BQ$, and $F'_T = -L$, $F'_Q = B$. Since the operator $F(T, Q)$ is linear in this case and $F''_{TT} = F''_{QT} = F''_{QQ} = 0$, the Hessian $\mathcal{H}$ acting on some $U = (w, \psi)^T$, $w \in L_2(D)$, $\psi \in L_2(\Omega \times (t_0, t_1))$ is defined by the successive solution of the following problems:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + L\phi = B\psi, \quad t \in (t_0, t_1) \\ \phi\big|_{t=t_0} = w, \end{cases} \tag{91}$$

$$\begin{cases} -\dfrac{\partial \phi^*}{\partial t} + L^*\phi^* = -Bm_0\phi, \quad t \in (t_0, t_1) \\ \phi^*\big|_{t=t_1} = 0, \end{cases} \tag{92}$$

$$\mathcal{H}U = (\beta w - \phi^*\big|_{t=0}, \alpha\psi - B^*\phi^*)^T. \tag{93}$$

To illustrate the above-presented theory, we consider the problem of sensitivity of functionals of the optimal solution T_0, Q to the observations T_{obs}. Let us introduce the following response function:

$$G(T) = \int_{t_0}^{t_1} dt \int_{\Omega} k(x, y, t) T(x, y, 0, t) d\Omega, \tag{94}$$

where $k(x, y, t)$ is a weight function related to the temperature field on the sea surface $z = 0$. For example, if we are interested in the mean temperature of a specific region of the sea ω for $z = 0$ in the interval $\bar{t} - \tau \le t \le \bar{t}$, then as k we take the function

$$k(x, y, t) = \begin{cases} 1 \Big/ (\tau \operatorname{mes} \omega) \text{ if } (x, y) \in \omega, \ \bar{t} - \tau \le t \le \bar{t} \\ 0 \qquad\qquad\qquad\qquad \text{else,} \end{cases} \tag{95}$$

where mes ω denotes the area of the region ω. Thus, the functional (94) is written in the form:

$$G(T) = \frac{1}{\tau} \int_{\bar{t}-\tau}^{\bar{t}} dt \left(\frac{1}{\operatorname{mes} \omega} \int_{\omega} T(x, y, 0, t) d\Omega \right). \tag{96}$$

Formula (96) represents the mean temperature averaged over the time interval $\bar{t} - \tau \le t \le \bar{t}$ for a given region ω. The response functions of this type are of most interest in the theory of climate change (Marchuk (1995), Marchuk et al. (1996)).

In our notations the functional (94) may be written as

$$G(T) = \int_{t_0}^{t_1} (Bk, T)dt = (Bk, T)_Y, \quad Y = L_2(D \times (t_0, t_1)).$$

We are interested in the sensitivity of the response function $G(T)$, obtained for T after data assimilation, with respect to the observation function T_{obs}.

By definition, the sensitivity is given by the gradient of G with respect to T_{obs}:

$$\frac{dG}{dT_{obs}} = \frac{\partial G}{\partial T}\frac{\partial T}{\partial T_{obs}}. \tag{97}$$

Since $\frac{\partial G}{\partial T} = Bk$, then according to the theory presented in Sect. 6, to compute the gradient (97) we need to perform the following steps:

1) For k defined by (95) solve the adjoint problem

$$\begin{cases} -\frac{\partial \tilde{\phi}^*}{\partial t} + L^*\tilde{\phi}^* = Bk, \quad t \in (t_0, t_1) \\ \tilde{\phi}^*\big|_{t=t_1} = 0 \end{cases} \tag{98}$$

and put $\Phi = (\tilde{\phi}^*\big|_{t=0}, B^*\tilde{\phi}^*)^T$.

2) Find $U = (w, v)^T$ by solving $\mathcal{H}U = \Phi$ with the Hessian defined by (91)–(93).

3) Solve the direct problem

$$\begin{cases} \frac{\partial P_2}{\partial t} + LP_2 = Bv, \quad t \in (t_0, t_1) \\ P_2\big|_{t=t_0} = w. \end{cases} \tag{99}$$

4) Compute the gradient of the response function as

$$\frac{dG}{dT_{obs}} = m_0 P_2\big|_{z=0}. \tag{100}$$

The last formula allows us to estimate the sensitivity of the functionals related to the mean temperature after data assimilation, with respect to the observations on the sea surface.

For numerical experiments have used the three-dimensional numerical model of the Baltic Sea hydrothermodynamics developed at the INM RAS on the base of the splitting method Zalesny et al. (2017) and supplied with the assimilation procedure Agoshkov et al. (2008) for the surface temperature T_{obs} with the aim to reconstruct the heat fluxes Q and the initial state T_0.

The parameters of the considered domain of the Baltic Sea and its geographic coordinates can be described as follows: σ-grid is $336 \times 394 \times 25$ (the latitude, longitude, and depth, respectively). The first point of the "grid C" Zalesny et al. (2017) has the coordinates 9.406° E and 53.64° N. The mesh sizes in x and y are constant and equal to 0.0625 and 0.03125 degrees. The time step is $\Delta t = 5$ minutes. The assimilation procedure worked only during some time windows. To start the assimilation procedure, the function $T^{(0)}$ was taken as a model forecast for the previous time interval.

The Baltic Sea daily-averaged nighttime surface temperature data were used for T_{obs}. These are the data of the Danish Meteorological Institute based on measurements of radiometers (AVHRR, AATSR and AMSRE) and spectroradiometers (SEVIRI and MODIS) Karagali et al. (2012). Data interpolation algorithms were used Zakharova et al. (2013) to convert observations on computational grid of the numerical model of the Baltic Sea thermodynamics. The mean climatic flux obtained from the NCEP (National Center for Environmental Prediction) reanalysis was taken for $Q^{(0)}$.

Using the hydrothermodynamics model mentioned above, which is supplied with the assimilation procedure for the surface temperature T_{obs}, we have performed calculations for the Baltic Sea area where the assimilation algorithm worked only at certain time moments t_0; in this case $t_1 = t_0 + \Delta t$. The aim of the experiment was the numerical study of the sensitivity of functionals of the optimal solution T_0, Q to observation errors in the interval (t_0, t_1).

We use the discretize-then-optimize approach, and for numerical experiments all the presented equations are understood in a discrete form, as finite-dimensional analogues of the corresponding problems, obtained after approximation. This allows us to consider model equations as a perfect model, with no approximation errors.

Let us present some results of numerical experiments.

The calculation results for $t_0 = 50$ h (600 time steps for the model) are presented in Fig. 1 showing the gradient of the response function $G(T)$ defined by (96) and related to the mean temperature after data assimilation, with respect to the observations on the sea surface, according to (98)– (100). Here $\omega = \Omega$, $\tau = \Delta t$, $\bar{t} = t_1$, $\alpha = \beta = 10^{-5}$ (Fig. 1).

We can see the sub-areas (in red) in which the response function $G(T)$ is most sensitive to errors in the observations during assimilation. The largest values of the gradient of $G(T)$ correspond to the points x, y with a small depth (cf. sea topography, Fig. 2). Thus, the considered functional $G(T)$ of the optimal solution turned out to be the most sensitive to observation errors at surface points near these regions. This result is confirmed by the direct computation of the response function $G(T)$ according to (96) obtained after assimilation, by introducing perturbations into the observation data T_{obs}.

The above studies allow to determine the sea sub-areas in which the response function related to the optimal solution is most sensitive to errors in the observations during variational data assimilation.

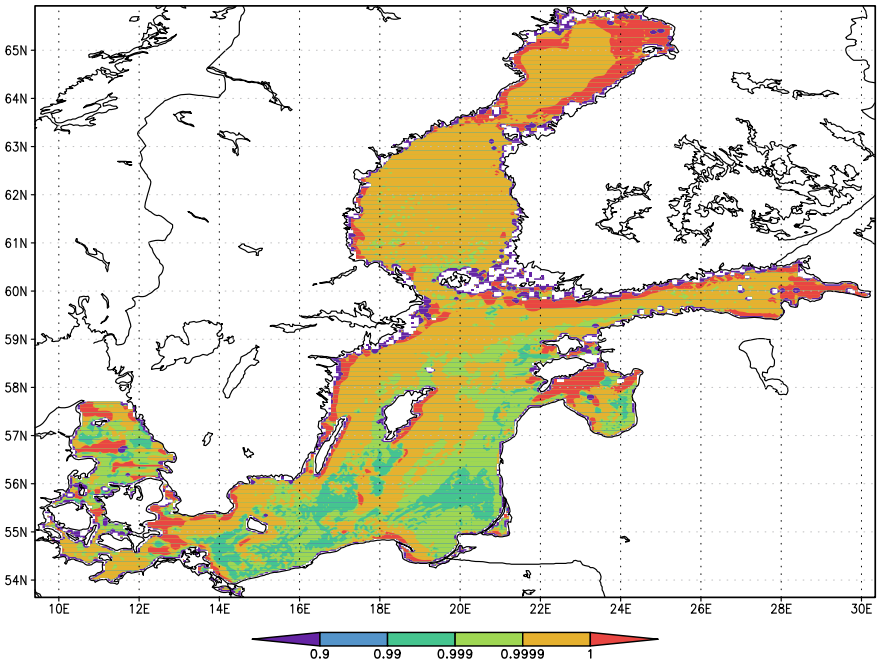

Fig. 1 The gradient of the response function $G(T)$

8 Conclusions

Variational data assimilation is an efficient method in modeling the large-scale geophysical flows, with the main difficulty being linked to the nonlinearity of the governing equations. This method allows to combine the observational data and the model forecasts. From the mathematical point of view, we have to deal with the initial-boundary-value control problems for a nonlinear evolution model governed by partial differential equations. The necessary optimality condition is defined by the optimality system which is based on the gradient of the cost function and involves forward and adjoint equations. To study the variational data assimilation problem as an optimal control problem and to develop efficient algorithms for its numerical solution, second-order information is needed. This is information about the Hessian of the cost functional. To construct the Hessian, it is necessary to differentiate the optimality system and derive a second-order adjoint problem. The investigation of the second-order adjoint equations and the Hessian of the cost function plays an important role in the study of the solvability of the variational assimilation problem, the construction of algorithms for its numerical solution based on the modification of Newton type methods, the identification of model parameters. The Hessian allows to study the sensitivity of the optimal solution and its functionals with respect to observations and uncertainties in parameters.

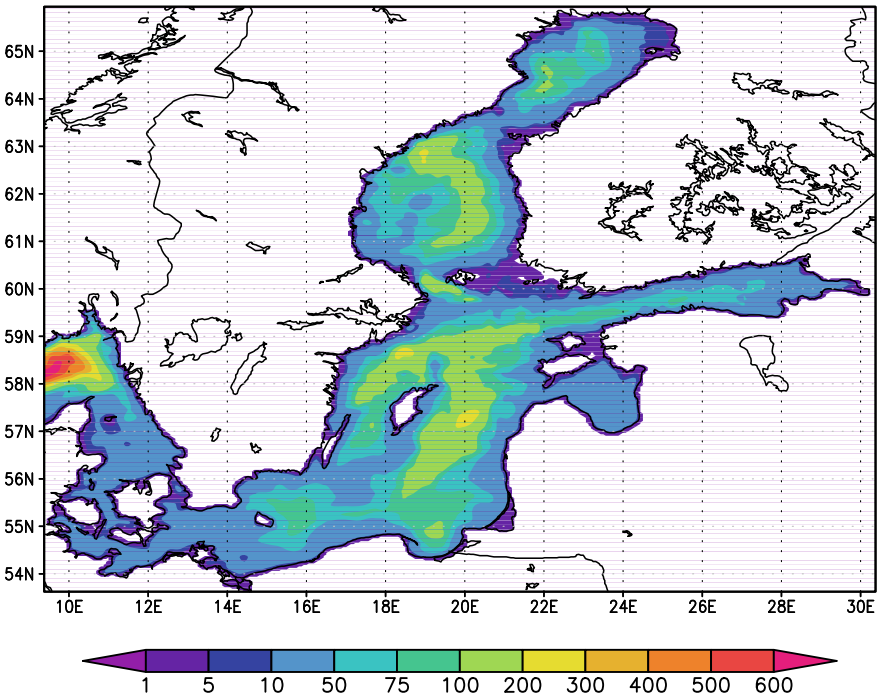

Fig. 2 Baltic Sea topography

Acknowledgements The research (Sections 5–7) was supported by the Russian Science Foundation (project No.20-11-20057).

References

Agoshkov VI, Sheloput TO (2017) The study and numerical solution of some inverse problems in simulation of hydrophysical fields in water areas with 'liquid' boundaries. Russ J Numer Anal Math Modelling 32:147–164

Agoshkov VI, Parmuzin EI, Shutyaev VP (2008) Numerical algorithm of variational assimilation of the ocean surface temperature data. Comp Math Math Phys 48:1371–1391

Agoshkov VI, Parmuzin EI, Zalesny VB, Shutyaev VP et al (2015) Variational assimilation of observation data in the mathematical model of the Baltic Sea dynamics. Russ J Numer Anal Math Modelling 30:203–212

Alifanov OM, Artyukhin EA, Rumyantsev SV (1996) Extreme methods for solving ill-posed problems with applications to inverse heat transfer problems. Begell House Publishers, Danbury

Asch M, Bocquet M, Nodet M (2016) Data assimilation: methods, algorithms, and applications. SIAM, Philadelphia

Bocquet M (2012) Parameter-field estimation for atmospheric dispersion: application to the Chernobyl accident using 4D-Var. Q J R Meteor Soc 138:664–681

Cacuci DG (1981) Sensitivity theory for nonlinear systems: II. Extensions to additional classes of responses. J Math Phys 22:2803–2812
Carrassi A, Bocquet M, Bertino L, Evensen G (2018) Data assimilation in the geosciences: an overview of methods, issues, and perspectives. WIREs Clim Change 9:1–80
Chavent G (1983) Local stability of the output least square parameter estimation technique. Math Appl Comp 2:3–22
Cioaca A, Sandu A, de Sturler E (2013) Efficient methods for computing observation impact in 4D-Var data assimilation. Comput Geosci 17:975–990
Daescu DN (2008) On the sensitivity equations of four-dimensional variational (4D-Var) data assimilation. Mon Weather Rev 136:3050–3065
Daescu DN, Langland RH (2013) Error covariance sensitivity and impact estimation with adjoint 4D-Var: Theoretical aspects and first applications to NAVDAS-AR. Q J R Meteorol Soc 139:226–241
Dontchev AL (1983) Perturbations, approximations and sensitivity analysis of optimal control systems. Lecture notes in control and information sciences, vol 52. Springer, Berlin
Fletcher SJ (2017) Data assimilation for the geosciences: from theory to application. Elsevier, Amsterdam
Gejadze I, Le Dimet FX, Shutyaev V (2008) On analysis error covariances in variational data assimilation. SIAM J Sci Comp 30:1847–1874
Gejadze I, Le Dimet FX, Shutyaev V (2010) On optimal solution error covariances in variational data assimilation problems. J Comput Phys 229:2159–2178
Gejadze IYu, Copeland GJM, Le Dimet FX, Shutyaev VP (2011) Computation of the analysis error covariance in variational data assimilation problems with nonlinear dynamics. J Comput Phys 230:7923–7943
Gejadze IYu, Shutyaev VP, Le Dimet FX (2013) Analysis error covariance versus posterior covariance in variational data assimilation. Q J R Meteor Soc 139:1826–1841
Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic Press, London
Godinez HC, Daescu DN, (2009) A second order adjoint method to targeted observations. In: Allen G, Nabrzyski J, Seidel E, van Albada GD, Dongarra J, Sloot PMA (eds) Computational Science - ICCS 2009. ICCS, (2009) Lecture notes in computer science, vol 5545. Springer, Berlin, Heidelberg, pp 332–341
Griesse R, Vexler B (2007) Numerical sensitivity analysis for the quantity of interest in pde-constrained optimization. SIAM J Sci Comp 29:22–48
Kalnay E, Ota Y, Miyoshi T et al (2012) A simpler formulation of forecast sensitivity to observations: Application to ensemble Kalman filters. Tellus 64A:18462
Karagali I, Hoyer J, Hasager CB (2012) SST diurnal variability in the North Sea and the Baltic Sea. Remote Sens Environ 121:159–170
Kolmogorov AN (1946) On the proof of the method of least squares. Uspekhi Mat Nauk 1:57–70
Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201
Le Dimet FX (1982) A general formalism of variational analysis. Report OK-73091-22-1, CIMMS, Norman
Le Dimet FX, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A 38:97–110
Le Dimet FX, Ngodock HE, Luong B, Verron J (1997) Sensitivity analysis in variational data assimilation. J Meteorol Soc Japan 75:245–255
Le Dimet FX, Navon IM, Daescu DN (2002) Second-order information in data assimilation. Mon Weather Rev 130:629–648
Lions JL (1968) Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles. Dunod, Paris
Marchuk GI, Penenko VV (1978) Application of optimization methods to the problem of mathematical simulation of atmospheric processes and environment. In: Modelling and optimization of complex systems. Proceedings of the IFIP-TC7 conference. Springer, Heidelberg, pp 240–252

Marchuk GI (1995) Adjoint equations and analysis of complex systems. Kluwer, Dordrecht
Marchuk GI, Dymnikov VP, Zalesny VB (1987) Mathematical models in geophysical hydrodynamics and numerical methods for their realization. Gidrometeoizdat, Leningrad
Marchuk GI, Agoshkov VI, Shutyaev VP (1996) Adjoint equations and perturbation algorithms in nonlinear problems. CRC Press, New York
Markov AA (1900) Ischislenie veroyatnostej. Imperial Academy of Sciences, St Petersburg
Navon IM (1998) Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. Dyn Atmos Oceans 27:55–79
Penenko V, Obraztsov NN (1976) A variational initialization method for the fields of the meteorological elements. Soviet Meteorol Hydrol (English transl) 11:1–11
Polak E (1997) Optimization: algorithms and consistent approximations. Appl Math Sci 124. Springer, New York
Pontryagin LS, Boltyansky VG, Gamkrelidze RV, Mishchenko EF (1964) The mathematical theory of optimal processes. International Series of Monographs in Pure and Applied Mathematics, vol 55. Pergamon Press, Oxford, New York
Sasaki Y (1958) An objective analysis based on the variational method. J Meteorol Soc Japan 36:77–88
Schirber S, Klocke D, Pincus R et al (2013) Parameter estimation using data assimilation in an atmospheric general circulation model: From a perfect toward the real world. J Adv Model Earth Syst 5:58–70
Shutyaev VP, Le Dimet FX, Gejadze IYu et al (2012) Optimal solution error covariance in highly nonlinear problems of variational data assimilation. Nonlin Processes Geophys 19:177–184
Shutyaev V, Le Dimet FX, Shubina E (2017) Sensitivity with respect to observations in variational data assimilation. Russ J Numer Anal Math Modelling 32:61–71
Shutyaev VP, Le Dimet FX, Parmuzin EI (2018) Sensitivity analysis with respect to observations in variational data assimilation for parameter estimation. Nonlin Processes Geophys 25:429–439
Smith PJ, Thornhill GD, Dance SL et al (2013) Data assimilation for state and parameter estimation: application to morphodynamic modelling. Q J R Meteor Soc 139:314–327
Storch RB, Pimentel LCG, Orlande HRB (2007) Identification of atmospheric boundary layer parameters by inverse problem. Atmos Environ 41:141–1425
Sun NZ (1994) Inverse problems in groundwater modeling. Kluwer, Dordrecht
Talagrand O, Courtier P (1987) Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. Q J R Meteor Soc 113:1311–1328
Thacker WC (1989) The role of the Hessian matrix in fitting models to measurements. J Geophys Res 94:6177–6196
White LW, Vieux B, Armand D, Le Dimet FX (2003) Estimation of optimal parameters for a surface hydrology model. Adv Water Resour 26:337–348
Yuepeng W, Yue C, Navon IM et al (2018) Parameter identification techniques applied to an environmental pollution model. J Ind Manage Optim 14:817–831
Zakharova NB, Agoshkov VI, Parmuzin EI (2013) The new method of ARGO buoys system observation data interpolation. Russ J Numer Anal Math Modelling 28:67–84
Zalesny V, Agoshkov V, Aps R, Shutyaev V et al (2017) Numerical modeling of marine circulation, pollution assessment and optimal ship routes. J Mar Sci Eng 5:1–20
Zhu Y, Navon IM (1999) Impact of parameter estimation on the performance of the FSU global spectral model using its full-physics adjoint. Mon Weather Rev 127:1497–1517

Statistical Parameter Estimation for Observation Error Modelling: Application to Meteor Radars

Elizabeth A. Satterfield, Joanne A. Waller, David D. Kuhl, Dan Hodyss, Karl W. Hoppel, Stephen D. Eckermann, John P. McCormack, Jun Ma, David C. Fritts, Hiroyuki Iimura, Gunter Stober, Chris E. Meek, Chris Hall, Christoph Jacobi, Ralph Latteck, Nicholas J. Mitchell, Patrick J. Espy, Guozhu Li, Peter Brown, Wen Yi, Na Li, Paulo Batista, Iain M. Reid, Eswaraiah Sunkara, Tracy Moffat-Griffin, Damian J. Murphy, Masaki Tsutsumi, and John Marino

Abstract Data assimilation schemes blend observational data, with limited coverage, with a short term forecast to produce an analysis, which is meant to be the best estimate of the current state of the atmosphere. Appropriately specifying observation error statistics is necessary to obtain an optimal analysis. Observation error can originate from instrument error as well as the error of representation. While representation error is most commonly associated with unresolved scales and processes, this term is often considered to include contributions from pre-processing or quality

E. A. Satterfield (✉)
Marine Meteorology Division, U.S. Naval Research Laboratory, Monterey, CA, USA
e-mail: elizabeth.satterfield@nrlmry.navy.mil

J. A. Waller
Met Office, Reading, UK

D. D. Kuhl · D. Hodyss · K. W. Hoppel · S. D. Eckermann · J. P. McCormack
U.S. Naval Research Laboratory, Washington, DC, USA

J. Ma
CPI, Springfield, VA, USA

D. C. Fritts · H. Iimura
GATS, Inc., Boulder, CO, USA

G. Stober
Institute of Applied Physics & Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

C. E. Meek
Institute of Space and Atmospheric Studies, University of Saskatchewan, Saskatoon, CA, USA

C. Hall
Tromsø Geophysical Observatory, University of Tromsø, Tromsø, Norway

C. Jacobi
Institute for Meteorology, University of Leipzig, Leipzig, Germany

R. Latteck
Leibniz Institute of Atmospheric Physics at the University of Rostock, Kuhlungsborn, Germany

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_8

control and errors associated with the observation operator. With a focus on practical operational implementation, this chapter aims to define the components of observation error, discusses their sources and characteristics, and provides an overview of current methods for estimating observation error statistics. We highlight the implicit assumptions of these methods, as well as their shortcomings. We will detail current operational practice for diagnosing observation error and accounting for correlated observation error. Finally, we provide a practical methodology for using these diagnostics, as well as the associated innovation-based observation impact, to optimize the assimilation of meteor radar observations in the upper atmosphere.

1 Introduction

Initial conditions for a numerical weather forecast are calculated using Data Assimilation (DA) which blends previous model forecast data with atmospheric observations to provide the best estimate of the current atmospheric state. In the assimilation process, the observations and prior model data, also known as the background, are weighted by their respective uncertainties; hence, the accurate specification of the

N. J. Mitchell
Centre for Space Atmosphere and Oceanic Science, Department of Electronic and Electrical Engineering, University of Bath, Bath, UK

P. J. Espy
Atmospheric and Environmental Physics at the Norwegian University of Science and Technology, Trondheim, Norway

G. Li
Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, CN, China

P. Brown
Department of Physics and Astronomy, University of Western Ontario, Ontario, CA, USA

W. Yi
CAS Key Laboratory of Geospace Environment, Department of Geophysics and Planetary Sciences, University of Science and Technology of China, Hefei, CN, China

N. Li
National Key Laboratory of Electromagnetic Environment, China Research Institute of Radiowave Propagation, Qingdao, CN, China

P. Batista
Instituto Nacional de Pesquisas Esaciais, São José dos Campos, SP, Brazil

I. M. Reid
School of Physical Sciences, The University of Adelaide, Adelaide, Australia

E. Sunkara
Space Science Laboratory, Department of Astronomy and Space Science, Chungnam National University, Daejeon, Korea

T. Moffat-Griffin
British Antarctic Survery, Cambridge, UK

observation and background error uncertainties is imperative if an optimal analysis is to be obtained. Until recently, more emphasis had been placed on understanding and correctly representing the background error statistics, with observation errors assumed to be uncorrelated. However, as observing systems become more complex, we are able to observe more complex variables from which meteorological information can be derived. As we need to assimilate observations with higher spatial and temporal frequencies and try to use observations previously deemed too difficult to assimilate, it is essential that the observation error statistics are also well estimated and correctly utilized in the assimilation scheme. This means moving away from the assumption of diagonal error covariance matrices that neglect observation error correlations. Instead, it will be necessary to use error statistics that accurately represent the error correlations, be they spatial, temporal, or inter-channel, along with consistent error standard deviations.

1.1 Definitions, Sources, and Characteristics

In data assimilation the observation error, ε^o, not only accounts for errors inherent in the observing instrument, ε^i, but also from the fact that the observation, $\mathbf{y}^o$, must be compared to a model equivalent of the observation which has been mapped from a discrete model using a forward operator, $\mathbf{H}$; hence, all observations will be subject to some form of uncertainty. In other words,

$$\varepsilon^o = \varepsilon^i + \varepsilon^r, \tag{1}$$

where ε^r denotes the error of representation which arises from the incompatibility between model grids and observation states: for example, observations of higher resolution than the model state estimate. This term must be included in the observation error due to the fact that we aim for an analysis that provides the best initial conditions in model space. If the analysis included information from unresolved scales, such information would result in noise once the model was integrated forward in time. Each observation type will have its own specific observation error sources, though

D. J. Murphy
Environment, Water, Population and Communities, Australian Antarctic Division of Sustainability, Kingston, TAS, Australia

M. Tsutsumi
National Institute of Polar Research, Tokyo, Japan

J. Marino
Colorado Center for Astrodynamics Research, Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, USA

J. P. McCormack · I. M. Reid
ATRAD Pty Ltd, 24 Now at Heliophysics Division, NASA Headquarters, Washington, DC, USA

in general, it is possible to assign each contribution to one of four main sources (Janjic et al. 2018): the instrument error, observation operator error, preprocessing (or quality control) error and the error due to unresolved scales and processes.

The instrument error is the error related to the precision and accuracy of the measuring device; this error is independent of how the observation is used in the data assimilation process. The observation operator, preprocessing and unresolved scales errors all contribute to the error of representation and will all be dependent on the specific settings of the assimilation system. The observation operator error arises due to approximations in the forward model used to map the model variables into observation space. These errors can be as simple as those introduced when interpolating from one location to another or may be far more complex due to the approximation of observation geometries or the misspecification of parameters in the forward models (e.g., the misrepresentation of gaseous constituents in a radiative transfer model.) The preprocessing, or quality control, error arises if an observation is not accurately handled by the quality control procedure or if the observation is processed prior to the assimilation. For example, a cloudy observation may somehow pass a cloud screening process and be assimilated as a cloud-free observation, even though it is not. Distinguishing between observation operator, and preprocessing error is not always straightforward, since for our example of the cloudy observation the failure could be seen to be the observation passing quality control, or the inability of the forward operator to simulate the cloudy observation. The final general source of error is the error due to unresolved scales and processes; this error arises when the observations and the model into which they are assimilated represent different spatial and temporal scales. The sum of the instrument, preprocessing, observation operator, and unresolved scales errors provides the total observation error. It is the statistics of these errors that are required for the data assimilation system.

1.2 Error Correlation

Each of the error sources has the potential to give rise to error correlations. It has long been assumed that the instrument error would be uncorrelated and unbiased. Although this may be true for conventional observations, which directly measure model variables, it has been shown that observational errors of remotely sensed observations, such as satellite observations can indeed be correlated (e.g. ATMS, IASI; e.g. Stewart et al. 2009, Bormann et al. 2010, Campbell et al. 2017). It is perhaps more intuitive that the error of representation can give rise to correlated observations errors, and that it is often the dominant contributor to error correlations (Hodyss and Satterfield 2017). Any approximation in the observation operator for a given observation type will result in a correlated error since the approximation is applied when calculating each of the model equivalent observations. For example, neglecting beam broadening in the forward operator for radar observations results in correlated observation errors; the error correlation length scale increases with distance from the radar as the observation operator becomes more approximate (Waller et al. 2016b;

Waller et al. 2019). Similarly, all observations that are subject to the same errors in quality control or preprocessing will exhibit error correlations (e.g., Bormann et al. 2003; Cordoba et al. 2017). The error due to unresolved scales is also expected to be correlated due to its dependence on the state of the geophysical system. It has been shown that the scale mis-match errors exhibit larger correlations when the scales in the observations are poorly resolved by the underlying model (Waller et al. 2014a). Prior to their use in the data assimilation system the observation uncertainties, including the error correlations must be specified (Gustafsson et al. 2018). While instrument errors are normally well characterized, with estimates given from instrument manufacturers, the contribution from the error of representation is complex and much harder to quantify. Furthermore, since we never have information about the true atmospheric state, the observation error statistics cannot be calculated directly. Instead, they must be estimated in a statistical sense. How these error statistics can be estimated is discussed in detail in Sect. 2.

1.3 Operational Treatment

Once the observation uncertainties have been specified, they must be incorporated in the data assimilation system correctly for the output of the assimilation system to be optimal. The observation error covariance matrix can be defined as

$$\mathbf{R} = E\left[\varepsilon^o (\varepsilon^o)^T\right]. \tag{2}$$

Until relatively recently the observation error covariance matrices used in operational data assimilation were assumed diagonal. Although the error correlations were ignored, there were attempts to improve the assumption of uncorrelated error by means of observation thinning or observation averaging (Daley 1991; Whitaker et al. 2008). If correlated errors were still expected to exist, then they were often compensated via the technique of variance inflation (Lahoz and Ménard 2010). However, since theoretical work has shown that even the inclusion of approximate observation error correlations can have a benefit (Healy and White 2005; Stewart et al. 2013), there has been a recent effort to improve the treatment of observation uncertainties in operational data assimilation. A review of the estimation and treatment of observation uncertainties in operational practice is given in Sect. 2. To date, the inclusion of observation error correlations in data assimilation has been shown to improve the analysis and the skill of operational forecasts (Weston et al. 2014; Bormann et al. 2016; Campbell et al. 2017). Furthermore, the accurate uncertainty specification can improve the representation of small scale features which is of particular importance for convective scale forecasting (Rainwater et al. 2015; Fowler et al. 2018; Simonin et al. 2019; Bédard and Buehner 2020).

1.4 Outlook

This chapter provides an overview of observation error statistics in data assimilation and how they may be practically estimated and included in an operational setting. Section 2 reviews methods that may be used to estimate observation error statistics. We highlight the assumptions of each of the methods along with their benefits and shortcomings. We provide an overview of the use of these methods to estimate error statistics for atmospheric observations; in addition, we discuss how the estimated error correlations are used in current operational practice. In Sect. 3 a practical application of observation uncertainty estimation is illustrated with Meteor Radar Assimilation. We describe the Meteor Radar observations and assimilation and provide the estimated error statistics. We then systematically investigate the estimated uncertainties by considering how the estimates vary by station and how inflating the background, **B**, and observation, **R**, error covariance matrices change the estimated statistics. The observation influence metrics and temporal correlations are also calculated. Finally, we summarize our findings and conclude in Sect. 4.

2 Diagnosing Observation Error Including Error of Representation

In this section, we review current statistical methods to recover observation error uncertainties as well as their underlying assumptions. We also provide an overview of current operational practice. Since this chapter is aimed at providing a practical methodology for operational systems, we focus mainly on the methods which have been implemented operationally. However, we briefly review more recent methods that have not been widely used in operations, either due to nascence, algorithm complexity, or computational cost.

2.1 Innovation Based Estimation Methods

The method introduced by Desroziers et al. (2005) has been used widely at operational centers, due to both the algorithm's simplicity and its computational efficiency. The method uses standard outputs of data assimilation (innovations, increments and analysis residuals) as inputs and provides estimates of the observation error covariance, as well as observation space estimates of both the analysis and background error covariance. First, we define the analysis residual, $\mathbf{d}_a^o$, as the difference between the observation and the analysis in observation space,

$$\mathbf{d}_a^o = \mathbf{y}^o - \mathbf{H}\mathbf{x}^a. \tag{3}$$

The innovation, $\mathbf{d}_a^o$, and increment, $\mathbf{d}_b^a$, can be defined similarly as the difference between the observation and the forecast in observation space and the difference between the analysis and the background in observation space, respectively,

$$\mathbf{d}_b^o = \mathbf{y}^o - \mathbf{H}\mathbf{x}^b, \tag{4}$$

$$\mathbf{d}_b^a = \mathbf{H}\mathbf{x}^a - \mathbf{H}\mathbf{x}^b. \tag{5}$$

This Desroziers method relies on the classical innovation relationship: the expected value of the outer product of the innovation vectors is equal to the sum of the observation space representation of the background error covariance matrix and the observation error covariance matrix. In other words,

$$E[\mathbf{d}_b^o(\mathbf{d}_b^o)^T] = \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}. \tag{6}$$

This relationship assumes linearity of the expectation operator and mutually uncorrelated observation and background errors. The Desroziers methods recovers the observation error covariance by taking the expected value of the outer product of the analysis residual ($\mathbf{d}_a^o$, observation minus analysis) and innovation ($\mathbf{d}_b^o$, observation minus forecast),

$$E[\mathbf{d}_a^o(\mathbf{d}_b^o)^T] = \mathbf{R}. \tag{7}$$

The background error covariance matrix is similarly obtained by taking the expected value of the outer product of the analysis increment and the innovation,

$$E[\mathbf{d}_b^a(\mathbf{d}_b^o)^T] = \mathbf{H}\mathbf{B}\mathbf{H}^T. \tag{8}$$

Although this method is easy to implement, it can result in erroneous estimates when the data assimilation system used to make the analysis uses inaccurately specified background and observation covariance matrices (e.g. Ménard 2016; Waller et al. 2016a; Waller et al. 2017, amongst others). The Desroziers method is often thought of as a "consistency check", as the method should return the same value of observation variance used in the data assimilation system if all covariance matrices are correctly specified. Therefore, this method will produce errors due to inaccuracies in the prescribed error covariance matrices and a computationally costly iterative procedure may be required. Waller et al. (2016a) showed that, although the output provided by this method is subject to prescribed background and observation covariance matrices, a useful solution can often be obtained in a single iteration even when iterative techniques cannot be expected to converge.

Another commonly used method to diagnose observation errors is that of Hollingsworth and Lönnberg (1986). This method bins innovation statistics by separation distance and forms a histogram of innovation covariance. This method also relies on the innovation relationship in (6) and assumes that observation errors are

spatially uncorrelated beyond zero separation distance. Under this assumption, the relative contributions of observation and background error covariance can be separated by fitting a correlation function and extrapolating to zero separation. This method has been extended, with additional assumptions, to provide estimates of inter-channel correlation for satellite radiances (e.g. Bormann et al. 2010, amongst others). Both the Desroziers algorithm and the Hollingsworth and Lönnberg algorithm are computationally efficient, enabling on-line or off-line variants.

The innovations used in these diagnostics can also be used to provide information on the observation impact. The actual and theoretical observation impact diagnostics are described in Fowler et al. (2020) and can be computed as,

$$OI^{TH} = trace\left(E\left[\hat{d}_b^a\left(\hat{d}_a^o\right)^T\right]\left(E\left[\hat{d}_b^o\left(\hat{d}_a^o\right)^T\right]\right)^{-1}\right), \tag{9}$$

$$OI^{AC} = trace\left(E\left[\hat{d}_a^o\left(\hat{d}_a^o\right)^T\right]\right), \tag{10}$$

where OI^{TH} and OI^{AC} represent the theoretical and actual observation impact respectively, and $\hat{d} = \mathbf{R}^{-1/2}\mathbf{d}$, here we use $\mathbf{d}$ generically and it can refer to either an increment or a residual These two measures provide us with an additional metric to assess the optimality of the data assimilation system and the specification of the observation and background error covariance matrices. To simplify the calculation and reduce the sensitivity of the estimate to sampling noise it is possible to assume that the errors are uncorrelated and calculate a 'diagonal only' approximation of Eq. (6).

2.2 Ensemble Methods

Ensemble DA systems rely on ensembles of short-term background forecasts to provide the background error covariance. Variants of the Desroziers method that also make use of the ensemble of background forecasts have been recently implemented. Miyoshi et al. (2013) and Waller et al. (2014b) introduced time variant inline estimates using the Desroziers method within ensemble transform Kalman filters. Karspeck (2016) suggested using the innovation relationship defined in (6), and estimating the observation error covariance by removing an ensemble based estimate of the background error covariance, after suitably tuning the ensemble. Satterfield et al. (2017) explored using the ensemble variance as a flow dependent predictor of observation error variance.

2.3 Representation Error

Hodyss and Nichols (2015) and Hodyss and Satterfield (2017) showed that innovation based methods would include a contribution from representation error when the observations were of higher resolution than the model state. However, it is often difficult to determine the contributions of errors stemming from incompatibility between model grids and observations. Several data driven methods have been introduced that aim to do just that. Oke and Sakov (2008) and Forget and Wunsch (2007) averaged observations to model resolution. After interpolating to observation location and subtracting the raw observations the associated error of representation could be estimated. Hodyss and Satterfield (2017) performed similar computations using high resolution model fields and spectrally truncated variants. Such data driven methods rely on assumptions of ergodicity and produce static estimates which are averaged over a specified time period. However, since representation error depends on the background flow, it is expected to show temporal variation. Frehlich (2006) used estimates of local turbulence to introduce a flow-dependent component of observation error. As mentioned in the previous subsection, Satterfield et al. (2017) used the ensemble variance as a predictor of observation error variance, under the assumption that the portion of observation error variance that varied as a function of the ensemble spread was due to representation error.

2.4 Sensitivity Diagnostics

A number of forecast sensitivity based methods have recently emerged and have been used to guide tuning of the observation and background error covariance matrices. Such sensitivity metrics are based on an error norm in a short term forecast and can be implemented in either variational or ensemble frameworks. Daescu and Todling (2010) explored using adjoint based sensitivity tools to estimate the forecast sensitivity to observation and background error covariance parameters. Daescu and Langland (2013) extended this work to the Naval Research Laboratory Atmospheric Variational Data Assimilation System–Accelerated Representer (NAVDAS-AR) system. Techniques to evaluate the observation sensitivity without an adjoint model in an Ensemble Kalman Filter framework were discussed in Liu and Kalnay (2008) and Hotta et al. (2017).

2.5 Other Methods

There are a variety of methods that have not been readily adapted to the operational framework, due to either newness or algorithm complexity. Many of these recent methods have been applied within an ensemble data assimilation framework.

A recent review paper by Tandeo et al. (2020) provides a detailed overview of some of these methods defining the key families of methods as moment based, using lag-innovations, or Maximum likelihood, which includes Bayesian inference and Maximization of the total innovation likelihood.

As discussed in Tandeo et al. (2020), the use of "lag", or current and past, innovations for covariance estimation was introduced by Mehra (1970) and Bélanger (1974). The key idea behind using lag innovations is that the innovation based diagnostics can be extended to the case of model error. Recent work using lag innovations includes Berry and Sauer (2013), Harlim et al. (2014), Zhen and Harlim (2015), amongst others. Various lag times have been explored to optimize the estimation within a toy model framework. As with online variants of the Desroziers algorithms, such methods also usually involve some level of temporal smoothing to avoid estimates changing too rapidly over time.

Maximum likelihood based methods are aimed at deriving parameters of observation error covariance matrices that will maximize the observation likelihood. Such methods include Bayesian inference or Maximization of the total likelihood. Bayesian methods, such as the one proposed by Ueno and Nakamura (2016), allow one to incorporate knowledge of the prior distribution. Typically, these parameters are estimated, through state augmentation, where the state and parameters are estimated jointly, or through Bayesian inference as a secondary filter. A recent review paper by Janjić et al. (2018) detailed state augmentation for including errors due to unresolved scales and processes and observation operator error in the Kalman filter algorithm. Additionally, one can maximize the total innovation likelihood using iterative procedures such as the Expectation-Maximization algorithm, as detailed in Dreano et al. (2017), Pulido et al. (2018) amongst others (see Tandeo et al. 2020 for a more complete discussion). Maximum likelihood approaches require assumptions about the prior and posterior distributions in order to make the problem practicable. In addition, typically the number of parameters to be estimated must be limited, which requires further assumptions about the accuracy of the prescribed parameters, which are not estimated.

2.6 Current Operational Practice

The innovation based methods discussed at the beginning of this section have proven the most popular for use in operational settings. They have been used to estimate and tune the observation error variances for a wide variety of observations including those from satellite instruments, aircraft, scatterometers and GNSS-RO (e.g. Lange and Janjić 2016; Valkonen et al. 2017; Bowler 2020). The use of innovation based methods has not been restricted to numerical weather prediction; they have been extensively used to estimate observation error variances in many other geophysical models: e.g. oceanography, hydrology, atmospheric chemistry and the carbon cycle (Hoffman et al. 2013; Pinnington et al. 2016; Waller et al. 2018; Merchant et al. 2020). However, it is only more recently that the methods have been used to provide

information on the observation error correlations and that these correlations have been utilized in operations. Below we discuss the operational use of inter-channel and spatial correlations; we do not include temporal correlations as, to date, they have received little attention and have not been incorporated in operational assimilation.

2.7 Inter-Channel Radiance Assimilation

Undoubtedly, to date, it is the inter-channel error correlations in radiances acquired from operational satellite nadir sensors that have received the most attention and hence have been successfully incorporated into operational NWP systems. The inter-channel correlations arise between neighboring channels that have similar spectral resolutions and similar sensitivities to temperature and atmospheric gases. Among the first instruments to have their error correlations characterized were the Atmospheric Infrared Sounder (AIRS) and Infrared Atmospheric Sounding Interferometer (IASI) (e.g. Garand et al. 2007; Bormann et al. 2010; Stewart et al. 2009, 2013). Results showed that water vapor sensitive channels, along with channels sensitive to the skin temperature exhibited some of the largest correlations. Due to the limited number of channels and single location of each observation profile the inclusion of these correlations in operational systems is fairly straight forward. Furthermore, the inclusion of inter-channel error correlations has been shown to improve forecast skill and hence a number of operational centers now use inter-channel error correlations for a variety of instruments including IASI (Weston et al. 2014; Bormann et al. 2016), the Cross-track Infrared Sounder (CrIS; Eresmaa et al. 2017), and the Advanced Technology Microwave Sounder (ATMS; Campbell et al. 2017). Following the success of accounting for inter-channel correlations it has become common for these uncertainties to be estimated as new observing instruments come online, such as Himawari-8 and the Chinese FengYun Microwave Temperature Sounder and Microwave Humidity Sounder (Honda et al. 2018; Wang et al. 2018).

As well as leading to increased analysis accuracy and improved forecast skill, the accurate specification of inter-channel correlations has improved the assimilation of cloudy observations. Geer and Bauer (2011) showed that the observation error variance is larger when a radiance observation is assimilated in a cloudy region in comparison to a clear sky observation. Subsequent work by Okamoto et al. (2019) and Geer (2019) shows that the inter-channel correlations are also much stronger in cloudy, compared to clear, skies suggesting that scene dependent error specification may be necessary for all-sky assimilation.

2.8 Spatial Correlations

There were some early attempts to estimate spatial observation error correlations (e.g. Bormann et al. 2003; Garand et al. 2007), though the complexity of using spatial

correlations meant they were never considered for use within the assimilation system. However, with the move towards higher resolution convection permitting data assimilation, there is a need to assimilate dense data with likely spatial correlations and hence the interest in spatial correlations has been reignited. Spatial error correlations have recently been estimated for the Doppler radar wind, SEVIRI radiance, atmospheric motion vector and GNSS Zenith Total Delay observations assimilated into the Met Office's high resolution limited area model and results have shown that spatial correlations can be significant (Waller et al. 2016b, 2016c; Cordoba et al. 2017; Bennitt et al. 2017). GNSS Zenith Total Delay observation uncertainties have also been estimated for observations assimilated into the central Europe AROME model (Mile et al. 2019); however, it has been shown that it is the temporal, rather than spatial, error correlations that are likely to be the most important for this observation type (Bennitt et al. 2017; Macpherson and Laroche 2019). In some cases, the estimated statistics have been used to inform spatial thinning distances to ensure that the assumption of uncorrelated observation errors is accurate (Cotton et al. 2018; Mile et al. 2019). However, if the correlations can be accounted for in the assimilation system then an understanding of the uncertainties allows denser observations to be assimilated. The work of Simonin et al. (2019) provided a strategy for incorporating spatially correlated observation errors into an operational data assimilation system. The technique was then applied to Doppler radar radial wind observations and results showed that the assimilation of dense data, whilst accounting for their correlated errors, leads to more small-scale information in the analysis and has a positive impact on the forecast skill. As a result, spatially correlated observation errors for Doppler radial winds are explicitly accounted for in the Met Office limited area model.

3 Practical Application: Application to Meteor Radar Assimilation

In this section we detail how in an operational system, inadequacies in prescribed error parameters can lead to suboptimal assimilation of observations. We outline a practical methodology to use diagnostic estimates to modify error parameters in an operational framework, employing the use of observation impact described in Sect. 2.1. Our ultimate goal is to improve the assimilation of meteor radar observations for prediction in the 70–100 km altitude range using a new high altitude global data assimilation framework.

3.1 Meteor Radar Observations

At upper atmospheric levels, where observations are sparse compared to the troposphere, meteor radars could potentially provide a valuable observational data set for a model correction. Since ionization trails produced by ablating meteors drift with local neutral winds at ~70–100 km altitude due to high plasma-neutral collision frequencies, radar tracking of meteor trails can be used to infer winds at these altitudes, year-round, both day and night. This technique has been refined over many years and extensively validated, to the point where meteor winds are now a widely accepted independent validation standard for other emerging wind measurement techniques at these altitudes (e.g., Jones et al. 2003; Wilhelm et al. 2017). For these reasons, new high-altitude reanalysis systems have used them as a validation standard for analyzed winds at these altitudes (e.g., McCormack et al. 2017; Eckermann et al. 2018; Stober et al. 2019) and are just beginning to explore their value for direct assimilation into research and ultimately into operational systems. Here we explore winds acquired from a network of radars, shown in Fig. 1 and initial efforts to assimilate these observations into a high-altitude version of NAVGEM. For the experiments that follow, we assimilated all available data from the stations shown in Fig. 1 except Davis and Buckland Park, which were inadvertently excluded owing to an internal error. Additionally, we note that there were data gaps for McMurdo and Cariri over the periods considered.

Several data processing algorithms are used to derive these winds. The Buckland Park, Beijing, Davis Station, King Sejong Island, Kunming, Mengcheng, Mohe, Svalbard and Wuhan sites use the method described in Holdsworth et al. (2004). Briefly, given a set of meteor echo radial drift velocities (V_r) and AOA direction

Fig. 1 Ground locations of the 23 meteor radars acquiring winds in the 70–100 km range, which we assimilate using NAVGEM and analyze using DA methods discussed in the text

cosines (l,m,n), the two-dimensional wind velocity (u,v) is estimated by applying a least squares fit of the form V_r = ul + vm, with the implicit assumption that the vertical velocity w = 0. The data are grouped into height/time bins of 2 km and 1 h. A minimum of six echoes in each height/time bin are required to compute a velocity estimate. An additional 15 quality controls are then applied as described in Holdsworth et al. (2004), who also describe the same basic radar type as all of these radars except for the Buckland Park radar. This radar is described by Spargo et al. (2019) and references therein.

The method described in Fritts et al. (2012) and Iimura et al. (2015) uses a least squares fit to the measured radial velocities of meteor trails when a minimum of 57 meteors are present in each time-altitude interval. Vertical profiles of hourly zonal and meridional winds are produced by assigning wind estimates with an altitude resolution of 2 km centered at 80, 82, 84, 86, 90, 92, 94, 96, 98 and 100 km to the middle of each time-altitude interval (altitude intervals are 75–80, 80–84, 84–86, 86–88, 88–90, 90–92, 94–96, and 96–100 km). This method is used for the Eureka, Esrange, Cariri, Cachoeira Paulist, King Edward Point, McMurdo, Trondheim, Bear Lake, Ascension Island, Tierra del Fuego and Rothera station sites shown in Fig. 1.

A third method is based on an updated version of Hocking et al. (2001). Details about the algorithm are presented in Stober et al. (2018). The fitting algorithm obtains the instantaneous three-dimensional wind vector V = (u, v, w) using a constrained least squares solution, where the vertical and time derivatives of each wind vector component are assumed to be constant between adjacent time and altitude bins. The vertical wind is assumed small, a good assumption for an observation volume of 400–600 km diameter. A minimum of 5 m within each time-altitude interval is required. This method is used to produce wind retrievals, at 2 km intervals, from the Andenes, Juliusruh, and Collm sites shown in Fig. 1 and from the Canadian Meteor Orbit Radar (CMOR), located near Tavistock, Ontario, Canada.

3.2 Assimilation System

The assimilation system used in this study is a high altitude NWP system based on the operational semi-Lagrangian, semi-implicit global spectral model, part of the Navy Global Environmental Model (NAVGEM) system described in Hogan et al. (2014), but extended to a 74-level (L74) configuration with a top pressure of 6×10^{-5} hPa (~116 km altitude) and a vertical resolution of ~2 km in the stratosphere, increasing to ~3–5 km in the 70–100 km altitude range spanned by the meteor wind observations (see Fig. 3b of Eckermann et al. 2018). We will refer to the high altitude version as NAVGEM-HA (for details see Eckermann et al. 2018).

For assimilation, we implement a hybrid 4DVAR algorithm, which linearly combines a static background error covariance estimate with covariances derived from an 80-member flow-dependent ensemble of instantaneous 6-h forecasts, as described in Kuhl et al. (2013). Ensembles for this system are generated with the ensemble transform method described in McLay et al. (2010). The relative lack of

observations in the stratosphere and mesosphere compared to the troposphere emphasizes the need for the improved background covariance estimates that this hybrid approach provides. For the experiments that follow we use a weighting parameter, $\alpha = 0.25$ for the ensemble covariance and $(1 - \alpha) = 0.75$ for the static component. We note that, since this system is relatively new, as a starting point, the ensemble methodology and the static background error component used for tropospheric assimilation were simply extended to upper atmospheric levels, but the error statistics have yet to be adequately tuned.

3.3 A First Look at Error Estimates

We applied the Desroziers diagnostics, described in Sect. 2.1, to the output of a cycling hybrid 4DVAR DA run (as described in the previous section) using a T119L74 version of NAVGEM. For this system, the Desroziers estimates of the observation and observation spaced background error standard deviation profiles are important for several reasons. First, the background error statistics of the system have not been fully tuned. Second, we know we will have additional components to the observation error statistics due to the observation operator, which interpolates the model background to observation space. In the horizontal, the observation volume for meteor radars is about 500 km in diameter. The smallest resolvable scales in the model are about 4 grid points. The spatial averaging of the meteor radars may remove the smallest scales that are resolved with NAVGEM T119 (~90 km), depending on diffusion (see, e.g., Fig. 13a of Eckermann et al. 2018 and Fig. 2 of Stober et al. 2019). The vertical resolution of the meteor radar observations is higher than that of the model. Additionally, variations in retrieval algorithms and vertical-temporal intervals will add components to the observation interval. For these reasons, simply assigning an instrument error is insufficient as such a value would be too low for practical use in an assimilation system. For the initial assimilation experiments, a vertically varying fraction of the background error variance was used as a proxy for observation variance. In this system, observation errors are currently treated as uncorrelated, which can also lead to the need for further inflation.

Prior to the Desroziers computation, a mean bias, computed as a function of pressure level, is removed from each innovation, increment and analysis residual. For our statistics, we include all meteor radar wind observations, but limit the data to only those that were assimilated and to measurements provided on the hour (half-hourly measurements are ignored). It is important to note that the innovation check, which acts to filter observations with innovations larger than a threshold value, was turned off for the initial control experiments. Figure 2 shows the estimated (solid lines) and prescribed (dashed lines) background and observation error standard deviations, in observation space, based on zonal wind observations and computed using Eqs. (7) and (8). Our findings indicate that the current prescribed observation error standard deviation is too low and should be increased (red curves) and that the current background error standard deviation should be reduced (blue curves). The static (dashed

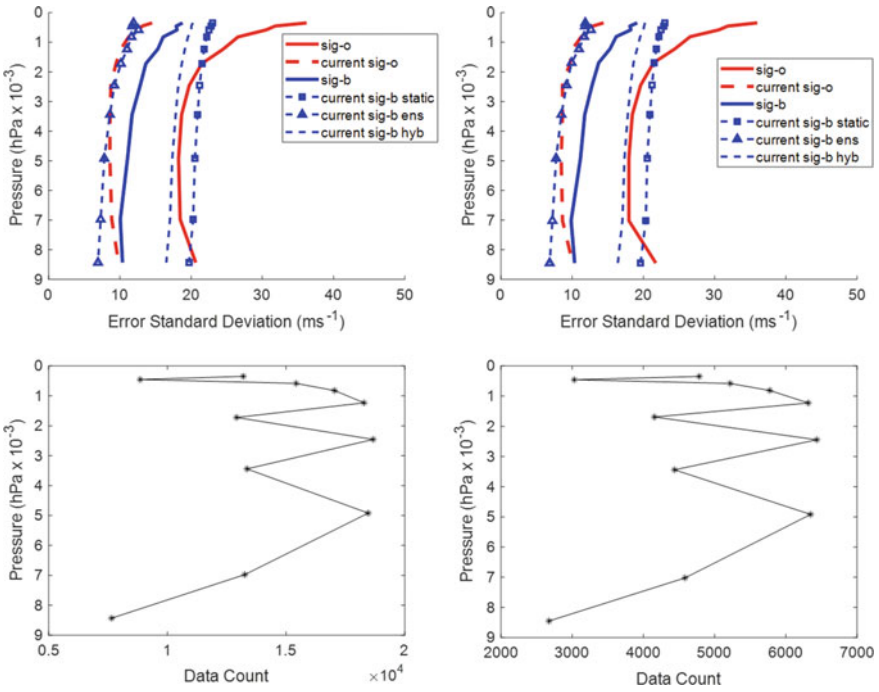

Fig. 2 Vertical profiles of the Desroziers diagnostic (top row) for SON 2017 (left) and October (2017) (right) for observations of zonal wind. The Desroziers estimates of observation (red) and background (blue) error variances are shown in solid lines, while the prescribed hybrid and observation error are shown as dashed lines. The static (dashed line with squares) and ensemble components (dashed line with triangles) of the background error variance are also shown. Data counts are shown on the bottom row

line with squares) and flow dependent (dashed line with triangles) components of the hybrid background error covariance are shown separately. While the static portion of the covariance needs to be reduced, the ensemble is under dispersive, which ultimately is acting to produce an effective hybrid weighting parameter which is much lower than the prescribed 0.25. Overall these results indicate that the analysis is drawing too closely to the meteor radar wind observations.

Focusing on zonal winds we looked at a seasonal comparison (not shown) between December–February (DJF) 2017–2018 and September–November (SON 2017). The results indicate only slight seasonal variation, so as a first approximation we can assume that the statistics will not have large seasonal variations. Since the error in the Desroziers based estimation can originate from a variety of sources, applying a standard error is not informative. Instead, we simply compare a one month estimation to a seasonal estimation to confirm stable statistics (Fig. 2). We find that the estimates for a single month (October 2017, right hand side of figure) are consistent with the other time periods (in particular SON 2017 shown on left hand side of

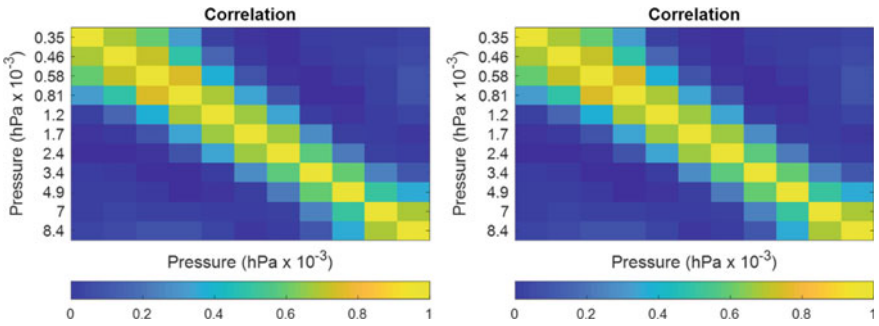

Fig. 3 Desroziers correlation matrix for SON 2017 (left) and October (2017) (right) for observations of zonal wind

figure), indicating that a single month provides sufficient data for stable estimation parameters.

In addition to investigating estimates of error variance, we also look at estimates of the full matrix. Currently, observation errors for meteor radars are treated as uncorrelated (diagonal matrix). Figure 3 shows that the correlation structure, by vertical level (x and y axes), has a strong block diagonal component, indicating the correlation between neighboring levels. This correlation may arise from the radar vertical range gate being wider than the 2 km resolution of the wind retrieval or from temporal smoothing resulting from the retrieval algorithm assigning wind estimates to the middle of each time-altitude interval. Implementing a block diagonal component in the assimilation of these observations would potentially offer benefit. Waller et al. (2016a) showed that inconsistencies in the prescribed covariance matrices, such as ignoring correlations or over/under specified error variances, can lead to inaccuracies in the Desroziers diagnostic. We proceed by comparing several experiments where corrections are made to the background error variances, the observation error variances and finally both the background and observation error variances to attempt to disentangle the contributions of these components.

3.4 Differences By Station

Many differences between stations may lead to differing observation errors. Such differences include differences in instrumentation, environment dependent instrument error, operator error, as well as differences in the wind fitting algorithms. Additionally, the prescribed error covariance matrices used in the data assimilation algorithm may be more or less optimal at different locations. As a control group, we use meteor winds from 3 radar stations shown in Fig. 1, Andenes, Juliusruh and Collm (red curves in the left panel of Fig. 4), the same stations used in McCormack et al. (2017), but limited to a single wind fitting algorithm to allow for homogeneous statistics. Figure 4 shows the Desroziers estimates of observation error variance by

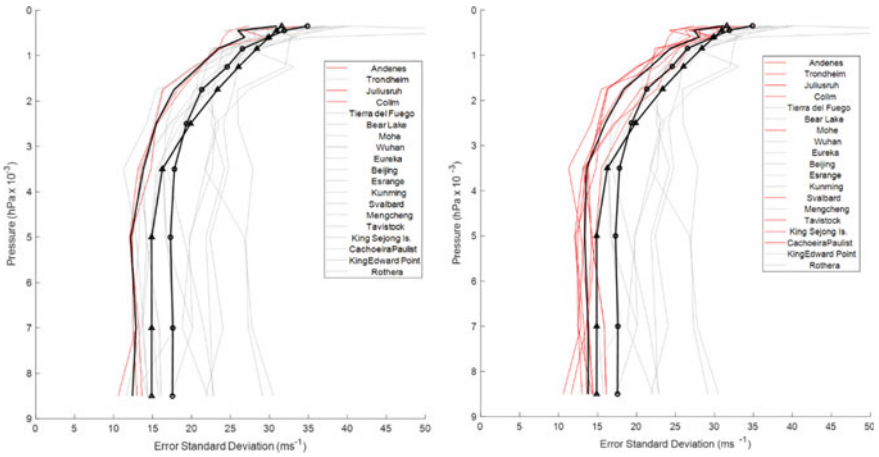

Fig. 4 Desroziers estimates of observation error standard deviation by station (red and grey lines) for SON 2017 for observations of zonal wind. The solid black lines shows the mean of the specific stations highlighted in red. The revised observation error (solid black line with triangles) and the mean over all stations (solid black line with circles) are also shown

the station. All stations with enough assimilated observations in the SON 2017 period to provide an error variance estimate at all atmospheric levels are plotted. Although in some cases/levels the data counts are low enough that the Desroziers estimate may not be reliable, it is useful to define a set of stations over which we find the estimates to be suitably homogeneous to define a subgrouping (red lines in the right panel of Fig. 4). The solid black curve in Fig. 4 shows the mean of all stations highlighted in red. The mean over all stations plotted is also shown (solid black line with circles). To revise the observation error variances, we created a functional fit based on pressure level to the Desroziers estimates. The function was assumed quadratic above the 0.004 hPa level (roughly 84 km in altitude), and constant below. The functional fit is shown by the solid black line with triangles in Fig. 4, for simplicity we will refer to this revised estimate as the 'quadratic model'. It is worth noting that, although there was some station to station variation, each station considered showed a similar block diagonal correlation structure as shown in Fig. 3 (not shown for individual stations).

3.5 Experiments with Inflated Ensemble Variance and Inflated Observation Error Variance

It is useful to explore the impact that the inaccuracies in the prescribed background error variances have on the Desroziers estimates. To this end, we performed a set of three additional experiments: the observation error variance is adjusted to match the quadratic fit shown in Fig. 4, the ensemble sample error covariance is inflated globally and multiplicatively by a factor of two, and combined observation and

ensemble covariance inflation. For all experiments with inflated observation error, all meteor radar observations are impacted. These three additional experiments are compared with the "control" experiment, which was discussed in the previous section. Since the innovation check is disabled, these adjustments do not impact the number of rejected observations. We also note that adding the block diagonal structure to **R** is left to future experiments. In these experiments, although a hybrid-4DVAR DA system is employed, we do not attempt to optimize the relative static and flow dependent contributions or their respective correlations.

Figure 5 shows the prescribed and estimated observation and background error variance for each of the four experiments considered for the period of October 2017. The control experiment shows an under dispersive ensemble; however, hybridization compensates, to some extent, for the static components which overestimate the error variance. When the ensemble error variance is inflated (Fig. 5, Panel 5b), the hybrid variance is now even more of an overestimate. The estimated background error variances increase with greater overestimation in the prescribed values and the estimates of the observation error variances are reduced to compensate. When we modify the observation error variance according to the quadratic model (Fig. 5, Panel 5c), we see that the prescribed values are still smaller than the Desroziers estimate, although they

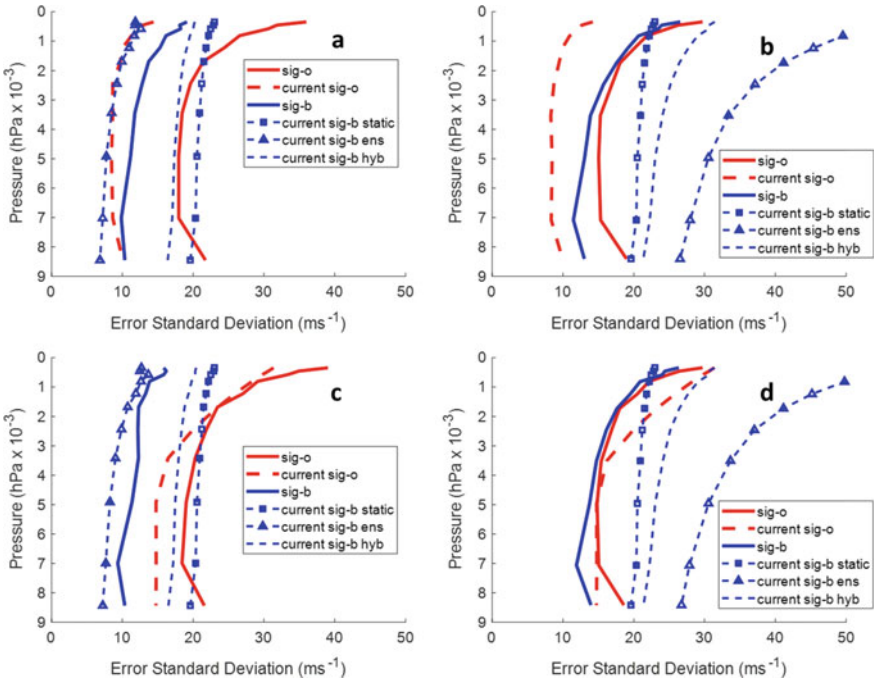

Fig. 5 Same presentation as in the top row of Fig. 2, but showing results for October 2017 for the control experiment (**a**), inflated ensemble (**b**), modified observation error (**c**) and modified observation error with ensemble inflation (**d**)

are an improved approximation. The Desroziers estimate remains fairly stable with only the observation error adjustment. However, since the prescribed background variances are too high, we expect the Desroziers estimate to be an underestimate. We see that when both the ensemble inflation and quadratic observation error model are used (Fig. 5, Panel 5d), the observation error is comparable to the Desroziers estimate. Since the impact of underestimating the observation error and overestimating the background error both lead to the analysis drawing too close to the observation, we expect the observation impact to be too large. We plan to implement a more reasonable ensemble inflation and appropriately handle inflation to the static component through hybrid weighting factors. After the adjustment to the background error covariance, we will be better able to adjust the observation error variance values based on station subgroupings.

3.6 Observation Impact

Figure 6 shows the results of the observation impact metrics (Eqs. 9 and 10) for the four experiments considered. In the control experiment the theoretical observation impact has (unitless) values ~0.1 with highest values in the upper levels, where there are very few observations. We note that the values ~0.1 are fairly typical and these values should not be interpreted as a measure of the quality of the observation, but rather the relative values of theoretical and actual observation impact that can, in turn, provide information on how these new observations can best be exploited within an assimilation system. In an optimal system, the actual and theoretical observation impacts would be in agreement. The diagonal only computation is shown in solid black line with triangles and indicates that the observation impact is reduced when off diagonal components are considered. For the control experiment, the actual observation impact is much higher, with the greatest values at upper levels. These findings are in agreement with Fig. 5, which shows that the observation error is underestimated, more so at upper levels. The overestimation of the hybrid error covariance adds to the issue of the actual impact being higher than the theoretical value for these observations. When the observation error is increased following the quadratic model (Fig. 6, Panels c and d), the actual impact is significantly reduced and the theoretical slightly decreased, likely due to an improved prescribed observation error. In the experiments which inflate the ensemble (Fig. 6, b and d), the actual impact is reduced, but this originates from the reduction of observation error due to greater overestimation in the hybrid background (shown in Fig. 5, panels b and d). The ratio of theoretical to actual impacts is much improved when the quadratic observation error model is used. Although the ensemble inflation also acts to improve the ratio, we find that improvement is stemming from errors in the prescribed variances.

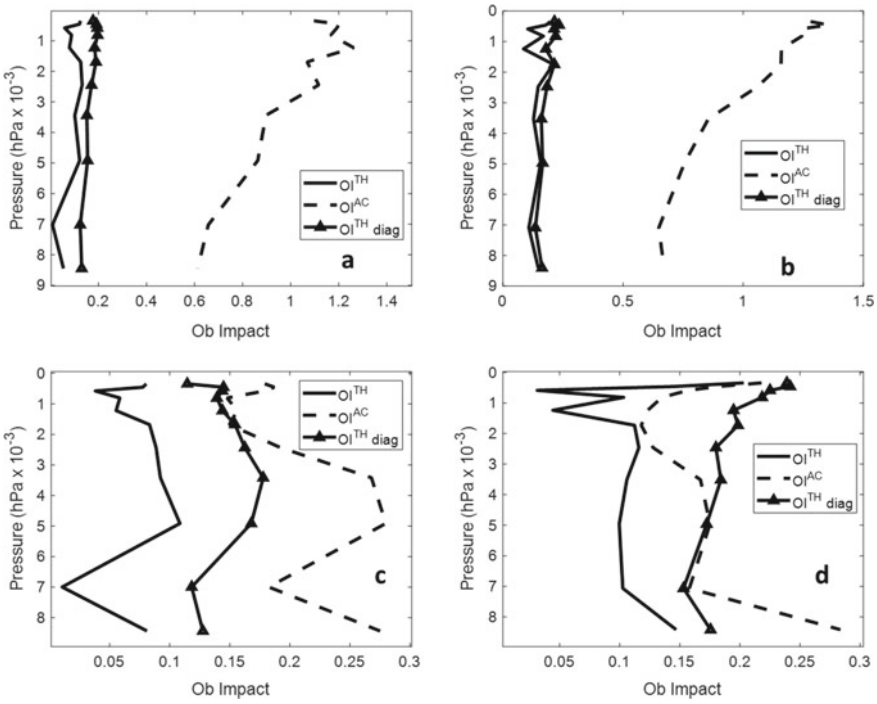

Fig. 6 Theoretical (solid black line) and actual (dashed black line) observation impact. The diagonal only estimates of the theoretical observation impact are also shown (solid black line with triangles). Results are shown for October 2017 for the control experiment (**a**), inflated ensemble (**b**), modified observation error (**c**) and modified observation error with ensemble inflation (**d**)

3.7 Root-Mean-Squared Error (RMSE)

Up to this point, we have focused on innovation-based diagnostics. Now we want to consider how well our innovation based covariance estimates and observation impact project onto forecast skill. To achieve this, we compare three experiments: the observation error variance increased following the quadratic model, inflated ensemble with original observation error variance and both increased observation error and inflated ensemble. Our findings thus far have indicated that the analysis is drawing too close to the observations, due to both the prescribed observation error variance being too low and the prescribed hybrid background error variance being too high. Our hypotheses are that (1) the quadratic observation error model will improve upon the originally prescribed error variances and (2) ensemble inflation will result in degradation due to a further increase in the hybrid variance pushing the analysis too close to the observations. We note that in the experiments which use ensemble inflation, the originally under variant ensemble becomes over variant at analysis time. The over variant analysis ensemble may act to optimize spread at longer lead times, but combined with an over-prescribed static component, further leads to the issue

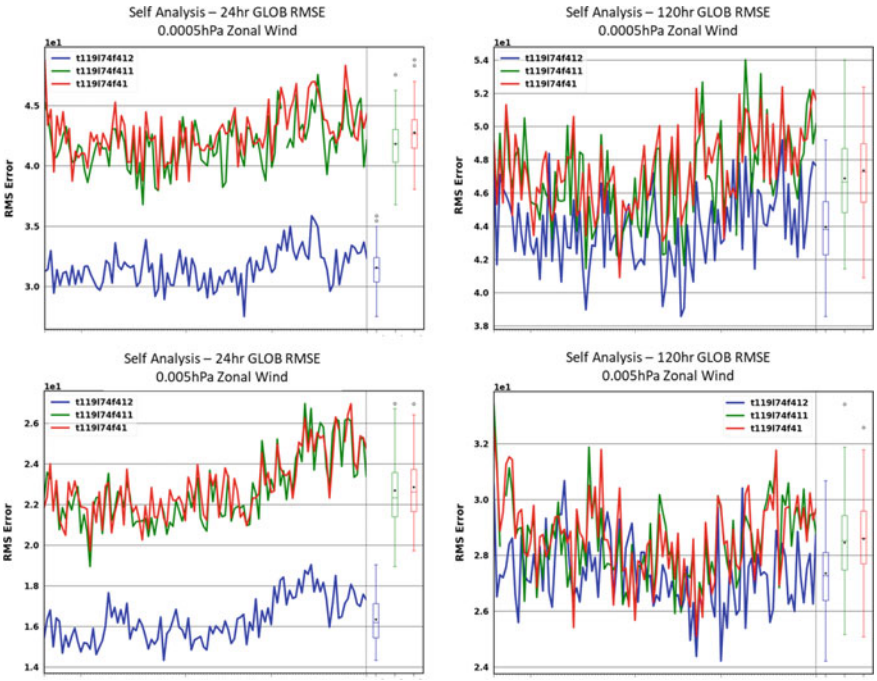

Fig. 7 Global RMSE computed for zonal wind (m s^{-1}) with forecast lead times of 24-h (left column) and 120-h (right column). The time series shows valid dates between 00Z 09 October 2017 and 00Z 30 November 2017. Plots are shown at 0.0005 hPa (top) and 0.005 hPa (bottom). Results are shown for experiments using the quadratic observation error model (blue line), ensemble inflation (red line) and both ensemble inflation and the quadratic error model (green line). The box and whisker plots to the right of each figure show the summary statistics over the time period, where the box represents the interquartile range (IQR)

of the analysis drawing to close to the observations. Figure 7 shows a time series of global RMSE computed for zonal wind with forecast lead times of 24 and 120-h. The time series shows valid dates between 00Z 09 October 2017 and 00Z 30 November 2017. Plots are shown at 0.0005 hPa and 0.005 hPa (we note that the highest and lowest levels in which meteor radar observations were assimilated were 0.00018 hPa and 0.0109 hPa respectively.) For verification we use self-analysis. For both the 24 and 120-h lead times, all levels we investigated showed significantly lower RMSE for the experiment without ensemble inflation. The experiments using the quadratic observation error variance model either showed improvement over the experiment with the originally prescribed observation error variance or no significant difference. These results were similar for meridional wind (not shown).

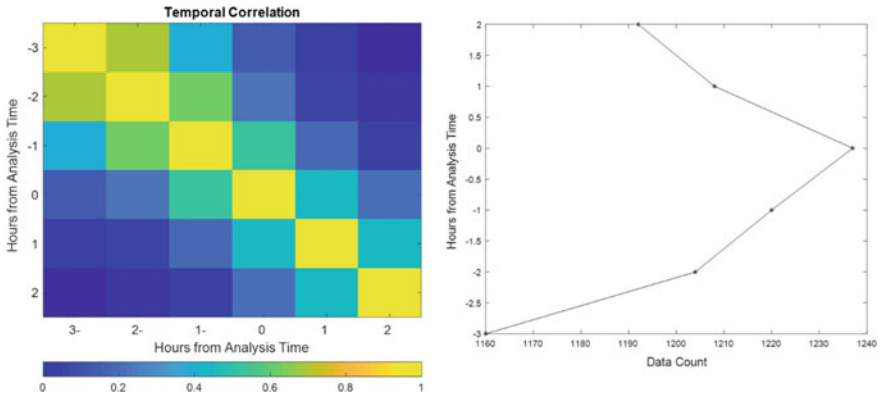

Fig. 8 Desroziers temporal correlation matrix for SON 2017 (left) and the corresponding data count (right) for observations of zonal wind for the stations highlighted in red in Fig. 4

3.8 Temporal Correlation

The previous section illustrated how the theoretical observation impact could be reduced when off diagonal terms were considered. In section d (Fig. 3) we explored the block diagonal structure of the pressure level based correlation matrix. We now turn our attention to temporal correlations, which are typically overlooked in data assimilation.

To apply the Desroziers algorithm to temporal bins, we define a subset of stations which were shown to have consistent error statistics (red curves on the right panel in Fig. 4). These ten stations were grouped together and innovation statistics were binned as a function of time within the 6-h data assimilation window, with the analysis centered in the window (e.g. 0-h). Figure 8 shows the temporal correlation structure by the hour. Since this wind fitting algorithm operated over temporal bins, it is not surprising that the temporal correlation also shows a block diagonal structure, with a correlation between neighboring hours. For data such as these, accounting for the temporal correlation within the observation error covariance matrix may prove beneficial.

4 Discussion and Conclusions

In this chapter, we have provided an overview of the methods used to estimate observation error statistics in operational data assimilation and discussed their underlying assumptions. We demonstrated the application of one of these methods to the assimilation of high-altitude (70–100 km) meteor radar wind observations, which provide important observational information at atmospheric levels where observations are sparse compared to the troposphere. To achieve this, meteor radar observations were

assimilated using a hybrid-4DVAR scheme recently implemented in NAVGEM. We systematically investigated the estimated errors by considering how the estimates varied by the station and how inflating the background, and observation error covariance matrices changed estimated statistics. Additionally, we explored vertical and temporal correlations in the observation errors. The recently developed observation influence metric was used to quantify the observation impact.

Our main findings indicated that the analysis was drawing too close to the observations. The issue was determined to be due to an under dispersive ensemble as well as under estimated observation error statistics. No strong seasonal variation was seen in the Desroziers diagnostic results and one month of data was found to provide stable reproducible statistics. The Desroziers estimates indicated block diagonal correlation structures between neighboring vertical levels and temporally between adjacent observation times. The Desroziers estimates indicated variability between stations, which could be attributed to include differences in instrumentation, environment dependent instrument error, operator error, as well as differences in the wind fitting algorithms. However, this diagnostic also indicated a set of stations over which the statistics were suitably similar to be treated homogeneously. Additionally, the block diagonal vertical correlation structure was consistent over all stations considered.

In future work we will implement a block diagonal correlation structure within the observation error covariance inflation and use the information provided from these statistics to guide ensemble inflation and to form a more appropriate static component of the background error covariance upper atmospheric assimilation. We also plan to use these diagnostics to guide optimal hybridization, following Satterfield et al (2018).

Acknowledgements NRL components of this research were supported by the Space Environment Exploitation (SEE) program of DARPA's Defense Sciences Office and by the Office of Naval Research through the NRL base 6.1 and 6.2 programs.

Funding Davis meteor radar was funded by Australian Antarctic Science project number 4445 and Syowa MF radar was funded by the National Institute of Polar Research (NIRP); Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (JSPS) KAKENHI under grant 17H02969. McMurdo meteor radar was supported by Scott E. Palo and Jeffrey M. Forbes and funded by the Office of Polar Programs of the National Science Foundation, Award #1543446. Meteor radars at Esrange, Bear Lake, King Edward Point and Rothera were funded by National Environment Research Council grant NE/R001391/1, and supported by NE/R001235/1 for King Edward Point and Rothera, and by Michael Taylor for Bear Lake. Carriri and Cachoeira Paulista meteor radars were supported by R. A. Buriti, I. Paulino, P. P. Batista, C. G Targon, and V. F. Andorioli, and funded by Fundação de Amparo àpesquisa do estado de São Paulo under 00/9510-1 and Conselho Nacional de Desenvolvimento Cientifico e Tecnológic (CNPQ) under grant PRONEX 76.97.1079.00. Buckland Park meteor radar was supported by Iain Reid and supported by the University of Adelaide and ATRAD Pty Ltd. King Sejong Island meteor radar was supported by Yongha Kim and Jeong-Han Kim, Korea Polar Research Institute (KOPRI). Eureka meteor radar and Saskatoon MF radar were supported by Alan Manson. Svalbard meteor radar was funded by University of Tromsø and NIPR. Andenes and Juliusruh meteor radars were supported by Jorge Chau and Collm meteor radar was supported by Deutsches Forchungsgemeinschaft (DFG), grant JA 836/38-1 (NOSTHEM). The operation of SAAMER in Tierra del Fuego is support by Diego Janches and NASA's SSO program and NESC assessment TI-17-01204. Trondheim meteor radar was supported by Robert Hibbins, Research Council of Norway/CoE under contract 223252/F50.

Mohe, Beijing, and Wuhan meteor radars were supported by You Yu, and funded by Solar-Terrestrial Environment Research Network (STERN) of Chinese Academy of Sciences and Chinese Meridian Project (CMP), with data archives in the Geophysics Center, National Earth System Science Data Center at Beijing National Observatory of Space Environment (BNOSE) and in the CMP data center. Mengcheng meteor radar wassupported by Xianghui Xue, and National Natural Science Foundation of China (41904135): B-type Strategic Priority of CAS Grant XDB41000000 and National Space Science Data Center, National Science & Technology Infrastructure of China.Kunming meteor radar was supported by National Key Laboratory of Electromagnetic Environment, China Research Institute of Radiowave Propagation.

References

Bédard, J, Buehne M (2020) A practical assimilation approach to extract smaller-scale information from observations with spatially correlated errors: an idealized study. Q J R Meteorol Soc 146 (726):468–482

Bélanger PR (1974) Estimation of noise covariance matrices for a linear time-varying stochastic process. Automatica 10(3):267–275

Bennitt GV, Johnson HR, Weston PP, Jones J, Pottiaux E (2017) An assessment of ground-based GNSS zenith total delay observation errors and their correlations using the Met Office UKV model. Q J R Meteorol Soc 143(707):2436–2447

Berry T, Sauer T (2013) Adaptive ensemble Kalman filtering of non-linear systems. Tellus, Ser A: Dyn Meteorol Ocean 65(20331):1–16

Bormann N, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. Q J R Meteorol Soc 136(649), 1036–1050

Bormann N, Bonavita M, Dragani R, Eresmaa R, Matricardi M, McNally A (2016) Enhancing the impact of IASI observations through an updated observation-error covariance matrix. Q J R Meteorol Soc 142(697):1767–1780

Bormann N, Collard A, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to airs and IASI data. Q J R Meteorol Soc 136(649):1051–1063

Bormann N, Saarinen S, Kelly G, Thépaut J-N (2003) The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. Mon Weather Rev 131(4):706–718

Bowler NE (2020) Revised GNSS-RO observation uncertainties in the Met Office NWP system. Q J R Meteorol Soc

Campbell WF, Satterfield EA, Ruston B, Baker NL (2017) Accounting for correlated observation error in a dual-formulation 4D variational data assimilation system. Mon Weather Rev 145(3):1019–1032

Cordoba M, Dance S, Kelly G, Nichols N, Waller J (2017) Diagnosing atmospheric motion vector observation errors for an operational high resolution data assimilation system. Q J R Meteorol Soc 143(702):333–341

Cotton J, Francis P, Heming J, Forsythe M, Reul N, Donlon C (2018) Assimilation of SMOS l-band wind speeds: impact on Met Office global NWP and tropical cyclone predictions. Q J R Meteorol Soc 144(711):614–629

Daescu DN, Todling R (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. Q J R Meteorol Soc 136:2000–2012. https://doi.org/10.1002/qj.693

Daescu DN, Langland RH (2013) Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. Q J R Meteorol Soc 139:226–241. https://doi.org/10.1002/qj.1943

Daley R (1991) Atmospheric data analysis. Cambridge University Press

Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background and analysis-error statistics in observation space. Q J R Meteorol Soc 131(613):3385–3396

Dreano D, Tandeo P, Pulido M, Chonavel T, AIt-El-Fquih B, Hoteit I (2017) Estimating model error covariances in nonlinear state-space models using Kalman smoothing and the expectation-maximisation algorithm. Q J R Meteorol Soc 143(705):1877–1885

Eckermann SD, Ma J, Hoppel KW, Kuhl DD, Allen DR, Doyle JA, Viner KC, Ruston BC, Baker NL, Swadley SD, Whitcomb TR, Reynolds CA, Xu L (2018) High-altitude (0–100 km) global atmospheric reanalysis system: Description and application to the 2014 austral winter of the Deep Propagating Gravity-Wave Experiment (DEEPWAVE). Mon Wea Rev 146:2639–2666. https://doi.org/10.1175/MWR-D-17-0386.1

Eresmaa R, Letertre-Danczak J, Lupu C, Bormann N, McNally AP (2017) The assimilation of cross-track infrared sounder radiances at ECMWF. Q J R Meteorol Soc 143(709):3177–3188

Forget G, Wunsch C (2007) Estimated global hydrographic variability. J Phys Oceanogr 37(8):1997–2008

Fowler AM, Dance SL, Waller JA (2018) On the interaction of observation and prior error correlations in data assimilation. Q J R Meteorol Soc 144(710):48–62

Fowler AM, Simonin D, Waller JA (2020) Measuring theoretical and actual observation influence in the Met Office UKV: application to Doppler radial winds. Geophys Res Lett 47. https://doi.org/10.1029/2020GL091110

Frehlich R (2006) Adaptive data assimilation including the effect of spatial variations in observation error. Quart J Roy Meteor Soc 132:1225–1257. https://doi.org/10.1256/qj.05.146

Fritts DC, Iimura H, Lieberman R, Janches D, Singer W (2012) A conjugate study of mean winds and planetary waves employing enhanced meteor radars at Rio Grande, Argentina (53.8°S) and Juliusruh, Germany (54.6°N), J Geophys Res 117(D5):D05117. https://doi.org/10.1029/2011JD016305

Garand L, Heilliette S, Buehner M (2007) Interchannel error correlation associated with AIRS radiance observations: Inference and impact in data assimilation. J Appl Meteorol Climatol 46(6):714–725

Geer AJ (2019) Correlated observation error models for assimilating all-sky infrared radiances. Atmos Meas Tech 12(7)

Geer AJ, Bauer P (2011) Observation errors in all-sky data assimilation. Q J R Meteorol Soc 137(661):2024–2037

Gustafsson N, Janjic T, Schraff C, Leuenberger D, Weissmann M, Reich H, Brousseau P, Montmerle T, Wattrelot E, Bučánek A, Mile M, Hamdi R, Lindskog M, Barkmeijer J, Dahlbom M, Macpherson B, Ballard S, Inverarity G, Carley J, Alexander C, Dowell D, Liu S, Ikuta Y, Fujita T (2018) Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. Q J R Meteorol Soc 144(713):1218–1256

Harlim J, Mahdi A, Majda AJ (2014) An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. J Comput Phys 257:782–812

Healy S, White A (2005) Use of discrete Fourier transforms in the 1d-var retrieval problem. Q J R Meteorol Soc: J Atmos Sci, Appl Meteorol Phys Ocean 131(605):63–72

Hocking W, Fuller B, Vandepeer B (2001) Real-time determination of meteor-related parameters utilizing modern digital technology. J Atmos Solar Terr Phys 63(2–3):155–169

Hodyss D, Nichols N (2015) The error of representation: Basic understanding. Tellus A 67(24):822–839

Hodyss D, Satterfield E (2017) The treatment, estimation, and issues with representation error modelling. In: Data assimilation for atmospheric, oceanic and hydrologic applications, vol III. Springer, pp 177–194

Hoffman RN, Ardizzone JV, Leidner SM, Smith DK, Atlas R (2013) Error estimates for ocean surface winds: applying Desroziers diagnostics to the cross-calibrated, multiplatform analysis of wind speed. J Atmos Oceanic Tech 30(11):2596–2603

Hogan TF, Liu M, Ridout JA, Peng MS, Whitcomb TR, Ruston BC, Reynolds CA, Eckermann SD, Moskaitis JR, Baker NL et al (2014) The navy global environmental model. Oceanography 27(3):116–125

Holdsworth DA, Reid IM, Cervera MA (2004) The Buckland Park all-sky interferometric meteor radar—description and first results. Radio Sci 39:RS5009. https://doi.org/10.1029/2003RS003014

Hollingsworth A, Lonnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. part I: The wind field. Tellus A 38A(2):111–136

Honda T, Miyoshi T, Lien G-Y, Nishizawa S, Yoshida R, Adachi SA, Terasaki K, Okamoto K, Tomita H, Bessho K (2018) Assimilating all-sky himawari-8 satellite infrared radiances: a case of typhoon Soudelor (2015). Mon Weather Rev 146(1):213–229

Hotta D, Kalnay E, Ota Y, Miyoshi T (2017) EFSR: Ensemble forecast sensitivity to observation error covariance. Mon Wea Rev 145:5015–5031. https://doi.org/10.1175/MWR-D-17-0122.1

Iimura H, Fritts DC, Janches D, Singer W, Mitchell NJ (2015) Interhemispheric structure and variability of the 5-day planetary wave from meteor radar wind measurements. Ann Geophys 33(11):1349–1359. https://doi.org/10.5194/angeo-33-1349-2015

Janjić T, Bormann N, Bocquet M, Carton JA, Cohn SE, Dance SL, Losa SN, Nichols NK, Potthast R, Waller JA, Weston P (2018) On the representation error in data assimilation. Q J R Meteorol Soc 144 (713):1257–1278

Jones GOL, Berkey FT, Fish CS, Hocking WK, Taylor MJ (2003) Validation of imaging Doppler interferometer winds using meteor radar. Geophys Res Lett 30:1743. https://doi.org/10.1029/2003GL017645

Karspeck AR (2016) An ensemble approach for the estimation of observational error illustrated for a nominal 1 global ocean model. Mon Weather Rev 144(5):1713–1728

Kuhl DD, Rosmond TE, Bishop CH, McLay J, Baker NL (2013) Comparison of hybrid ensemble/4dvar and 4dvar within the navdas-ar data assimilation framework. Mon Weather Rev 141(8):2740–2758

Lahoz BKW, Ménard R (eds) (2010) Data assimilation: Making sense of observations. Springer

Lange H, Janjić T (2016) Assimilation of Mode-S EHS aircraft observations in COSMO-KENDA. Mon Weather Rev 144(5):1697–1711

Liu J, Kalnay E (2008) Estimating observation impact without adjoint model in an ensemble Kalman filter. Q J R Meteorol Soc 134:1327–1335. https://doi.org/10.1002/qj.280

Macpherson S, Laroche S (2019) Estimation of ground-based GNSS zenith total delay temporal observation error correlations using data from the NOAA and e-gvap networks. Q J R Meteorol Soc 145(719):513–529

McCormack J, Hoppel K, Kuhl D, de Wit R, Stober G, Espy P, Baker N, Brown P, Fritts D, Jacobi C et al (2017) Comparison of mesospheric winds from a high-altitude meteorological analysis system and meteor radar observations during the boreal winters of 2009–2010 and 2012–2013. J Atmos Solar Terr Phys 154:132–166

McLay J, Bishop CH, Reynolds CA (2010) A local formulation of the ensemble transform (ET) analysis perturbation scheme. Weather Forecast 25(3):985–993

Mehra RK (1970) On the identification of variances and adaptive Kalman filtering. IEEE Trans Autom Control AC-15(2):175–184

Ménard R (2016) Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks. Q J R Meteorol Soc 142:257–273

Merchant CJ, Saux-Picart S, Waller J (2020) Bias correction and covariance parameters for optimal estimation by exploiting matched in-situ references. Remote Sens Environ 237:111590

Mile M, Ben´a˘cek P, R´ozsa S (2019) The use of GNSS zenith total delays in operational arome/hungary 3d-var over a central European domain. Atmos Meas Tech 12(3)

Miyoshi T, Kalnay E, Li H (2013) Estimating and including observation error correlations in data assimilation. Inverse Problems in Science and Engineering 21(3):387–398

Okamoto K, Sawada Y, Kunii M (2019) Comparison of assimilating all-sky and clear-sky infrared radiances from himawari-8 in a mesoscale system. Q J R Meteorol Soc 145(719):745–766

Oke PR, Sakov P (2008) Representation error of oceanic observations for data assimilation. J Atmos Oceanic Tech 25(6):1004–1017

Pinnington EM, Casella E, Dance SL, Lawless AS, Morison JI, Nichols NK, Wilkinson M, Quaife TL (2016) Investigating the role of prior and observation error correlations in improving a model forecast of forest carbon balance using four-dimensional variational data assimilation. Agric for Meteorol 228:299–314

Pulido M, Tandeo P, Bocquet M, Carrassi A, Lucini M (2018) Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. Tellus A: Dyn Meteorol Ocean 70(1):1442 099

Rainwater S, Bishop CH, Campbell WF (2015) The benefits of correlated observation errors for small scales. Q J R Meteorol Soc 141(693):3439–3445

Satterfield E, Hodyss D, Kuhl DD, Bishop CH (2017) Investigating the use of ensemble variance to predict observation error of representation. Mon Weather Rev 145(2):653–667

Satterfield EA, Hodyss D, Kuhl DD, Bishop CH (2018) Observation informed generalized hybrid error covariance models. Mon Weather Rev 146(11):3605–3622

Simonin D, Waller JA, Ballard SP, Dance SL, Nichols NK (2019) A pragmatic strategy for implementing spatially correlated observation errors in an operational system: an application to Doppler radial winds. Q J R Meteorol Soc 145(723):2772–2790

Spargo AJ, Reid IM, MacKinnon AD (2019) Multistatic meteor radar observations of gravity-wave–tidal interaction over southern Australia. Atmos Meas Tech 12:4791–4812. https://doi.org/10.5194/amt-12-4791-2019

Stewart LM, Cameron J, Dance SL, English S, Eyre JR, co-authors. Technical report, University of Reading. Mathematics reports series. Observation Error Correlations in IASI Radiance Data. 2009. http://www.reading.ac.uk/web/FILES/maths/obs_error_IASI_radiance.pdf

Stewart LM, Dance SL, Nichols NK (2013). Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. Tellus A 65

Stober G, Chau JL, Vierinen J, Jacobi C, Wilhelm S (2018) Retrieving horizontally resolved wind fields using multi-static meteor radar observations. Atmos Meas Tech 11:4891–4907. https://doi.org/10.5194/amt-11-4891-2018

Stober G, Baumgarten K, McCormack JP, Brown P, Czarnecki J (2019) Comparative study between ground-based observations and NAVGEM-HA reanalysis data in the MLT region. Atmos Chem Phys Discuss https://doi.org/10.5194/acp-2019-1006, in review

Tandeo P, Ailliot P, Bocquet M, Carrassi A, Miyoshi T, Pulido M, Zhen Y (2020) A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. Mon Wea Rev, Accepted

Ueno G, Nakamura N (2016) Bayesian estimation of observation error covariance matrix in ensemble based filters. Q J R Meteorol Soc 142:2055–2080. https://doi.org/10.1002/qj.2803

Valkonen T, Schyberg H, Figa-Saldana J (2017) Assimilating advanced scatterometer winds in a high-resolution limited area model over northern Europe. IEEE J Sel Top Appl Earth Obs Remote Sens 10(5):2394–2405

Waller JA, Dance SL, Lawless AS, Nichols NK, Eyre JR (2014) Representativity error for temperature and humidity using the Met Office high-resolution model. Q J R Meteorol Soc 140(681):1189–1197

Waller JA, Dance SL, Lawless AS, Nichols NK (2014) Estimating correlated observation error statistics using an ensemble transform Kalman filter. Tellus A: Dyn Meteorol Ocean 66(1):23294

Waller JA, Dance SL, Nichols NK (2016) Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. Q J R Meteorol Soc 142:418–431

Waller JA, Simonin D, Dance SL, Nichols NK, Ballard SP (2016) Diagnosing observation error correlations for Doppler radar radial winds in the Met Office UKV model using observation-minus-background and observation-minus-analysis statistics. Mon Weather Rev 144(10):3533–3551

Waller JA, Ballard SP, Dance SL, Kelly G, Nichols NK, Simonin D (2016) Diagnosing horizontal and inter-channel observation error correlations for SEVIRI observations using observation-minus-background and observation-minus-analysis statistics. Remote Sensing 8(7):581

Waller J, Dance S, Nichols N (2017) On diagnosing observation-error statistics with local ensemble data assimilation. Q J R Meteorol Soc 143(708):2677–2686

Waller JA, Garcıa-Pintado J, Mason DC, Dance SL, Nichols NK (2018) Technical note: Assessment of observation quality for data assimilation in flood models. Hydrol Earth Syst Sci 22(7):3983–3992

Waller JA, Bauernschubert E, Dance SL, Nichols NK, Potthast R, Simonin D (2019) Observation error statistics for Doppler radar radial wind superobservations assimilated into the DWD COSMO_KENDA system. Mon Weather Rev 147(9):3351–3364

Wang T, Fei J, Cheng X, Huang X, Zhong J (2018) Estimating the correlated observation-error characteristics of the chinese fengyun microwave temperature sounder and microwave humidity sounder. Adv Atmos Sci 35(11):1428–1441

Weston PP, Bell W, Eyre JR (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. Q J R Meteorol Soc 140(685):2420–2429

Wilhelm S, Stober G, Chau JL (2017) A comparison of 11-year mesospheric and lower thermospheric winds determined by meteor and MF radar at 69° N. Ann Geophys 35:893–906. https://doi.org/10.5194/angeo-35-893-2017

Whitaker JS, Hamill TM, Wei X, Song Y, Toth Z (2008) Ensemble data assimilation with the NCEP global forecast system. Mon Weather Rev 136(2):463–482

Zhen Y, Harlim J (2015) Adaptive error covariances estimation methods for ensemble Kalman filters. J Comput Phys 294:619–638

Observability Gramian and Its Role in the Placement of Observations in Dynamic Data Assimilation

S. Lakshmivarahan, John M. Lewis, and Sai Kiran Reddy Maryada

Abstract A method of data assimilation that is complementary to traditional 4D-Var (4D-Var) has been developed. 4D-Var has appealed to scientists because of the efficiency with which it determines the cost function gradient with respect to control and available observations. Then through use of any of the gradient-based optimization algorithms, the minimum is iteratively found. The alternate methodology does not depend on available observations; rather, the methodology determines placement of observations that avoid flatness of the cost functional about the operating point in control space. Avoidance of flat patches by bounding the norm of the gradient away from zero fundamentally depends on the dynamics of forecast sensitivities to control that are found through differentiation of the governing constraint equations and coupled solution to these equations and the basic constraint equations. These sensitivities are used to define a linear transformation which turns out to be the observability Gramian (symmetric positive semi-definite matrix) G that maps control error (initially unknown) to the cost-function gradient (as a function of space and time and an arbitrary starting operating point). With observations taken at optimal locations defined by (a) the maxima of the diagonal elements of G or (b) that of the trace of G, gradient-based optimization schemes are used to locate cost-function minimum. The methodology is tested on an air-sea interaction model where results indicate that judicious placement of observations avoiding flatness in control space give good results whereas placement that leads to small absolute-valued gradients produce poor results. The theory also gives guidance on the minimum number of observations necessary to achieve success in locating the cost-function minimum.

S. Lakshmivarahan (✉) · S. K. Reddy Maryada
School of Computer Science, University of Oklahoma Norman, Norman, OK 73019, USA
e-mail: varahan@ou.edu

J. M. Lewis
National Severe Storms Laboratory, Norman, OK and Desert Research Institute, Reno, NV, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_9

1 Introduction

Estimation of the unknown initial condition and/or model parameters that control the evolution and behavior of the solution of a deterministic, dynamic model of a physical phenomenon has been the central theme of the 4-dimensional variational (4-D VAR) approach to dynamic data assimilation (DDA). Basic to any estimation is the availability of "relevant" data—a data set containing information about the unknowns being estimated. In the context of the 4-D VAR framework, it is assumed that we are given a collection of noisy observations at N distinct times. An observation, in general, is a (scalar or vector valued) function of the true state of the physical phenomenon being modeled. The function that maps the state to an observable is called the forward operator in geosciences and output function in Systems and Control theory. Its form—linear/nonlinear, scalar/vector valued, depends on the type of sensors—satellite, radar, lidar, thermometer, pressure gauge, to name a few, used by the observing system.

Given the functional form of the forward operator, by running the model forward in time starting from an arbitrary (incorrect) value of the control consisting of the initial conditions and parameters, we can generate the model counterpart of the observations. The difference between the given noisy observation and its model counterpart at a given time is called the innovation or the forecast error. Following the well-established practice in feed-back and adaptive Control Theory where a function of the error is used to achieve the desired goal (refer to Appendix for an overview of the principles of Control Theory), we seek to minimize a cost functional which is the weighted sum of squared forecast errors, using the model equations as the equality constraint since the above minimization is carried out along the model solution.

This equality constrained minimization can be formulated in one of two ways—either as a strong constrained problem solved using the classical Lagrangian multiplier method (LeDimet and Talagrand 1986; Lewis and Derber 1985) or as a weak constrained problem (Lakshmivarahan 2016). In this paper, we follow along the classical 4-D VAR method that relies on the strong constrained formulation. The utility and the strength of this approach lies in the resulting recursive framework, called the adjoint method, for efficiently computing the gradient—also known as the adjoint gradient/sensitivity of the cost functional with respect to the control. Once the adjoint sensitivity is available, it can be used in conjunction with one of many well-known minimization algorithms (Chaps. 10–12, Lewis et al. 2006) to obtain an improved estimate of the control. These two steps—computing the adjoint gradient and the subsequent update of the control are repeated until a desired accuracy, measured, say, by the square of the norm of the current forecast error, is achieved. Refer to Chaps. 22–24, (Lewis et al. 2006) for more details on the mechanics of this methodology.

4-D VAR based method for forecasting has been the workhorse of the meteorological prediction centers around the world for well over three decades and has enjoyed great success in improving the quality of short-term weather forecasting (Lewis and Lakshmivarahan 2008). However, the cost functional is quadratic in the

unknown control and unimodal only in a special case when the model and the forward operator are both linear. When the model and the forward operator are nonlinear, we virtually do not know anything about the shape—number and the distribution of critical points including the maxima, minima, and saddle points, of the cost functional in the control space. The only computable information about the cost functional in this general case is the local information given by the adjoint gradient. Consequently, 4-D VAR based methodology can only ensure convergence to a local minimum that is close to the initial (incorrect) controls. Further, it is well documented that the magnitude of the adjoint sensitivity essentially control the march towards the local minimum and these control iterates can become marooned in the region of the control space, known as flat patches, where the magnitude of the gradient is very small. To date, the important question of how to avoid the flat patches in the control space by selecting the number and the distribution of observations in time (and space) has not received any attention and the 4-D VAR methodology itself is silent on this question.

In a recent paper, (Lakshmivarahan et al. 2020a) (hereafter referred to as LLH (2020a)), have developed an off-line strategy based on the forward sensitivity of the model solution to address this question. In this and in the companion paper (Lewis et al. 2020b) we further analyze and illustrate this new methodology.

Following LLH (2020a), in this paper we examine the factors affecting the shape of the cost functional in two steps. In the first step we derive a new class of linear transformations defined by a matrix G that directly maps the error in the control into the adjoint gradient. It is shown that this matrix G is, in fact, the observability Gramian developed in the control literature. Refer to Appendix A for a short summary of the role of observability in state and parameter estimation in dynamic systems. In the second step we examine the conditions under which this observability Gramian is positive definite which in turn guarantees the existence and uniqueness of the optimal estimates of the control. For a given model map and forward operator pair, this condition naturally leads to the minimum number of observations needed for estimation. It is further shown that we can control the condition number of G by distributing these observations in locations close to the maxima of the certain functions of the forward sensitivity of the model solution.

Historical remarks: Within the framework of the classical observability developed by Kalman (1960a), a system is either observable or not depending on whether the observability Gramian is positive definite or not. However, theoretically a matrix can be positive definite while its smallest eigenvalue may be very small and positive. In such cases, we will encounter major difficulty to numerically invert the Gramian. To describe and distinguish such cases, Krener (2008a, b), Krener and Ide (2009) introduced the notion of strong and weak observability depending on when the smallest eigenvalue of the Gramian is large or small. Subsequently, in a series of papers Kang and Xu (2012, 2014), King et al. (2015) examined the the variation of the smallest eigenvalues of the Gramian resulting from the distribution of observations in space and time to make the system as strongly observable as is feasible. In this paper, however we, advocate placing the observations at locations where (a) the diagonal elements, G_{ii} attain their maxima or (b) the trace of G (sum of the diagonal elements of G attains its maxima. The basic goal of these strategies is to place the

observation that will render the columns of the observability matrix, G to be linearly independent. To avoid weak observability, we can further examine the variation of the condition number of G with respect to the distribution of observations to pin down that distribution that has a lower value for the condition number as illustrated in Sect. 6.

From the early 1990s there is a growing body of work within the meteorological literature on targeted or adaptive observations. The primary goal in this area is to adaptively decide on the number and placement of new observations that will further reduce the analysis error and hence the subsequent forecast error. This literature examines a variety of tools including statistical experimental design, adjoint sensitivity to observations, ensemble transform filter-based approaches, ensemble sensitivity and more recently, observability Gramian (Yoshimura et al. 2020). A discussion of this body of literature is contained in the introduction to LLH (2020a). For more details, refer to the literature cited in LLH (2020a).

This paper is an extension of LLH (2020a) and differs from those in the literature, both in the methodology and goal. Our goal is to find out the minimum number of observations and their distribution for effective estimation of both the initial condition and model parameters. While most of the papers in this literature deal with estimation of initial states, we treat the initial state and parameter estimation on the same footing.

A summary of the model equations—linear and nonlinear, observations described by linear and nonlinear forward operators, innovation/forecast error, cost functional and the statement of the problem along with the basic notations are given in Sect. 2. Section 3 provides a short summary of the dynamics of evolution of forward sensitivities. Section 4 examines the structure of the linear transformation given by the matrix G and the conditions for it to be positive definite when the model map, forward operator- (M, H) pair is linear, and its nonlinear counterpart is covered in Sect. 5. These two sections develop the basic conditions for the positive definiteness of the matrix G and further examines the role of distribution of observations and its impact on controlling the condition number of G. In Sect. 6 we illustrate the key points of the approach using a simple 1-D problem known as the air-sea interaction. Concluding summary along with the guidelines for the distribution of observations is given in Sect. 7. Appendix A provides an overview of the role of observability in state and parameter estimation and Appendix B contains some of the results from matrix theory that are critical to developments in this paper and Appendix C develops a set of conditions for observability. Further illustrative examples are contained in the companion paper by Lewis et al. (2020b) in this volume.

2 Notations and Statement of Problem

In this section we start by describing the key players in the game of dynamic data assimilation—model, observation, innovation and the cost functional. Refer to Table 1 for details.

2.1 Model

Let R^n be the state space and $x(k) = (x_1(k), x_2(k), \ldots, x_n(k))^T \in R^n$ be the state of a deterministic, discrete time, dynamic model where the time index k takes values over the non-negative integers starting at $k = 0$, the initial time and $x(0) \in R^n$ is the initial condition. Let R^p be the parameter space and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T \in R^p$ be a vector of model parameters.

We consider two versions of the model equations-linear and non-linear. Let $M : R^n \times R^p \to R^n$ be the given one step state transition map of the non-linear model where $M(x, \alpha) = (M_1(x, \alpha), M_2(x, \alpha), \ldots M_n(x, \alpha))^T \in R^n$. Let $M \in R^{n \times n}$ be the given one step state transition matrix for the linear model. The structure of the governing equations for both the models are given in Table 1.

The solution of the non-linear model starting from $x(0)$ as the initial state and α as the parameter is denoted by $x(k) = x(k, x(0), \alpha)$. In the linear case, for simplicity it is assumed that the matrix M is known and the solution in this case is denoted by $x(k) = x(k, x(0))$. If some of the elements of M are not known in advance, we can indeed treat these as the elements of the unknown parameter vector α. Since the evolution of the solution $x(k)$ depends on the initial condition $x(0)$ and/or the parameter α, the pair $c = (x^T(0), \alpha^T)^T \in R^n \times R^p$ is called the control and $R^n \times R^p$ is called the control space. By definition, each point in the control space defines a particular instant of the model and so a model, by definition, represents a class of models, one for each allowed pair c.

The need and the rationale for parameterizing a model stems from the nature and type of the physical phenomenon it is meant to capture. Certain natural phenomenon such as the dynamics of motion of the planets around the sun has remained invariant over time and so these models enjoy the luxury of having a small set of known parameters. On the other hand, there are examples of natural phenomenon that exhibit natural variation depending on the season and/or geographical location. The problem of hurricane prediction is an example of the second kind. Hurricanes in the south China sea, sea of Japan, Bay of Bengal and in the mid-north Atlantic have their differences based on the configuration of land masses adjoining the respective oceans and differing general circulation patterns. While barotropic vorticity equation is the basis of all hurricane track prediction, it is often advisable to account for smaller-scale non-conservative process by parameterizing them in terms of large scale forcing. Henceforth, it is assumed that we are given a class of models faithful to the phenomenon whose parameters can be fine-tuned to capture the evolution of the process at a given season and location in question.

The first step in the development of the model based predictive science is to 'instantiate' the model by estimating the unknown control. This is accomplished by using the observations from the phenomenon of interest.

Table 1 Model, observation and cost function

Model	Linear	$x(k+1) = Mx(k), x(0)$ - I.C.
		$x(k) = M^k x(0)$, control c = x(0)
	Non-linear	$x(k+1) = M(x(k), \alpha), x(0)$ - I.C.
		control c = $(x^T(0), \alpha^T)^T$
Observation	Linear	$z(k) = H\bar{x}(k) + \xi(k)$
	Non-linear	$z(k) = h(\bar{x}(k)) + \xi(k)$
Observation noise	$\xi(k)$	$\sim N(0, R)$, temporally uncorrelated
	Covariance	$R \in R^{m \times m}$ is a known SPD matrix
Innovation $\bar{e}(k)$	Linear	$\bar{e}(k) = z(k) - Hx(k)$
	Non-linear	$\bar{e}(k) = z(k) - h(x(k))$
Initial error in control	Linear	$\delta c = \delta x(0) = \bar{x}(0) - x(0)$
	Non-linear	$\delta c = (\delta x^T(0), \delta \alpha^T)^T$
		$\delta x(0) = \bar{x}(0) - x(0), \delta\alpha = \bar{\alpha} - \alpha$
Induced variation $\delta x(k)$	Linear	$\delta x(k) = \bar{x}(k) - x(k) = U(k)\delta x(0)$
		$= M^k \delta x(0)$
	Non-linear	$\delta x(k) = \bar{x}(k) - x(k)$
		$\delta x(k) = U(k)\delta x(0) + V(k)\delta\alpha$
Cost functional $J_e(c)$		$J_e(c) = \frac{1}{2}\sum_{k=1}^{N} \bar{e}^T(k) R^{-1} \bar{e}(k)$

2.2 Observations

Let $\bar{c} = (\bar{x}^T(0), \bar{\alpha}^T)^T$ be the true but unknown control and let $\bar{x}(k)$ be the resulting true state of the phenomenon under study. More often than not, we may not be able to observe this true state directly, but only a (scalar or vector valued) function of it. Let $z(k) = h(\bar{x}(k))$ be the observation where $z(k) \in R^m$ and $h : R^n \to R^m$ is called the forward operator and $h(x) = (h_1(x), h_2(x), ..., h_m(x))^T$. In the special case, when $h(x)$ is a linear function, then $z(k) = H\bar{x}(k)$ where $H \in R^{m \times n}$. In general, observations are corrupted by additive, temporally uncorrelated, Gaussian noise with zero mean and known covariance matrix $R \in R^{m \times m}$ which is assumed to be symmetric and positive definite. Refer to Table 1 for further details.

2.3 Innovation/Forecast Error

Given the model equation and a finite set of observations, our goal is to estimate the true but unknown control $\bar{c}$. To this end, we pick an arbitrary control c and run the model forward to obtain the solution $x(k)$. Let $h(x(k))$ be the model counterpart of the observation. The difference between $z(k)$ and $h(x(k))$ is called the innovation/forecast error (refer to Table 1) which is a measure of the difference between the chosen control c and the true control $\bar{c}$.

2.4 Cost Functional

Since the components of the innovation vector $\bar{e}(k)$ can be positive or negative, in the spirit of the principles of feedback control (Appendix A), we define a scalar valued cost functional which is the weighted sum of squared forecast errors as shown in Table 1. Consequently, we have reduced the problem of estimating the true control $\bar{c}$ to one of minimizing the cost functional, $J_e(c)$ in Table 1.

2.5 Statement of Problem

Given a model-forward operator combination—either (M, h) or (M, H) pair and the second order properties of the noise corrupting the observations, an important design question is: how to control the shape of the cost functional in the control space to avoid flat patches by bounding the magnitude of the adjoint sensitivity (gradient of the cost functional) away from zero by suitable placement of observations in the spatio-temporal domain of the model.

In a recent paper, LLH (2020a), we have answered this question in the affirmative by developing a two step procedure. In the first step, we derive a basic linear map that maps the error in control to the adjoint sensitivity. Since this matrix G is additive in the number of observations, it will be shown that the minimum requisite number of observations is determined by the positive definiteness of G—that is, by sequentially adding observations one at a time, there comes a point when the G matrix becomes positive definite and that determines the minimum number of observations. Further, placement of these observations following one of the two strategies outlined in Sect. 5 and demonstrated in Sect. 6 leads to the determination of the strategy with the smallest condition number.

Remark 1 It is important to identify the philosophical difference between the method proposed in this study and the traditional 4D-VAR. The 4-D VAR approach concentrates on efficiently computing the adjoint gradient that serves as input to any of the gradient based optimization schemes which in turn iteratively proceed toward the minimum of the cost functional. It is predicated on the key assumption

that we are given a set of observations with a prespecified distribution in the spatio-temporal domain that contains the model solution. In this paper, however, we are interested in the complementary design question: how to distribute the observations so as to avoid the occurrences of flat patches in the control space. Consequently, in the following analysis there is no need to have any observations on hand. All that is needed are (1) the knowledge of the (M, h) or (M, H) pair and (2) the covariance matrix, R of the observation noise.The output of our analysis is a recommendation for the number and placement of observations that will guarantee that norm of the adjoint gradient is bounded away from zero. Once the number and the distribution are known, we can then commence the iterative process of finding the minimum of the cost functional using the forward sensitivity since all the relevant quantities - forecast sensitivity to control, covariance of observation error, and the innovations are in hand (Lakshmivarahan et al. 2017).

In the following sections, we further elaborate on the results in LLH (2020a) and extend it to linear dynamics and relate the properties of G matrices to the observability Gramian described in Appendix A.

3 Dynamics of Evolution of Forward Sensitivities

Our approach to the problem of placement of observations is closely linked to the properties of the (forward) sensitivity of the model solution $x(k)$ with respect to both the initial condition and parameters. Refer to Table 2 for the definition of $U(k)$, the forward sensitivity of $x(k)$ with respect to the initial condition $x(0)$ and $V(k)$ that with respect to the parameter α. This Table 2 also contains the definitions of all related Jacobians that control the evolution of $U(k)$ and $V(k)$. The linear, time varying, discrete time dynamics of $U(k)$ and $V(k)$ are given in Table 3. For a derivation of the dynamics in Table 3, refer to Lakshmivarahan and Lewis (2010) and the monograph by Lakshmivarahan et al. (2017).

4 Relation Between Adjoint Sensitivity and Initial Control Error: Linear Case

Consider a linear model with known one step state transition matrix, $M \in R^{n\times n}$ and the observation defined by the linear, forward operator, $H \in R^{m\times n}$ defined in Table 1. This combination is called the (M, H) pair. Recall that since M is assumed to be known, the solution of the model is controlled only by the initial condition, $c = x(0)$.

The model solution $\bar{x}(k)$ and $x(k)$ starting from the unknown true state $\bar{c} = \bar{x}(0)$ and an arbitrary initial state $c = x(0)$ are given by

$$\bar{x}(k) = M^k \bar{x}(0) \text{ and } x(k) = M^k x(0). \tag{1}$$

Table 2 Jacobians of model, observations and state

Model Jacobian w.r.t x	Non-linear	$D_M(k) = [\frac{\partial M_i(x(k),\alpha)}{\partial x_j(k)}] \in R^{n\times n}$
	Linear	$D_M(k) = M \in R^{n\times n}$
Model Jacobian w.r.t α	Non-linear	$D^{\alpha}{}_M(k) = [\frac{\partial M_i(x(k),\alpha)}{\partial \alpha_j}] \in R^{n\times p}$
Jacobian of forward operation	Non-linear	$D_h(k) = [\frac{\partial h_i(x(k))}{\partial x_j(k)}] \in R^{m\times n}$
	Linear	$D_H(k) = H \in R^{m\times n}$
Forward sensitivity w.r.t $x(0)$	Non-linear	$U(k) = [\frac{\partial x_i(k)}{\partial x_j(0)}] \in R^{n\times n}$
	Linear	$U(k) = [\frac{\partial x_i(k)}{\partial x_j(0)}] = M^k$
Forward sensitivity w.r.t α	Non-linear	$V(k) = [\frac{\partial x_i(k)}{\partial \alpha_j}] \in R^{n\times p}$

Table 3 Dynamics of evolution of $U(k)$ and $V(k)$ where I_n is the identity matrix of order n

Non-linear	$U(k+1) = D_M(k)U(k), I.C., U(0) = I_n$ $U(k) = D_M(k-1)D_M(k-2)\ldots D_M(0) = D_M(k-1:0)$ $V(k+1) = D_M(k)V(k) + D^{\alpha}{}_M(k), I.C., V(0) = 0$
Linear	$U(k+1) = MU(k), I.C., U(0) = I_n$ $U(k) = M^k$

The difference $\delta c = \bar{c} - c =$ is called the initial control error. From (1), the induced error $\delta x(k)$ in the state $x(k)$ at time k is given by

$$\delta x(k) = \bar{x}(k) - x(k) = M^k \delta c. \tag{2a}$$

Alternatively, referring to Tables 1 and 3, since $U(k) = M^k$, from the first principles, we obtain the same expression

$$\delta x(k) = \left[\frac{\delta x(k)}{\delta x(0)}\right] \delta x(0) = U(k)\delta c = M^k \delta c. \tag{2b}$$

Further, the expression for the innovation/forecast error in Table 1 takes the form

$$\bar{e}(k) = z(k) - Hx(k) = H(\bar{x}(k) - x(k)) + \xi(k) = HU(k)\delta c + \xi(k). \tag{3}$$

Substituting (3) in the expression for the cost functional in Table 1, we obtain

$$J_e(c) = \frac{1}{2}\sum_{k=1}^{N} (\delta c)^T \left[(U(k)^T \bar{H} U(k)\right] \delta c = \frac{1}{2}(\delta c)^T G \delta c \tag{4}$$

where the matrix G is given by

$$G = \sum_{k=1}^{N} \left[(U(k)^T \bar{H} U(k) \right] \in R^{n \times n} \tag{5}$$

and $\bar{H} = H^T R^{-1} H$. Comparing this expression with that in (76) in Appendix A, it follows that this G matrix is indeed the observability Gramian for the (M, H) pair when the observations are noisy. Since $\delta c = \bar{c} - c$, by direct differentiation of (4) we get a new expression for the adjoint gradient given by

$$\nabla_c J_e(c) = -G\delta c. \tag{6a}$$

By way of simplifying the notation, define

$$g = -\nabla_c J_e(c) \text{ and } f = \delta c. \tag{6b}$$

Then, (6) becomes

$$g = Gf \tag{7}$$

which is the sought after relation between the control error f and the negative of the adjoint gradient, g. Recall that a Gramian by definition is a symmetric, positive, semi-definite (SPSD) matrix (Meyer 2000). Consequently, if G is SPSD, then it has a null space, $Null(G)$ such that

$$dim(Null(G)) = n - dim(Range(G)) > 0$$

and for any vector $f \in Null(G)$, $Gf = 0$ and the adjoint gradient vanishes. Thus, $J_e(c)$ is identically zero in the $Null(G)$. Hence, a necessary and sufficient condition for the existence and uniqueness of the minimizer of $J_e(c)$ is that the Gramian G in (7) must be positive definite, that is, G is a SPD matrix. We now explore the condition on (M, H) pair that will guarantee that G is SPD. Since the observation covariance R is assumed to be positive definite, so is R^{-1} and it is well known that there exists an upper triangular matrix W called the square root of R^{-1} such that $R^{-1} = W^T W$ (Chap. 9, Lewis et al. (2006). Accordingly, the expression (5) for G can be written as

$$G = \sum_{k=1}^{N} (EU(k))^T (EU(k) \tag{8}$$

where for simplicity in notation $E = WH$ and $\bar{H} = E^T E$. Define, since $U(k) = M^k$,

$$L = \begin{bmatrix} EM \\ EM^2 \\ \vdots \\ EM^N \end{bmatrix} \in R^{Nm \times n}. \tag{9}$$

It can be verified that

$$G = L^T L. \tag{10}$$

Again, from Appendix A, it is well known that G is SPD if and only if L is a full rank matrix with $Rank(L) = n$. Stated in the other words, for a given R^{-1} and W, this rank condition on L translates into conditions on the (M, H) pair.

This condition is further examined in Appendix C. It immediately follows from Corollary C.3 in this Appendix that L is of full rank exactly when the energy in E as measured by its Frobenius norm, $||E^T||_F$ is distributed across all the eigen directions of the matrix M^T. A demonstration of this condition is contained in the following:

Example 1 Consider the (M_2, H_2) pair from the Appendix A. In this case $n = 2$, $m = 1$, $R^{-1} = I = W$ and $N = 2$ where

$$M_2 = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} and\ H_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

It is easy to check that the eigenvalues of M_2^T are 1 and 2 and the corresponding eigen vectors are $(1, 0)^T$ and $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$. Since H_2^T lies along the eigenvector $(1, 0)^T$, it follows from Appendix C that this (M_2, H_2) is not observable. We further examine the properties of the corresponding L and G.

It can be verified that $L = \begin{bmatrix} H_2 M_2 \\ H_2 M_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ and $G = L^T L = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ which is a rank one matrix. Let $\bar{c} = (\bar{c}_1, \bar{c}_2)^T$ and $c = (c_1, c_2)^T$. Then $\delta c = (\bar{c}_1 - c_1, \bar{c}_2 - c_2)^T$ and $Q = (\delta c)^T G \delta c = 2(\bar{c}_1 - c_1)^2$. Hence $\nabla_c Q = (4(\bar{c}_1 - c_1), 0)^T$ and Q is a constant along the second dimension and consequently this pair (M_2, H_2) is not observable. We leave it to the reader to verify that if we consider a new pair (M_2, H_3) with $H_3 = (0, 1)$, then

$$L = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \text{ and } G = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix}$$

which is positive definite. Hence (M_2, H_3) is observable.

It is instructive to examine the interplay between M and H to see why (M_2, H_2) is not observable. The model equations with M_2 as the state transition matrix are

$$x_1(k+1) = x_1(k)$$

$$x_2(k+1) = x_1(k) + 2x_2(k).$$

It then follows that $x_2(k+1)$ being a linear combination of $x_1(k)$ and $x_2(k)$ has information on both the components but $x_1(k+1)$ does not have any information on $x_2(k)$. Consequently, for the (M_2, H_2) pair with the observation $z(k) = H_2 x(k) = x_1(k)$, is not observable. But for the (M_2, H_3) pair, the with observation $z(k) = H_3 x(k) = x_2(k)$ is observable.

We now explore the consequences of G being positive definite besides being symmetric.

1. From (4) it follows that $J_e(c)$ is quadratic and is unimodal in c with the minimum located at $\bar{c}$, the unknown true state. Further, it can be verified that G is the Hessian of $J_e(c)$ at $\bar{c}$, and G is additive in the number N of observations. As the number of observations increases, while the location, $\bar{c}$ of the minimum remains the same, the curvature of $J_e(c)$ at $\bar{c}$ as measured by the Hessian increases and function $J_e(c)$ as a whole is increasing pointwise. However, more observations also increases the computational time.
2. A visual presentation of the intrinsic relation in (7) in given in Fig. 1. Referring to this Fig. 1, let $\hat{f}$ and $\hat{g}$ be the unit vectors in the direction of f and g and are given by

$$\hat{f} = \frac{f}{||f||}, \hat{g} = \frac{g}{||g||} = \frac{Gf}{||Gf||}. \tag{11}$$

Let $f^{\perp}$ denote the direction that is orthogonal to f. Now we can resolve g into two components, g^p and $g^{\perp}$ along $\hat{f}$ and $f^{\perp}$ respectively where

$$g = g^p + g^{\perp} \tag{12}$$

that is, g^p and $g^{\perp}$ are the orthogonal projection of g along $\hat{f}$ and $f^{\perp}$. Thus

$$g^p =< \hat{f}, g > \hat{f} =< \hat{f}, Gf > \hat{f} =< \hat{f}, G\hat{f} > ||f||\hat{f} =< \hat{f}, G\hat{f} > f \tag{13}$$

where the Rayleigh coefficient

$$< \hat{f}, G\hat{f} >> 0 \tag{14}$$

since G is SPD. Let θ be the angle between g and f. Then

$$cos\theta =< \hat{f}, \hat{g} >=< \hat{f}, \frac{Gf}{||Gf||} >= \frac{< \hat{f}, G\hat{f} >}{(\frac{||Gf||}{||f||})}. \tag{15}$$

Since G is symmetric

$$\frac{||Gf||^2}{||f||^2} = \frac{< f, G^2 f >}{||f||^2} =< \hat{f}, G^2\hat{f} > . \tag{16}$$

Combining, referring to Sect. B.4 in Appendix B, it can be verified that

$$cos\theta = \frac{< \hat{f}, G\hat{f} >}{< \hat{f}, G^2\hat{f} >^{\frac{1}{2}}} \leq 1. \tag{17}$$

Hence, $|\theta| \leq 90^0$ and $||g^p|| \geq 0$. That is, g has a non-negative projection along f.

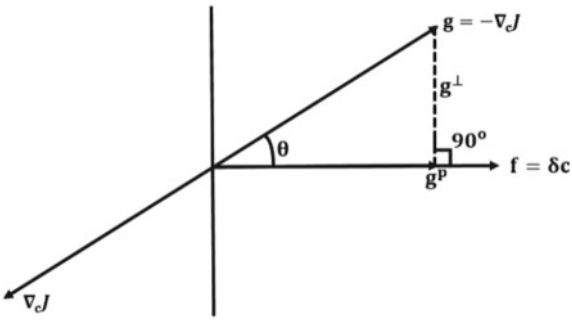

Fig. 1 A pictorial view of the relation between f and g where g^p is the orthogonal projection of g onto f. That is, $g^p = \hat{f} < \hat{f}g >= P_{fg}$

3. The goal of DDA is to move the current control c towards the minimum $\bar{c}$ iteratively by moving along the negative of the adjoint gradient. Recall that $g = Gf$ is the negative of the adjoint gradient and the new control c^n is given by

$$c^n = c + \beta G f \tag{18}$$

where $f = \bar{c} - c$ and the scalar, $\beta > 0$ is the step length parameter. Then,

$$f^n = \bar{c} - c^n = (\bar{c} - c) - \beta G f = (I - \beta G) f. \tag{19}$$

Since $J_e(c)$ is quadratic, the optimal value of the step length β (refer to Chap. 10, Lewis et al. 2006) is given by

$$\beta = \frac{g^T g}{g^T G g} = \frac{f^T G^2 f}{f^T G^3 f}. \tag{20}$$

Referring to Sect. B.5 in Appendix B, it can be verified that the spectral radius of $(I - \beta G)$ are less than 1 and hence

$$||f^n|| < ||f||. \tag{21}$$

That is, c^n is closer to $\bar{c}$ than c and convergence is guaranteed. We can also rewrite (17) as follows:

$$c^n = c + \beta G(\bar{c} - c) = (I - \beta G)c + \beta G\bar{c} \tag{22}$$

which in turn implies that c^n lies in the line segment joining c and $\bar{c}$ and c^n is a convex combination of c and $\bar{c}$

5 Relation Between Adjoint Sensitivity and Initial Control Error: Non Linear Case

Consider a given (M, h) pair where M is a non-linear one step state transition map that defines a discrete time dynamic model and h is a non-linear forward operator. Refer to Table 1. It is assumed that the model map is known up to a set of p unknown model parameters represented by the vector $\alpha \in R^p$. In addition, the initial condition, $x(0) \in R^n$ is also not known in advance. The combined vector $c = (x^T(0), \alpha^T)^T \in R^{n+p}$ is called the control vector. We are given a set $S = \{z(k_i) : 1 \leq i \leq N\}$ of N observations at times $k_1 < k_2 \ldots k_N$. Our goal is to estimate the unknown control based on S.

Assumption 1 Non-degeneracy of (M, h) pair: Referring to Table 2, it is assumed that the pair (M, h) is non-degenerate in the sense that their Jacobians $D_M(k) \in R^{n\times n}$, $D_M^{\alpha}(k) \in R^{n\times p}$ and $D_h(k) \in R^{m\times n}$ respectively are full rank matrices along the model solution for all initial conditions and allowed set of parameter values.

Let $\bar{c} = (\bar{x}^T(0), \bar{\alpha}^T)^T$ be the true but unknown control from which the true states, $\bar{x}(k), k > 0$ are generated. Let $x(k)$ be the model solution starting from an erroneous control $\delta c = (x^T(0), \delta\alpha^T)^T$ where

$$\delta x(0) = \bar{x}(0) - x(0) \text{ and } \delta\alpha = \bar{\alpha} - \alpha. \tag{23}$$

The induced error in $x(k)$ is then given by

$$\delta x(k) = \bar{x}(k) - x(k). \tag{24}$$

The first step is to relate $\delta x(k)$ to δc. To this end, recall that from first principles, we get (refer to Table 1).

$$\begin{aligned}\delta x(k) &= \left[\frac{\partial x(k)}{\partial x(0)}\right]\delta x(0) + \left[\frac{\partial x(k)}{\partial \alpha}\right]\delta\alpha \\ &= U(k)\delta x(0) + V(k)\delta\alpha,\end{aligned} \tag{25}$$

where $U(k) \in R^{n\times n}$ and $V(k) \in R^{n\times p}$ are forward sensitivities of the model solution with respect to $x(0)$ and α respectively.

Referring to the dynamics of evolution of forward sensitivities in Table 3, the non-degeneracy assumption guarantees that U(k) and V(k) are full rank matrices for all k.

Consequently, the exact expression for the innovation or forecast error is given by (refer to Table 1)

$$\bar{e}(k) = Z(k) - h(x(k)) = h(\bar{x}(k)) - h(x(k)) + \xi(k). \tag{26}$$

Expanding $h(\bar{x}(k))$ in the first-order Taylor series around the known forecast x(k) using (24), we get a first-order approximation

$$e(k) = D_h(k)\delta x(k) + \xi(k) \tag{27}$$

to $\bar{e}(k)$ where $D_h(k)$ is the Jacobian of h (Refer to Table 2) which by Assumption 5.1 is a full rank matrix for all k. We hasten to add that when M and h are both linear $e(k) = \bar{e}(k)$ and (27) is exact. (Refer to Sect. 4). But, when M and h are strongly nonlinear (as measured by the curvature), we can improve the approximation in (27) by adding the second-order correction term (Lakshmivarahan et al. 2003), Chapter 7 in Lewis et al. (2006).

Substituting (25) in (27) and simplifying, we get a fundamental expression that directly relates the innovation $e(k)$ to the initial control error δc:

$$e(k) = D_h(k)F(k)\delta c \tag{28}$$

where the combined forward sensitivity matrix $F(k)$ is given in the partitioned form as

$$F(k) = [U(k), V(k)] \in R^{n\times(n+p)}. \tag{29}$$

Now replacing $\bar{e}(k)$ by e(k) in $J_e(c)$ in Table 1, we get a corresponding local first-order approximation, $J(c)$ to $J_e(c)$ given by

$$J(c) = \frac{1}{2}\sum_{k=1}^{N}[D_h(k)F(k)\delta C]^T R^{-1}[D_h(k)F(k)\delta c] = \frac{1}{2}(\delta C)^T G(\delta c) \tag{30}$$

where the Gramian G is a sum of the Gramians $G(k)$ given by

$$\begin{aligned} G &= \sum_{k=1}^{N} G(k), \\ G(k) &= F^T(k)\bar{H}F(k) \in R^{(n+p)\times(n+p)}, \\ \bar{H}(k) &= D_h^T(k)R^{-1}D_h(k) \in R^{n\times n} = E^T(k)E(k) \end{aligned} \tag{31}$$

with $E(k) = WD_h(k) \in R^{m\times n}$ where $W \in R^{m\times m}$ is the square root of R^{-1} defined in (8)

Remark 2 It is interesting to note that in (31) the component matrix $F(k)$ depends purely on the model through the forward sensitivities, $U(k)$ and $V(k)$ of the model solution and the matrix $\bar{H}$ depends purely on the observation system through the Jacobian of the forward operator and covariance of the observation noise. In other words, the product structure of the Gramian $G(k)$ brings out the inherent natural separability of the effect due of the model and the observation system. Also notice that

in computing $G(k)$ we do not need to have any knowledge of the actual observations. The role of actual observation comes into play only at the time of data assimilation.

The fine structure of the Gramian $G(k)$ in (31) is given by

$$G(k) = F^T(k)\bar{H}F(k) = \begin{bmatrix} U^T\bar{H}U & U^T\bar{H}V \\ V^T\bar{H}U & V^T\bar{H}V \end{bmatrix}, \tag{32}$$

where we have dropped the time k to avoid cluttering. Clearly, G is a sum of matrices of the type given in (32).

Since $\delta c = \bar{c} - c$, by direct differentiation of (30), we get sought after expression for the adjoint gradient.

$$g = Gf \tag{33}$$

where, as in Sect. 4,

$$g = -\nabla_c J(c) \text{ and } f = \delta c. \tag{34}$$

We now examine the conditions for the existence and uniqueness of the minimum of $J(c)$ in (30). To this end, it is useful to divide the discussions into three cases.

Case 1: $\alpha = \bar{\alpha}$ and x(0) is not known: In this case, $\delta\alpha = 0$ but $\delta x(0) \neq 0$. Hence, the error in control $\delta c = \delta x(0)$. Further, $F(k) = U(k)$ and from (5.9).

$$G(k) = [E(k)U(k)]^T[E(k)U(k)].$$

Now, define,

$$L = \begin{bmatrix} E(k_1)U(k_1) \\ E(k_2)U(k_2) \\ \vdots \\ E(k_N)U(k_N) \end{bmatrix} \in R^{Nm\times n}. \tag{35}$$

It can be verified that

$$G = L^T L \in R^{n\times n}. \tag{36}$$

Clearly, G in positive definite if and only if $Rank(L) = n$, for some $N \geq \lceil \frac{n}{m} \rceil$.

Since (M, h) is non-degenerate, both $E(k)$ and $U(k)$ are full rank matrices. Referring to Appendix B,

$$Rank(E(k)) = m \wedge n = min\ (m, n)$$

Hence, when $m > n$, using N = 1 observation $z(k) \in R^m$, it can be verified that

$$Rank(L) = Rank(E(k_1)U(k_1)) = n \tag{37}$$

and hence $G = G(k_1)$ is positive definite. On the other hand, when $m < n$, referring to the results in Sects. B.2 and B.3 in Appendix B, with $N \geq \lceil \frac{n}{m} \rceil$ observations,

depending on the functional form of M and , we can guarantee the positive definiteness of G. Refer to Example 4.1 and Appendix A to get a flavor of this dependence on the functional form of M and h.

In fact, we can develop an analogous version of the Corollary C.3 in Appendix C which may be stated as follows. If the energy in $E^T(k)$ is distributed across all the eigen directions of $U^T(k)$, uniformly for all k, then L in (35) will be of full rank and G in (36) will be positive definite.

Against this backdrop, we now describe a method for distributing the observations. To this end, consider the column partition of $U(k)$ given by

$$U(k) = [U_1(k), U_2(k), \ldots, U_n(k)] \tag{38}$$

where by definition (refer to Table 2)

$$U_i(k) = \frac{\partial x(k)}{\partial x_i(0)} \in R^n \tag{39}$$

is the forward sensitivity of the model solution with respect to $x_i(0)$, $0 \leq i \leq n$. Then, $G(k)$ can be expressed as an outer product matrix:

$$G(k) = \begin{bmatrix} [E(k)U_1(k)]^T \\ [E(k)U_2(k)]^T \\ \vdots \\ [E(k)U_n(k)]^T \end{bmatrix} \Big[E(k)U_1(k), E(k)U_2(k), \ldots, E(k)U_n(k)\Big]. \tag{40}$$

Consequently, the diagonal element of G(k) is given by

$$G_{ii}(k) =< E(k)U_i(k), E(k)U_i(k) >= ||E(k)U_i(k)||^2. \tag{41}$$

It can be verified that

$$tr(G(k)) = \sum_{i=1}^{n} G_{ii}(k) = \sum_{i=1}^{n} ||E(k)U_i(k)||^2 = ||EU(k)||_F^2 \tag{42}$$

where $||A||_F^2$ is the Frobenius norm of A.

Accordingly, we suggest two strategies for the distribution of observations:

Strategy I – Coarse granularity: Place the required number N of observations at times where the $tr(G(k))$ attains maximum values, for $1 \leq k \leq N$.

Strategy II – Finer granularity: Place the required number N of observations at times where the diagonal elements $G_{ii}(k)$ attains a maximum 'individual' as shown: 'the individual diagonal'.

The use and effectiveness of these strategies are illustrated in Sect. 6. We hasten to add that Strategy II in addition to guaranteeing that the columns are linearly

independent, it has an additional effect of controlling the condition number of G as illustrated in Sect. 6.

Case 2: $x(0) = \bar{x}(0)$ **and** α **is not known:** In this case $\delta x(0) = 0$, $\delta\alpha \neq 0$, $\delta C = \delta\alpha$ and $F(k) = V(k) \in R^{n\times p}$. Hence,

$$G(k) = [E(k)V(k)]^T[E(k)V(k)]. \tag{43}$$

Define

$$L = \begin{bmatrix} E(k_1)F(k_1) \\ E(k_2)F(k_2) \\ \vdots \\ E(k_N)F(k_N) \end{bmatrix} \in R^{Nm\times p}. \tag{44}$$

Then

$$G = L^T L \in R^{p\times p}. \tag{45}$$

From Corollary C.4 in Appendix C, it can be verified that

$$(m \wedge n) + (n \wedge p) - n \leq Rank(E(k)V(k)) \leq m \wedge n \wedge p. \tag{46}$$

Assume that $p < (m \wedge n)$. Then, when $m > n$, we get

$$Rank(E(k)V(k)) = p$$

and in this case with $N = \lceil \frac{n}{m} \rceil = 1$ observation, L is of full rank and $G = G(k_1)$ is positive definite. On the other hand, when $m < n$, it follows from (41) that

$$p - (n - m) \leq Rank(E(k)V(k)) \leq p. \tag{47}$$

In this latter case, in view of the results in Sects. B.2 and B.3 in Appendix B and depending on the functional form of M and h, with $N \leq \lceil \frac{n}{m} \rceil$ observations, we can achieve positive definiteness of G. As we did in case 1, let

$$V(k) = [V_1(k), V_2(k), \ldots, V_p(k)] \tag{48}$$

be the column partition of V(k) where

$$V_i(k) = \frac{\partial x(k)}{\partial \alpha_i} \in R^n \tag{49}$$

is the forward sensitivity of the solution $x(k)$ with respect to α_i, $1 \leq i \leq p$. Then, the elements of $G(k)$ in (39) is given by the outer product.

$$G(k) = \begin{bmatrix} [E(k)V_1(k)]^T \\ [E(k)V_2(k)]^T \\ \vdots \\ [E(k)V_n(k)]^T \end{bmatrix} \left[E(k)V_1(k), E(k)V_2(k), \ldots, E(k)V_p(k)\right]. \tag{50}$$

The diagonal element $G_{ii}(k)$ is then given by

$$G_{ii}(k) =< E(k)V_i(k), E(k)V_i(k) >= ||E(k)V_i(k)||^2. \tag{51}$$

Again,

$$tr(G(k)) = \sum_{i=1}^{p} G_{ii}(k) = \sum_{i=1}^{p} ||E(k)V_i(k)||^2 = ||EV(k)||_F^2. \tag{52}$$

We can now consider the analog of the two strategies described above to place N observations in this case.

Case 3: $x(0) \neq \bar{x}(0)$ **and** $c \neq \bar{c}$: In this case $\delta x(0) \neq 0$ and $\delta\alpha \neq 0$ and $F(k) = [U(k), V(k)] \in R^{n\times(n+p)}$. Hence

$$G(k) = [E(k)F(k)]^T[E(k)F(k)] \in R^{(n+p)\times(n+p)} \tag{53}$$

$$L = \begin{bmatrix} E(k_1)V(k_1) \\ E(k_2)V(k_2) \\ \vdots \\ E(k_N)V(k_N), \end{bmatrix} \in R^{Nm\times(n+p)} \tag{54}$$

and

$$G = L^T L \in R^{(n+p)\times(n+p)}. \tag{55}$$

From Appendix C, it can be verified that (since p>0)

$$Rank(E(k)F(k)) = m \wedge n. \tag{56}$$

Let

$$F(k) = [F_1(k), F_2(k), \ldots, F_{n+p}(k)] \tag{57}$$

be the column partition of F(k) where

$$F_i(k) = U_i(k), \ if \ 1 \leq i \leq n, \tag{58}$$

$$F_i(k) = V_i(k), \ if \ n+1 \leq i \leq n+p. \tag{59}$$

Then, G(k) in (43) is given by the outer product

$$G(k) = \begin{bmatrix} [E(k)F_1(k)]^T \\ [E(k)F_2(k)]^T \\ \vdots \\ [E(k)F_{n+p}(k)]^T \end{bmatrix} \left[E(k)F_1(k),\, E(k)F_2(k), \ldots, E(k)F_{n+p}(k)\right]. \quad (60)$$

The diagonal elements

$$G_{ii}(k) =< E(k)F_i(k),\, E(k)F_i(k) >= ||E(k)F_i(k)||^2 \quad (61)$$

and

$$tr(G(k)) = \sum_{i=1}^{n+p} ||E(k)F_i(k)||^2 = ||E(k)F(k)||_F^2. \quad (62)$$

Again, in the light of the discussions in Sects. B.2 and B.3 in Appendix B, we can invoke the two strategies described above to place the required number of observations such that G is SPD.

6 Air-Sea Interaction Example

An important practical problem in air sea interaction is the return-flow phenomenon that takes place over the Gulf of Mexico in the cool season (November – April). During that time, cold fronts pass over the Gulf and the cold continental air is modified, warmed and moistened, through interaction with the warm sea surface.This modified air returns to the US coastal plain over a period of 2–3 days in response to low-level wind-direction changes associated with the eastward movement of the attending large scale anticyclone. A good example of the low-level air modification is seen in Fig. 2 that tracks cold air from its entry into the Gulf and its anticyclonic turning as it heads back to the coastal plain over a period of 2 d. The data displayed here came from a return flow episode in March 1988. The relatively cold continental air entered the Gulf just east of New Orleans, LA, and 2 d later it resided over the western Gulf just east of Brownsville, TX. The air temperature rose from 47°F to 69°F in the presence of sea surface temperatures that rose from 69°F to 77°F just south of 25°N latitude before it began to decrease to 72°F at the terminal point of the trajectory.

The example we choose is a simplification of the process displayed in Fig. 2, but it contains essential physics–namely, a boundary condition related to sea-surface temperature, an initial condition related to low-level air temperature when the continental air begins its journey over the Gulf, and the turbulent transport of heat from ocean to air at the interface. In this model, the independent variable is time and the dependent variables are air temperature and sea-surface temperature, x(t) and θ (a constant), respectively, along with a constant turbulent-transfer coefficient κ. These quantities are made non dimensional with scaling for temperature ($1°C$) and scaling

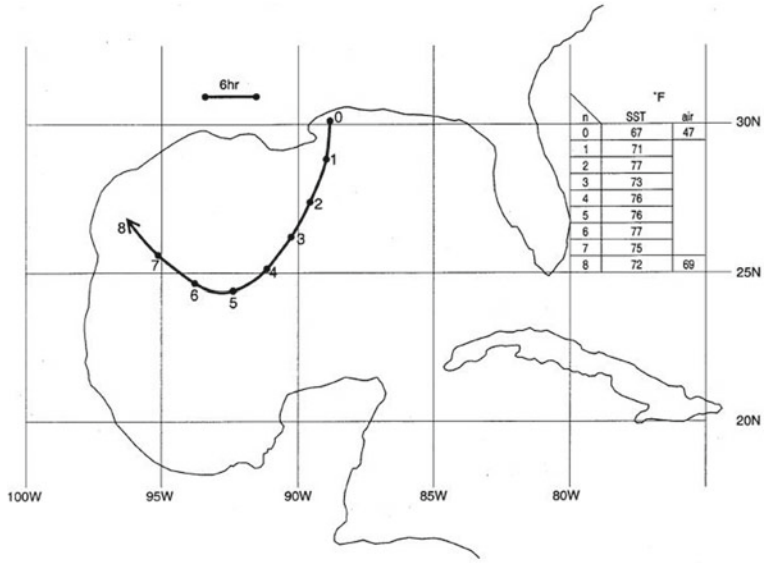

Fig. 2 A trajectory of low-level air that enters the Gulf just east of New Orleans and ends up off the coast of southern Texas 2-days later is representative of air movement during return flow events. Measurements of air temperature and seas surface temperature (SST) were made at 6-hour intervals and the location of these measurements are denoted by sequentially numbered dots along the trajectory. This return-flow event took place during the field phase of project GULF-MEX (Gulf of Mexico Experiment)

for time $(1\ hour)$. The governing non dimensional equation takes the form

$$\frac{dx(t)}{dt} = k\,(\theta - x(t)) \tag{63}$$

where $x_0 = x(0)$ is the initial temperature, turbulent transfer coefficient $\kappa > 0$, and the physical basis for air temperature change is a forcing governed by the positive air/sea temperature difference, $\theta - x(t)$. The solution is given by

$$x(t) = (x_0 - \theta)\, e^{-kt} + \theta \tag{64}$$

where the control elements are $(x_0,\ \theta,\ \kappa) = (10,\ 20,\ 0.25\ )$. In dimensional form, these elements of control are $10°C,\ 20°C,$ and $0.25\,h^{-1}$, respectively. The forecast is based on non dimensional controls $\left(x_0',\ \theta',\ \kappa'\right) = (11,\ 19,\ 0.30\ )$. The difference between true-state evolution and forecast evolution, i. e., forecast error, is shown in Fig. 3. Both states asymptotically approach the separate sea surface temperature controls.

The sensitivities of the air temperature to control are given by the formulas listed below:

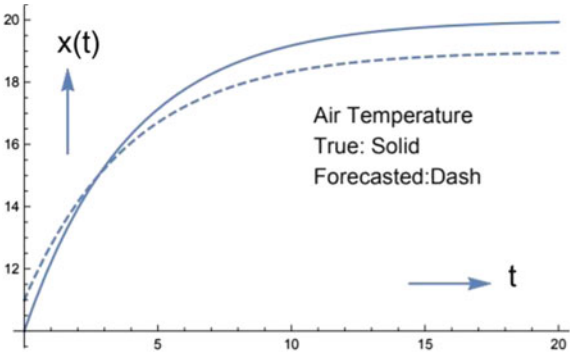

Fig. 3 The true and forecasted state of low-level air temperature as a function of time

$$\frac{\partial x}{\partial x_0} = e^{-kt},\ \frac{\partial x}{\partial \theta} = 1 - e^{-kt},\ \frac{\partial x}{\partial k} = t(\theta - x_0)e^{-kt}. \tag{65}$$

The structure of these sensitivities based on forecast control are displayed in Fig. 4.

In practice, true controls are unknown and sensitivities are expressed in terms of forecast controls. The strongest temperature sensitivity to x_0, θ, and k occur at early times, late times and intermediate times, respectively (relative to steady-state at $t \cong 15 - 20h$).

Based on an overlay of forecast error and the various sensitivities, those places in time where sensitivities coincide with large error are ideal places to make observations. This assumes that the model is faithful to the phenomenon and errors are not due to the absence of important physical processes in the governing constraint. It is clear from evolution of sensitivities that correction of controls through data assimilation will benefit from an observation at an early time (to recover the initial condition), an observation at the intermediate time to recover the turbulent transfer coefficient, and an observation at a late time to recover the sea surface temperature.

A significant advantage of G-matrix method for choosing observation sites is the unification of sensitivities in the norm of G rather than trying to consider sensitivities separately. However, the primary advantage of the G-matrix method is that determination of observation sites that bound the gradient away from zero at the operating point in the space of control is its independence of available observations. The method strictly depends on forecast sensitivity to controls in a unified manner. Another characteristic is that since G has an additive structure, the potential observation sites in space and time can be considered separately. That is, the forecast sensitivity at a particular point in control space and at a specific time determines one row in the sensitivity matrix, the F matrix. Thus, the G matrix can be constructed sequentially, one row at a time, or it can be constructed with all rows considered collectively.

Data Assimilation Experiments The Gramian G matrix is formed from the sensitivity matrix F and its transpose F^T. The forecast sensitivities depend on forecast control and time and were displayed earlier in Fig. 4. These sensitivities determine the F matrix, in this case an arbitrary observation location in time given by a 1×3

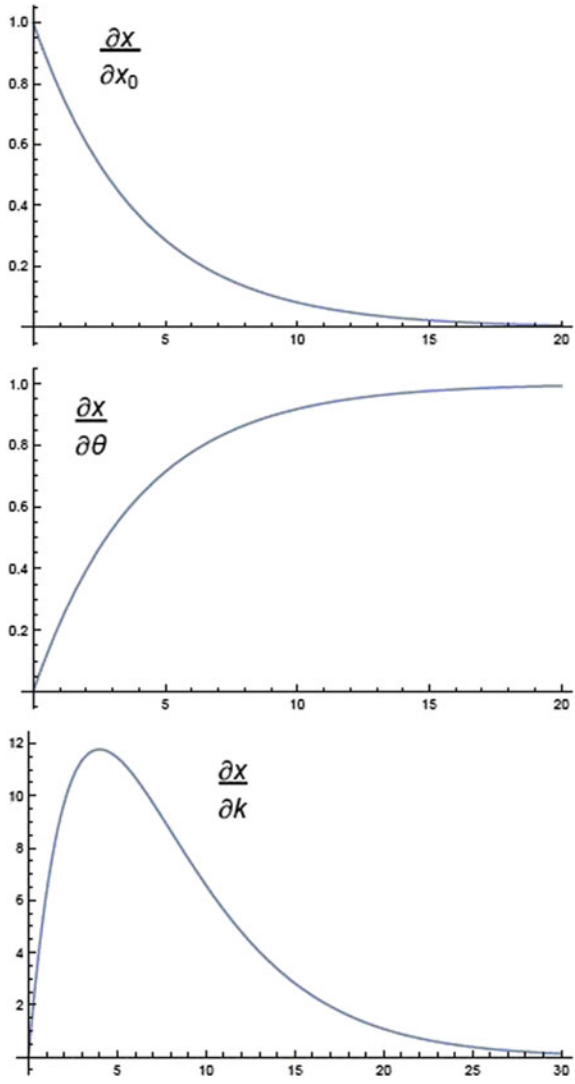

Fig. 4 Forecast sensitivity to the three elements of control: x_0 (initial temperature), θ (sea surface temperature), and turbulent transfer coefficient (κ)

row vector with elements given by the three sensitivities at this arbitrary time. The G matrix is the product of $F^T F$, a 3×1 column vector multiplied by a 1×3 row vector, and G is therefore a 3×3 matrix, each element a function of time alone. The diagonal elements are the squares of the sensitivities–$(\frac{\partial x(t)}{\partial x_0})^2$, $(\frac{\partial x(t)}{\partial \theta})^2$, $(\frac{\partial x(t)}{\partial \kappa})^2$. From this matrix, we consider two strategies to choose observation sites: Strategy I: observations are chosen in the zone where Norm of G is large. In Strategy II, observations are chosen where the diagonal elements of matrix G are separately large. For each experiment, three observation locations at times t_1, t_2, t_3 are chosen. From

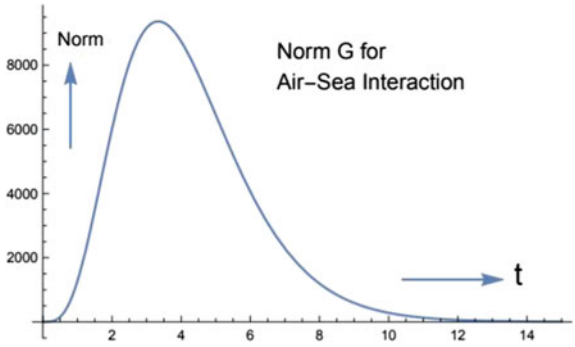

Fig. 5 G-matrix Norm as a function of time

these observation locations, a 3×3 sensitivity matrix F and its transpose F^T can be found for each strategy–each row of F associated with sensitivities at one of the three times. And from the resulting 3×3 G matrix, its positive definiteness and associated condition number can be found. As developed in foregoing theory of matrix G and its role in locating observation sites, these characteristics of the matrix are central to finding the observation locations that bound the cost-function gradient away from zero.

The air temperature observations at each location site are found by using true control in the analytic solution with added noise–normally distributed random error with mean value zero and standard deviation 0.1 (non-dimensional). The cost function is one-half the sum of squared differences between the forecasts (functions of forecast control once the three times are specified) and the numerical values of observations denoted by $Z_1(t_1),\ Z_2(t_2),\ Z_3(t_3)$. The cost function takes the form

$$J(x_0, \theta,\ \kappa) = 0.5 \sum_{i=1}^{3} (\, x_f\,(t_i) - Z\,(t_i)\,)^2)$$

where $x_f\,(t_i)$ and $Z\,(t_i)$ are the forecast and observation, respectively, at $t = t_i$. The cost function is minimized using gradient-based optimization–in particular, the Newton method that uses both gradient and Hessian at the sequence of operating points. Minimization of this cost function leads to forecast control adjustment, iterative adjustment starting at the forecast control point.

Strategy I: Norm of G used to locate observation sites The norm of G based on the arbitrary time version of $F^T F$ is shown in Fig. 5. This unification of the three separate sensitivity functions is intuitively expected. The structure basically cautions against choosing all observations at the very early times and/or very late times (that is, don't choose all observations at $t \leq 0.1$ *or* $t \geq 15$, for example). Accordingly, we choose observations at t=1, 4, 8. In this case, the eigenvalues of the G matrix are 163, 0.43, and 0.09, giving a condition number $\cong 2.10^3$. The 3-D cost function is displayed in Fig. 6 and the gradient-based optimization locates the minimum at

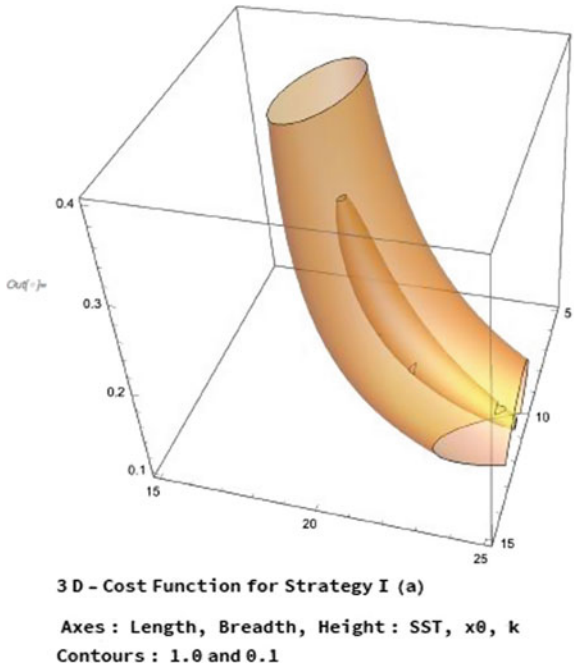

Fig. 6 Cost Function for Strategy I (a), Observations at t = 1, 4, 8

$(x_0,\ \theta,\ \kappa) = (10.3,\ 21.0,\ 0.20)$ which reduced the cost function value from 0.584 to 5.10^{-6}, essentially zero.

When observations were taken at t = 12, 13, and 14, the eigenvalues of the associated G matrix were 17, 0.08, and 6.10^{-8} giving a condition number of 3.10^8. The 3-D structure of the cost function is shown in Fig. 7 where there is no obvious minimum in the presence of contours in the form of nearly parallel sheets (Fig. 8).

Interestingly, if the observation site at $t = 12$ is replaced by one at $t = 4$, the result is improved significantly. In this case, the eigenvalues are 101, 1.32, and 5.10^{-5} so the condition number is only smaller by two orders of magnitude, but the structure of the cost function shown in Fig. 9 clearly indicates a minimum at $(x_0,\ \theta,\ \kappa) = (14.7,\ 31.2,\ 0.03)$ and a reduction in cost function value from 0.84 to 0.23.

Strategy II: Observation locations based on diagonal elements of G Plots of the diagonal elements of the arbitrary time version of $F^T F$ are shown in Fig. 10.

The striking difference in the plots is the magnitude of $(\frac{\partial x(t)}{\partial \kappa})^2$ compared to the magnitudes of $(\frac{\partial x(t)}{\partial x_0})^2$ and $(\frac{\partial x(t)}{\partial \theta})^2$–a 4 order of magnitude difference. One can work with a different scaling of parameters to bring the squares of the sensitivities more in line with one another, but the end result from the data assimilation step will not change with a change of scaling. And there is no loss of information in the present scheme where location sites are based on structure of diagonal elements separately. Thus, we choose sites at t= 0.5 (for the structure of $(\frac{\partial x(t)}{\partial x_0})^2$), t =20 (for

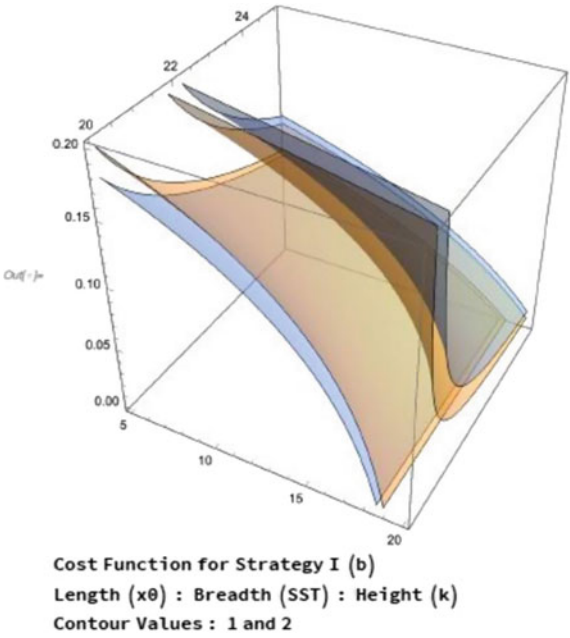

Fig. 7 Cost Function for Strategy I (b), Observations at t = 12, 13, 14

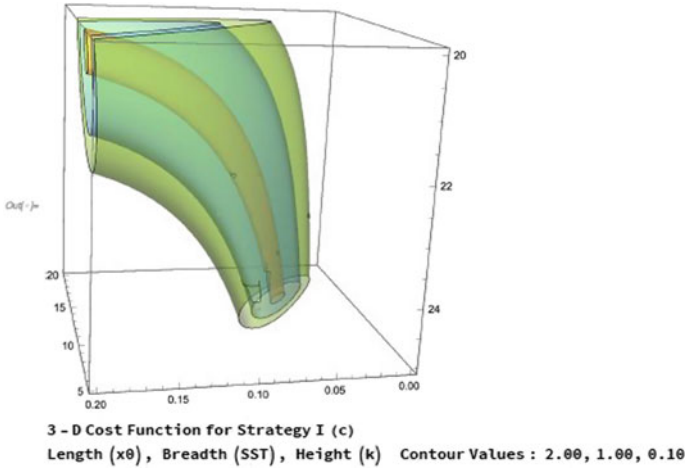

Fig. 8 Cost Function for Strategy 1 (c), Observations at t = 4, 13, 14

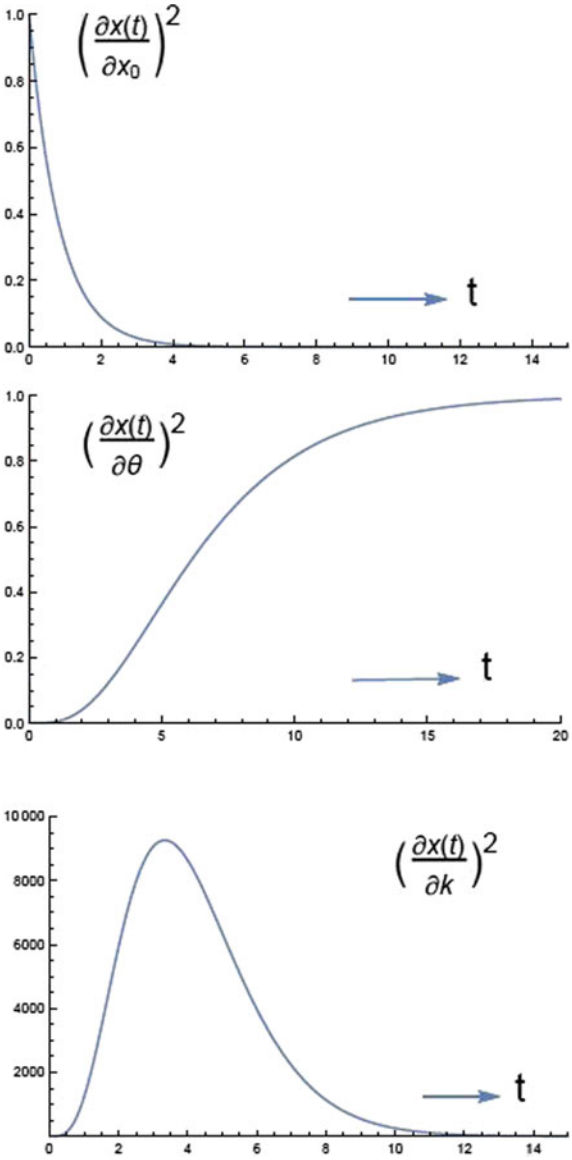

Fig. 9 Time evolution of matrix G's diagonal elements

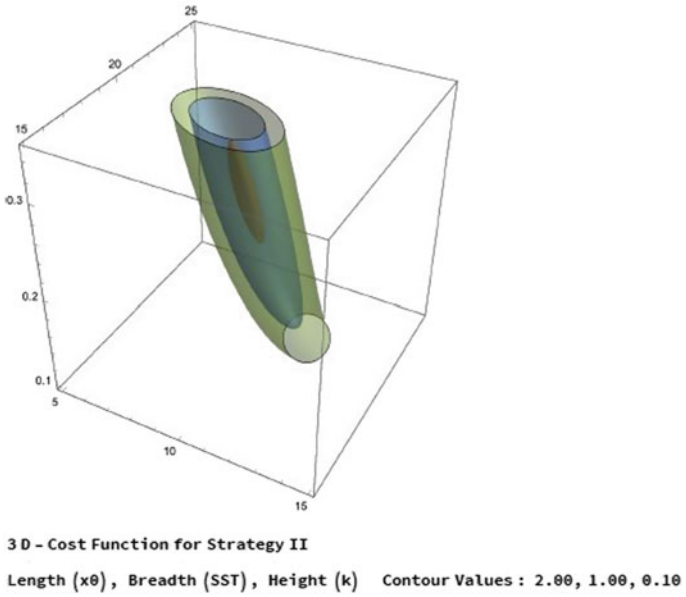

Fig. 10 Cost Function for Strategy II, Observations at t = 0.5, 3.0, 20.0

the structure of $(\frac{\partial x(t)}{\partial \theta})^2$) and t = 3 (for the structure of $(\frac{\partial x(t)}{\partial \kappa})^2$). The choices of t are $t_1,\ t_2,\ t_3 = 0.5,\ 20.0\ and\ 3.0$. The eigenvalues of the G matrix are 108.09, 0.96, and 0.45 that yields a condition number of 240.

The plot of the 3-D cost function in this case is found in Fig. 10. where the gradient-based optimization procedure located the minimum at $(x_0,\ \theta,\ \kappa) = (9.95,\ 19.89,\ 0.27)$. The cost-function reduction went from 0.800 to 0.002.

7 Conclusions

The earlier work by Lakshmivarahan et al. (2020a) developed a theory for observation placement in dynamic data assimilation—a sensitivity-based methodology. The current work is an extension of this earlier study. The difference between 4D-Var (Four-Dimensional Data Assimilation) and this sensitivity-based approach is essentially philosophical where the mechanics of the two methods have much in common and should be viewed as complementary as opposed to competitive.

In 4D-Var, all available observations are used to determine the cost function gradient in a most efficient manner based on the adjoint method. Once the gradient is found at a guessed operating point in control space, the standard gradient-based optimization schemes are used to iterate toward the cost function minimum.

In the sensitivity-based methodology, observation locations are first identified through examination of the G-matrix norm where elements of this matrix are functions of control elements, space, and time. Specification of the forecast control elements (the first operating point in the space of control) delivers a G-Norm that is a function of space and time. Those locations in space and time where the norm exhibits the largest values are the best locations for observations. Note that actual observations, if available, are not used in this first step. In the second step, observations are chosen sequentially, one at a time since the construction of the G matrix is additive–all possible observations can be considered simultaneously, or one at a time. The advantage of examining the G matrix in the additive mode, i.e., one at a time, is that the condition number and positive-definiteness of the matrix can be examined sequentially. When the matrix becomes positive definite, the minimum required number of observations is determined. Thus, the sensitivity-based methodology delivers placement of observations and the minimum requisite number of observations that hold promise for finding the minimum of the cost functional through gradient-based optimization as used in 4D-VAR or another minimization-search method. The complementarity of 4D-Var and the sensitivity-based method is established.

The principles underlying this sensitivity-based approach to data assimilation are tested on an air-sea model where the interplay between an initial condition, a boundary condition, and an empirical parameter gives substance to the assimilation process. In a series of numerical experiments that choose observation sites with large and small gradients about the operating points, the results convincingly show that location choices that avoid flat zones in the gradient of the cost function deliver good results whereas choices that allow small gradients around the operating point face difficulty that can only be alleviated by including an observation that contributes to a larger gradient. Further numerical investigations of the theory developed here are contained in the companion paper by Lewis et al. (2020b) in this volume. Based on these encouraging results, there is a current plan to use this methodology in real-time forecasting of return flow in the Gulf of Mexico. Preliminary results for a historical case of return flow has produced excellent results where the optimal observation locations are found to be those over the warmest SST's in the Gulf (Lewis et al., 2020a, b).

In addition to the numerical experiments, this study also included a theoretical component that investigated the intrinsic connection between the concept of Observability in control theory and dynamic data assimilation.

Acknowledgements We are grateful to an anonymous reviewer for bringing the papers by Krener (2008a, b); Krener and Ide (2009); Kang and Xu (2012, 2014), King et al. (2015) and Yoshimura et al. (2020) related to a classification - strong and weak observability, and their applications to our attention.

Appendix A Role of observability in Estimation

A.1 Historical Background

Kalman is a series of papers (Kalman 1960a, 1963) and Kalman et al. (1969) laid the foundations of the state space approach to modern Control/Systems Theory by introducing several basic concepts - controllability/reachability, observability/constructability, realizability and stability, all related to the analysis, design and (optimal) control of engineering systems. This Appendix provides a short summary - a bird's eye view, of the role of observability in state/parameter estimation problem that is critical to both Control Theory and Dynamic Data Assimilation. For a more elaborate treatment of these concepts and their applications refer to the two volumes by J. L. Casti - Casti (1977) for linear analysis and Casti (1985) for the nonlinear counterpart.

It is useful to broadly divide the problems in Control Theory into two classes: open-loop and closed-loop/feedback control. Many of the household appliances - washer/dryer, microwave owen, light bulb, bread toaster, to name a few, implement the open loop control strategy where the control action is limited to a simple on or off switch to execute a preprogrammed task.

Feedback control, on the other hand, involves comparing the current state of a system with a prespecified reference value. If the error = (reference - current state) is positive, the controller generates an extra input/forcing that forces the current state towards the reference. If the error is negative, then the controller lets the system relax to the reference, without any extra forcing. Examples of feedback controlled devices are too numerous-the fly ball governor in Steam Engines, the pressure cooker in the kitchen, cruise control in automobiles, thermostat control of temperature in a building, sophisticated avionics in aircraft flight control, etc.

From the above discussion, it should now be obvious that a fundamental requirement in the design of feedback control relates to the ability to measure the current state of a system being controlled. However, except in special cases, the current state may not be directly observed but can measure only certain (scalar or vector valued) functions - called output in Engineering and observations in geosciences, of the state in question. Roughly speaking, observability relates to the ability to estimate/reconstruct a past state from the future observations or outputs.

Despite its origin in Control/Systems theory, observability plays an important role in the estimation of the initial conditions and parameters of a dynamical model that arise within the context of the 4-dimensional variational (4-D VAR) approach to dynamic data assimilation, which is our primary interest.

A.2 Observability: Linear, Deterministic, Time Invariant Model

We follow the notations laid out in the main body of the paper. Consider a linear, deterministic, time invariant, discrete time model given by

$$x(k+1) = Mx(k) \tag{66}$$

with $x(0)$ is the unknown initial condition, where $M \in R^{n\times n}$ is the one step state transition matrix, assumed to be non-singular. Solving (66), it is obvious that

$$x(k) = M^k x(0). \tag{67}$$

Let $H \in R^{m\times n}$ and

$$z(k) = Hx(k) = HM^k x(0). \tag{68}$$

be the noiseless observations of the state $x(k)$.

It is assumed that we have a set $S = \{z(1), z(2), \ldots, z(N)\}$ of N outputs and our goal is to estimate $x(0)$ using S. To this end, we stack the N output vectors in a column to create a new vector $z(1:N) \in R^{Nm}$ given by

$$z(1:k) = \begin{bmatrix} z(1) \\ z(2) \\ \vdots \\ z(N) \end{bmatrix} = \begin{bmatrix} HM \\ HM^2 \\ \vdots \\ H(M^N) \end{bmatrix} x(0) = Lx(0) \tag{69}$$

where $L \in R^{Nm\times n}$. By Cayley-Hamilton theorem, since M^n can be expressed as linear combinations of M^k for $0 \leq k < n$, we only need to consider $N < n$.

A necessary and sufficient condition for the existence and uniqueness of the solution $x(0)$ satisfying (69) is that the matrix L must be a full rank matrix, that is, rank$(L) = n$. In this case, we say that the matrix pair (M, H) is observable and we can recover $x(0)$ exactly by solving

$$G(N)x(0) = L^T z(1:k) \tag{70}$$

where $G(N) = (L^T L) \in R^{n\times n}$ is called the observability Gramian and is given by

$$G(N) = \sum_{k=1}^{N} (M^T)^k (H^T H) M^k. \tag{71}$$

Indeed, this Gramian $G(N)$ is symmetric and positive definite when L is of full rank, and the solution, $x(0)$ is given by

$$x(0) = (L^T L)^{-1} L^T Z(1:k) \tag{72}$$

We leave it to the reader to verify the following claims by computing $H = \begin{bmatrix} HM \\ HM^2 \end{bmatrix}$

Claim 1: Let $M_1 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $H_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Then $H_1 M_1 = (1, 1)$ and $H_1 M_1^2 = (1, 3)$ and $\begin{bmatrix} H_1 M_1 \\ H_1 (M_1)^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$ which is of rank = 2 and hence (M_1, H_1) is observable.Refer to Example 4.1 in Sect. 4 for more details.

Claim 2: Let $M_2 = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$ and $H_2 = (1, 0)$. Then $H_2 M_2 = (1, 0)$ and $H_2 M_2^2 = (1, 0)$. Then $\begin{bmatrix} H_2 M_2 \\ H_2 (M_2)^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ which is of rank = 1. Hence, (M_2, H_2) is not observable. We leave it to the reader to verify that with $H_3 = (0, 1)$, the pair (M_2, H_3) is observable.

A.3 Generalizations

For completeness, we now enlist several extensions of the above result as Remarks with citations to the appropriate literature.

Remark 3 Linear time invariant model with noisy observation: Consider the model in (66) but the observations are subjected to additional Gaussian noise where

$$z(k) = Hz(k) + \xi_k \tag{73}$$

where $\xi_k \sim N(0, R_k)$ and ξ_k is temporally uncorrelated. In this case. the least squares solution is obtained by minimizing the weighted sum of the squared errors, given by

$$J(x(0)) = \sum_{k=1}^{N} [z(k) - HM^k x(0)]^T R_k^{-1} [z(k) - HM^k x(0)]. \tag{74}$$

It can be verified (Chap. 5, LLD (2006)) that the minimizer $\hat{x}(0)$ is obtained as the solution of the linear system.

$$G(N)x(0) = \sum_{k=1}^{N} M^k H^T R_k^{-1} Z(k) \tag{75}$$

when the observability Gramian G(N) is given by

$$G(N) = \sum_{k=1}^{N} (M^T)^k)[H^T R_k^{-1} H] M^k \tag{76}$$

is positive definite where $(M^k)^T = (M^T)^k$.

Remark 4 Linear, deterministic, time varying system with noiseless observations: This case is treated in full in Chap. 4, Casti (1977) where several examples are also given. Extension to noisy observations can be easily obtained by following along the strategy described in Remark 3.

Remark 5 Non-linear deterministic systems–Local observability: While the observability analysis of linear model is intrinsically global (no constraint on $x(0)$), that of nonlinear system can be viewed from a local or a global point of view. Both of these cases are treated in Chap. 5 of Casti (1985). Local analysis relies on the observability of the first-order variational equation and applying the conditions in Chap. 4 of Casti (1977) referred to in Remark 4. To wit, let

$$x(k+1) = M(x(k)) \tag{77}$$

be the nonlinear model with $x(0)$ as the initial condition and

$$z(k) = h(x(k)) \tag{78}$$

be the observation.

One approach is to linearize (77)—(78) about a base trajectory starting from an arbitrarily chosen initial state $y(0)$. Let $\delta x(0) = x(0) - y(0)$ be the perturbation super imposed on $y(0)$. Then the dynamics of $\delta x(k) = x(k) - y(k)$ is given by variational equation which is linear, time varying dynamics:

$$\delta x(k+1) = D_M(k)\delta x(k) \tag{79}$$

with $\delta x(0)$ as its initial condition and the induced variation in $z(k)$ is given by

$$\delta z(k) = D_h(k)\delta x(k) \tag{80}$$

where $D_M(k)$ and $D_h(k)$ are the Jacobians of $M(y(k))$ and $h(y(k))$. It can be verified that $\delta x(0)$ can be estimated by minimizing a sum of squared error criterion similar to (74). The resulting $\delta x(0)$ is obtained by solving

$$G(N)\delta x(0) = \sum_{k=1}^{N} D_h^T(k) D_M(k) \delta z(k) \tag{81}$$

where

$$G(N) = \sum_{k=1}^{N} D_M^T(k)[D_h^T(k) D_h(k)] D_M(k) \tag{82}$$

is the required Gramian. Indeed, we can recover $\delta x(0)$ provided $G(N)$ in (82) is positive definite.

Remark 6 Nonlinear deterministic system: Global observability: Analysis of global observability of a non linear model is considerably more involved and requires concepts and tools from differential geometry. An exquisite expose' of this topic is contained in Casti (1985) and in Chap. 7, Vidyasagar (2020). For more detailed treatment, refer to Isidori (1985) and Nijmeijer and van der Schaft (1990).

Remark 7 Linear and non-linear filtering: Kalman in another epoch making paper (Kalman (1960b)) developed a sequential method of estimating the state of a stochastic, linear, dynamical model when the observations are linear but noisy, called the Kalman filtering. Extensions to non-linear stochastic models with noisy nonlinear observations are known since the early 1960s. Refer to Kushner (1964a, b, 1967) and Chap. 5, Casti (1985) for a quick summary of results in nonlinear filtering. The handbook on"Nonlinear Filtering" by Crisan and Rozovskii contains a comprehensive treatment of this and related topics.

Appendix B Results from matrix theory

For completeness and ease of reference, we collect a set of results from Matrix Theory that are basic to the developments in this paper. For detailed proofs refer to Meyer (2000); Bapat (2012).

B.1 Solution of linear systems: Let $A \in R^{m\times n}$ be a linear map from R^n to R^m. The range of A, denoted by $Range(A)$, is the subspace of R^m generated by the linear combination of the columns of A. Thus, $Range(A) \subseteq R^m$ and

$$Range(A) = \{y \in R^m | y = Ax\,, x \in R^n\}. \tag{83}$$

The null space of A, denoted by $Null(A)$ is the set of vectors in R^n annihilated by A. That is, $Null(A) \subseteq R^n$ and

$$Null(A) = \{x \in R^n | Ax = 0\}. \tag{84}$$

The rank of A, denoted by $Rank(A)$, is the number of linearly independent columns or equivalent by the number of linearly independent rows of A. Clearly,

$$Rank(A) \leq m \wedge n = min\ (m, n) \tag{85}$$

If equality holds in (85), then A is said to be of full rank, otherwise, it is rank deficient. In the following, we catalog the conditions for the existence and uniqueness of the solution of the linear system.

$$Ax = b \tag{86}$$

where $x \in R^n$ and $b \in R^m$. The system (86), given A and b, is said to be consistent if there exists a vector $x \in R^n$ that satisfies (86), otherwise, it is inconsistent. For

example, the homogeneous system.

$$Ax = 0 \tag{87}$$

is always consistent, since $x = 0$ satisfies it. But the non-homogeneous system in (86), depending on the properties - relative location of b in R^m, may or may not be consistent.

If the system is consistent (b ∈ Range (A)), then we can talk about the solution in the traditional sense where the residual, $r(x) = b - Ax = 0$. On the other hand, if (86) is inconsistent (b ∉ Range (A)), then we have to contend with the so called least squares solution—that minimizes the square of the length of the non-zero residual vector $r(x)$.

The functional form and uniqueness of the solution of (86) critically depends on two factors: (a) relative values of m and n and (b) the rank of A. For brevity, we only consider the case when A is of full rank.

Case B.1.1: Let $m = n$ and $Rank(A) = n$. Then, A is non-singular and the solution of (86) is given by

$$x = A^{-1}b. \tag{88}$$

Case B.1.2: Let $m > n$ and $Rank(A) = n$. In this case unique solution of (86) is given by

$$x = A^{+}b \tag{89}$$

where

$$A^{+} = (A^T A)^{-1} A^T \in R^{n \times m} \tag{90}$$

called the generalized or Moore-Penrose inverse of A that satisfies the following conditions:

$$\begin{aligned} AA^{+}A = A, (A^{+}A)^T = A^{+}A \\ A^{+}AA^{+} = A^{+}, (AA^{+})^T = AA^{+}. \end{aligned} \tag{91}$$

The matrices $A^T A \in R^{n \times n}$ and $AA^T \in R^{m \times m}$ are called Gramians of A. When $Rank(A) = n$, $(A^T A)$ is a symmetric and positive definite matrix. It can be verified that $A^{+}A = I_n$ and $AA^{+} = A(A^T A)^{-1} A^T$ is the orthogonal projection matrix onto the range of A.

Case B.1.3: Let $m < n$ and $Rank(A) = m$. In this case, there are infinitely many solutions of (86) and the one with minimum norm is given by

$$x = A^{+}b \tag{92}$$

where

$$A^{+} = A^T (AA^T)^{-1} \in R^{n \times m} \tag{93}$$

is the generalized inverse of A that satisfy (91). It can be verified $AA^+ = I_m$ and $A^+A = A^T(AA^T)^{-1}A$ is the orthogonal projection onto the range of A^T.

B.2: Rank of the partitioned matrix G: We start by stating a general result relating to partitioned symmetric matrices. Let

$$S = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{94}$$

be a symmetric matrix with both $A \in R^{q\times q}$ and $C \in R^{r\times r}$ symmetric and $B \in R^{p\times q}$. Let A be non-singular. If I_K denotes an identity matrix of order K, then

$$P = \begin{bmatrix} I_p & 0 \\ -B^T A^{-1} & I_q \end{bmatrix} \tag{95}$$

is non-singular, since $Det(P) = 1$. By direct multiplication, it can be verified that

$$PSP^T = \begin{bmatrix} A & 0 \\ 0 & C - B^T A^{-1} B. \end{bmatrix} \tag{96}$$

where $C - B^T A^{-1} B$ is called the Schur Complement of A in S. The following claim is easily proved. (Chap. 3, Bapat (2012))

Claim

1. If S is SPD, then so is $C - B^T A^{-1} B$.
2. Let S be symmetric. If S is positive definite then so are A and $C - B^T A^{-1} B$
3. $Det(S) = Det(A)Det(C - B^T A^{-1} B)$.

Now consider the symmetric matrix $G \in R^{(n+p)\times(n+p)}$ in its partitioned form given by

$$G = \begin{bmatrix} U^T \bar{H} U & U^T \bar{H} V \\ V^T \bar{H} U & V^T \bar{H} V. \end{bmatrix}. \tag{97}$$

Recall that $\bar{H} = D_h^T R^{-1} D_h$ is symmetric where $D_h \in R^{m\times n}$ and $R^{-1} \in R^{m\times m}$ is assumed to be non-singular. Under the assumptions $m \geq n$ and D_h is of full rank, it follows that $\bar{H}$ is non-singular. If in addition, U is non-singular then, $A = U^T \bar{H} U$ is symmetric, positive definite and hence non-singular. Then identifying $B = U^T \bar{H} V$ and $C = V^T \bar{H} V$, it can be verified that the Schur Complement of $A = U^T \bar{H} U$ in G reduces to a zero matrix of size $r \times r$. That is, the matrix on the right hand side of (B.14), becomes

$$PGP^T = \begin{bmatrix} U^T \bar{H} U & 0 \\ 0 & 0 \end{bmatrix} \tag{98}$$

which is a matrix of rank n. That is, $G(k) \in R^{(n+p)\times(n+p)}$ of the forward sensitivity matrices $U(k)$, $V(k)$ and the Jacobian $D_h(k)$ at time k, is a rank deficient matrix of rank n.

B.3. Rank of the sum $\sum_{k=1}^{t} G(k)$, for some integer t>0:

Let $G \in R^{(n+p)\times(n+p)}$ be a symmetric matrix of rank n. Then there exists are orthogonal matrix $Q \in R^{(n+p)\times(n+p)}$ such that

$$Q^T G Q = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{99}$$

where $QQ^T = Q^T Q = I_{n+p}$, $D = diag(\alpha_1, \alpha_2, \ldots, \alpha_n)$ with

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n > 0 \tag{100}$$

Now we can build a matrix $\bar{G}$ as

$$\bar{G} = Q \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix} Q^T. \tag{101}$$

Then, it can be verified that

$$Q^T (G + \bar{G}) Q = \begin{bmatrix} D & 0 \\ 0 & I_p \end{bmatrix}$$

Hence,

$$(G + \bar{G}) = Q \begin{bmatrix} D & 0 \\ 0 & I_p \end{bmatrix} Q^T$$

is a full rank matrix of rank$(n + p)$. Stated in other words, by adding a suitably designed matrix $\bar{G}$ of rank p to the matrix G, we can create a matrix of full rank.

Recall that Gramian G by definition is SPSD and in general $Null(G) \neq \emptyset$ and $DIM(Range(G)) + DIM(Null(G)) = n$.

B.4. Verification of (4.17):

If G is SPD, then there exists an eigen decomposition of G given by

$$G = QDQ^T \tag{102}$$

where Q is an orthogonal matrix of eigenvectors and $D = diag(d_1, d_2, d_3, \ldots, d_n)$ is a diagonal matrix of the corresponding eigenvalues of G where $Q^T Q = QQ^T = I$ and

$$d_1 \geq d_2 \ldots \geq d_n > 0. \tag{103}$$

It then follows from (102) that for $k \geq 1$

$$G^k = QD^k Q^T. \tag{104}$$

Now define $\eta = Q^T \hat{f}$. Then, since Q is orthogonal, we get

$$\sum_{i=1}^{n} \eta_i^2 = ||\eta||^2 = ||Q^T \hat{f}||^2 = ||\hat{f}||^2 = 1. \tag{105}$$

Thus, we can interpret $\{\eta_i^2\}$ as the probability distribution of a random variable d where

$$Prob[d = d_i] = \eta_i^2 \tag{106}$$

Consequently,

$$< \hat{f}, G^k \hat{f} >=< \hat{f}, QD^k Q^T \hat{f} >=< \eta, D^k \eta >= \sum \eta_i^2 d_i^k = \mu_k \tag{107}$$

the k^{th} (non-central) moment of the random variable d. Since

$$Var(d) = \mu_2 - \mu_1^2 \geq 0,$$

we get $\mu_2^{1/2} \geq \mu_1$. Consequently, from

$$\frac{\mu_1}{\mu_2^{1/2}} = \frac{< \hat{f}, G\hat{f} >}{< \hat{f}, G^2 \hat{f} >}^{1/2} \leq 1,$$

claim (17) follows.

B.5. Spectral Radius of $(I - \beta G)$ in (19)

From (20), using (104) and (107), we get

$$\beta = \frac{f^T G^2 f}{f^T G^3 f} = \frac{\hat{f}^T G^2 \hat{f}}{\hat{f}^T G^3 \hat{f}} = \frac{\sum \eta_i^2 d_i^2}{\sum \eta_i^2 d_i^3}. \tag{108}$$

Hence, using (108), from

$$[I - \beta G] = [I - \beta Q D Q^T] = Q[I - \beta D]Q^T,$$

it can be verfied that the eigenvalues of $[I - \beta G]$ are

$$1 - \beta d_i = 1 - d_i \frac{\sum \eta_i^2 d_i^2}{\sum \eta_i^2 d_i^3} \tag{109}$$

The spectral radius, inview of (103), is

$$\rho(I - \beta G) = max\{1 - d_i \frac{\sum \eta_i^2 d_i^2}{\sum \eta_i^2 d_i^3}\} \leq 1 - d_n \frac{\sum \eta_i^2 d_i^2}{\sum \eta_i^2 d_i^3} \leq 1 \tag{110}$$

since

$$d_n \sum \eta_i^2 d_i^2 \leq \sum \eta_i^2 d_i^3.$$

Appendix C Conditions for the Matrix L in (9) to be of full rank

Let $A \in R^{n\times n}$ be a non-singular, diagonalizable matrix. Let $Q = [q_1, q_2, \ldots, q_n] \in R^{n\times n}$ and $D = Dia(d_1, d_2, \ldots, d_n) \in R^{n\times n}$ be the matrices of eigenvectors and the corresponding eigenvalues of A. Then, by definition

$$AQ = QD \tag{111}$$

and the columns of Q are linearly independent and constitute a basis for R^n.

Let $b \in R^n$ and define the Krylov sequence.

$$K_p(A, b) = \{b, Ab, A^2b, \ldots, A^{p-1}b\} \tag{112}$$

for $1 \leq p \leq n$. The space generated by the columns in $K_p(A, b)$ is called Krylov subspace and is denoted by $Span\{K_p(A, b)\}$. Let

$$[K_p(A, b)] = [b, Ab, A^2b, \ldots, A^{p-1}b] \in R^{n\times p} \tag{113}$$

be corresponding Krylov matrix. Let

$$S_k = \{q_{i_1}, q_{i_2}, \ldots, q_{i_k}\} \tag{114}$$

be a k-subset of eigenvectors of A, for $1 \leq k \leq n$. Then, $DIM(Span(S_k)) = k$ and the $Span\{S_k\}$ is an invariant subspace of A. That is, if $b \in Span\{S_k\}$ then so is Ab. It can be verified that if $y \in S_k$, that is, y an eigenvector of A then, from

$$Span\{K_p(A, y)\} = Span\{y\} \tag{115}$$

it follows that the dimension of the Krylov subspace is one. Stated in words, if b is an eigenvector of A, since the vector Ab is a constant multiple of b, the the dimension of the Krylov subspace in (115) is one.

Let the energy in a vector $b \in R^n$, be measured by the square of its norm:

$$||b||^2 = \sum_{j=1}^{n} b_j^2. \tag{116}$$

Let $\bar{b} \in R^n$ be the new coordinates representation of b in the new basis defined by the eigenvectors of A. That is

$$b = Q\bar{b} \tag{117}$$

If, for some $1 \leq j \leq n$, $\bar{b}_j = 0$, then we say that b has no energy along the j^{th} eigen direction q_j of A. That is, b belongs to the invariant subspace of dimension $(n - 1)$

defined by the rest of all eigenvectors, $q_i \neq q_j$. Stated in other words, if $\bar{b}$ has no zero (row) element, then the energy in b is distributed across all the eigen directions of A. This discussion leads to the following:

Property C.1: Expanding Krylov Subspace:
If $b = Q\bar{b}$ is such that $\bar{b}$ has no zero (row) element, then for $1 \leq p \leq n$

$$DIM\{Span\{K_p(A, b)\}\} = p. \tag{118}$$

By way of generalizing the above property, now consider

$$B = [b_1, b_2, \ldots, b_m] \in R^{n\times m} \tag{119}$$

a full rank matrix for some $1 \leq m \leq n$. Then, we can extend the Krylov subspace using B in place of b as

$$K_p(A, B) = \{B, AB, A^2B, \ldots, A^{p-1}B\}. \tag{120}$$

Clearly

$$Span\{K_p(A, B)\} = \cup_{j=1}^{m} Span K_p(A, b_j). \tag{121}$$

Let $\bar{B} \in R^{n\times m}$ be such that

$$B = Q\bar{B} \tag{122}$$

and let $\bar{B}$ has no rows of zeros. Then, it can be easily verified that the total energy in B as measured by the Frobenius norm

$$||B||_F^2 = \sum_{j=1}^{m} ||b_j||^2$$

is distributed across all eigen directions of A. This leads to the following:

Property C.2: Expanding Krylov subspace: If $B \in R^{n\times n}$ is such that $B = Q\bar{B}$ and $\bar{B}$ has no zero rows, then

$$DIM\{Span\{K_p(A, B)\}\} = mp \tag{123}$$

for $1 \leq p \leq \lceil \frac{n}{m} \rceil$. That is, for $p = \lceil \frac{n}{m} \rceil$, the dimension of the Krylov subspace $K_p(A, b)$ is n.

Observability of (M, H)- pair: Now consider the observability matrix $L \in R^{Nm\times n}$ given in (4.9). Then L^T is related to a Krylov matrix given by

$$L^T = M^T K_N(M^T, E^T) \tag{124}$$

where

$$K_N(M^T, E^T) = [E^T, M^T E^T, (M^T)^2 E^T, \ldots, (M^T)^{N-1} E^T]. \tag{125}$$

Then Property C.2 immediately suggests an answer to the question: when is L in (4.9) a full rank matrix?.

Corollary 1 *Let E^T be such that its total energy is distributed across all of the eigen directions of M^T. Then, setting $p = N$, $A = M^T$, and $B = E^T$ in (123), it follows from (123) that*

$$Rank(K_N(M^T, E^T)) = Nm \tag{126}$$

for some N in the range $1 \leq N \leq \lceil \frac{n}{m} \rceil$

Corollary 2 *Rank of product matrices (Meyer (2000), Chap. 4) If $B \in R^{m\times n}$ and $C \in R^{n\times p}$ then*

$$Rank(B) + Rank(C) - n \leq Rank(BC) \leq Rank(B) \wedge Rank(C). \tag{127}$$

Using the fact that $Rank(M^T) = Rank(M)$, it immediately follows that the Rank of $K_N(M^T, E^T)$ for $N = \lceil \frac{n}{m} \rceil$ is n. Hence, by the Property C.4, the observability matrix L^T and its transpose, L are of full rank. Consequently, the observability Gramian $G = L^T L$ in (10) is symmetric and positive definite (SPD). Stated in other words, the condition for G to be SPD rests entirely on the choice of the E (with respect to M) in the sense that the distribution of the total energy in all of the columns of E^T must be spread across all the eigen directions of M^T. Clearly the choice of E depends on the forward operator H and the noise covariance, R as defined in (8)-(9). From (9) and (10), recall that

$$G = L^T L = \sum_{k=1}^{N} G(k) \tag{128}$$

where

$$G(k) = (EM^k)^T (EM^K). \tag{129}$$

Now, column partition EM^K as

$$EM^K = [\eta_1(k), \eta_2(k), \ldots, \eta_n(k)] \tag{130}$$

where $\eta_i(k) \in R^m$. Then, it can be verified that the $(i, j)^{th}$ element of the outer product matrix $G(k)$ is given by

$$[G(k)]_{ij} = < \eta_i(k), \eta_j(k) > . \tag{131}$$

Consequently, the diagonal elements of G are given by

$$[G]_{ii} = \sum_{k=1}^{N} ||\eta_i(k)||^2. \tag{132}$$

Special Case: Setting $H = I_n$, and $R = I_n$, we get $E = I$ and $G(k) = (M^T)^k M^k = U^T(k)U(k)$. In this case, $\eta_i(k) = U_i(k)$, the ith column of the forward sensitivity matrix $U(k)$. Consequently,

$$[G]_{ii} = \sum_{k=1}^{N} ||U_i(k)||^2. \tag{133}$$

Now recall that $U_i(k) = \frac{\partial x(k)}{\partial x_i(0)} \in R^n$ is the vector of sensitivity of $x(k)$ with respect to the ith component of the initial condition. From (133) it immediately follows that, by placing the observations where the sum in (133) is a maximum with respect to k, we can indeed control the condition number of G.

References

Bapat RB (2012) Linear algebra and linear models, 3rd edn. Hindustan Book Agency, New Delhi

Casti JL (1977) Dynamical systems and their applications: linear theory. Academic Press, New York

Casti JL (1985) Nonlinear system theory. Academic Press, New York

Crisan D, Rozovskii B (2011) The Oxford handbook of nonlinear filtering. Oxford University Press, London

Isidori A (1985) Nonlinear control systems. Springer, New York

Kalman RE (1960a) On the general theory of control systems. In: Proceedings of the first IFAC congress, Moscow, pp 481-492

Kalman RE (1960b) A new approach to linear filtering and prediction problem. Trans Am Soc Mech Eng, J Basic Eng Ser D 83:35–45

Kalman RE (1963) Mathematical description of linear dynamical system. SIAM J Control 1:152–192

Kalman RE, Falb P, Arbib M (1969) Topics in mathematical system theory. McGraw Hill, New York

Kang W, Xu L (2012) Optimal placement of mobile sensors for data assimilation. Tellus A 64:17133

Kang W, Xu L (2014) Partial observability for some distributed parameter systems. Int J Dyn Control 2(4):587–596

King S, Kang W, Xu L (2015) Observability for optimal sensor locations in data assimilation. Int J Dyn Control 3:416-424

Krener AJ (2008a) Observability of vortex flows. In: Proceedings of the forty seventh IEEE conference on decision and control, Cancun, Mexico

Krener AJ (2008b) Eulerian and Lagrangian observability of point vortex flows. Tellus A 60:1089–1102

Krener AJ, Kayo I (2009) A quantitative measures of obserability. In: Proceedings of the IEEE on decision and control, Shanghai, China, pp 6413–6418

Kushner HJ (1964a) On the dynamical equations of conditional probability density functions with applications to optimal stochastic control. J Math Anal Appl 8:332–344

Kushner HJ (1964b) On the differential equations satisfied by conditional probability densities of Markov process. SIAM J Control 2:106–119

Kushner HJ (1967) Dynamical equations for optimal nonlinear filtering. J Diff Equations 3:179–190

Lakshmivarahan S, Honda Y, Lewis JM (2003) Second-order approximation to 3-D VAR cost function: applications to analysis/forecast. Tellus 55A:371–384

Lakshmivarahan S, Lewis JM (2010) Forward sensitivity method for dynamic data assimilation. Advances in meteorology, vol 2010, Article ID 375615, 12 pp
Lakshmivarahan S (2016) Convergence of a class of weak solutions to the strong solution of linear constrained quadratic minimization problem: a direct proof using matrix identities. In: ED, (ed) Data assimilation for atmospheric, oceanic and hydrologic applications, vol III. Seon Ki Park and Liang Xu, pp 115–119
Lakshmivarahan S, Lewis JM, Jabrzemski R (2017) Forecast error correction using dynamic data assimilation. Springer, New York
Lakshmivarahan S, Lewis JM, Hu J (2020a) On controlling the shape of the cost functional in dynamic data assimilation: guidelines for placement of observations and application to Saltzman's model of convection. J Atmos Sci 77:2969–2989
LeDimet FX, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations. Tellus 38A:97–110
Lewis JM, Derber J (1985) The use of adjoint equations to solve a variational adjustment problem with advective constraints. Tellus 37A:309–322
Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach, vol 104. Encyclopedia of mathematics and its applications. Bridge University Press, London
Lewis JM, Lakshmivarahan S (2008) Sasaki's PivotalContributions: calculus of variations applied to weather map analysis. Mon Weather Rev 136:3553–3567
Lewis JM, Lakshmivarahan S, Hu J, Rabin R (2020a) Placement of observations to correct return flow forecasts. E-J Severe Storms Meteorol 15(4):1–20
Lewis JM, Lakshmivarahan S, Maryada S (2020b) Placement of observations for variational data assimilation: application to Burgers' Equation and Seiche Phenomenon, This volume
Meyer CD (2000) Matrix analysis and applied linear algebra. SIAM Publications, Philadelphia
Nijmeijer H, van der Schaft A (1990) Nonlinear dynamical control systems. Springer, New York
Vidyasagar M (2002) Nonlinear system analysis, 2nd edn. SIAM Publication, Philadelphia
Yoshimura R, Yakeno A, Misaka T, Obayashi S (2020) Application of observability Gramian to targeted observation in WRF data assimilation. Tellus A 72(1):1–11

Placement of Observations for Variational Data Assimilation: Application to Burgers' Equation and Seiche Phenomenon

John M. Lewis, S. Lakshmivarahan, and S. K. R. Maryada

Abstract Observation placement in variational data assimilation determines cost function structure in the space of control. The presence of flatness in the cost function's gradient presents problems in the iterative passage to the cost function's minimum. Determination of observation placement that avoids these flat zones generally permits expeditious passage to the cost function minimum. A contribution to this volume (Lakshmivarahan S, Lewis JM, Maryada SKR (2020b) Observability Gramian and its role in the placement of observations in dynamical data assimilation. In: Data assimilation for atmospheric, oceanic, and hydrologic applications. Springer Pub. Co., New York) has theoretically determined methodology that identifies observation placement that avoids these flat zones. The placement relies on the norm of a semi-definite positive Gramian matrix G—a matrix derived from forecast sensitivity to control. Two dynamical systems are tested with this methodology: (1) Burgers' Equation, and (2) Seiche phenomenon, the normal mode oscillations in lakes. Analytic solutions to both constraints have been found. For each dynamical system, two sets of observation placement are considered: one where observations locations correspond to places where the norm of G is large, and one where the norm is small. Results indicate that observation placement where the norm of G was large led to well-defined structure of the cost function at the operating point in control space, a structure where the cost-function gradient was bound away from zero, whereas choices for observation locations where the norm of G was small-magnitude led to troublesome cost function structure, a structure where small-magnitude gradient presented difficulty in advancing toward the cost-function minimum.

1 Introduction

In Lakshmivarahan et al. (2020a), guidance on placement of observations in data assimilation has been presented with the goal of correcting model controls which

J. M. Lewis (✉) · S. Lakshmivarahan · S. K. R. Maryada
National Severe Storms Laboratory (NSSL), Desert Research Institute (DRI), Reno, NV 89512, USA
e-mail: jlewis@dri.edu

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_10

in turn minimize the squared distance between forecast and observations. Central to the theme is realization that observation placement directly determines cost function structure in the space of control. And a structure that avoids flatness in the cost-function gradient about the operating point in control space allows for more assured passage to the cost-function minimum.

Determination of these optimal placements depends on forecast sensitivity to control. In fact, overlay of forecast error and forecast sensitivity to control intuitively identifies places where observations are most valuable. But in the absence of sufficient observations to identify these places, a method that depends only on forecast sensitivity is most valuable. In effect, an overlay of forecast sensitivity to control and forecast error in the form of a first-order Taylor expansion about control has been developed in Lakshmivarahan et al. (2020a). The process relies on the structure of a positive semi-definite Gramian matrix labeled G that is derived from the forecast sensitivity matrix and its transpose. It is shown that the cost function gradient in control space is the product of G and δc where δc is the distance between the yet unknown optimal control and the forecast control. The norm of G identifies locations in space and time that avoid flatness in the cost function gradient about the operating point in control space and thereby supports passage toward the cost-function minimum by gradient-based optimization techniques. The mathematical foundation underpinning the methodology is found in Lakshmivarahan et al. (2020a, Sect. 4).

The strategy described above is applied to: (1) Burgers' Equation where a sinusoidal wave steepens to an infinite slope in a relatively short period of time, and (2) the normal mode oscillation of gravity waves in a lake (seiche) where nonlinear interaction between control elements presents a challenge.

2 Burgers' Equation

Johannes Burgers (1895–1981) was a physicist who studied under Paul Ehrenfest at the University of Leiden and completed a dissertation on the Rutherford–Bohr model of the atom in 1918 (Burgers 1975). He professionally migrated from atomic physics into fluid dynamics in October 1918 when he accepted a newly created faculty position in fluid mechanics at Technical University of Delft. He became especially interested in turbulence, both theoretical and experimental, and is credited with using simplified models of turbulence to more fundamentally understand its behavior (Burgers 1939). The well-known equation that bears his name takes the form

$$\frac{\partial q}{\partial t} + q\frac{\partial q}{\partial x} = v\frac{\partial}{\partial x}\left(\frac{\partial q}{\partial x}\right), \tag{1}$$

an equation that describes nonlinear advection of the quantity q while being subjected to turbulent diffusion. This dynamical equation has enjoyed widespread use in fluid-flow problems and applied mathematics (for example, see Benton and Platzman 1972). Burgers immigrated to the United States from the Netherlands in 1955 and worked at the Institute for Fluid Mechanics and Applied Mathematics, University of Maryland, for the remainder of his life. A photograph of Burgers working at the institute is shown in Fig. 1.

In our study of observation placement to correct errors in dynamical prediction, we use the form of Burgers' Equation that follows:

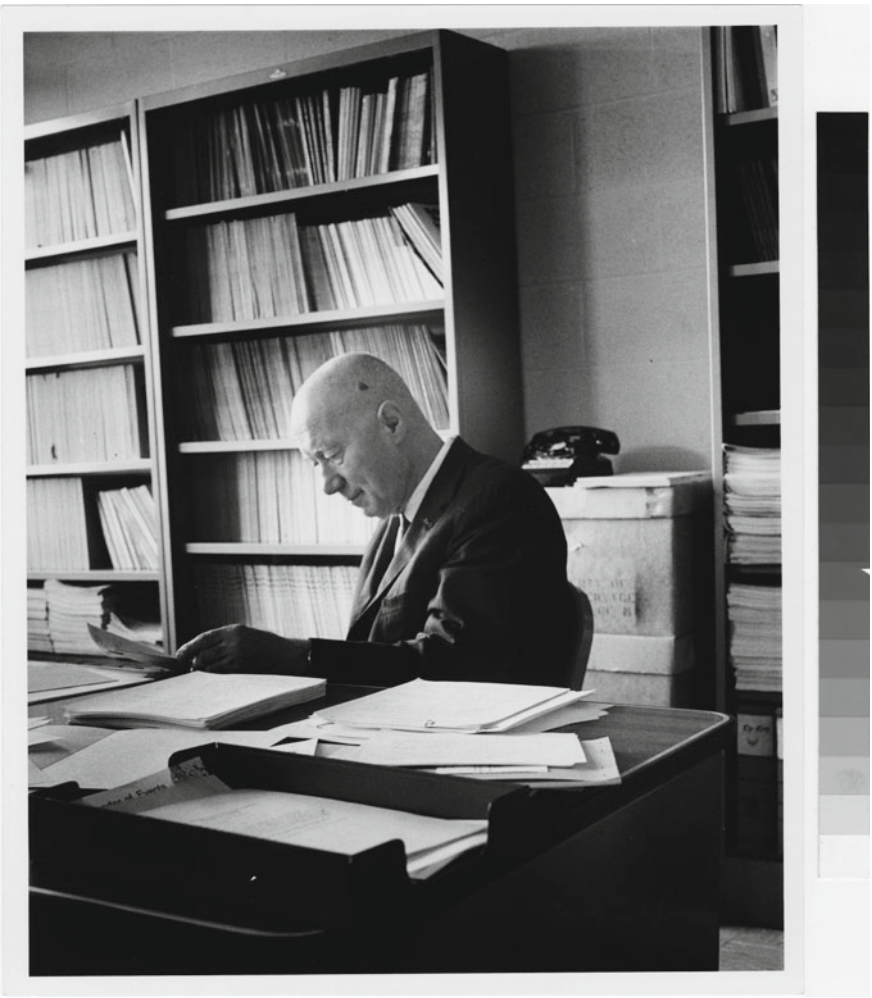

Fig. 1 Photograph of J. M. Burgers working at the Institute for Fluid Dynamics and Applied Mathematics, University of Maryland (Courtesy Univ. Maryland)

$$\frac{\partial q}{\partial t} + b \sin(x) \frac{\partial q}{\partial x} = a\, q. \tag{2}$$

where a *and* b are positive constants and the solution domain is $0 \leq x \leq 2\,\pi$ and $t \geq 0$. The linear term on the right-hand-side of this equation is consistent with Eq. (1) when the solution is a sinusoidal function, but $a > 0$ leads to amplification instead of damping. This equation is quasi-linear in view of its advection velocity $b \sin(x)$, a function of the spatial coordinate x instead of a strict function of $q(x,\ t)$. The analytic solution to this equation is found by the method of characteristics (Carrier and Pearson 1976) and takes the form

$$q(x,\ t) = \frac{2\, b \sin(x) e^{(a+b)t}}{1 + e^{2bt} + \cos(x)\left(e^{2bt} - 1\right)} \tag{3}$$

where the initial condition is $q(x_0,\ 0) = b\ \sin(x_0)$, $0 \leq x_0 \leq 2\pi$, and the characteristics are given by

$$\frac{1 + \cos(x)}{1 - \cos(x)} e^{2bt} = \frac{1 + \cos(x_0)}{1 - \cos(x_0)}. \tag{4}$$

A plot of five characteristics is shown in Fig. 2 when $a = 0.1$ and $b = 1$. These characteristics emanate from $x_0 = \frac{\pi}{2},\ \frac{3\pi}{4},\ \pi,\ \frac{5\pi}{4},\ and\ \frac{3\pi}{2}$. They present a physical view of the problem where, for example, the characteristic emanating from $x_0 = \frac{\pi}{2}$

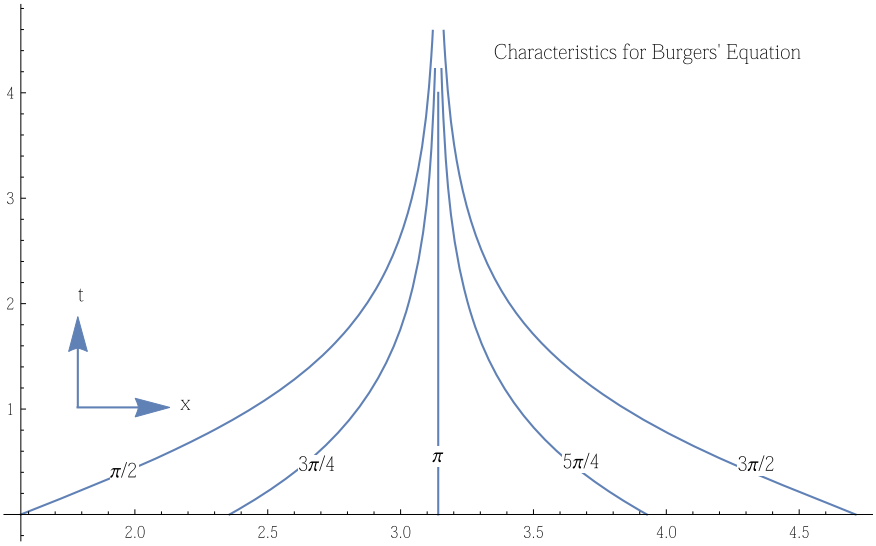

Fig. 2 Five characteristics for the Burgers' equation constraint

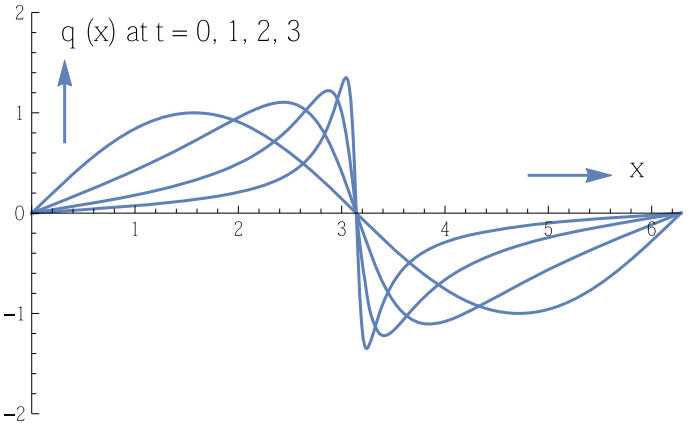

Fig. 3 Evolution of $q(x)$ at 4 times proceeding from the sine curve at t = 0 to the progressively steeper and amplifying curves at t = 1, 2, and 3

indicates that the crest of the sine wave with amplitude $b \sin\left(\frac{\pi}{2}\right) = b$ at t = 0 moves to a value of $x \to \pi$ by t = 4. Analogously, the characteristic emanating from $x_0 = \frac{3\pi}{4}$ with amplitude $b \sin\left(\frac{3\pi}{4}\right) = \frac{\sqrt{2}}{2} b$ moves just to the right of the crest position at t = 4. And along each characteristic the value of $q(x, t) = b \sin(x_0) e^{at}$ indicating amplification of its value at t = 0. The interpretation is progression and steepening of the initial sine wave as it moves from left and right toward $x = \pi$ through the action of advection and amplification where the slope of the wave at $x = \pi \to \infty$ as time $\to \infty$. This effect can be seen with the plot of the solution at various times as shown in Fig. 3.

3 Data Assimilation Experiment with Burgers' Equation

Using Eq. (2) and it closed-form solution Eq. (3) as constraint, a variational data assimilation problem is posed where the central theme is choice of observation locations that hold promise for creating a cost function whose gradient is bounded away from zero.

In the experiments to follow, the true controls are chosen to be $(a*, b*) = (0.10, 1.00)$ while forecast controls are $(a, b) = (0.20, 0.80)$. Observations are created from true controls with added normally distributed noise with zero mean and standard deviation equal to 10% of the true state. Determination of observation placement is based on locating the values of x and t where the norm of matrix G is greatest. For this task the sensitivities are needed. These are displayed in Fig. 4. The elements of this 2 × 2 G matrix take the form

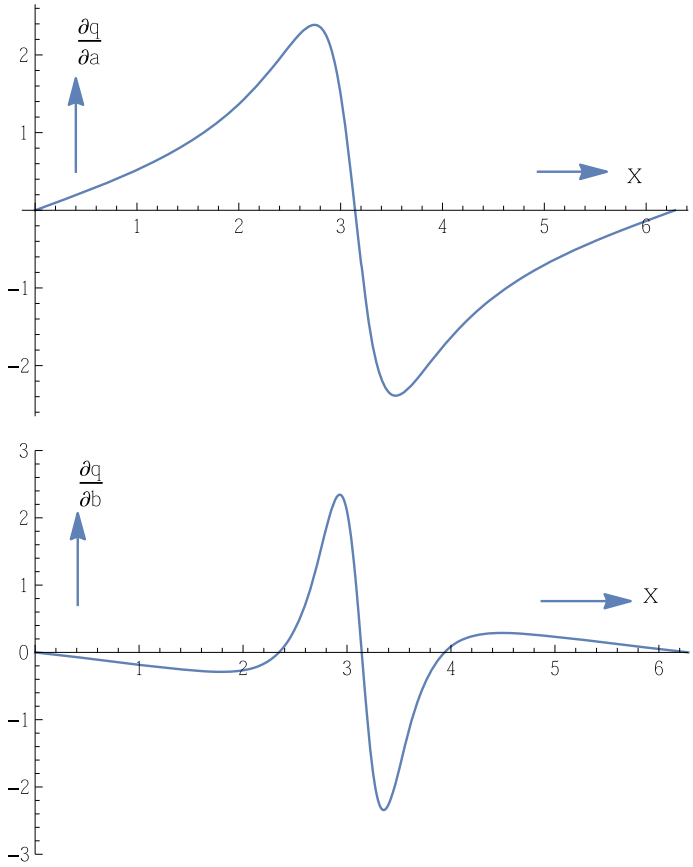

Fig. 4 Forecast sensitivity to controls *a and b* at t = 2

$$G = \begin{bmatrix} \left(\frac{\partial q}{\partial a}\right)^2 & \frac{\partial q}{\partial a}\frac{\partial q}{\partial b} \\ \frac{\partial q}{\partial b}\frac{\partial q}{\partial a} & \left(\frac{\partial q}{\partial b}\right)^2 \end{bmatrix}, \tag{5}$$

functions of controls *a and b*, space (x), time (t). G's norm is the sum of the squared elements. The G norm is a function of x, t after forecast controls a, b are substituted into the expression and it is displayed in Fig. 5. This norm is symmetric about $x = \pi$ with lobes of largest values aside the symmetry point. The values are very small in the vicinity of $x = 0, \pi, and\ 2\pi$.

We create observations at t = 2, a time when the crest of the initial sine curve has steepened, amplified and moved from $x = \pi$ to $x = 2.74$. The structure of norm G at t = 2 is displayed in Fig. 6. The two spikes in this plot are positioned very close to the crest and trough of the wave at t = 2. The G-norm values drop off rapidly in both directions away from these spike maxima. Norm values are small at

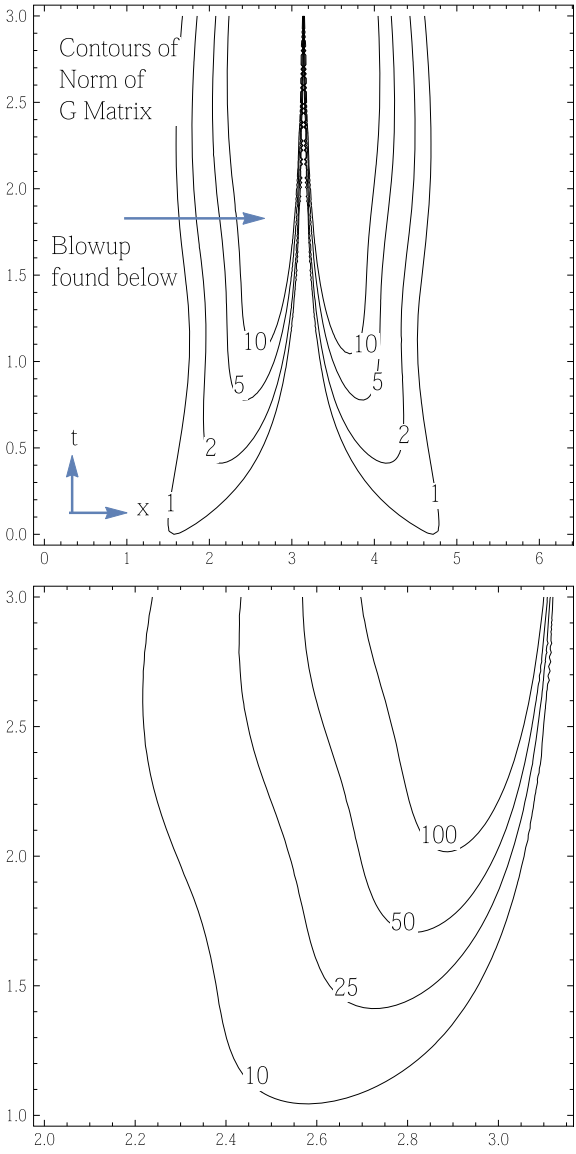

Fig. 5 G-Norm for Burgers' equation as function of (x, t) where details of contour structure are shown in the lower panel

values $x < 2$, $x > 4.5$ and values of x adjoining the symmetry point. Two sets of observations are separately examined: set I in a range of x-values where G-norm is very large, and set II in a zone where G-norm has small values. Since observations will be made at t = 2, forecasts at that time are functions of x and the controls *a and b*. Three observations are created for each set. The forecast takes the form

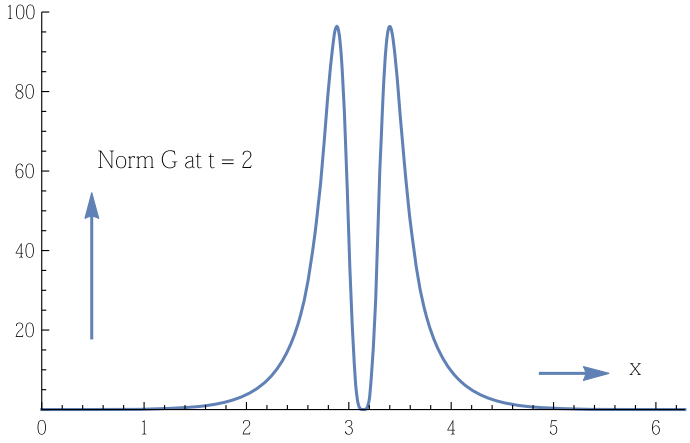

Fig. 6 The G-Norm for Burgers' equation is shown as a function of x at t = 2

$$q(x,2) = \frac{2b\ \sin(x)e^{2(a+b)}}{1+e^{4b}+\cos(x)\left(e^{4b}-1\right)} \tag{6}$$

The cost functional J follows

$$J(a,b) = \left\{ (q(x_1,2) - Obs_1)^2 + (q(x_2,2) - Obs_2)^2 + (q(x_3,2) - Obs_3)^2 \right\} \tag{7}$$

where observations Obs_i $1 \leq i \leq 3$ are known from the true control and added noise.

Observations for set I are chosen to be $x_1 = \frac{4\pi}{5}$, $x_2 = \frac{9\pi}{10}$, $x_3 = \frac{19\pi}{20}$. Observations for set II are located at $x_1 = 0.57\ \pi\ (= 1.8)$, $x_2 = 0.64\ \pi\ (= 2.0)$, $x_3 = 0.70\ \pi\ (= 2.2)$. These locations reside in that zone where G-norm is relatively small but on the rise. Contours of J when observation set I is used in the data assimilation exercise are shown in Fig. 7. Gradient-base optimization, specifically the Newton method, is used to locate the minimum at $(a, b) = (0.121, 0.928)$ relatively close to the true controls. For set set II observations, the minimum is found at $(a, b) = (0.293, 0.577)$ quite far from true control. When observations are chosen in the zone below x = 2, the results are physically unreasonable with very large values of both parameters.

4 Seiche Dynamics

Lake water set in motion due to strong winds and downdrafts from squall line passage or strong gusty winds accompanying a front can lead to a phenomenon called seiche,

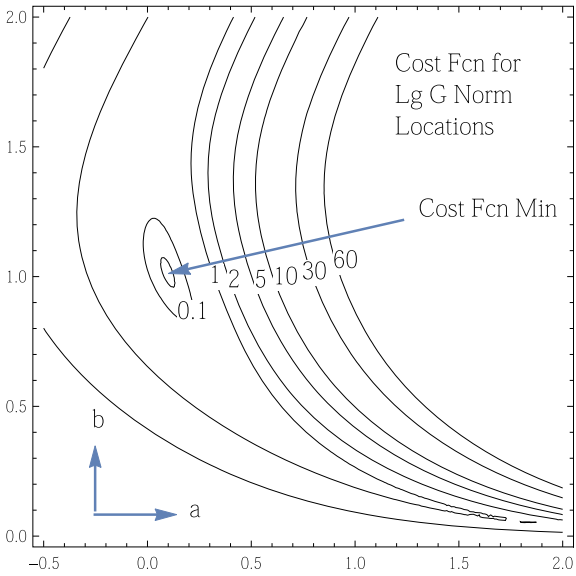

Fig. 7 Cost function contours for Burgers' equation constraint when observations were chosen at (x, t) = (x, 2) locations where G-Norm had large values

a word in the French language meaning "to sway back and forth". It is a most appropriate word to describe the oscillations of lake water after being disturbed by external forces such as those mentioned. The wavelengths and periods of oscillation fundamentally depend on lake dimensions. Once started, these oscillations continue for relatively long periods of time just like a pendulum.

Investigation of these oscillations in an elongated rectangular lake (length ≫ breadth) is a good way to gain fundamental understanding of the phenomenon. The dynamics underpinning these oscillations has been clearly presented in the dynamic oceanography textbook (Proudman 1963, Chap. XI). Let the x-axis measure distance along the lake's length (x = 0 and x = L at lake's beginning and end). No lake is in the exact shape of an exact rectangular solid, but we assume it can be well approximated by mean dimensions: mean length L, mean breadth b, and mean depth $\bar{D}$ where we will assume that the length and breadth can be determined with great accuracy but where the depth varies significantly along the x-axis and is subject to error in its measurement. This inaccuracy will come into play in the data assimilation component of the problem. Governing dynamic equations for the seiche, coupled equations for momentum and mass conservation take the form

$$\frac{\partial Q}{\partial t} = -gb\,\bar{D}\frac{\partial h}{\partial x} \tag{8}$$

$$\frac{\partial h}{\partial t} = -\frac{1}{b}\frac{\partial Q}{\partial x}, \tag{9}$$

respectively, where physical variables are
Q (x, t): flux of water through the cross section of area b$\bar{D}$

u (x, t): horizontal current related to mass flux by Q = b$\bar{D}$u (x, t)

h (x, t): perturbation in water level relative to the mean depth $\bar{D}$.

T_1: primary period of oscillation (called the Merian period), $T_1 = 2L\Big/\sqrt{g\bar{D}}$, g the acceleration of gravity.

Since no transport of water takes place at the lake's end points in the x-direction, the boundary conditions on flux are $Q(0) = Q(L) = 0$ and the initial condition is $h(x, 0) = h_0$. It is assumed that the initiating external force causes the lake's surface to deviate from the mean height, and in turn this creates a horizontal pressure gradient force within the water mass that initiates a horizontal current. Thus, at t = 0, there is an initial height perturbation h_0 that generates currents at times t > 0.

The solution to the coupled equations takes the form

$$\begin{aligned} Q &= b\sqrt{g\bar{D}}h_0 \sin\frac{r\pi x}{L} \sin\frac{2\pi t}{T} \\ or\ u(x,t) &= h_0\sqrt{\frac{g}{\bar{D}}} \sin\frac{r\pi x}{L} \sin\frac{2\pi t}{T} \\ and\ h(x,t) &= h_0 \cos\frac{r\pi x}{L} \cos\frac{2\pi t}{T} \end{aligned} \tag{10}$$

where $T = T_r \equiv \frac{T_1}{r}$, r an integer that determines the various modes of oscillation; $r = 1$ a unimodal solution (Merian mode), r = 2 a bimodal solution etc. For each mode, there are specific places where no rise or fall of water at the lake's surface takes place—along lines perpendicular to the x-axis. There are r such lines for the rth mode and they are found at values of x where $\cos\frac{r\pi x}{L}$ vanishes. Graphs of time-dependent heights and currents for the unimodal ($r = 1$) and bimodal ($r = 2$) oscillations are shown in Figs. 8 and 9. Generally all modes of oscillation are active once the lake is disturbed, and since the governing equations are linear, the complete solution is the sum of the various modes. However, the initial height of each mode must be determined by a Fourier decomposition of the lake's spatial structure at time of disturbance initiation. This problem bears a strong similarity to the vibrating string problem encountered in mechanics. The boundary conditions, for the string connected at both ends, and in this problem where the flux vanishes at both ends, permits only a select set of wavelengths and periods called normal modes. Since all circumstances of the r-mode motion are assumed to repeat after intervals of time $T = T_r$, it is sufficient to consider one interval of time from t = 0 to T_r for the rth mode.

When solutions to the seiche equations are used as constraints in the variational data assimilation problem, the variables must be made non-dimensional so that minimization of the cost function is not biased. That is, the squared terms in the cost function should have the same order of magnitude; otherwise, the largest terms undergo more adjustment than the smaller terms in a minimization process. The hat symbol $\widehat{(\,)}$ will be used to denote non-dimensional forms of variables. The non-dimensional forms follow:

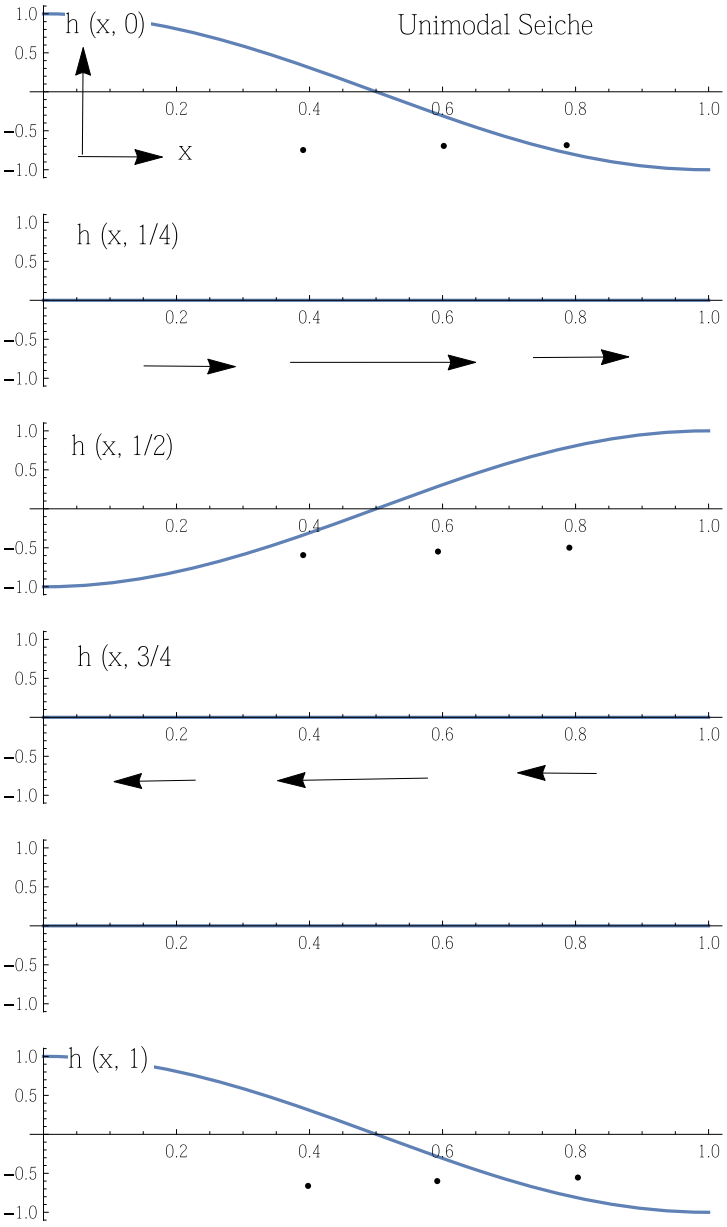

Fig. 8 The evolution of unimodal water height and currents in a rectangular lake that extends from $\hat{x} = 0$ to $\hat{x} = 1$ (nondimensional length scale) at equally spaced intervals of time $\left(\Delta\hat{t} = 1/4\right)$ covering the period of seiche oscillation from $\hat{t} = 0$ to $\hat{t} = 1$ (dimensional period = 3 h). Currents are displayed as vectors pointing in the direction of water movement and magnitude proportional to arrow length where the dot $(\cdot)$ indicates current = 0. The dimensional height of the wave is 4 ft and the dimensional current is 4 $ft \cdot s^{-1}$

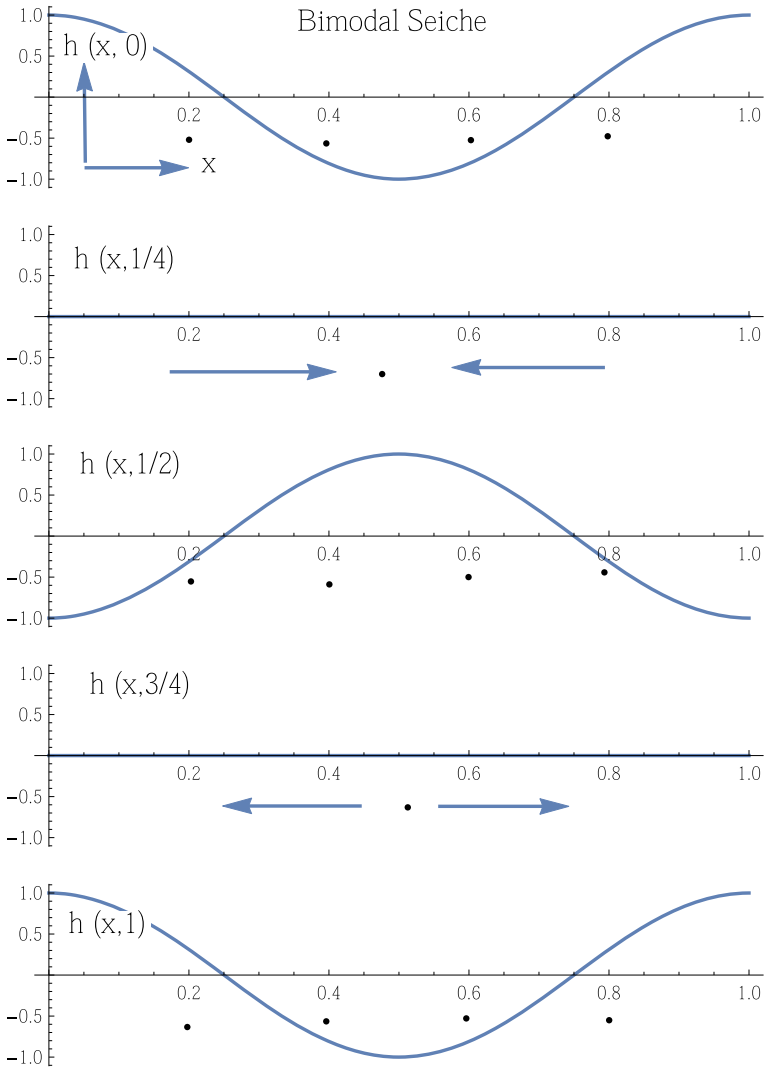

Fig. 9 Same description as found for Fig. 8 except the dimensional period is 3/2 h

$$
\begin{aligned}
h(x,t) &= \bar{H}\hat{h}(\hat{x},\hat{t}) \\
u(x,t) &= \bar{U}\hat{u}(\hat{x},\hat{t}) \\
D &= \bar{D}\hat{d} \\
h_0 &= \bar{H}\hat{h}_0 \\
x &= L\hat{x} \\
t &= T_r\hat{t}
\end{aligned}
\tag{11}
$$

where

$$T_r = \frac{2L}{r\sqrt{g\bar{D}}}$$

When the estimated mean depth of water is different from $\bar{D}$, call it D as defined above with $\hat{d} \neq 1$, then the period of oscillation is not T_r, but

$$T = \frac{2L}{r\sqrt{g\bar{D}\hat{d}}} = \frac{T_r}{\sqrt{\hat{d}}} \tag{12}$$

Governing equations for height and current become

$$\begin{aligned} \hat{h}(\hat{x}, \hat{t}) &= \hat{h}_0 \cos(r\pi\hat{x}) \cos\left(2\pi\hat{t}\sqrt{\hat{d}}\right) \\ \hat{u}(\bar{x}, \bar{t}) &= N\frac{\hat{h}_0}{\sqrt{\hat{d}}} \sin(r\pi\hat{x}) \sin\left(2\pi\hat{t}\sqrt{\hat{d}}\right) \end{aligned} \tag{13}$$

where the non-dimensional coefficient $N = \frac{\bar{H}}{\bar{U}}\sqrt{\frac{g}{\bar{D}}}$.

4.1 Data Assimilation for Seiche

We assume the constraint is the unimodal oscillation ($r = 1$). Let us specify the mean values used in the non-dimensional process:

$$\begin{aligned} &\bar{H} = 4\ ft \\ &\bar{U} = 4\ ft \cdot s^{-1} \\ &\bar{D} = 50 \text{ fathoms (ftm)} = 300\ ft \\ &L = 100\ mi \\ &T = \frac{2L}{\sqrt{g\bar{D}}} = 3\ hours \end{aligned} \tag{14}$$

The equations governing the oscillation take the form of Eq. (13) with $r = 1$.

Errors in height and current prediction are due to inaccuracy in the lakes mean depth and amplitude of the initial height perturbation. All other parameters are known exactly. Thus, the control vector takes the form $\left(C = \hat{h}_0,\ \hat{d}\right)$. Both control-vector components are the order of 1. We will assume the true control vector is $C_{true} =$ (1, 1). Thus, the dimensional true initial height amplitude is 4 feet and the true mean

depth is 50 ftm. The forecast control is erroneous and given by $C_{fcst} = (0.9, 1.1)$ so that the initial dimensional estimates of forecast control are 3.6 feet and 55 ftm.

To determine the optimal placement of observations, we need to calculate the forecast sensitivity to the elements of control, viz., $\frac{\partial \hat{h}}{\partial \hat{h}_0}$, $\frac{\partial \hat{u}}{\partial \hat{h}_0}$, $\frac{\partial \hat{h}}{\partial \hat{d}}$, $\frac{\partial \hat{u}}{\partial \hat{d}}$. And from these sensitivities we can calculate the 2×2 Gramian matrix G using forecast control C_{fcst} as follows:

$$G = \begin{bmatrix} \left[\left(\frac{\partial \hat{h}}{\partial \hat{h}_0}\right)^2 + \left(\frac{\partial \hat{u}}{\partial \hat{h}_0}\right)^2\right] & \left[\frac{\partial \hat{h}}{\partial \hat{h}_0}\frac{\partial \hat{h}}{\partial \hat{d}} + \frac{\partial \hat{u}}{\partial \hat{h}_0}\frac{\partial \hat{u}}{\partial \hat{d}}\right] \\ \left[\frac{\partial \hat{h}}{\partial \hat{d}}\frac{\partial \hat{h}}{\partial \hat{h}_0} + \frac{\partial \hat{u}}{\partial \hat{d}}\frac{\partial \hat{u}}{\partial \hat{h}_0}\right] & \left[\left(\frac{\partial \hat{h}}{\partial \hat{d}}\right)^2 + \left(\frac{\partial \hat{u}}{\partial \hat{d}}\right)^2\right] \end{bmatrix} \tag{15}$$

The G-norm as a function of (x, t) is displayed in Fig. 10. The norm is symmetric about $\hat{x} = 0.5$, lake's midpoint 50 miles from each end of the lake, and it features maxima near the lakes end points $\hat{x} = 0,\ 1$ ($x = 0\ and\ 100$ miles dimensionally) and at $\hat{t} = 3/4$ (2.25 h dimensionally). Based on these results, one set of observations will be taken at $\hat{t} = 3/4$ and near one end of the lake, near $\hat{x} = 0$, and the associated data assimilation experiment is labeled Experiment I. The other set of observations will be taken at the same time but near the center of the lake where small values of G-norm are found and the associated data assimilation experiment is labeled Experiment II. To get a better view of G-norm at the planned observation time $\hat{t} = 3/4$, Fig. 11 shows G-norm over the length of the lake at this time.

Experiment I takes observations at x = 0.1, 0.2, and 0.3 (height observations at 0.1 and 0.3 and a current observation at t = 0.2). In Experiment II, observations were taken at x = 0.48, 0.50, and 0.52, again with height observations at bounding values of x and a current observation at the mid-point in x. Normally distributed random error is added to observations—zero mean and standard deviation equal to 10% of the typical values of height and current. The cost function for Experiment I is shown in Fig. 12 where the Newton method found the minimum at control point $C_{Exp\ I} = (1.106, 0.998)$. The cost function for Experiment II is shown in Fig. 13 where the Newton method found the minimum at control point $C_{Exp\ II} = (1.050, 1.189)$. In this case, the location of observations at small values of norm G gave reasonable results, not expected. Further investigation revealed that the negative gradient of the cost function at the operating point was indeed very small, but it was precisely directed toward the minimum.

5 Conclusions

The two dynamical constraints used to explore the value of the G-matrix method of locating observation sites to be used in data assimilation and thereby correct forecasts have validated the usefulness of this methodology. Burgers' Equation was a quasi-linear constraint and the seiche dynamics were linear, but the forecast sensitivities

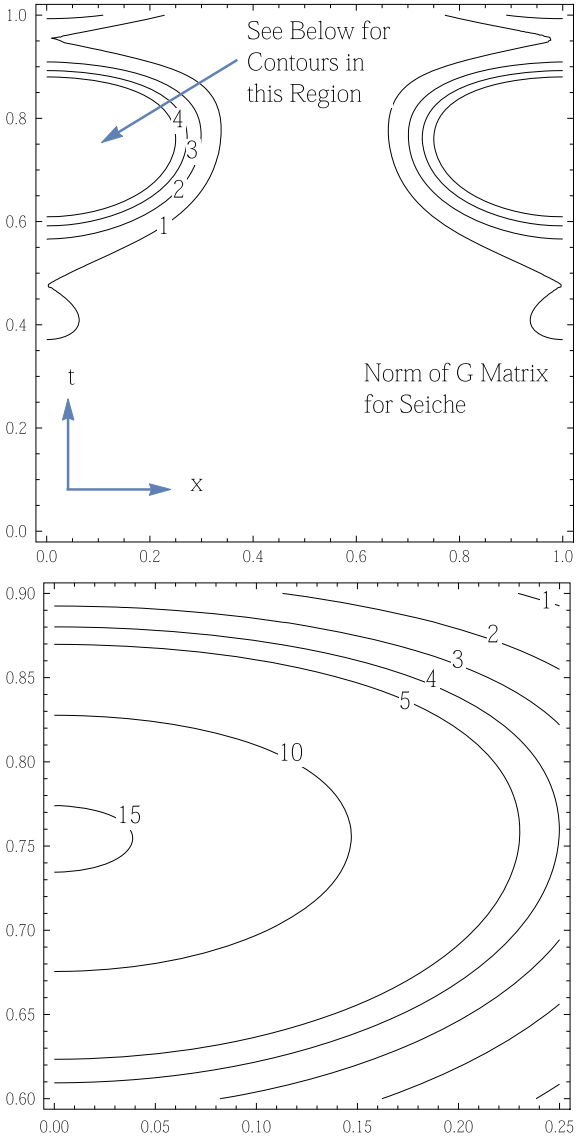

Fig. 10 G-Norm for seiche dynamics as function of nondimensional space $(\hat{x})$ and time $(\hat{t})$ where details of contour structure are shown in the lower panel

to control in both cases were highly nonlinear which is often a barrier or difficulty in finding cost-function minima.

The G-norm structures for both dynamical systems were appealing—the structures made sense physically. In the case of Burgers' Equation, the choice sites were near the symmetry point but not at the symmetry point—close to the steep gradient zone where sensitivity is large and forecast errors are relatively large and away from the spatial end points where sensitivity is small and wave amplitude becomes vanishingly

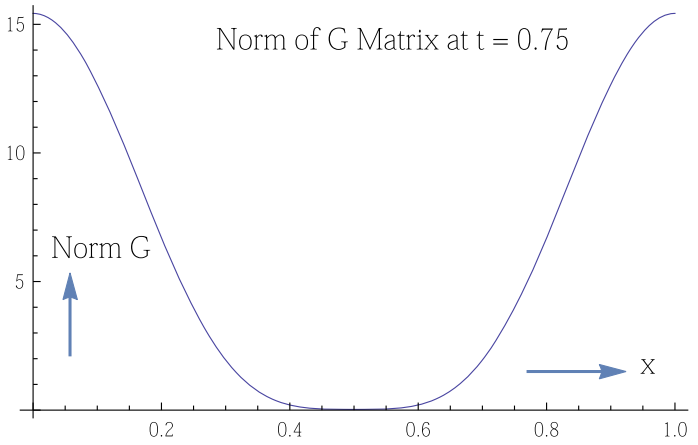

Fig. 11 Profile of G-Norm at nondimensional time $(\hat{t}) = 0.75$

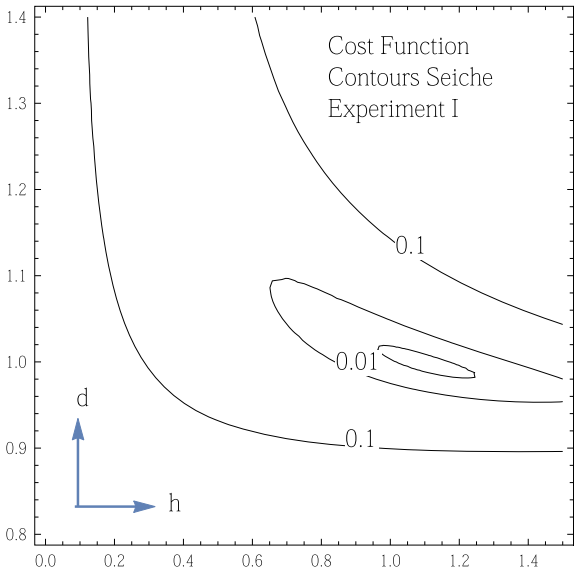

Fig. 12 Cost function for seiche constraint when observations were chosen at locations $(\hat{x}, \hat{t}) = (\hat{x}, 0.75)$ where G-Norm had large values

small as time increases. For the seiche dynamics, the choice points were removed from that point in space where the water level never changed (at the mid-point of the lake for the unimodal mode). The choice points were near the lake's end points where the oscillations were largest. There was a revealing fact that emerged in the seiche dynamics case. That is, even if the G-matrix method identifies observation locations where the cost-function gradient is vanishingly small, it is not necessarily true that choosing observations at these points will produce very poor results. They did not produce poor results in our Experiment II for seiche dynamics. On the other

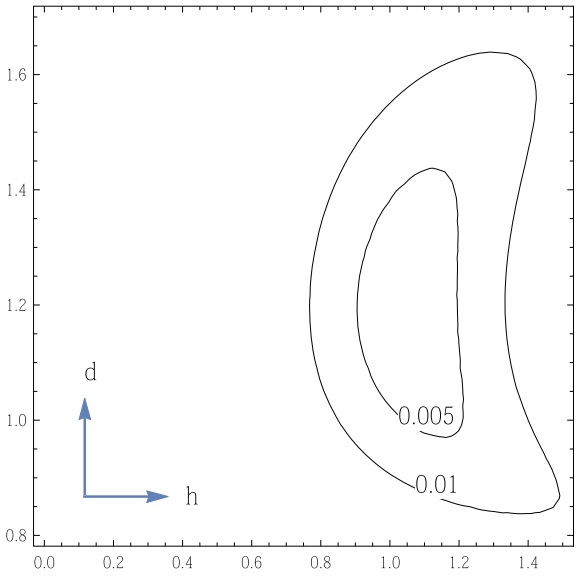

Fig. 13 Cost function for seiche constraint when observations were chosen at locations $(\hat{x}, \hat{t}) = (\hat{x}, 0.75)$ where G-Norm had small values

hand, the G-matrix method steers one away from observation sites that would be associated with extremely small gradients at the operating point, and by so doing it is faithful to its claim that it is best to choose sites where the G-norm is large.

Acknowledgements Course material on the dynamical constraints used in this data assimilation study were admirably taught to the lead author by Professor George Platzman at the University of Chicago (Seiche phenomenon) and Professor Yoshikazu Sasaki at University of Oklahoma (Burgers' Equation) in the early and mid-1960s, respectively, while a graduate student at these institutions.

References

Benton ER, Platzman GW (1972) A table of solutions of the one-dimensional Burgers' equation. Q Appl Math 195–212

Burgers JM (1939) Mathematical examples illustrating relations occurring in the theory of turbulent fluid motion. Kon Ned Akad Wet Verh (Eerste Sectie) D1. XVII 2:1–53

Burgers JM (1975) Some memories of early work in fluid mechanics at the Technical University of Delft. Annu Rev Fluid Mech 7:1–11

Carrier GF, Pearson CE (1976) Partial differential equations (theory and technique). Academic Press, Cambridge, p 320

Lakshmivarahan S, Lewis JM, Hu J (2020a) On controlling the shape of the cost functional in dynamic data assimilation: guidelines for placement of observations and application to Saltzman's model of convection. J Atmos Sci 77:2969–2989. http://0.0.0.10:1175/JAS-D-19-0329.1

Lakshmivarahan S, Lewis JM, Maryada SKR (2020b) Observability Gramian and its role in the placement of observations in dynamical data assimilation. In: Data assimilation for atmospheric, oceanic, and hydrologic applications. Springer Pub. Co., New York

Proudman J (1963) Dynamical oceanography. Methuen & Co., LTD, London, p 409

Analysis, Lateral Boundary, and Observation Impacts in a Limited Area Model

Clark Amerault

Abstract Impacts from analysis and lateral boundary updates as well as the assimilation of observations are investigated using the forecast sensitivty to observation impact framework in a limited-area atmospheric model. High temporal frequency estimates of forecast error are produced using aircraft observations for validation. Using these estimates, forecast error reduction between background and analysis trajectories is shown to decrease through the first 24 h of forecast time. The increasing importance of lateral boundary updates in decreasing forecast error with forecast lead time is presented. However, the ability of the adjoint forecast model to attribute forecast error reduction to analysis and lateral boundary updates decreases as forecast length increases. The relative distributions of the largest observation impacts for different lead times are similar. This means that impacts for shorter forecast lengths are a good proxy for impacts on longer forecasts, thereby mitigating some of the problems in long adjoint model integrations. Finally, a metric that measures forecast error against radiosondes is introduced and produces different distributions of observation impact importance.

1 Forecast Sensivity to Observation Impact

A robust procedure for quantitatively evaluating the impact of an observation's assimilation on short term forecast error utilizing the adjoint observation sensitvity framework Baker and Daley (2000) was developed by Langland and Baker (2004). The terminology for this procedure has evolved to be known as forecast sensitivity to observation impact (FSOI) and has been implemented for global atmospheric modeling systems Langland (2005); Gelaro and Zhu (2009); Cardinali (2009) as well as limited-area models Amerault et al. (2013); Jung et al. (2013); Zhang et al. (2015). The FSOI framework developed by Langland and Baker (2004) is the focus of this

C. Amerault (✉)
Naval Research Laboratory, Monterey, CA, USA
e-mail: clark.amerault@nrlmry.navy.mil

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_11

chapter, so its details are summarized below along with additional information unique to its application in limited-area models.

The error of two forecasts of lengths f and g can be measured against an analysis $\mathbf{x}_t$ available at verification time t in an inner product $\langle , \rangle$ using the following two equations,

$$e_f = \left\langle (\mathbf{x}_f - \mathbf{x}_t), \mathbf{C}(\mathbf{x}_f - \mathbf{x}_t) \right\rangle, \tag{1}$$

and,

$$e_g = \left\langle (\mathbf{x}_g - \mathbf{x}_t), \mathbf{C}(\mathbf{x}_g - \mathbf{x}_t) \right\rangle. \tag{2}$$

The coefficients in $\mathbf{C}$ weight the model fields so that the error is measured in an energy norm (the inner product has units of energy). The forecast for g begins at an earlier time than f, and a short term field from the g forecast serves as the background field $\mathbf{x_b}$ in the analysis procedure to produce $\mathbf{x_a}$ for the f forecast. The value of e_f is generally less than e_g due to the assimilation of observations $\mathbf{y}$ to update $\mathbf{x_b}$. In the remainder of this chapter, the combination of $\mathbf{x_a}$ and the resulting forecast $\mathbf{x_f}$ will also be referred to as the analysis trajectory. Conversely, the background trajectory will refer to the combination of $\mathbf{x_b}$ and $\mathbf{x_g}$.

To quantify the value of observations in reducing forecast error, an equation for the difference in e_f and e_g is defined,

$$\Delta e_f^g = e_f - e_g. \tag{3}$$

The quantity Δe_f^g is known as the forecast error reduction (FER). Using the adjoint forecast model, Δe_f^g can be mapped backward in time to analysis space. To do this, two cost functions are defined along with their corresponding first derivatives, which will serve as input for two adjoint model integrations along the forecast and background trajectories,

$$J_f = \frac{1}{2} e_f, \tag{4}$$

$$J_g = \frac{1}{2} e_g. \tag{5}$$

$$\frac{\partial J_f}{\partial \mathbf{x}_f} = \mathbf{C}(\mathbf{x}_f - \mathbf{x}_t), \tag{6}$$

$$\frac{\partial J_g}{\partial \mathbf{x}_g} = \mathbf{C}(\mathbf{x}_g - \mathbf{x}_t). \tag{7}$$

Eqs. 1-2 and 6-7 can be used to rewrite Eq. 3 as,

$$\Delta e_f^g = \left\langle (\mathbf{x}_f - \mathbf{x}_g), \frac{\partial J_f}{\partial \mathbf{x}_f} + \frac{\partial J_g}{\partial \mathbf{x}_g} \right\rangle. \tag{8}$$

The difference between forecast trajectories f and g at the analysis time is the increment $(\mathbf{x_a} - \mathbf{x_b})$. The adjoint model maps $\frac{\partial J_f}{\partial \mathbf{x}_f}$ to $\frac{\partial J_f}{\partial \mathbf{x_a}}$ and $\frac{\partial J_f}{\partial \mathbf{x}_g}$ to $\frac{\partial J_f}{\partial \mathbf{x_b}}$. Assuming that the analysis increment evolves approximately tangent linearly, then an estimate of Δe_f^g in analysis space δe_f^g can be written as,

$$\delta e_f^g = \left\langle (\mathbf{x_a} - \mathbf{x_b}), \frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right\rangle + \left\langle (\mathbf{l_a} - \mathbf{l_b}), \frac{\partial J_f}{\partial \mathbf{l_a}} + \frac{\partial J_g}{\partial \mathbf{l_b}} \right\rangle, \tag{9}$$

where $\mathbf{l_a}$ and $\mathbf{l_b}$ are lateral boundary conditions along the analysis and background trajectories, respectively. The first term on the right hand side of Eq. 9 is due to the analysis impacts and is present in all FSOI systems. The second term arises from the lateral boundary conditions and is only found in FSOI for limited-area models Amerault et al. (2013). As the adjoint forecast model marches backward in time, some of the gradient information is passed to the lateral boundaries. This term must be included, especially at longer forecast lead times, to better account for the FER (more explantion of lateral boundary impacts are provided in Sect. 2.3). Together, the analysis and lateral boundary impacts are known as model space impacts. Even after accounting for the lateral boundaries, these impacts in model space δe_f^g are not an exact match to the FER (Δe_f^g) because the adjoint forecast model is linear and is usually lacking some of the physical processes of the nonlinear model.

The impacts in observation space are determined by using only the analysis impacts (first term on the right hand side of Eq. 9) and replacing the analysis increment with the data assimilation solver $\mathbf{K}$ and the innovation vector $(\mathbf{y} - \mathbf{H}\mathbf{x_b})$ in the following manner,

$$\delta e_f^g = \left\langle \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x_b}), \frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right\rangle. \tag{10}$$

Using the properties of an adjoint operator in an inner product, the following expression in observation space,

$$\delta e_f^g = \left\langle (\mathbf{y} - \mathbf{H}\mathbf{x_b}), \mathbf{K^T} \left(\frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right) \right\rangle, \tag{11}$$

is obtained. The observation impacts are a product of the innovation vector components and the vector obtained from the adjoint data assimilation process $\mathbf{K^T}$. The inner product in Eq. 11 gives a total estimate for all observations, but the inner product can be paritioned into any particular subset of interest.

Unique aspects of FSOI in a limited-area model pertaining to lateral boundary conditions were previously presented Amerault et al. (2013). Here, we continue the investigation of FSOI in a limited-area model. The modeling system and domain is presented in Sect. 2. Forecast error reduction (Sect. 3), model space impacts (Sect. 4), and observation impacts (Sect. 5) follow. A summary is provided in Sect. 6.

2 COAMPS and NAVDAS

Results presented in this chapter were obtained from an FSOI system developed for the Naval Research Laboratory's (NRL) limited-area model. The system includes the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS®)[1] atmospheric model and its accompanying data assimilation component, the NRL Atmospheric Variational Data Assimiliation System (NAVDAS). Brief descriptions of these components and the FSOI system are provided below.

2.1 COAMPS Atmospheric Model

The COAMPS atmospheric model is a limited-area, relocatable, grid point model. The model is non-hyrdostatic and contains predictive equations for zonal wind u, meridional wind v, vertical velocity w, the dimensionless Exner pressure function π, the potential temperature θ, water vapor q_v, and turbulent kinetic energy e. The bulk cloud microphysics scheme calculates the source and sink terms in the prognostic equations for cloud droplets q_c, cloud ice q_i, rain water q_r, snow q_s, and graupel q_g. The other parameterizations in the model for subgrid scale processes are turbulent mixing, surface fluxes, cumulus convection, and radiation. The vertical coordinate of the model is a terrain following σ_z defined as

$$\sigma_z = \frac{z_\mathrm{t}(z - z_\mathrm{s})}{z_\mathrm{t} - z_\mathrm{s}}, \tag{12}$$

where the constant z_t is the depth of the model domain and z_s is the terrain height. Lateral boundary conditions are provided from the Navy Global Environmental Model (NAVGEM Hogan et al. (2014)). Tendencies are computed from the NAVGEM fields and applied to the COAMPS forecast fields on the outermost gridpoints throughout the model's integration. This process forces the COAMPS forecast fields to the same values as NAVGEM on these outermost grid points. A more detailed description of COAMPS is given in Hodur Hodur (1997).

2.2 NAVDAS

COAMPS atmospheric analysis fields are produced by NAVDAS Daley and Barker (2001). It is a three dimensional variational system that includes a geostrophic balance constraint and uniform analysis length scale. NAVDAS assimilates conventional and aircraft observations, cloud feature track satellite winds, satellite total precipitable water and temperature retrievals, and scatterometer and passive microwave derived

[1] COAMPS® is a registered trademark of NRL.

Table 1 Observation types assimilated by NAVDAS. The identifier used in the impact figures in this chapter are given in the left column and descriptions are provided in the right column

Observation type	Description
RADIOSONDE	Radiosonde measurement
CLD_WIND	Polar satellite feature track wind
leogeo WIND	Composite satellite feature track wind
AMDAR	Aircraft instrument
MDCRS	Aircraft instrument
ACARS	Aircraft instrument
AIREP	Aircraft instrument
MIL ACAR	Military aircraft instrument
LandSfc	Surface observation over land
ShipSfc	Surface observation over water
SSMIS SFC WIND	Surface wind from SSM/I
SSMI-PRH	Moisture retrieval from SSM/I
ASCAT SFC WIND	Surface wind vector from ASCAT
WINDSAT SFC WIND	Surface wind from WNDST
WINDSAT-PRH	Moisture retrieval from WNDST
HDOB	High density hurricane hunter aircraft
TC Synth	Synthetic TC data
ATOV RET	ATOVS temperature retrieval
SFMR	SFMR wind
DROPSONDE	Dropsonde measurement
PIBAL	Piloted balloon measurement
AMSUA	AMSUA radiances (not assimilated)

surface marine winds. Satellite sounding radiances are not currently assimilated in the system. A full list of assimilated intrument types is given in Table 1. The observations will be grouped by these types in Sect. 5 when discussing their impact.

2.3 COAMPS FSOI and Lateral Boundary Impacts

The COAMPS FSOI system includes adjoint components of the COAMPS atmospheric forecast model Amerault et al. (2008) and NAVDAS Amerault et al. (2013). As was discussed in Sect. 1, analysis and lateral boundary impacts are produced by the COAMPS atmospheric adjoint model. Lateral boundary impacts are unique to limited-area models that use global model forecast fields (not analyses) to compute lateral boundary tendencies (LBTs). For each forecast cycle, the LBTs are updated using the latest global model fields which will have less error than the fields used in

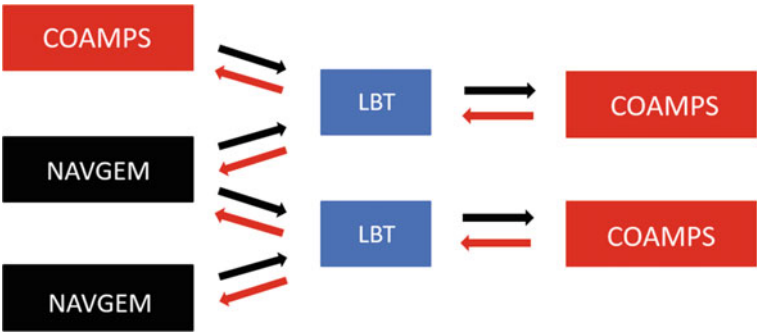

Fig. 1 Schematic of the flow of information between the global model (NAVGEM) and COAMPS atmosperic model through the first two LBTs. The black (red) arrows indicate the flow of information in the nonlinear (adjoint) model. Time increase from top to bottom

the previous cycle. Therefore, the updates to the LBTs can contribute to FER along with the assimilation of observations.

Figure 1 is a schematic of the flow of information between COAMPS and NAVGEM in the nonlinear and adjoint models. For simplicity, only the first two lateral boundary tendency (LBT) calculations are shown. In this study, the LBT is updated every 3 h, so this schematic is valid for the first 6 h of the forecast. The first lateral boundary tendency (LBT) 0–3 h is calculated from the COAMPS analysis and a NAVGEM forecast. The inputs to all subsequent LBTs from 3–6 h and beyond are NAVGEM forecasts. In the forward nonlinear integrations, the LBTs are applied to the COAMPS forecast fields on the outermost grid points. In the adjoint model integrations, gradient information flows from the interior of the model's domain onto the outermost grid points and then onto the adjoint LBT variable. This information is then pushed into NAVGEM space, except during the LBT window closest to the analysis time where some of the infomation remains in COAMPS space. The information in NAVGEM space comprises the lateral boundary impacts, while all the information in COAMPS space makes up the analysis impacts, which are then passed to the adjoint of NAVDAS to produce observation impacts.

2.4 Forecast Domain

The forecast domain for the experiment is shown in Fig. 2. All experiments were performed on a single domain with 45 km horizontal grid spacing and 60 vertical levels because the adjoint model does not use nests in the FSOI framework. The inner box indicates the area over which the forecast error was calculated. A dry energy norm was used to compute Δe_f^g. Forecasts and impact calculations were conducted from March 1–14 2020.

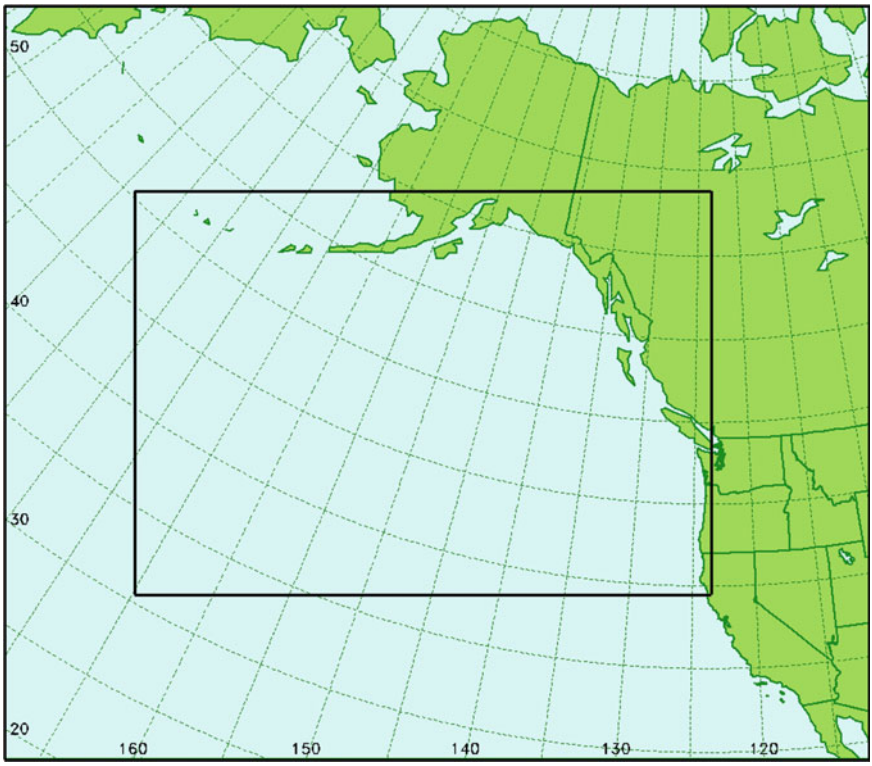

Fig. 2 COAMPS atmospheric forecast domain used in this study. The inner box indicates the area where forecast error was calculated

3 Forecast Error Reduction

As outlined above, the driver of an FSOI system is the difference in forecast error between analysis and background trajectories. An artistic rendition of forecast error temporal evolution is provided for these trajectories in Fig. 3 of Langland and Baker (2004). In this sketch, the initial difference at the analysis time is relatively small, it then grows slightly with time, followed by a contraction. This means that the analysis impacts would be greatest during the intermediate forecast hours if this sketch is an accurate representation of reality. Here, we will estimate the temporal evolution of the forecast error along the analysis and background trajectories to better understand the relative impact of analysis and lateral boundary updates as forecast lead time increases.

The root mean square of forecast error with respect to aircraft observations (MDCRS, ACARS, AIREP) normalized by the observation error assigned by NAVDAS is shown in Fig. 3 for the analysis and background trajectories over the two weeks of interest. To calculate this error, the forecast trajectories are output every

15 minutes and the NAVDAS observation operators are used to compute the observation minus forecast values. Aircraft observations were used because they are well distributed with respect to forecast lead time (bottom panel of Fig. 3). The initial time bin spans from 15 min before to 15 min after the analysis time. All other time bins are 15 min in length. Therefore, the number of observations in the initial bin is roughly two times as large as any other bin. Vertical lines were drawn at 6, 12, and 24 h to highlight the estimate of FER at these leads times for which observation impacts will be calculated.

The largest difference in error between the analysis and background trajectories is at the analysis time. The error of the analyis trajectory grows relatively rapidly with respect to the background during the first 6 h of the forecast and then only slightly faster through 24 h. This means that the difference in forecast error between the two trajectories decreases during the first 24 h. If this is a good proxy for FER, then FER would also decrease during this time. Therefore, observation impacts would be expected to decrease with forecast lead time. Past 24 h, the difference in error is roughly constant to slightly increasing. In other words, the assimilation of observations has its greatest impact in reducing forecast error at the analysis time, and this impact generally decreases through the first 24 h of the forecasts.

Aircraft observations primarily sample the atmosphere at flight levels above 30,000 ft. Measurements are also taken during ascents and descents that provide some information on error throughout the depth of the troposphere. Although this analysis is not an optimal recreation of FER, it is an improvement on the hypothetical sketch provided in Langland and Baker (2004).

4 Model Space Impacts

Analysis and lateral boundary impacts were previously compared for various areas over which the forecast error was calculated Amerault et al. (2013). Here, we will investigate how model space (analysis and lateral boundary) impacts vary with forecast lead time (6, 12, and 24 h).

The ratio of analysis and lateral boundary impacts to FER for 6, 12, and 24 h (vertical lines in Fig. 3) are shown in Fig. 4. The red portion of the bars correspond the reduction in forecast error due to analysis updates, while the black portion is due to lateral boundary updates. A combined value of 1.0 for the red and black bars (analysis and lateral boundaries) would mean that the COAMPS atmospheric adjoint model was able to properly attribute all FER to the analysis and lateral boundary fields. In general, these ratios are less than 1.0 because the linear adjoint model is not able to account for all of the nonlinear model's error reduction.

For 6 h forecast error (Fig. 4a), the combined impact of analysis and lateral boundary updates generally accounts for more than 80% of the FER. This value is similar to the ratio of analysis impacts to FER in global FSOI systems which do not have lateral boundary considerations. The analysis impacts are much greater than those for the lateral boundaries, which are generally below 10% of the the FER.

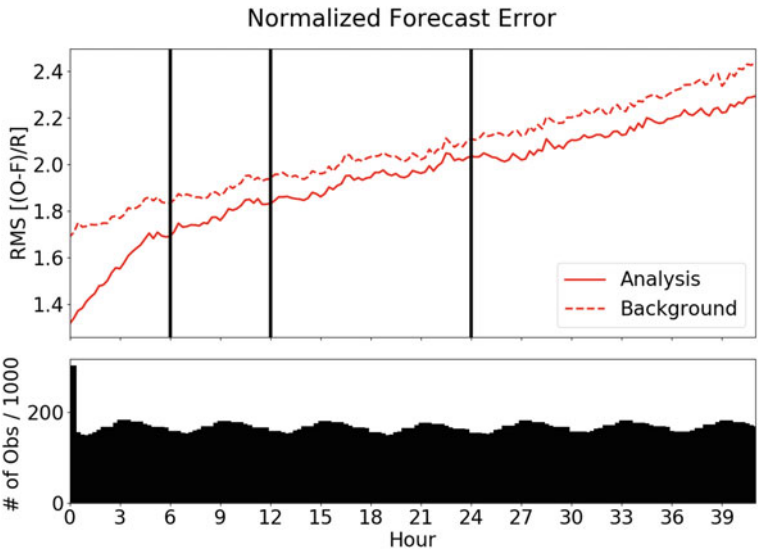

Fig. 3 Average model forecast error out to 42 h measured against aircraft observations for March 1–14 2020. The number of observations used in the error calculation for every 15 min interval is shown in the bottom panel

At 12 h, the ratio of the combined impacts to the FER is generally below 80%. The relative impact of lateral boundaries is greater when compared to 6 h and generally ranges from 10–30% of the FER. The range of ratios shown in the figure fall between 0 and 1. However, for some dates the ratios fall outside of this range. For ratios greater than 1, the FER is usually positive. This means the error in the backgroud trajectory is less than the error in the analysis trajectory and the adjoint estimate grows. For ratios less than 1, the FER is typically a relatively small value and the loss of information in the adjoint model causes the impact estimate and ratio to change sign. These are not uncommon occurrences in a limited-area model where the error is calculated over a much smaller area compared to a global system.

Finally for 24 h, even more ratios fall outside of the figure range, meaning there are more cases with little to no error reduction between trajectories when compared to the 6 and 12 h verification times. However, for those dates where the ratios do fall within the figure bounds, the lateral boundary impacts can be just as large and for some forecast cycles, larger than the analysis impacts. The area and time of interest (midlatitudes winter for this experiment) is expected to effect the ratio of analysis to lateral boundary impacts. Furthermore, the impact based on location of the boundary (north, south, east, or west) with respect to the area over which the error was calculated is interesting, but beyond the scope of this study.

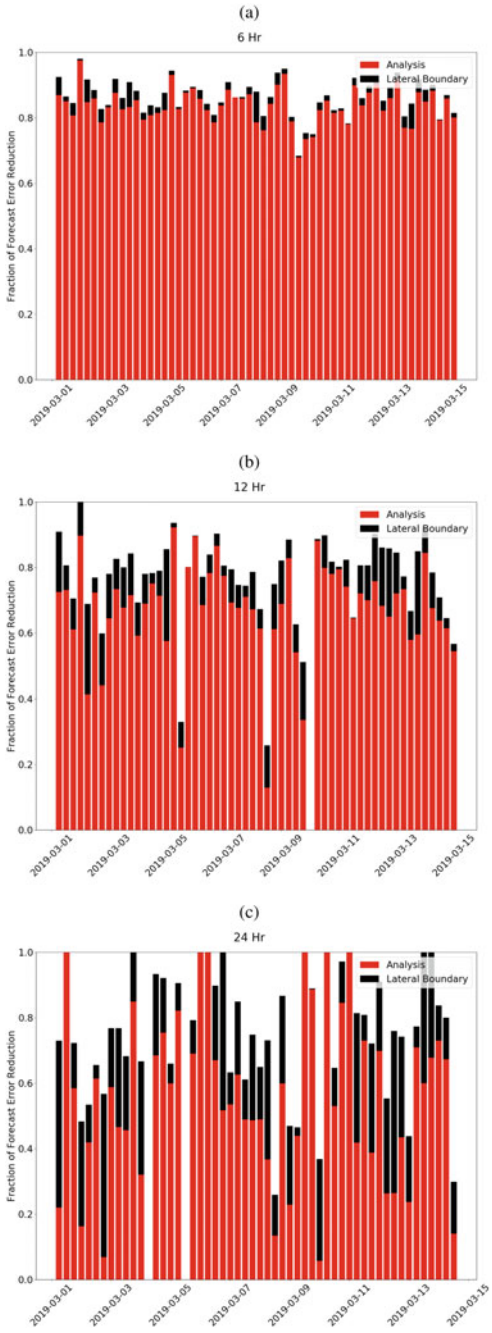

Fig. 4 Ratio of δe_f^g to Δe_f^g for each forecast cycle from 1–14 March 2020. The red (black) portion of the bar indicates the analysis (lateral boundary) impact

5 Observation Impacts

Observations impacts were calculated for the three cases discussed in Sect. 4 by feeding the analysis impacts into the adjoint of the NAVDAS solver. The total impacts grouped by observation type for the 6, 12, and 24 h verification times are shown in Fig. 5. Aircraft observations are the most impactful for each verification time followed by satellite feature track winds. These are also the most plentiful types of observations in NAVDAS over the open ocean area covered by the verification box (Fig. 2). ATOVS retrievals and radiosondes measurements are the next most important types, but their relative rank depends on the verification time. Beyond that, the impacts are substantially lower, owing to lower observation counts. The observation types with zero impact also have zero observations counts over this domain and time period. Furthermore, there is more variability in the relative ranks of observation types below the top four. However, the relative ranks for the bulk of the impacts are similar between verification times. In other words, observations types that are important in reducing error at 6 h, are also important in doing the same at 12 and 24 h. Therefore, 6 h impacts can be as informative as 24 h impacts with less concern for the model space impacts being able to account for the nonlinear FER.

Additionally, the total impacts (sum of the blue bars in each figure) decreases as the verification time increases. This validates what was seen in the FER (Sect. 3). As the forecast time increases (through 24 h), the FER decreases and the corresponding observation impacts also decrease. Although it is important to note that some of the observation impact decrease can also be attributed to the decreasing ability of model space impacts to match FER as forecast lead time increases.

5.1 Radiosonde Verification

As noted above, the relative rank of observation impacts largely reflect the relative number of an observation type in a given domain of a limited-area model. To mitigate some of this effect, the ability to calculate the forecast error against radiosonde observations (as opposed to self-analyses as was done in the previous experiments) was added to the COAMPS FSOI system. To do this, Eqs. 1 and 2 were recast in observation space using the NAVDAS operators to go from COAMPS model space to observation space. The forecast error was computed at all radiosonde locations inside the forecast domain. The metric was computed in terms of dry energy using only the wind and temperature observations. The adjoint of the NAVDAS operators were used to cast the error information from observations space back onto COAMPS model space.

The impacts for this new metric along with the traditional metric are shown in Fig. 6. Radiosondes are only available at 0000 and 1200 UTC so impacts were only calculated every 12 h, as opposed to every 6 h in the previous results. The difference between Fig. 6(b) and Fig. 5(b) is that the impacts at 0600 and 1800 UTC have been

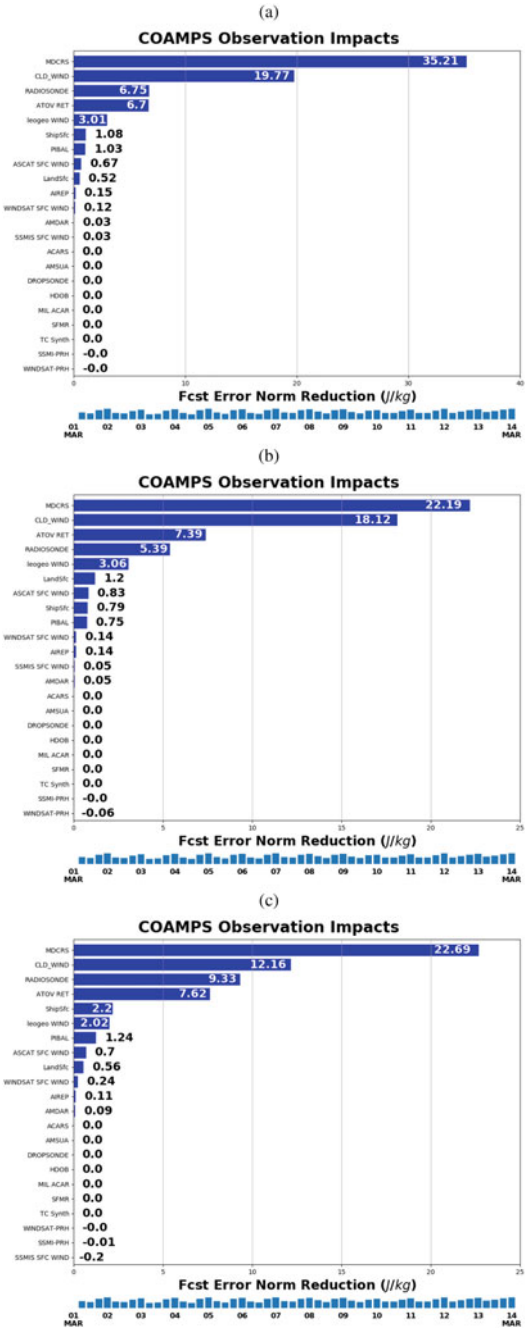

Fig. 5 Observation impacts for (**a**) 6, (**b**) 12, and (**c**) 24 h lead times. The number of assimilated observations for each cycle is indicated in the histogram below each plot

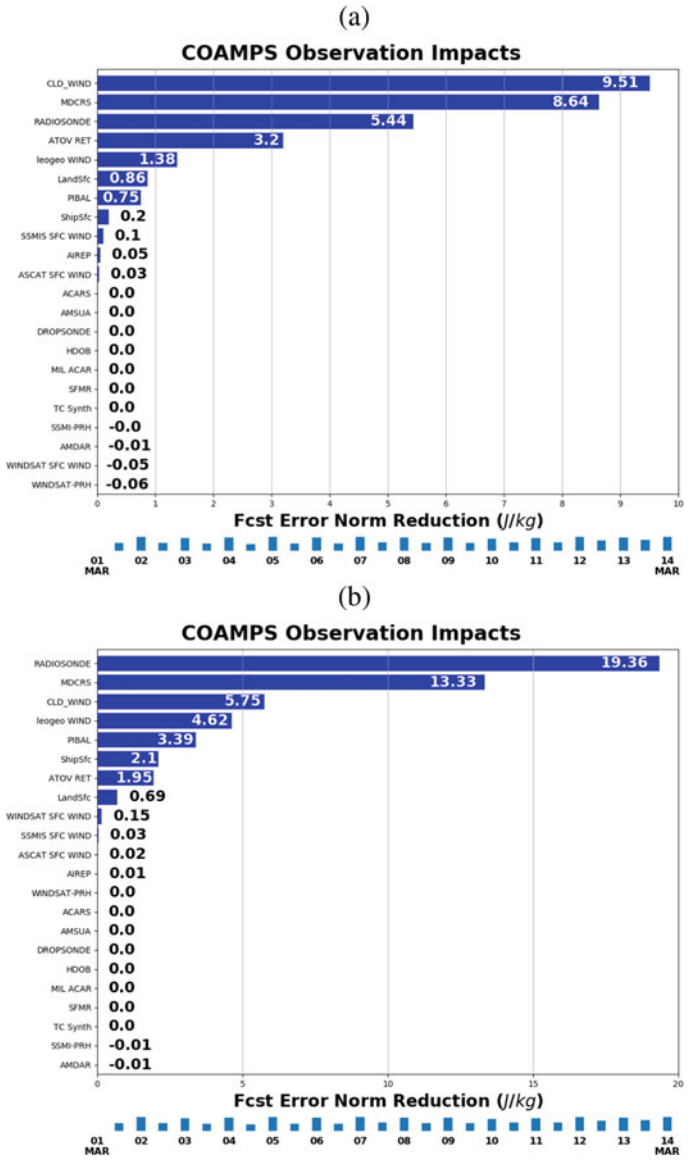

Fig. 6 Observation impacts using (**a**) radiosonde observations and (**b**) self-analyses for verifcation. The number of assimilated observations for each cycle is indicated in the histogram below each plot

removed. Radiosondes have the greatest impact in reducing forecast error when they are also used as truth. Aircraft and satellite feature track winds continue to be important and rank second and third respectively. ATOVS retrievals, which are roughly as important as radiosondes in the self-analysis case, fall to seventh when radiosondes are used as truth. Also to note, in the self-analysis case, when the off times are removed, satellite feature track winds are slightly more impactful than aircraft observations.

6 Summary

Impacts in analysis, lateral boundary, and observation space were investigated with a limited-area FSOI system. As forecast lead time increases, FER decreases and the relative impact of the lateral boundaries increases. The ability of the COAMPS atmospheric adjoint model to capture the FER also decreases with increasing lead time. However, it may not be necessary to compute observation impacts for longer lead times, because those observations that are important to reducing forecast error for shorter lead times are also important at longer lead times. The additional ability to compute forecast error against radiosondes was presented and resulted in a different ranking of observation types. This ability will be useful in future studies for different model domains.

Acknowledgements Computational resources from the Department of Defense's High Performance Computing Modernization Program were vital to this work. I also thank an anonymous reviewer who provided valuable comments and suggestions to improve this chapter.

References

Amerault C, Sasheygi K, Pauley P, Doyle J (2013) Quantifying observation impact for a limited area atmospheric forecast model. In: Park S, Xu L (eds) Data assimilation for atmospheric, oceanic, and hydrologic applications, vol II. Springer, pp 125–145. https://doi.org/10.1007/978-3-642-35088-7_6

Amerault C, Zou X, Doyle J (2008) Tests of an adjoint mesoscale model with explicit moist physics on the cloud scale. Mon. Wea. Rev. 136:2120–2132. https://doi.org/10.1175/2007MWR2259.1

Baker N, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation-targeting problem. Q. J. R. Meteorol. Soc. 126:1431–1454

Cardinali C (2009) Monitoring the observation impact on the short-range forecast. Q J R Meteorol Soc 135:239–250

Daley R, Barker E (2001) NAVDAS: Formulation and diagnostics. Mon Wea Rev 129:869–883. https://doi.org/10.1175/1520-0493(2001)129<0869:NFAD>2.0.CO;2(2001)129<0869:NFAD>2.0.CO;2

Gelaro, R., Zhu, Y.: Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. Tellus 61A, 179–193 (2009)

Hodur R (1997) The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). Mon Wea Rev 125:1414–1430. https://doi.org/10.1175/1520-0493(1997)125<1414:TNRLSC>2.0.CO;2

Hogan T, Liu M, Ridout J, Peng M, Whitcomb T, Ruston B, Reynolds C, Eckermann S, Moskaitis J, Baker N, McCormack J, Viner K, McLay J, Flatau M, Xu L, Chen C, Chang S (2014) The Navy Global Environmental Model. Oceanography 27:116–125. https://doi.org/10.5670/oceanog.2014.73

Jung BJ, Kim HM, Auligne T, Zhang X, Zhang X, Huang XY (2013) Adjoint-derived observation impact using wrf in the western north pacific. Mon Wea Rev 141(11):4080–4097. https://doi.org/10.1175/MWR-D-12-00197.1

Langland R (2005) Observation impact during the North Atlantic TReC-2003. Mon. Wea. Rev. 133:2297–2309

Langland, R., Baker, N.: Estimation of observation impact using the NRL atmospheric variational data assimilation system. Tellus 56A, 189–201 (2004)

Zhang X, Wang H, Huang XY, Gao F, Jacobs N (2015) Using adjoint-based forecast sensitivity method to evaluate tamdar data impacts on regional forecasts. Advances in Meteorology 2015:13. https://doi.org/10.1155/2015/427616

Assimilation of In-Situ Observations

Patricia M. Pauley and Bruce Ingleby

1 Introduction

In-situ observations have been used in the initialization of model forecasts since the inception of operational NWP in the 1950s (see the Appendix for acronym definitions). While in those early days in-situ observations were the only ones available for use, in the present day remote sensing provides roughly 90% of the observations assimilated in operational global models. Even so, in-situ observations provide a larger impact in NWP forecasts than one might expect. For example, radiosonde, surface, and aircraft data together comprised 9% of the observations assimilated in ECMWF's IFS (11% for the U.S. Navy's global model NAVGEM) in January 2020, but provided 26% of the error reduction in 24-h IFS forecasts (29% for NAVGEM 24-h forecasts), based on Forecast Sensitivity to Observation Impact (FSOI; Langland and Baker 2004; Cardinali 2009). The impact of in-situ observations can also be evaluated through Observing System Experiments (OSEs); a recent OSE performed at ECMWF found that denying in-situ observations (together with Doppler wind profiler observations) had the greatest effect for the Northern Hemisphere out of all the observation groups examined, with the day 3 forecast error increased by 10% (Bormann et al. 2019). Similar results were obtained for the Arctic in an OSE where only polar in-situ data were denied, despite the comparative paucity of such observations poleward of 60° (Lawrence et al. 2019). Although some limited-area

P. M. Pauley (✉)
Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA
e-mail: patricia.pauley@nrlmry.navy.mil

B. Ingleby
Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_12

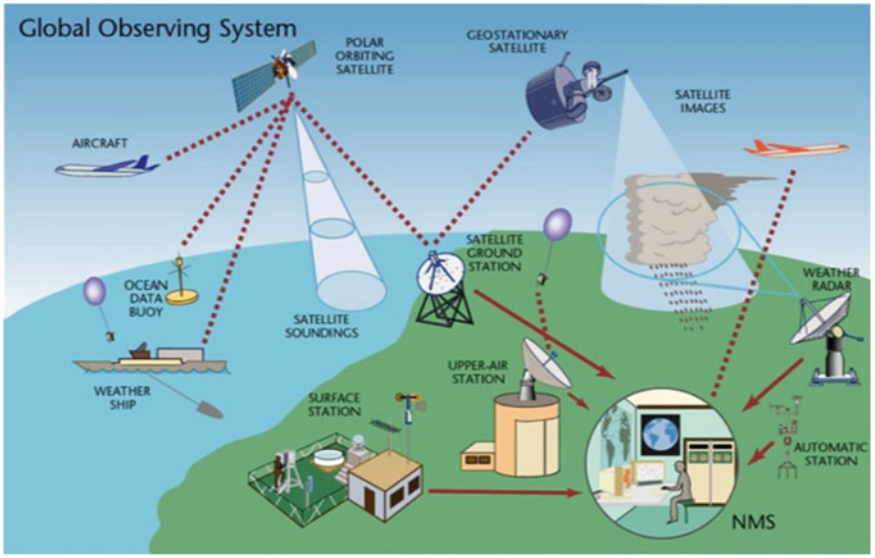

Fig. 1 Observation types available for use in NWP. In-situ observations come from radiosondes (from land and ship platforms), surface platforms (automatic and manual land surface stations, ships, and buoys), and aircraft (WMO)

atmospheric data assimilation systems use extra information from surface reports and ocean data assimilation systems use additional surface and sub-surface information from buoys and other ocean-going platforms, this chapter will focus on the use of in-situ observations in current operational atmospheric global NWP. Figure 1 shows the wide variety of observations currently available for use in NWP.

In addition, in-situ observations are important to NWP for reasons that extend beyond their direct impact in data assimilation. They are widely used in the verification of NWP forecasts, and they also serve as anchor observations (Eyre 2016) in variational bias correction (VarBC) schemes (Derber and Wu 1998) used to counter biases present in satellite radiances. While FSOI only measures the direct impact, OSEs can measure the total impact including bias correction effects. Bormann et al. (2019) found that denying in-situ observations led to not only the increase in random forecast error mentioned above but also changes in mean analyses and in the bias corrections produced by VarBC.

Assimilating in-situ observations is straightforward in some respects since the instruments measure model or analysis control variables (such as temperature, humidity, and wind) more or less directly, as detailed in subsequent sections. This means that the observation operator (using the nomenclature from Ide et al. 1997) is often a simple interpolation in space and in time. Converting wind speed and direction to u and v wind components and converting humidity from the observed to the analysis variable are typically done as part of the observation pre-processor rather than the observation operator. In addition, observation errors must be specified in the

data assimilation system, since they are used in weighting observations from a given instrument type relative to observations from other types. The observation error must account for instrument error as well as representation error, which is often greater in magnitude than instrument error (Janjic et al. 2018). Representation error includes error due to unresolved scales/processes as well as error due to inadequacies in the observation operator and in preprocessing/quality control. The observation error is typically assumed to be Gaussian, unbiased, and uncorrelated in time and space and with the model background. Difficulties with these assumptions as applied to specific in-situ observations will be discussed later in this chapter.

In-situ observations themselves can be very messy, in part a result of approximately 200 NMSs providing in-situ reports. The same observation can be disseminated in multiple formats on the GTS (Global Telecommunication System, a WMO-governed service that enables the sharing of near-real-time meteorological observations among member countries (WMO 2020a), now regarded as part of the WMO Information System (WIS).) The older WMO formats are commonly referred to as TAC (Traditional Alphanumeric Codes) and are forms of packed ASCII, while the newer WMO format is BUFR (Binary Universal Format for the Representation of meteorological data) and is a binary format as its name specifies. At present, many in-situ observations are encoded in both TAC and BUFR with differing precision and vertical/temporal resolution, necessitating some form of duplicate checking that extends beyond searching for exact matches. In addition, individual countries can differ in their interpretation of the WMO rules governing data formats, even sometimes misinterpreting the rules and so introducing errors in the observations. Originally, all in-situ observations were measured, encoded, and transmitted manually, but there has been increasing automation of these processes so that most aircraft and radiosonde reports are now automated. Some countries have largely automated their surface observations whereas others still rely heavily on human observers. Even well trained and motivated observers make occasional mistakes such as mistyping a digit or transposing two digits. There are even stories of observers fabricating a sequence of observations at the start of their shift so as to do something else (such as sleep during a night shift). In general, automation has led to improved quality of observations, but checks are still needed to detect errors.

Observation quality control and monitoring

Effective quality control (QC) is critical for in-situ observations given the issues mentioned above. One important quantity used in QC is Observation minus Background (O-B) departures (also referred to as observation increments or innovations). The background here is a short-term forecast from the NWP model used in the data assimilation system. Defining background values can range from simply interpolating a six-hour (or twelve-hour) forecast to the observation location, to including time interpolation using three-hourly forecasts, all the way to using the background value determined by 4DVAR with even finer time resolution (e.g., 15–30 min at ECMWF).

One use of O-B departures is to estimate the observation error. Assuming that specified observation and background errors are correct, uncorrelated with each other, and

unbiased, the root-mean-square (RMS) difference can be estimated as $\sqrt{<(O\text{-}B)^2>} \cong \sqrt{<(\sigma_o{}^2 + \sigma_b{}^2)>}$, where σ_o is the standard deviation of the observation error, σ_b is the standard deviation of the background error, and <…> indicates an average over a suitable sample. This provides an upper bound for σ_o; sometimes a simple fraction of the RMS difference will be used for the estimate. Hollingsworth and Lönnberg (1986) provided a more sophisticated method for use with a network of observations that can be assumed to have the same error characteristics, but to be uncorrelated with each other. Cross-correlations (or cross-covariances) are calculated and then extrapolated to zero distance: the spatially correlated part is taken to be the background error and the uncorrelated part the observation error. Another approach is based on the work of Desroziers et al. (2005) who used the covariances of O-B departures with O-A (Observation minus Analysis) departures to improve the estimates of observation error covariances. Even when the assumption that observations are given the correct weight in the assimilation only approximately holds, this diagnostic can give useful information (Waller et al. 2016).

Another use of O-B departures is to reject (or downweight) observations that are far from the background value. A simple background check might reject a value if $|O\text{-}B|^2 > k^2(\sigma_o{}^2 + \sigma_b{}^2)$, where k is a constant, often between 3 and 5. There is often also a "buddy check" comparing an observation with its neighbors, largely in case the background is worse than usual. This can be put on a more theoretical basis by assuming that with probability $(1\text{-}P_G)$ the observation error comes from a Gaussian distribution with standard deviation σ_o and that with probability P_G it comes from a much broader distribution. In the latter case, it is said to have a gross error and should be rejected/ignored in the assimilation step. This construct makes it is possible to use a variational framework to iteratively improve QC decisions (Ingleby and Lorenc 1993; Andersson and Järvinen 1999). While variational QC is very useful, it cannot easily cope with correlated observation errors. Tavolato and Isaksen (2015) took a further step, incorporating the Huber norm where observations with large O-B departures have a gradually reduced weight rather than a sharp accept/reject cutoff. The actual proportion of gross errors amongst in-situ observations is only about 1% or less, but decisions on 'borderline' cases can be quite important—a large departure might indicate that an observation is bad or that the forecast has underestimated a cyclogenesis event.

In the 1980s, it was realized that statistics from data assimilation systems could be used to detect systematic problems from certain observing stations (Hollingsworth et al. 1986). One notable problem revealed with this method was a wind direction bias of over 12° from a very isolated radiosonde station. The station operators were contacted and confirmed the error and "the necessary corrective action was taken". This has been the model for feedback of observation quality problems since then, for all types of observations. Typically, statistics are produced for a calendar month, particular stations with data that seem problematic are noted, and two actions may be taken: (1) the problematic data are added to a reject list within the data assimilation system (e.g., Haiden et al. 2018), and (2) the data producer may be provided with a summary of the evidence and asked to investigate the issue. Of course, NWP systems also have biases/problems, and so sometimes results are compared from two or more

NWP systems before contacting the observation provider. However, it can be difficult to know who to contact when a problem has been identified, and there is a wide variety of responses, from helpful cooperation (and even then it can take a while for a problem to be resolved) to no response at all (possibly language problems or a reluctance to acknowledge observation errors). EUMETNET (a consortium of European NMSs) has been providing NWP feedback on observation errors and missing observations for over a decade, and this has been successful in maintaining or even improving the observation quality. More recently, the WIGOS Data Quality Monitoring System (WDQMS) has been working on a similar global system, and WMO is establishing Regional WIGOS Centers (RWC) to help with the process. The authors have also been involved with reporting errors more informally.

In addition to using monthly O-B statistics as the basis for defining station reject lists, these statistics can be used to define both a list of platforms having a significant bias and a list of corrections that can be applied to handle that bias. This technique has been successfully applied to surface pressure observations, where the bias often results from incorrect metadata for the station elevation (e.g., Ingleby 1995). Details about these and other QC considerations are discussed for each type of in-situ observations in the following sections.

Metadata and documentation

While nearly all TAC reports from land surface and radiosonde stations do not include position metadata (i.e., latitude, longitude, and station elevation), the corresponding BUFR reports do include the position metadata but with occasional errors, bringing up yet another problem for QC. OSCAR/Surface (https://oscar.wmo.int/surface/#/) is the official WMO repository of metadata for all surface-based observing stations and platforms (replacing WMO Publication 9 Volume A). (In this context, "surface-based" includes platforms such as radiosondes and aircraft that originate at and return to the surface but make upper-air observations, in contrast with space-based platforms (i.e., satellites) that are described in OSCAR/Space (https://www.wmo-sat.info/oscar/spacecapabilities#/).) In practice, NWP centers maintain local lists of station positions based on OSCAR/Surface (possibly supplemented with other information) to provide the positions of TAC reports and, in some cases, to check the positions reported in BUFR. Wrong/inconsistent positions are more of a problem than one might expect, sometimes due to rounding or mistyping (including wrong sign of latitude/longitude), sometimes due to a station relocation (or station identifier reassignment) not reflected in the metadata, and at other times due to a surface station and radiosonde station sharing the same WMO identifier but being more than a few km apart or having elevations that differ by 5 m or more. Problems can also exist when OSCAR/Surface is not updated in a timely fashion or has incomplete information. At times, correct values for station metadata can be virtually impossible to determine, making the data from that station unusable. Examples of problems associated with metadata errors are discussed in Appendix 2.

While documentation about in-situ observations can in some cases be difficult to find, the WMO has made available a lot of useful information from the WMO Instruments and Methods of Observation Program (IMOP) (https://com

munity.wmo.int/activity-areas/imop). In particular, the Guide to Instruments and Methods of Observation, published as WMO-No. 8 (WMO 2018a and other volumes, which are updated every two years), discusses instrument principles and achievable accuracy for all sorts of in-situ instruments. The IOM Report Series (https://community.wmo.int/activity-areas/imop/publications-and-iom-reports) includes instrument/system intercomparisons and presentations from technical conferences, which may also be of interest. The WMO Information System (WIS) (https://community.wmo.int/activity-areas/wmo-information-system-wis) provides information on WMO codes and other data management aspects. And finally, the WMO Operational Information Service (OIS) (https://community.wmo.int/activity-areas/operational-information-service) provides information on operational aspects of the World Weather Watch (WWW). Documentation specific to particular observing systems is mentioned in the sections below.

The remainder of this chapter is organized in sections that describe the primary types of in-situ observations used in operational NWP—radiosonde data, surface land and marine data, and aircraft data—as well as considerations to take into account in their assimilation. A final section will look at the importance of in-situ observations in NWP systems, including the impact of the recent COVID-19 drop in aircraft data.

2 Radiosonde Observations

Radiosondes are a major source of in-situ profile data for NWP, providing observations of geopotential height, wind speed and direction, temperature, and humidity as a function of pressure (or height) by means of a small balloon-borne instrument package. There are about 800 land stations worldwide providing near real-time reports; 175 of these are designated as GUAN (GCOS upper air network) stations, which have a commitment to certain standards and long-term operation. About 20 are GRUAN stations (GCOS reference upper air network, https://www.gruan.org/); these also produce delayed mode reports including uncertainty profiles. These soundings include data from the earth's surface through the lower half of the stratosphere, often up to 10 hPa (approximately 30 km). Most stations report nominally at 0000 and 1200 UTC, although some report only at one of those times and a few make reports at 0600 and 1800 UTC as well. Unlike some in-situ observations, radiosonde measurements are filtered or corrected in various ways before they are provided to users; Dirksen et al. (2014) provide insight into the types of processing applied and also into the accuracy achievable. Radiosonde data are also widely used in verification and validation, not only of NWP analyses and forecasts but also of other observation types, such as satellite-derived winds; they are also valuable for climate change monitoring. And, radiosonde data together with GNSS radio-occultation data are the primary sources of anchor observations for the bias correction of satellite radiance data (e.g., Cucurull et al. 2014).

While radiosondes have been in widespread operational use since the 1940s, they have undergone considerable improvement over the intervening years, not only in

instrumentation but also in automation. Pettifer (2009) describes the entirely manual process used in the 1940s and 1950s, where after the radiosonde launch, an analyst plotted the incoming data, converted the units, selected significant levels, and encoded the observation as the radiosonde ascended, all with the goal of having the message ready to transmit within a few minutes of balloon burst 1–2 h after launch. Most of the computational tasks were automated by the 1980s (Madonna et al. 2020), but until recently personnel were still required to perform the radiosonde ground checks and physically launch the balloon. Three manufacturers now make automatic radiosonde launchers that operate hands-off except for stocking the launcher with radiosondes and balloons once or twice per month. These are in operational use at roughly 90 sites and have a performance similar to manually launched radiosondes (Madonna et al. 2020). However even with these technological advances, radiosondes in current operational use have characteristics that vary widely and that affect their use in NWP (e.g., Ingleby 2017).

The following section describes the characteristics of each of the measurements made by a radiosonde, as well as related observation types and aspects about processing radiosonde data. To keep the discussion more focused, statistics are mainly given for the Vaisala RS92 radiosonde and its successor, the Vaisala RS41. These radiosondes are in use by many countries across the world, accounting for a large fraction of radiosondes currently in use, including by GRUAN stations, and have been more carefully studied than other radiosonde types. However, many different radiosonde types are in use across the world, some of which are used only in certain countries or regions (Fig. 2). Details of O-B statistics by radiosonde type are discussed in Ingleby (2017), with temperature statistics for Northern Hemisphere midlatitudes shown in Fig. 3 as an example of the variability among radiosonde types.

Measurement characteristics

Pressure/height: The vertical coordinate in radiosonde profiles has historically been supplied by pressure measurements, with geopotential height computed from measured pressure, temperature, and humidity via the hydrostatic equation. Solid-state pressure sensors are widely used in current radiosondes with an uncertainty,[1] for example, of 1 hPa at 850 hPa decreasing to 0.4 hPa at 10 hPa for the Vaisala RS92 (see Nash 2015 for a comparison of uncertainty for various pressure sensors). However, many radiosondes do not have pressure sensors, including the RS41 in its most commonly used configuration; geometric height is measured instead and converted to geopotential height. Pressure can then be computed hydrostatically from the geopotential height, temperature, and humidity, or alternatively, height can be used as the vertical coordinate for NWP.

[1] "Uncertainty" is defined in the WMO Guide to Meteorological Instruments and Methods of Observation (WMO 2018a, Sect. 1.6.2) as a quantitative measure of accuracy related to both random variations and any systematic error left after corrections have been applied. For a large number of observations, the uncertainty is expressed as $k\,\sigma/\sqrt{n}$, where σ is the standard deviation, n is the number of observations, and k = 2 for a 95% confidence interval. The true value of the observation is expected to be within ± the uncertainty of the mean of the measured values.

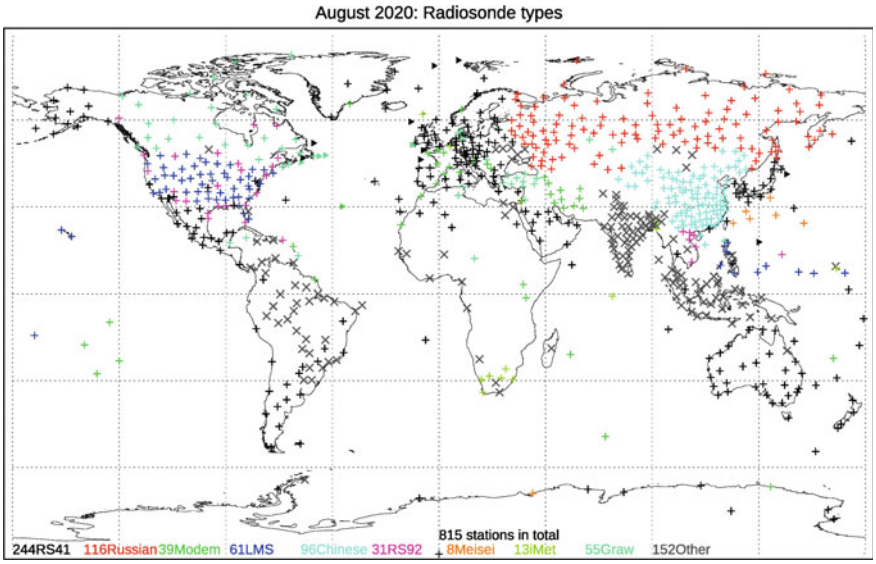

Fig. 2 Distribution of radiosonde types reporting in August 2020. The color coding indicates the radiosonde type (or group of related types) as indicated in the legend, which also gives the number of stations in each group

The two main ways of measuring height in current use are secondary radar and GPS (more generically termed GNSS). Secondary radars are in widespread use in Russia and China (although Chinese radiosondes typically also include a pressure sensor). These radars send a pulse that triggers a response from a transponder on the radiosonde. The time between the transmission of the pulse and the receipt of the response is used to calculate the slant range to the radiosonde, which is combined with the measured elevation angle to compute height (Zaitseva 1993), typically with a constant correction for refractivity and the earth's curvature (Ingleby 2017). However as the wind becomes strong, the elevation angle becomes small, and the radar height becomes less accurate. A 1989 radiosonde intercomparison (Ivanov et al. 1989) found geopotential height differences (relative to the Vaisala RS80) for two Russian systems of 30–40 m below 100 hPa and 100–200 m at 10 hPa, consistent with height O-B statistics shown in Ingleby (2017). Radar height measurements are also sensitive to alignment/leveling errors, which can lead to systematic height errors (WMO 2018a, Sect. 12.3.7.2).

A more accurate means of measuring the geometric height of a radiosonde is through GPS technology. However, the GPS height is given relative to a reference ellipsoid and so must be converted first to height above mean sea level and then to geopotential height (WMO 2018a, Sect. 12.3.6.2; Dirksen et al. 2014); differences between geopotential and geometric height can exceed 200 m at heights above 30 km. In an intercomparison of data from RS92 and RS41 radiosondes flown together, Edwards et al. (2014) found that flight-by-flight differences in GPS-derived heights

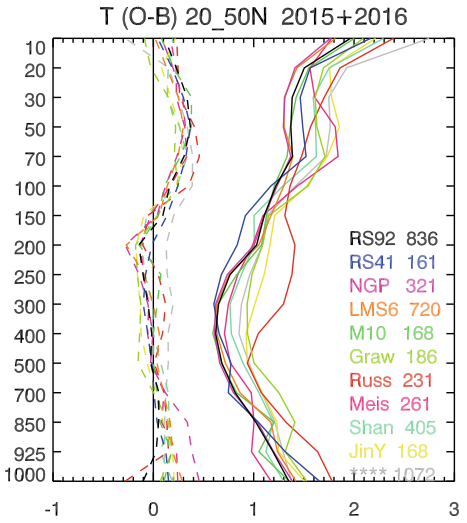

Fig. 3 ECMWF temperature (°C) O-B mean (dashed lines) and RMS (solid lines) statistics for 2015 and 2016 at mandatory pressure levels (hPa) for major radiosonde types for the Northern Hemisphere midlatitudes. The numbers of reports in hundreds are listed with the radiosonde types; only types with at least 2000 reports are shown. The types shown include Vaisala RS92 and RS41, Lockheed Martin LMS6, Modem M10, Graw DFM-09, Shanghai GTS1-1 ("Shan"), and JinYang ("JinY"). Various Russian types are combined and labeled "Russ", and various Meisei types are combined and labeled "Meis". Vaisala RS92 radiosondes using the U.S. NWS ground stations are labeled "NGP", and stations not reporting a radiosonde type ("***") are shown in grey (From Ingleby 2017, Fig. 3.1)

had standard deviations less than 1.4 m for the RS92 and less than 0.8 m for the RS41. In contrast, flight-by-flight differences in RS92 pressure-derived heights increased with height to yield standard deviations of 20 m at 20 km and 140 m at 33 km, a rather dramatic increase in the stratosphere. Accurate GPS heights require using the correct height for the ground station's GPS antenna relative to the surface pressure sensor (e.g., Fig. 9.2.1 in Nash et al. 2011). The accuracy of GPS heights is also affected by the quality of the GPS receiver, the availability of GPS signals, and the degree to which GPS signals are reflected by the ground and delayed in the ionosphere (Vaisala 2017).

GPS-derived heights together with measured temperature and humidity can be used to compute pressure hydrostatically. Uncertainty analyses for pressure derived from GPS height from Nash (2015) for the RS92 and Lehtinen et al. (2016) for the RS41 gave similar results, with uncertainties of 1 hPa or greater in the lower troposphere (mainly due to uncertainty in height), decreasing to 0.05 hPa at 10 hPa. However, even larger errors in GPS-derived pressure can occur under highly non-hydrostatic conditions. For example, Lehtinen et al. (2016) show 4.5–5.5 hPa peak differences in the mid-troposphere between RS92 sensor pressure and GPS-derived pressure for soundings from Taiwan during a typhoon passage. Given these measurement characteristics, the processing for RS92s used in the GRUAN reference

radiosonde network (Dirksen et al. 2014) combines heights derived from pressure at low levels with GPS heights to form an optimal height profile that is then converted to geopotential height and pressure.

Temperature: Radiosonde temperatures are important in global NWP not only as assimilated data, but also as anchor observations in VarBC, as verification for forecasts, and for deducing and monitoring climate change. However, the measurement of temperature is not without problems. One of the key factors in accurately measuring temperature is the rate of response of the sensor (Nash 2015); the sensor must respond quickly as the balloon ascends at approximately 5 m/s. The platinum resistor used in the Vaisala RS41 has time constants [2] similar to the RS92's thermocapacitor (ranging from about 0.5 s at 1000 hPa, to 3.0 s at 10 hPa). Note that some radiosonde ground stations apply proprietary time-lag corrections, including that used for the RS41 (Vaisala 2017).

A greater factor in measuring temperature accurately is either accounting for or preventing radiative heating/cooling of the sensor, which is primarily an issue in the stratosphere. Most radiosondes use an exposed temperature sensor to minimize sensor lag, but this also exposes the sensor to solar and infrared radiation, which can lead to significant temperature errors when uncorrected (Nash 2015). Sensor coatings can essentially eliminate infrared cooling (although not all radiosondes use this technology). For example, the uncertainty for nighttime temperature for the RS92's thermocapacitor and the RS41's resistive platinum sensor is less than 0.6 K (Nash 2015; Jauhiainen et al. 2014). However, solar heating of the temperature sensor is a problem in the stratosphere regardless of sensor type; therefore, the uncertainty of stratospheric radiosonde temperature is larger in daylight than at night (Dirksen et al. 2014, Fig. 10). Many manufacturers make proprietary solar radiation corrections, typically as a function of pressure (or altitude) and solar elevation angle, although factors such as ventilation speed, angle of exposure of the sensor, and clouds play a role. Nash (2015) lists the daytime uncertainty for RS92 corrected temperatures as up to 0.9 K. The Vaisala RS41 proprietary correction is larger than the RS92's, increasing from 0.44 K at 100 hPa to 1.14 K at 10 hPa for a solar elevation angle of 60° (Vaisala 2017).

Raw temperature profiles can contain spikes that are usually warmer than the baseline temperature and are more common in daylight. These are generally removed by processing (e.g., Dirksen et al. 2014) and are likely due to the radiosonde passing through the wake/shadow of the balloon as it swings in pendulum motion on its ~30 m tether, thermal contamination from the radiosonde packaging, and the change in illumination as the radiosonde spins on its tether (Shimizu and Hasebe 2010; Dirksen et al. 2014). While changing the sensor mount can mitigate the effect of the radiosonde packaging, Shimizu and Hasebe (2010) found that a radiation correction that relies only on the solar zenith angle cannot completely correct for the other two effects. In addition, cloud effects (shadowing beneath/within clouds or backscattered radiation from clouds below the radiosonde) can only be handled in an average sense

[2] Sensor lag is typically phrased in terms of a time constant, the time the sensor needs to respond to a sudden change by $1 - 1/e$ (approximately 63%) of its asymptotic value.

in radiation correction schemes (Dirksen et al. 2014). The inability of manufacturer-specified radiation corrections to counteract solar heating completely as well as the use of country-specified corrections for some radiosonde types has led some NWP centers to use bias correction techniques to make further adjustments to radiosonde temperatures.

As an example, the operational ECMWF radiosonde bias correction scheme (Agustí-Panareda et al. 2009; Ingleby 2017) uses O-B statistics for the previous year to calculate bias corrections as a function of radiosonde type, pressure, and solar elevation angle and are updated once per month. Currently, nighttime data for RS92s and RS41s are assumed to be without bias and are used to estimate the model bias, which is then subtracted from the O-B statistics for each radiosonde type. For good modern radiosondes, the temperature and humidity corrections are quite small; in fact, ECMWF switched off humidity bias corrections for RS41s, since they seemed to be doing more harm than good. This scheme is also used in the ERA5 reanalysis for recent years, although different methods are used for earlier years (Hersbach et al. 2020). An examination of the ECMWF results shows that one particular radiosonde type has a relatively large temperature bias correction related to a small proportion of profiles with a large bias that often increases with height. There is some circumstantial evidence that mishandling during the launch is the main cause of the bad profiles.

Humidity: Measuring humidity is much more difficult than measuring temperature. Specific humidity/mixing ratio varies over a large dynamic range—four to five orders of magnitude between observed values in the lower troposphere and the stratosphere. Relative humidity constrains this variation, but introduces a temperature dependence that brings its own set of problems. And, humidity is typically reported as dewpoint temperature or dewpoint depression, so problems can arise in converting from the measured variable to the reported variable. Note that relative humidity measurements are reported with respect to saturation over a planar water surface regardless of temperature. WMO (2018a, Chapter 12) states, "The saturation with respect to water cannot be measured much below -50 °C, so manufacturers should use one of the following expressions for calculating saturation vapor pressure relative to water at the lowest temperatures—Wexler (1976, 1977), Hyland and Wexler (1983), or Sonntag (1994)." We recommend that NWP centers also use one of these formulae when converting reported dew point temperatures to another humidity variable.

Sensor lag is an even greater issue for humidity than for temperature. Time constants for humidity sensors are generally given as a function of temperature, with the response of the sensor slowing down at very cold temperatures. Widely used capped thin-film capacitors have a time constant of 15–20 s at -40 °C and 150–300 s at -70 °C. The exposed twin thin-film capacitor used in the Vaisala RS92 and the RS41's heated single thin-film capacitor are faster, with time constants of 10 s at -40 °C, and 80 s at -70 °C (Nash 2015; Edwards et al. 2014). Time-lag corrections for humidity sensors have been extensively studied (e.g., Miloshevich et al. 2004; Dirksen et al. 2014) and are widely used in operational radiosonde systems, but should be used only for time constants less than 2–3 min (Nash et al. 2011).

Wetting/icing of the temperature and humidity sensors in a cloud layer can lead to sensor failure (e.g., Wang et al. 2003; Liu and Tang 2014) or more commonly to "wet-bulbing"—evaporative cooling decreasing the temperature to the wet-bulb temperature and the presence of liquid water or ice increasing the relative humidity (Nash 2015). In some cases, wet-bulbing can lead to superadiabatic lapse rates right above the cloud layer (Hodge 1956; Dirksen et al. 2014). Vaisala's use of a hydrophobic coating on the RS92's sensors allows the temperatures and humidity to recover more quickly than the other radiosonde types examined in Nash et al. (2011). The Vaisala RS41 has even better performance under these conditions (Edwards et al. 2014). The impact of wetting/icing on temperature and humidity is typically not corrected—even the detailed RS92 GRUAN processing lacks such a correction (Dirksen et al. 2014). However, QC checks can be used to mitigate sensor failure and excessive relative humidity above a cloud layer (Wang et al. 2003; Liu and Tang 2014; Ingleby and Edwards 2015, Appendix S1).

Various biases in humidity measurements have also been documented. A dry bias for the RS92 results from daytime solar heating of the sensor leading to a decrease in relative humidity (Vömel et al. 2007). Dirksen et al. (2014) and Wang et al. (2013) have devised corrections for this bias, but residual biases can still exist when conditions differ from those assumed in the correction (e.g., a residual moist bias when cloudiness decreases the solar heating). The RS41 requires no solar radiation correction, since it uses temperature from an on-chip sensor in its relative humidity calculation (Jauhiainen et al. 2014). On the other hand, a moist bias is often seen in the upper troposphere and lower stratosphere, in some cases resulting from the slow response of humidity sensors at cold temperatures, but in other cases reflecting adjustments used by some manufacturers to artificially reduce relative humidity to a "reasonable" value above the temperature-derived tropopause (Nash et al. 2011). As a result, ECMWF (and other centers such as FNMOC) only use radiosonde humidity for temperatures warmer than -40 °C or pressures greater than 300 hPa by default, but use humidity for Vaisala RS92 and RS41 down to -80 °C, reflecting that humidity from the best radiosondes is usable up to the tropopause. Jauhiainen et al. (2014) cite an uncertainty of 4% RH for the RS41 and 5% RH for the RS92. At low temperatures, it also matters which saturation vapor pressure equation is used. ECMWF recently changed to use the Sonntag equation (one of those recommended by WMO); this reduced O-B biases for the RS41 in the upper troposphere (Ingleby 2017).

Winds: Traditionally, winds have been determined by measuring the drift of the balloon over time and assuming that the balloon is drifting with the wind. This is done using tracking devices to determine the balloon's azimuth, elevation, and range, either following the balloon visually using an optical theodolite, by following the radiosonde's radio signals using a radiotheodolite, or by following the balloon/radiosonde itself using radar as previously described (WMO 2018a). While radar techniques explicitly calculate the slant range to the radiosonde, optical theodolites and radiotheodolites infer slant range from geopotential height computed from thermodynamic measurements and measured elevation angle (WMO 2018a).

Tracking errors vary among these methods. Using a radiotheodolite provides advantages over optical tracking in that radio direction finding is typically automated, more precise, and can occur in the presence of clouds that obscure visual tracking of the radiosonde. However at low elevation angles, multipath interference can lead to wind errors when reflections of the radiosonde's signal from nearby surfaces are received in addition to the direct signal (WMO 2018a). In this case, the strongest signal does not point to the radiosonde and wind errors can exceed 10 m/s, sometimes leading to large outliers that must be removed (Houchi et al. 2015). The theoretical vector wind error increases with height and with the ratio of wind speed to rate of ascent for both radar and radiotheodolite systems, but the increase is much larger for radiotheodolites; consequently, vertical resolution in the upper-troposphere and lower stratosphere is often decreased by increasing the time interval to avoid having the measurement swamped by error (WMO 2018a). Estimates of typical random vector errors obtained in the 1989 WMO radiosonde intercomparison were 2.6 m/s for a Russian secondary radar and 4.8 m/s for a portable radiotheodolite, both at 18 km for relatively light winds—only three times the balloon ascent rate. In addition to tracking errors, both radiotheodolite and radar winds can have a bias in wind direction if the radar itself is not correctly aligned (WMO 2018a).

GPS (or generically GNSS) windfinding is now used by almost 70% of radiosonde stations worldwide (Ingleby, 2017, Fig. 5.1). The process of determining winds using GPS for most modern radiosondes is fundamentally different from tracking the radiosonde/balloon. Typically, GPS signals from a minimum of four satellites are received by both the radiosonde itself and by the ground station. While some signal processing is performed on board the radiosonde, the main signal processing is performed in the ground station, where the GPS information from the radiosonde is differenced from similarly processed information from the ground station for each satellite. This removes errors that are common to both sets of signals and isolates the Doppler shifts associated with the motion of the radiosonde, which are then used to compute wind speed and direction (WMO 2018a; Kaisti et al. 1994). Highly accurate latitudes, longitudes, and heights are also provided by the differential GPS calculations. GPS windfinding can be done at a much higher resolution than traditional methods and so is able to capture the 10–15 s period of the radiosonde's pendulum motion (WMO 2018a). Differences in algorithms to filter out this motion lead to small differences among GPS radiosonde winds (Nash et al. 2011); some of these algorithms may oversmooth. An example of raw and filtered wind data is shown in Fig. 4. The "noise" varies not only within the profile but also day-to-day. Sometimes gravity wave activity in the stratosphere can add to the representation error.

Errors for GPS windfinding are smaller than those for windfinding by radiotheodolite or radar and have much less variation with height. GPS windfinding is not subject to errors at small elevation angles as viewed from the launch point and can give high quality winds at the large wind speeds that give small elevation angles, so long as no RF interference is present and a sufficient number of GPS satellites are available. However, poor GPS reception does occasionally lead to missing winds, at times with bad winds near the gaps. The estimated random vector error for GPS winds in the 2010 WMO radiosonde intercomparison using 2-min averaging (2 km)

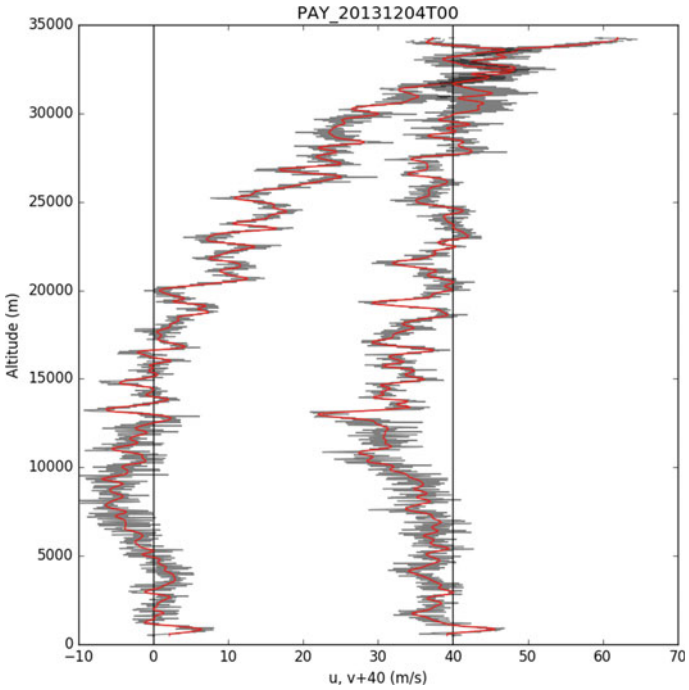

Fig. 4 Profiles of u- and v-component winds from a GRUAN Vaisala RS92 sounding from Payerne, Switzerland (06010) for 0000 UTC 4 December 2013. Note that 40 m/s was added to the v-component to separate the profiles on the graph. Raw GPS-derived winds are shown in gray, while the filtered wind is shown in red

was 0.15–0.4 m/s above 17 km (Nash et al. 2011; WMO 2018a). This is roughly an order of magnitude smaller than the values for secondary radar and radiotheodolite from the 1989 WMO radiosonde comparison. This estimated random vector error increased with decreasing averaging period, with 30-s averaging (300 m) roughly doubling and 10 s averaging (100 m) roughly tripling the values for 2-min averaging above 17 km (WMO 2018a). The UK Met Office comparison of RS92 and RS41 radiosondes (Edwards et al. 2014) found that flight-by-flight differences were only 1–2 cm/s with standard deviations generally less than 10 cm/s.

Related observing systems

Ship radiosondes: There are usually between 3 and 10 ships launching radiosondes on any given day, mostly in the North Atlantic through the E-ASAP (EUMETNET Automated Shipboard Aerological Program) ships. As of 2018, there were 18 E-ASAP ships that provided 85% of ship soundings (Krockauer and Prates 2018). While specialized equipment is often used to launch the radiosondes, the radiosondes themselves are the same types as used for land radiosondes. As a result of constraints imposed by satellite communications from the ships, BUFR messages are available on the GTS typically at medium resolution with levels reported every 10 s,

compared to typically 2 s for land radiosondes. While inexperienced crew members and challenging weather conditions contribute to higher failure rates than are seen for land-based radiosondes, ship-based soundings have proven to be valuable in NWP and research for providing soundings in oceanic regions.

Radiosonde descents: Most of the time "radiosonde data" refers to radiosonde/balloon ascents. However, the Vaisala RS41 and a couple of other radiosonde systems can be configured to provide data after balloon burst as the radiosonde descends. Real-time descent reports are currently available from a few countries, mainly in Europe (Ingleby and Edwards 2018). Descents occur more rapidly than ascents, especially right after balloon burst where the air density is low and descent rates can exceed 50 m/s. Depending on the balloon weight, tropospheric descent rates are generally about 10 m/s or less when a parachute is used, but 13–15 m/s with no parachute (compared to a typical ascent rate of 5 m/s). Ingleby and Edwards (2018) found that the O-B statistics in descents were similar to those for ascents, except at the highest levels where descent data are closest in space and time to ascent data and so could be easily excluded (a practice in use at FNMOC and ECMWF). They also found that descent winds were smoother than ascent winds. The digital filter designed to remove pendulum motion from ascent winds is applied to descent winds as well, where it might be less appropriate. The data from descent soundings can end well above the surface as the radiosonde descends behind the horizon/mountains (Ingleby and Edwards 2018). More recent work suggests that descents with a parachute and a pressure sensor are higher quality than those without either. Descents are offset in time and to some extent in space from their corresponding ascents, in effect providing two soundings for the cost of one.

Pilot balloons: Wind-only profiles can be made using pilot balloons and are sometimes referred to as pibals. There is no instrument package suspended beneath the balloon so no thermodynamic observations are made; wind observations are made by tracking the balloon, usually with an optical theodolite. Since no height observations are available, heights are estimated by assuming an ascent rate for the balloon and measuring the time elapsed from balloon launch. Departures from the assumed ascent rate therefore lead to errors in the estimated heights (WMO 2018a), and wind errors result from limitations in the tracking technique as described above for radiosondes. As a result, these wind profiles often do not extend as high as radiosonde wind profiles. Pilot balloon observations were first routinely made in the early twentieth century (Douglas et al. 2008); at present there are approximately 100 stations mainly in Asia providing wind-only profiles from pilot balloons.

Dropsondes: Another radiosonde variant is the dropsonde—an instrumented package that takes observations as it falls beneath a parachute after being ejected from an aircraft. These are deployed mostly for operational monitoring and forecasting from reconnaissance aircraft near or in tropical cyclones threatening landfall, especially along the U.S. East and Gulf Coasts. They are also used in field experiments such as recent campaigns to study North Pacific atmospheric rivers (e.g., Ralph et al. 2017). The unmanned NASA Global Hawk was outfitted with a dropsonde

launcher developed by the National Center for Atmospheric Research (NCAR) for special campaigns in 2011–2017 to fly high-altitude missions primarily over hurricanes (Holger Vömel, personal communication). Other countries such as Taiwan, Hong Kong, Japan, Germany, and the UK also have aircraft capable of deploying dropsondes.

Although the sensor housing for a dropsonde is tube-shaped rather than box-like, the same sensors are used in dropsondes as in radiosondes. NCAR has developed dropsondes since the 1960s, first using radar tracking and then radio navigation (Holger Vömel, personal communication), but using GNSS since the mid-1990s (Hock and Franklin 1999). The "RD94" dropsonde was developed in 2008 using sensors mostly from the Vaisala RS92 (Ikonen et al. 2010); a miniaturized version was used in the NASA Global Hawk campaigns (Wick et al. 2018). The latest generation of NCAR dropsondes is designated "RD41", which is built by Vaisala using the same sensors as their RS41. NCAR has also developed the Airborne Vertical Atmospheric Profiling System (AVAPS—https://www.eol.ucar.edu/content/avaps-aircraft-data-system), which includes the aircraft-based equivalent of a radiosonde ground station that receives and processes the data. The Atmospheric Sounding Processing Environment (ASPEN) is the component of AVAPS that performs data quality control and generates formatted reports. High-resolution native BUFR reports were first generated and disseminated in real-time on the GTS in Fall 2018 and were put into operational use at ECMWF in September 2019 (Ingleby et al. 2020b).

The U.S. National Hurricane Operations Plan (OFCM 2020a) outlines operational flight patterns used by the NOAA and USAF hurricane reconnaissance aircraft; these include synoptic surveillance missions to measure the large-scale environment as well as flight patterns designed to locate the tropical cyclone center and determine its intensity. While dropsondes from synoptic surveillance missions are typically made in a large-scale grid around the storm and are quite helpful for global NWP, dropsondes near the hurricane center have a large representation error as shown by very large O-B values, and so are challenging to use in data assimilation. Bonavita et al. (2017) show an example of a dropsonde from Hurricane Matthew that was released in the eyewall at 700 hPa and was carried roughly one-quarter of the way around the eye. The use of BUFR dropsonde data should help address this problem—having locations at each level rather than assimilating the data at its launch location should provide a more dynamically consistent analysis. But Bonavita et al. (2017) take a more fundamental approach, similar to Tavolato and Isaksen (2015), using an adaptive observation error model subsequently implemented at ECMWF. This model accounts for the greater representation errors for tropical cyclone observations as well as the larger observation operator error in mapping background values to observation locations in or near the extreme conditions in a tropical cyclone; this is done by using the background error variance derived from the Ensemble of Data Assimilations (EDA, Isaksen et al. 2010). This approach downweights the observations in the immediate vicinity of a tropical cyclone center.

Driftsondes: NCAR has also developed a balloon-borne gondola capable of deploying dropsondes when suspended below a long-duration stratospheric balloon

such as developed by the French Centre National d'Études Spatiales (CNES). This balloon/gondola system can carry as many as 54 miniaturized dropsondes and has been deployed for field experiments such as the African Monsoon Multidisciplinary Analysis (AMMA, Redelsperger et al. 2006) and Concordiasi (Rabier et al. 2010). The superpressure balloons used for driftsondes are made of polyester rather than latex and are designed to become neutrally buoyant at approximately 20 km, where they drift with the wind for days to months (Cohn et al. 2013). Observations are transmitted from the sonde back to the gondola and from there to the ground using satellite communications, allowing the driftsonde observations to be put on the GTS in near real-time. Since driftsondes are essentially dropsondes deployed from stratospheric balloons rather than aircraft, the characteristics of the observations are the same as for dropsondes. In addition, GPS measurements of the balloon's motions provide valuable in-situ wind observations.

Data formatting

TAC formatting considerations: The alphanumeric TEMP code (designated by WMO as FM-35) represents radiosonde data as a series of numeric groups without decimal points or negative signs. It is a direct descendent of the code developed for the era of 300 baud teletypes (e.g., U.S. Weather Bureau 1949; WBAN 1957, 1963). Current details about formatting radiosonde data in TEMP messages are given in Vol. I.1 of the WMO Manual on Codes (WMO 2019a). While these details are not important in the current context, there are aspects of the code that do affect their use in NWP.

First of all, a full radiosonde profile from a particular station requires four TEMP messages (or parts). Parts A and B include mandatory- and significant-level data, respectively, below 100 hPa; Parts C and D include mandatory- and significant-level data, respectively, above 100 hPa. Mandatory levels (also called standard levels) are pressure levels that are required to be reported for each sounding. WMO (2019a) specifies these as 1000, 925, 850, 700, 500, 400, 300, 250, 200, 150, and 100 hPa for Part A and 100, 70, 50, 30, 20, and 10 hPa for Part C; the surface level, tropopause level, and the level of maximum winds are also required levels. Mandatory-level data are widely used in model verification since each sounding includes these levels.

Many countries routinely extrapolate heights below ground to mandatory levels, which is done by assuming some sort of hypothetical (virtual) temperature distribution below ground, the details of which are not reported. While these heights can be useful if the extrapolation is only over a short distance, they can be a problem if the extrapolation is over a significant depth. Ingleby et al. (2016b, Fig. 8) show an example for station 04417 in central Greenland where the surface pressure is close to 700 hPa. The 925 hPa heights have roughly a 60 m negative bias compared to the ECMWF background; the 1000 hPa heights have a bias so large that it is off-scale in these diagrams. These fictitious heights should not be used for model verification!

Significant levels are pressure levels that give the change points in a piecewise linear depiction of the sounding, typically linear in the logarithm of pressure. The line segments are chosen to represent the sounding within a certain tolerance (e.g., 1 °C for temperature below 100 hPa). This leads to a representation of a sounding

as the 16 or so mandatory levels plus typically 30–100 significant levels. However, using significant levels directly in an NWP system without interpolation is less than optimal, since these levels are chosen to represent local extrema in the sounding (Ingleby et al. 2016a). Comparing TEMP observations from German radiosondes to ECMWF background values, significant levels have RMS O-B values that are greater than those for mandatory values by more than 0.5 K near the surface and in the stratosphere. Ingleby and Edwards (2015, Appendix S1) describe the practice at the UK Met Office of averaging values over the model layer to better match the observations to the model variables.

Secondly, not all variables are reported at all levels. Geopotential height (or surface pressure), temperature, dewpoint depression (temperature minus dewpoint temperature), wind speed, and wind direction are reported for mandatory levels. The data reported at significant levels only include pressure, temperature, and dewpoint depression; a separately chosen set of significant level winds and corresponding pressure values can also be included at the end of the TEMP Part B and Part D messages.

Third and perhaps most important from an NWP perspective, observations are truncated or rounded to fit within the format—only 2 or 3 digits are used for each variable. In the case of geopotential height, leading digits are also omitted to fit the value in 3 digits. The details of the rounding are given in Table 1. Ingleby et al. (2016b) show a jump in RMS O-B for heights at 500mb in TEMP reports which results from the change in precision from 1 to 10 m at that level. Also, note that the sign of temperature is inferred from the tenths digit, with even tenths used for positive and odd tenths for negative temperatures. The "rounding down" prescribed by the WMO leads to an average offset of -0.05 °C for temperatures measured to one decimal place. But, the average offset is -0.095 °C for radiosonde systems (e.g., the RS92 DigiCORA III) that measure temperature to two decimal places and truncate to one decimal place before "rounding down" to achieve the 0.2 °C precision (Ingleby and Edwards 2015). These offsets appear as a bias in TAC-reported temperatures.

Table 1 Reporting precision used in TEMP and BUFR radiosonde Messages

Variable	TEMP precision	"Native" BUFR precision
Pressure	1 hPa	0.1 hPa (0.01 hPa in new template)
Geopotential height	1 m at and below 500 hPa 10 m at and above 500 hPa	1 m
Temperature	0.2 °C (even tenths => positive values; odd tenths => negative values)	0.01 K
Dewpoint	Dewpoint depression (DD) 0.1 °C for DD <= 5.0 °C 1.0 °C for 6 °C <= DD <= 49 °C (DD > 49 °C not representable)	Dewpoint temperature 0.01 K
Wind direction	5°	1°
Wind speed	1.0 m/s or 1.0 kts, as specified	0.1 m/s

Finally, station metadata (i.e., latitude, longitude, and elevation) are not included in radiosonde messages from land stations, although they are in messages from mobile stations, ships, and dropsondes. NWP centers must therefore maintain a local station list to provide the station metadata, which is yet another source of error. This is discussed in more detail in Appendix 2.

One other TAC code form requires some attention—the "PILOT" code (designated by WMO as FM-32) is used for wind profiles. PILOT messages are defined in the same four parts as TEMP messages and are used for pilot balloon data, where significant level winds are typically presented at geopotential height levels in 300m increments. The PILOT code form is also used by some countries such as the U.S. and Canada to disseminate radiosonde significant-level winds. It should be noted that even though a 300 m increment is listed in reports for U.S. radiosonde winds, a 1000 ft increment is actually used, requiring a correction (Table 5.2, OFCM 1997). The use of PILOT messages for significant level winds from radiosondes is an unnecessary complication for NWP, since these winds could be included in TEMP messages. However, this practice will (thankfully!) die out as the migration to BUFR proceeds. In NWP systems where the vertical coordinate is a function of pressure, the reported geopotential height must be converted to pressure. While this often uses the standard atmosphere relationship between geopotential height and pressure, the conversion can instead use the model background field to provide a local geopotential height-pressure relationship, as is done at ECMWF, FNMOC, and the UK Met Office (Ingleby and Edwards 2015, Appendix S1).

BUFR formatting considerations: The WMO undertook a migration from TAC code forms such as TEMP and PILOT to the binary code form BUFR,[3] with the original goal of completing the migration in November 2010 and stopping parallel distribution of data in TAC formats in November 2014 (see various documents describing the migration at https://community.wmo.int/activity-areas/wmo-codes/mtdcf and the summaries given in Ingleby et al. 2016a, b). Although progress has been slow, well-formatted BUFR reports are now being produced at the majority of radiosonde stations (Fig. 5). Parallel distribution of TEMP messages is ongoing at most stations although some countries (notably China and some European countries) have discontinued TAC. The BUFR format is described in Vol. I.2 of the WMO Manual on Codes (WMO 2019b).

BUFR has the disadvantage of not being human readable, although given the complexity of the TEMP format, only the well-initiated can actually "read" those messages. The advantages of well formatted BUFR for radiosonde data are: (1) the full sounding is available in a single message, (2) all variables (geopotential height, temperature, dewpoint temperature, wind speed, and wind direction) are provided

[3] A distinction needs to be drawn between WMO BUFR and NCEP prepBUFR. While both are variants of the BUFR format originally developed in the 1980s, WMO's implementation of BUFR is used for the international exchange of "raw" observations, typically using standardized templates as described in WMO (2019b). NCEP's prepBUFR, on the other hand, is an implementation of BUFR used to hold decoded conventional observations before and after quality control, with the post-QC files used not only for operational data assimilation at NCEP but also as a source of conventional data used widely in the U.S. research community.

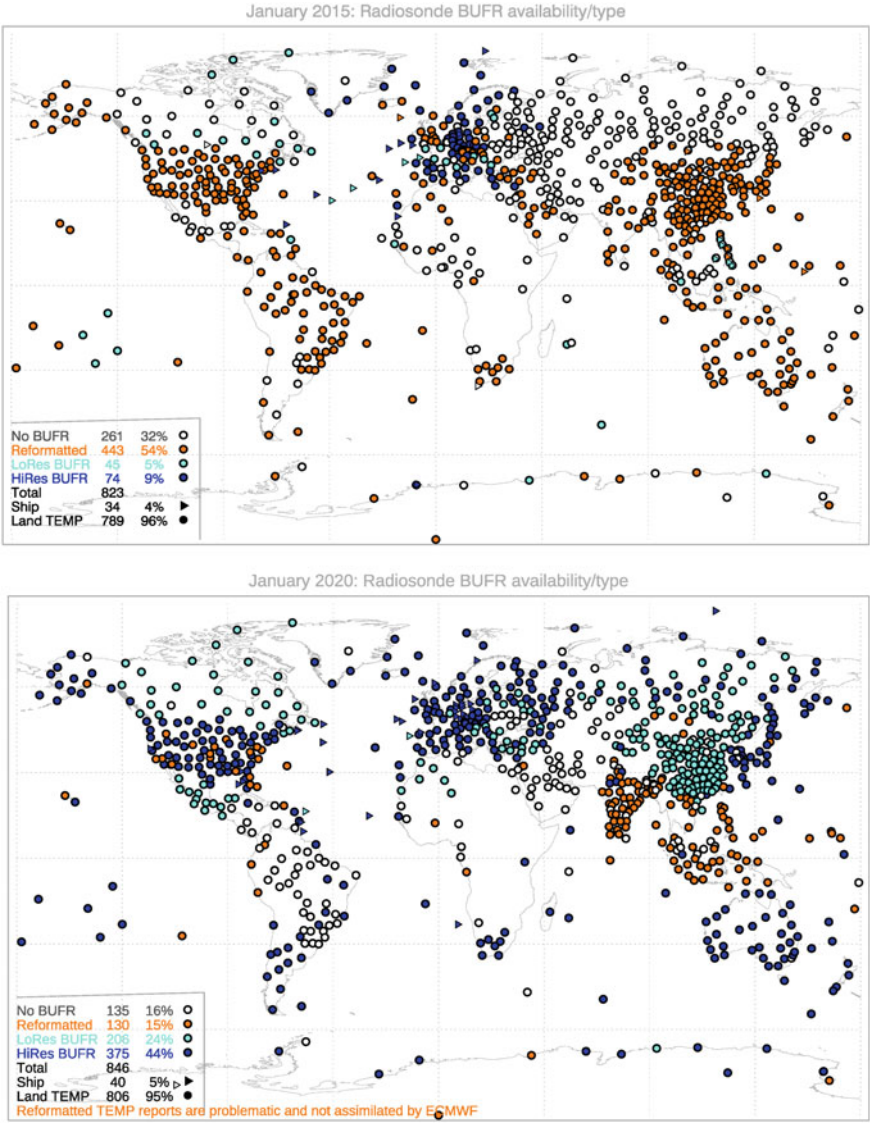

Fig. 5 Progress of TAC to BUFR migration for January 2015 (top) and January 2020 (bottom). Filled circles are shown at station locations, with coloring indicating the report format–TEMP reports only (white), reformatted BUFR (orange), high-resolution native BUFR (blue), and low-resolution native BUFR (cyan)

at all levels, (3) the reporting precision for a given variable is the same throughout the profile and is typically better than the TAC (Table 1), and (4) metadata including latitude, longitude, station elevation, radiosonde release height, radiosonde type, and radiation correction type are given in each message. (See Appendix 2 for a discussion of issues with BUFR station metadata, including the upcoming switchover to WIGOS Station Identifiers.) Note that two messages are distributed for each radiosonde release: an "early" message that contains the data from launch to 100 hPa, and a "late" message that contains the data from launch to balloon burst. The latter message is the only one needed for NWP, so long as the data cutoff time (when the data preprocessors are run) is sufficiently late to allow receipt of these messages.

Well formatted BUFR reports as described above are what Ingleby et al. (2016a) term "native" BUFR reports—those generated directly from the more-or-less raw radiosonde data,[4] with this process performed in the radiosonde ground station. Native BUFR reports can be high resolution, with observation intervals ranging from 1 to 2 s up to 10 s (~400 to ~6000 levels), or low resolution, with the number of levels similar to TEMP. In May 2020, roughly 45% of stations were providing high-resolution native BUFR and another 27% were providing low-resolution native BUFR. Figure 6 shows a time series of the availability of high-resolution and low-resolution native BUFR radiosonde temperatures at ECMWF. The jump in high-resolution data in early 2019 reflects the newly available Russian data, which are currently monitored at ECMWF with Russian TAC data used operationally. The larger jump in low-resolution data in late 2019 reflects the newly available Chinese data, which were put in operational use at ECMWF a few months later prior to the discontinuation of Chinese TAC data in January 2020. Regardless of resolution, native BUFR nearly always includes balloon drift information, with time, latitude, and longitude offsets (relative to the launch time and location) provided for essentially every level. When soundings are encoded in both TAC and BUFR, the native BUFR version should generally be given preference over the TAC version (as is done at FNMOC), perhaps after monitoring the native BUFR from a given station/country to assess its quality (as is done at ECMWF).

However, not all BUFR radiosonde data are given in well-formatted messages."Reformatted" BUFR reports are generated from TEMP reports and so have the same levels and usually the same precision and no balloon drift information. Since TEMP reports for land radiosonde stations lack station metadata, the station latitude, longitude, and elevation in reformatted BUFR must be copied from a station list, which can be a significant source of error (Appendix 2). Some of these reformatted BUFR reports use a message-by-message translation, with each TEMP part having a separate BUFR part. Such "BUFR-by-parts" was common in the early days of the BUFR migration but does not conform to WMO BUFR regulations

[4] "Native" BUFR messages are generated after previously discussed corrections are made. These corrections include such things as time-lag and radiation corrections for temperature and RH, filtering to remove pendulum effects on winds, and the conversion of radar or GPS heights from geometric heights to geopotential heights.

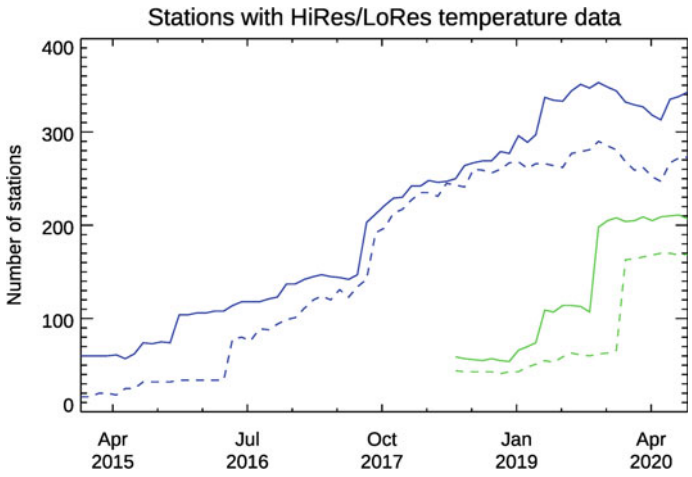

Fig. 6 Time series of stations providing high-resolution (blue) and low-resolution (green) native BUFR temperature observations. Solid lines are the number of stations available; dashed lines are the number of stations assimilated operationally at ECMWF. The lag between when data become available and when they are assimilated reflects the period during which data are monitored for quality. Recent changes include the provision of high-resolution native BUFR from Russia in early 2019 (which is still being monitored) and the provision of low-resolution native BUFR from China in late 2019 (which went into operational use a few months later). The decrease in the number of stations reporting high-resolution BUFR in Spring 2020 resulted from the COVID-19 crisis

(Ingleby et al. 2016a). Since reformatted BUFR has no advantage over TEMP and may contain errors, TEMP should generally be given preference in this case.

The availability of drift offsets in "native" BUFR messages both simplifies and provides more accuracy in accounting for the change in time and location. A radiosonde ascent takes roughly one hour to reach 100 hPa and nearly another hour to reach 10 hPa, during which time the balloon can travel 250 km or more when winds are strong (Seidel et al. 2011). While many NWP centers simply assimilated radiosonde data as instantaneous vertical profiles prior to the availability of native BUFR reports, some NWP centers such as NCEP accounted for balloon drift based on data from TEMP/PILOT messages. In this process, the reported heights and winds are used to estimate both the mean wind vector in a layer and the time for the balloon to travel through the layer, assuming an ascent rate of 5 m/s (Keyser 2000). The product of the mean wind vector and the estimated time gives a displacement vector that can be converted to latitude and longitude offsets. Similarly, Environment and Climate Change Canada (ECCC) estimated drift using the mean wind vector and the mean elapsed ascent time, but as a function of pressure rather than height (Larouche and Sarrazin 2013); they still use this method for TAC soundings (Judy St. James, personal communication). The relatively coarse vertical resolution and rounding used in TEMP and PILOT messages, plus any departure of the balloon ascent rate from 5 m/s all contribute to errors in the drift offsets and therefore the location and time assigned to each level of observations. At present, ECMWF, the UK Met Office, and

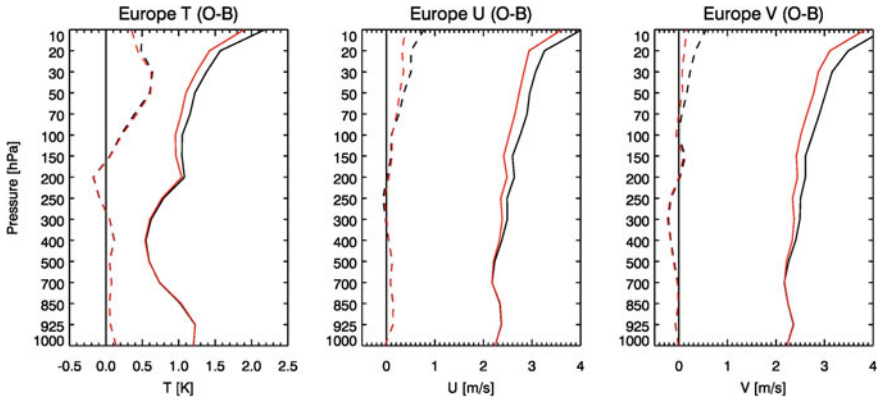

Fig. 7 Standard deviation (solid lines) and mean (dashed lines) radiosonde observation-minus-background (O–B) statistics for European stations with (red) and without (black) drift processing for (a) temperature (b) zonal wind and (c) meridional wind, for November 2016 to February 2017 (From Ingleby et al. 2018)

FNMOC among others use BUFR-reported drift time and location offsets in their operational systems. FNMOC applies the drift offsets prior to thinning the number of levels, while ECMWF splits radiosonde profiles into 15-min intervals in which observations are treated as vertical and instantaneous. The UK Met Office uses drift locations in determining where the radiosonde profile crosses model levels in the interpolation of the model background (Ingleby and Edwards 2015, Appendix S1). Tests at ECMWF showed that using BUFR-reported drift improved O-B standard deviations by 5–10% at stations reporting drift and that biases were also improved especially for stratospheric winds (Fig. 7). ECMWF's treatment of radiosonde drift went into their operational system in June 2018; FNMOC's went into use December 2019.

Preprocessing and quality control for NWP

Collins (2001) describes three categories of data errors: generally small random errors that are essentially independent in space and time; "rough" errors that are relatively infrequent but have a particular cause; and systematic errors that are usually small but persist in time or space. Random errors are handled by the formulation of the data assimilation system itself and reflected in the assigned observation errors. If necessary, systematic errors are handled by bias correction schemes as discussed earlier. Therefore, quality control is designed primarily to handle rough/gross errors, ideally in such a way as to minimize both the number of good observations that are rejected and the number of bad observations that are retained (Eskridge et al. 1995).

While much radiosonde QC hinges on O-B tests as described in the introduction, there are other means of testing for large errors with a low risk of rejecting a good observation. A physical limits test is often performed to screen out any "unphysical" observations (e.g., wind directions greater than 360°, relative humidity greater than

100% or less than zero). While O-B checks would usually reject the same observations, those checks are often applied inside the data assimilation system rather than as part of the data preprocessor. It is typically necessary to take physically unrealistic observations out of consideration before other preprocessor calculations, such as converting wind speed and direction to u and v components or performing humidity conversions.

Reject listing can be an effective if extreme measure; it is useful for stations with long-term issues that have no quick fix, but it has a higher risk of rejecting good observations. Some stations need to be rejected because of siting issues such as radar alignment. Problems with station metadata can also become intractable when BUFR and OSCAR/Surface values conflict or appear to be in error but no alternate source for the metadata is found. The use of a reject list should be accompanied by regular evaluations to look for stations that should be removed from the list since they no longer produce bad data, as well as stations that should be added to the list since they have started producing bad data.

In contrast to these all-or-nothing checks, the Complex Quality Control (CQC) algorithm developed at NCEP by Collins and Gandin (e.g., Collins 2001) makes use of hydrostatic and other checks to detect (and in some cases correct) errors in TEMP messages, especially those that were manually encoded. The results of the checks are combined in a decision-making algorithm to make a final accept/reject/correct determination that is less susceptible to rejecting good observations or accepting bad ones than individual checks. The CQC system and its predecessors have been used operationally at NCEP since 1988 and at other NWP centers, including at FNMOC since the 1990s. An interesting case showing the impact of CQC in the NCEP-NCAR 50-year reanalysis is described in Kistler et al. (2001). They show a sharp increase in the number of hydrostatic error corrections in 1973 that coincided with the introduction of the "Office Note 29" (ON29) internal format used to store observations. The operational forecast skill had deteriorated for a few years after the introduction of ON29, while bugs were worked out of that code (Kalnay 2003, p. 201). However, the use of the CQC system in the reanalysis counteracted many of the early ON29 problems; Kistler et al. (2001) show that skill improved during those same years as a result of corrections by the CQC system.

While these CQC and similar UK Met Office checks (Ingleby and Edwards 2015, Appendix S1) are primarily useful for TEMP soundings, native BUFR soundings still benefit from QC measures. For example, FNMOC received corrupted soundings from a handful of stations along with uncorrupted duplicates during a period in 2016. The corrupted files contained many nonsensical values that were rejected by physical limits checks, as well as disconnected geopotential heights and pressures that a hydrostatic check flagged. A useful process for dealing with issues like this is to reject all values for a particular variable when more than a certain percentage are rejected. Nevertheless, well formatted data from high-quality radiosondes like the RS92 and RS41 seldom have "bad" data, although they can at times have unrepresentative data.

A final aspect of the radiosonde preprocessing for high-resolution BUFR radiosonde data is thinning or averaging the data to reduce the number of levels and the possibility of correlated error between levels. Using thinned/averaged levels

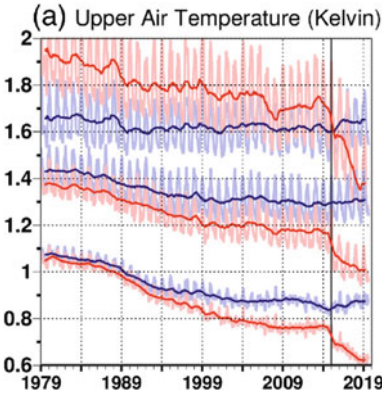

Fig. 8 30-day mean (light colors) and 360-day mean (dark colors) for the standard deviation of O-B for radiosonde temperature data used in ERA-Interim (blue) and ERA5 (red) within ±25 hPa of 50 hPa (top), 850 hPa (middle), and 400hPa (bottom). BUFR radiosonde temperature data were used in ERA5 beginning with 1 January 2015 (vertical black line) (From Fig. 14a, Hersbach et al. 2020, ©2020 Quarterly Journal of the Royal Meteorological Society, licensed under CC)

rather than significant levels should also provide a more representative profile, since as previously described, significant levels are chosen at local extrema in temperature, humidity or winds. Figure 8 (from Hersbach et al. 2020, Fig. 14a) shows a decrease in the standard deviation of O-B temperatures when BUFR was introduced in the ERA5 reanalysis in 2015. Before 2015, ERA5 assimilated standard/mandatory and significant level data, similar to the pre-BUFR practice at ECMWF. The UK Met Office averages temperatures and winds within a model layer (Ingleby and Edwards 2015, Appendix S1), ECMWF thins BUFR profiles to roughly 400 levels keeping a roughly constant number of levels within a model layer (Ingleby et al. 2016a), and FNMOC thins to approximately 130 m spacing.

3 Surface Observations

Surface observations include data from both land-based and marine platforms and have provided the earliest quantitative meteorological observations, dating back as far as the seventeenth century (Eden 2009). In a sense, the modern era began in the mid-1800s, with a conference of meteorologists in Cambridge in 1845 that discussed meteorological data exchange, the founding of the International Meteorological Organization in 1873, the establishment of networks of surface stations, and the distribution of their data in common formats via teletype starting in the 1870s (Thorne et al. 2017; Eden 2009, Fiebrich 2009). Ship observations also gained a more quantitative footing when the Brussels Maritime Conference of 1853 standardized logbook entries and observation practices (Brohan et al. 2009). Today's

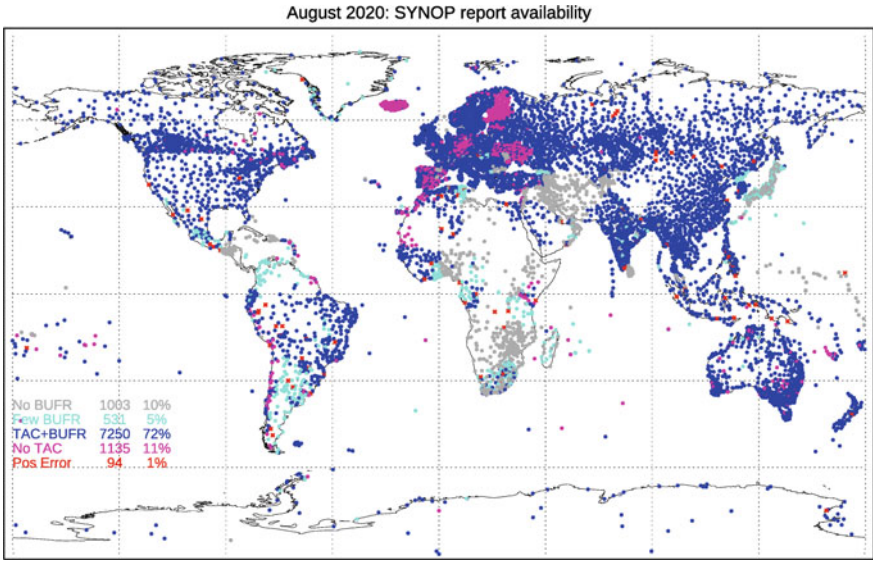

Fig. 9 Availability of SYNOP reports at ECMWF for August 2020 (10013 stations in total) showing the status of migration to BUFR. Counts and percentages for each category are shown in the legend. Position differences between TAC and BUFR are shown in red. Note that current plots of data coverage are available from https://www.ecmwf.int/en/forecasts/quality-our-forecasts/monitoring-observing-system

near real-time surface observations used in global NWP include land-based observations for meteorology (SYNOP) and aviation (METAR), ship-based observations, and various networks of buoys, including near-shore and deep sea moored buoys as well as drifting buoys.

For global NWP the most important variable from surface reports is the surface pressure (Psfc), also sometimes called station pressure, which gives the mass of the atmosphere in a column above the station. Psfc is an assimilated variable in most data assimilation systems, typically after an adjustment to account for the difference between the station elevation and the model representation of terrain at that point (e.g., Ingleby 2015). Reasonably successful reanalyses are possible just using surface pressure reports with monthly averages of sea surface temperature and sea ice extent (Compo et al. 2011; Poli et al. 2016). In contrast, mean sea level pressure (Pmsl) is the pressure hydrostatically extrapolated to sea level from the surface pressure and station elevation (e.g., Ingleby 1995; Pauley 1998) and is widely used to show the evolution of synoptic weather patterns in the extratropics. As such, charts of Pmsl (with other surface data) were the keystone of weather forecasting in the pre-NWP era (e.g., Benjamin et al. 2018; Mitchell and Wexler 1941). Surface reports can contain many other variables: wind, temperature, humidity, and precipitation, as well as cloud and visibility data (although these latter two are not strictly in-situ).

Figures 9, 10 and 11 show typical data coverage; the inhomogeneity of the data density is immediately apparent. Some continental areas are very well observed but

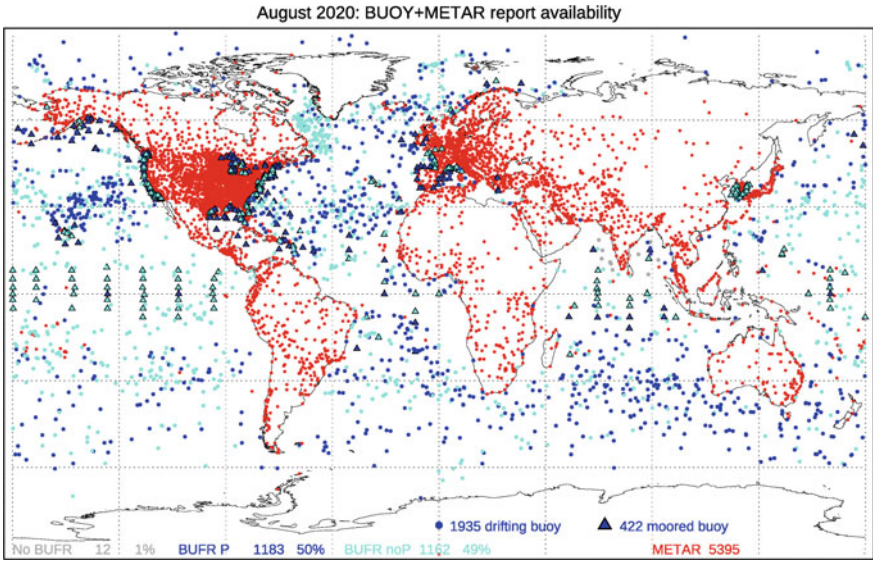

Fig. 10 Availability of BUOY and METAR reports for August 2020. Buoys are plotted at their last reported position in the month and blue/cyan symbols indicate the availability of pressure or not

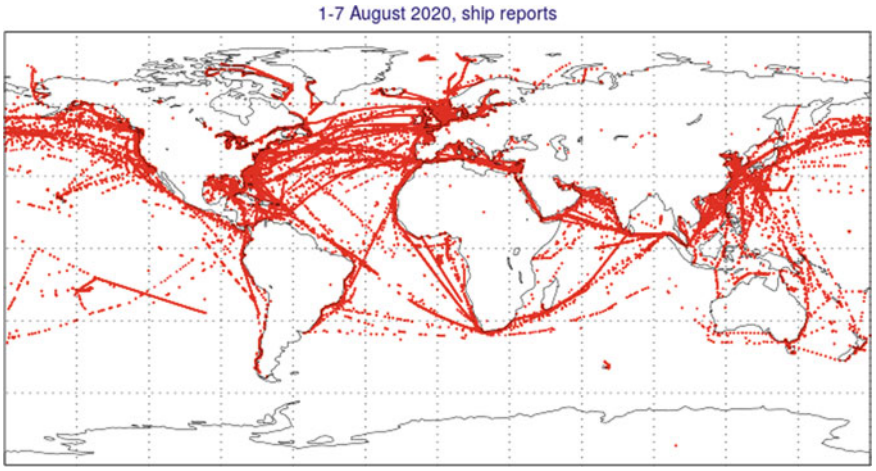

Fig. 11 Location of SHIP reports (BUFR and TAC) for a 7-day period. There were 1916 identifiers reporting in this period, 2537 for the whole month. There are some position errors—evident in reports that are well inland

there are gaps in some desert and mountain regions and also some areas with poor infrastructure (notably parts of Africa). For the USA the coverage of METAR reports is better than that of SYNOP reports (these types are described below). Figure 9 also shows that the SYNOP migration to BUFR is incomplete, some stations do not report BUFR (grey dots) whereas some report BUFR at lower frequency (cyan dots, often 6-hourly BUFR versus 3-hourly TAC). Also clear is that data coverage over the oceans is generally worse than over land. The BUOY migration to BUFR is almost complete (Fig. 10), although Canadian and UK moored buoys still report in TAC SHIP code. Because ships move (and considerably faster than drifting buoys) SHIP positions are shown for a week (Fig. 11)—distinct 'shipping lanes' are clearly visible and about half of the reports are relatively close to land. There is a seasonal cycle in ship reports because of the cycle in sea ice and the generally rougher weather in winter.

Data sources

Land stations: Surface land stations generally consist of a group of sensors that vary according to the requirements of their parent network, with siting requirements also varying among networks. In general, sensors include a barometer to measure surface pressure (often located inside a building and sometimes above the ground floor), temperature and humidity sensors (usually at a height of 2 m above ground level) and wind speed and direction sensors (usually at a height of 10 m above ground level). Airport stations typically report altimeter setting rather than station pressure, and add visibility and ceiling (the height of the cloud base); their sensors are sited to give the best measurement of runway winds–typically near large areas of concrete. Other networks are deployed for specialized needs that might be less useful or less available for global NWP. For example, agricultural stations may also measure evaporation and solar radiation and are sited in grassy areas. Fire weather stations such as RAWS (Remote Automated Weather Stations) can measure fuel temperature and fuel moisture in addition to standard meteorological parameters and are typically located in complex terrain. And, automated mesoscale networks ("mesonets") are being increasingly deployed, often by local governments (e.g., the 120 station Oklahoma mesonet, Brock et al. 1995). The discussion here will focus on synoptic and airport stations as shown in Figs. 9 and 10, which are the best suited to global NWP both in terms of global coverage and data availability.

There are about 10000 active synoptic stations reporting using WMO SYNOP code and about 5000 airport stations reporting in ICAO METAR format. METAR reports are sent hourly or more frequently; SYNOP reports can be 6-hourly, 3-hourly, or hourly (with some manned stations having a gap overnight). In many cases, hourly SYNOP reports are available within a particular country; the WMO is now encouraging the global exchange of hourly reports, which would improve NWP forecasts. A comparison of units and precisions for SYNOP and METAR variables of interest to global NWP is given in Table 2. Rounding is not necessarily performed uniformly; the altimeter setting in U.S. METARs is currently rounded to the nearest 0.01 in

Table 2 Units and precision used in SYNOP and METAR reports

Variable	TAC SYNOP	BUFR SYNOP	TAC METAR
Identifier	5-digit WMO block-station number	Block-station number or WIGOS identifier	4-letter ICAO identifier
Pressure	Psfc or Pmsl in 0.1 hPa (leading "1" omitted for pressure >= 1000 hPa)	Psfc or Pmsl in 0.1 hPa	Altimeter setting (QNH) (a) 0.01 inch Hg, or (b) 1.0 hPa rounded down
Temperature	0.1 °C	0.01 K	1.0 °C
Dewpoint temperature	0.1 °C	0.01 K	1.0 °C
Wind direction	10°	1°	10°
Wind speed	1.0 knots or m/s	0.1 m/s	1.0 knots or m/s

Hg but will be truncated (i.e., rounded down) in the next surface observing system deployment (Chet Schmitt, personal communication).

Almost half of METAR stations also produce SYNOP reports (Ingleby 2015), but not necessarily independently. For example, the U.S. (and likely other countries) generate SYNOP reports by reformatting METAR reports. This conversion requires a station list that matches the WMO five-digit block-station numbers used in SYNOP reports with the ICAO four-letter identifiers used in METAR reports and provides station (field) elevations that are used in converting METAR altimeter settings ("QNH") to SYNOP surface pressures. Altimeter setting is a form of sea-level pressure, with the temperatures in the below-ground hydrostatic extrapolation provided by standard atmosphere values. The formula for computing surface pressure from altimeter setting therefore only requires the specification of the station (field) elevation (e.g., Pauley 1998). Errors in surface pressure resulting from errors in either station identifier match-ups or station elevation are difficult to discern since the station list used in the METAR-to-SYNOP converter is unknown; even so, these errors have the potential to yield surface pressure biases. (See Appendix 2 for an example of this type of error.)

Both SYNOP and METAR codes indicate whether the reports are automated or not; in practice a range of semi-automation is possible, with some variables automated and others, such as cloud and visibility, input manually. ECMWF and some other NWP centers specify a lower observation error for pressure from automated stations reflecting an overall feature of the O-B statistics. Over time automation is increasing; one factor in this is the Minamata convention to phase out the use of mercury in civil applications. The WMO International Conference on Automatic Weather Stations (ICAWS-2017) (https://community.wmo.int/activity-areas/imop/icaws-2017) is a good starting point to learn more about automation.

In addition to the standard FM-12 SYNOP format that does not include station metadata (latitude, longitude, and elevation), relatively small numbers of stations report in FM-14 'MOBILE SYNOP' format which does include station metadata (WMO 2019a). These stations typically use alphanumeric identifiers and are often

not truly mobile; this includes, for example, automatic weather stations in Antarctica that may be relocated from time to time. MOBILE SYNOP reports are also generated and put on the GTS by stations temporarily set up for field experiments. Ingleby (2015) provides more details of reporting practices and assimilation. The impact of migration to BUFR as well as details about the different variables are discussed below.

Buoys and ships: For global NWP the most important marine data are the surface pressure reports from drifting buoys, because they sample some very data sparse areas. It is unfortunate that about 50% of drifting buoys do not have a pressure sensor (Ingleby and Isaksen 2018), as shown in Fig. 10. As the name suggests these buoys drift with the near-surface ocean currents; they are not usually deployed in semi-enclosed seas because of the increased risk of beaching. Typically, a drifting buoy uses GNSS for location and will report data for about two years. Early systems suffered from occasional position errors, but since about the year 2000 these have been less of a problem.

Moored buoys come in two main categories: coastal (North America, Europe and some other areas) and open-ocean tropical arrays (TAO/TRITON, PIRATA and RAMA). The tropical buoys also usually report ocean variables at several depths, although unfortunately they typically are not equipped with a barometer. More information is available from the following websites: Global Drifter Program (https://www.aoml.noaa.gov/phod/gdp/); Global Tropical Moored Buoy Array (https://www.pmel.noaa.gov/gtmba/); and the National Data Buoy Center, both for details about global moored buoys (https://www.ndbc.noaa.gov/) and metadata for global moored buoys (https://www.ndbc.noaa.gov/station_metadata.txt).

Ship reports have a long history (e.g., Brohan et al. 2009), although the number of Volunteer Observing Ship (VOS) reports has declined by half from a peak in the late 1980s (Kent et al. 2007). Over time the ships have also become bigger and measurement heights have increased as a consequence (Kent et al. 2007), which in some cases makes the observations less representative of the adjacent sea areas. There is also a trend towards automation; as for land stations, this generally improves quality. For manual reports, position errors are still a problem, making track checking an important component of quality control (e.g., the Ingleby and Huddleston (2007) track check algorithm originally developed for sub-surface reports). A problem that has been increasing in recent years is the number of reports that either use "SHIP" as their identifier (and so cannot be track-checked) or use anonymized identifiers; in either case, the metadata describing the ship cannot be retrieved from standard ship metadata lists.[5] The use of external metadata lists is needed for reports in FM-13 SHIP format, since that format does not provide a means of reporting barometer or anemometer height (WMO 2019a). In principle, BUFR SHIP format allows these sensor heights to be reported, but they often are not. However, if the ship reports both surface pressure and sea level pressure, those can be used to hydrostatically estimate

[5] Ship metadata used to be provided by WMO Publication 47 (https://www.wmo.int/pages/prog/www/ois/pub47/pub47-home.htm) but are now available from JCOMMOPS (http://sot.jcommops.org/vos/resources.html).

a barometer height. Ingleby (2010) looked at the quality of reports as a function of ship type and provides more background. It should be noted that coastal moored buoys and oil rigs also provide reports using either FM-13 or BUFR SHIP formats (Ingleby 2010).

One problem area with ship reports is the reporting of winds. Traditionally the wind was estimated from the sea state using the Beaufort scale (e.g., NWS 2010a), but more often now it is measured by an anemometer (after subtracting the ship velocity). For assimilation, either the speed is adjusted to a standard height (usually 10 m) or the wind is used at the (actual or estimated) anemometer height. For ships, the anemometer can be up to about 60 m high; for oil rigs it can be more than 100 m but some adjust to 10 m. For most moored buoys, the anemometer is between 3 and 5 m, but large buoys have it at 10 m. Drifting buoys are too small to have a conventional anemometer.

Observations and assimilation considerations

Pressure: Most SYNOPs include both Pmsl and pressure at the barometer height (Psfc). In principle, it is better to assimilate the measured quantity, Psfc, but one has to know the barometer height to do this. Unfortunately, some heights are incorrect, giving rise to pressure biases (Ingleby 1995). The measurement uncertainty should be around 0.3 hPa, and so as not to increase this, the height should be accurate to 1 or 2 m. With modern GNSS systems, accurate surveying of land stations should be easier. However two problems exist: a) a simple GNSS fix is probably not accurate enough, although time averaging or other methods can be used to improve the height measurement; and b) GNSS systems use a reference ellipsoid, and so the 'undulation of the geoid' must be taken into account to get heights relative to sea level. (Appendix 2 describes a nearly 40 m error at a radiosonde site where this was not taken into account.) Some NWP centers apply pressure bias corrections to a subset of stations to compensate. Operationally ECMWF uses the system described by Vasiljevic (2006) although changes are planned.

As described above, METAR reports provide altimeter setting rather than surface pressure, and so a conversion is required, which only requires knowledge of the elevation (e.g., Pauley 1998). While the barometer height is the appropriate elevation to use when assimilating SYNOP surface pressure, the airport field elevation is more appropriate for use both in converting METAR altimeter setting to surface pressure and in assimilating that converted surface pressure. This follows from the basic definition of altimeter setting ("QNH") as the value that is entered into a pressure altimeter to cause the altimeter to read the field elevation (height above mean sea level) while the aircraft is on the ground. However, this choice is much less critical than the correct choice of elevation for SYNOP Psfc.

Winds: At SYNOP stations the standard anemometer height is 10 m; for marine stations it varies as described above. Some global NWP systems have assimilated marine winds but not winds over land. Ingleby (2015) modified the UK Met Office global system to use winds over land but found various problems with speed biases, notably over the Indian subcontinent. Over land, mean 10 m wind speeds are about

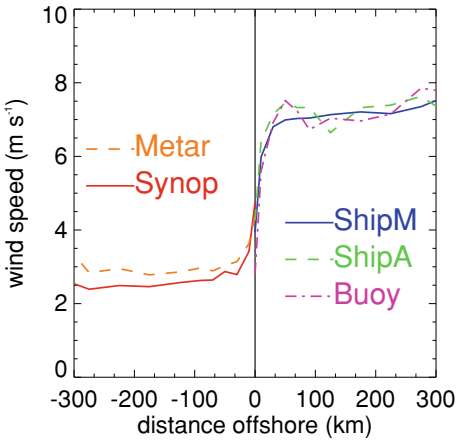

Fig. 12 Mean wind speed as a function of distance offshore for April 2011 to March 2013 for SYNOPs, METARs, manual/automated ship reports, and buoys reporting in SHIP code. Marine speeds were adjusted to 10 m as appropriate. A ~10 km land-sea mask was used with negative values over land. Statistics are given in 20 km bins within 100 km of the coast, then in 50 km bins out to 300 km. Reports further from the coast are plotted at ± 300 km and reports over the 'wrong' surface are plotted at 0 km (From Fig. 7 of Ingleby (2015) © British Crown copyright, Met Office)

2.5 m s^{-1}, and over the ocean they are about 7 m s^{-1} as shown in Fig. 12. One should therefore not assimilate land winds at a model ocean grid-point or vice versa; this is one form of representation uncertainty caused by approximations in forecast models. U.S. C-MAN (Coastal-Marine Automated Network) stations are typically located right at the shoreline or even on small offshore islands; their winds are therefore subject to this sort of representation error. Even with various exclusions, it can be difficult to get much of a positive impact from 10 m winds over land. Physically one would prefer winds at 100 m or more, which are less affected by surface friction, but observations from wind farms are not generally available. Bédard et al. (2015) used a geo-statistical observation operator to improve the impact but this adds complexity.

Screen temperature: Temperature is usually measured in a screen between 1.25 and 2 m above the land surface. The screen temperature is a major climate variable. In most conditions, this works well, but in calm, sunny conditions the air in the screen can be warmer than the air outside by up to one or two degrees. Some stations (e.g., in the U.S.) use a fan to draw the air across the sensor (aspirated sensors) to avoid this issue, but it seems that most SYNOP stations rely on natural ventilation. Ingleby (2015) found a clear benefit from assimilating screen temperature and humidity in the UK Met Office global system. Recent research at ECMWF has given a more mixed signal, with some problems at high latitudes in winter, particularly at night. At nighttime in winter there can be large surface inversions, and in these conditions one would want very little vertical spreading of the increments from screen temperature. In neutral or unstable conditions, temperature will be reasonably well mixed through the boundary layer and the analysis increments should reflect this.

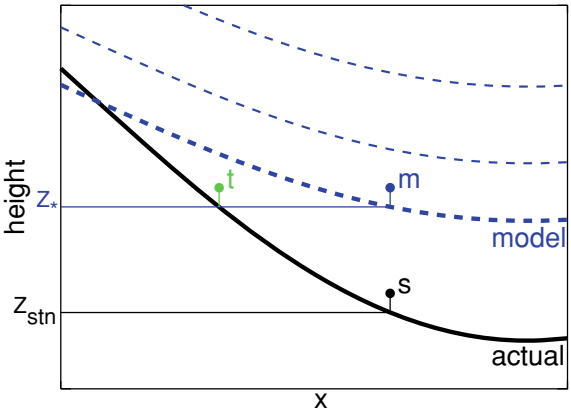

Fig. 13 Schematic of the adjustments needed in going from model values—on blue levels with thick blue line the model surface—to the actual screen value (at 's') just above the real surface—thick black line (Fig. 5 of Ingleby (2015) © British Crown copyright, Met Office)

In general, it is necessary to adjust the temperature for the difference between the station height and the model height at that location. Ingleby (2015) used a climatological lapse rate of 6.5 °C/km but noted that this was slightly higher than the data suggested. Dutra et al. (2020) suggest a lapse rate of 4.5 °C/km. Figure 13 shows a schematic of the situation. Note that there is also a model diagnostic to derive the temperature at a height of 2 m above the model surface from the lowest model level (typically at 10 m or more) and also the skin temperature of the model surface. This diagnostic can be considered as part of the observation operator.

Screen humidity: Humidity is usually measured in the same screen as temperature. Traditionally a wet bulb thermometer was used, but this is not well suited to automated systems. As a result, there is a trend towards the use of capacitive humidity sensors, with some implications for non-homogeneity in the records (Ingleby et al. 2013). Individual capacitive sensors generally drift towards reporting higher values over a year or so and consequently need replacing or recalibrating; drift seems to be worse in damp climates. ECMWF assimilates daytime screen humidity values in its atmospheric 4DVAR. At night, relative humidity values tend to bunch up towards 100% RH (e.g., Fig. 3 of Ingleby 2015) which tends to reduce the difference between observations and forecast.

Surface analysis and soil moisture: Soil moisture affects medium range forecasts, so there has been work at various centers to improve its modelling and initialization. There are some local measurements of soil moisture, but these are not currently included in SYNOP reports. Several centers perform a surface analysis using screen temperature and humidity and use this, sometimes combined with satellite measurements, to update the model soil moisture (see review by de Rosnay et al. 2014). Note that the 2 m temperature and humidity fields available for ERA5 (Hersbach et al. 2020) come from the surface analysis.

Snow depth: Snow makes a large difference to surface energy exchanges; de Rosnay et al. (2014, 2015) outline work on snow depth analysis at ECMWF and elsewhere.

Since about 2010 there have been efforts to improve snow depth reporting—it can be reported in SYNOP messages, and there is also now a separate BUFR template that just contains snow depth. Traditionally, snow depth is measured manually using a snow board or similar device (e.g., NWS 2013), although there has been some work with ultrasonic snow depth sensors (Ryan et al. 2008). Satellites also provide an indication of snow cover (in the absence of clouds) and this is merged with the in situ information in snow data assimilation schemes.

Rainfall: Over the U.S., ECMWF assimilates a field of 6-h rainfall accumulations that contain information from both rain gauges and radar precipitation estimates (Lopez 2011). High-resolution limited area assimilation systems tend to use radar reflectivity more directly (e.g., Benjamin et al. 2016; Zhao et al. 2008).

Clouds and visibility: Cloud and visibility data are used in the UK Met Office high-resolution UKV assimilation system (see Milan et al. 2019 and references). Cloud information is taken from both surface cloud base reports and cloud tops from geostationary satellite data (in an earlier version of the system Dow and Macpherson (2013) found that the impact of the cloud data was mixed). Visibility is linked to both aerosol concentration and humidity with the humidity dependence being particularly marked near 100 % RH. Clouds and visibility are also assimilated in NCEP's High-Resolution Rapid Refresh (HRRR) system (Benjamin et al. 2016).

Ground based GNSS observations: The time delay of GNSS signals at fixed stations depends on the surface pressure and on the integrated water vapor (IWV) above the station. The latter is assimilated in many limited area forecasting systems and some global ones (e.g., Poli et al. 2007)—these data are monitored at ECMWF but not assimilated. There is an assimilation issue in that the IWV must be split into a profile of humidity increments; this can be problematic with sharp vertical gradients of humidity, notably at the boundary layer top, especially if combined with biases from the observations or from the model.

Other considerations

Verification: Surface reports are used extensively in the verification of NWP forecasts (e.g., Haiden et al. 2018). In addition to the variables already mentioned, wind gusts and solar radiation are also used. Wind gusts are linked to damage to buildings/infrastructure and closure of transport routes. Unfortunately, different countries use different reporting practices for wind gusts (and a few don't report gusts) making verification more difficult. Verification of forecast solar radiation has increased in importance as its use in electricity generation has increased. In the past this has used the research BSRN network—with about eight stations in Europe and data only available in delayed mode. Hundreds of SYNOP stations, mainly in Europe, now report hourly solar radiation measurements in their real-time SYNOP messages and some use is being made at ECMWF. We would encourage other countries to report radiation measurements if they are made.

Migration to BUFR and WIGOS identifiers: As previously discussed, a migration from alphanumeric to binary (BUFR) codes is ongoing. In August 2020, about

90% of SYNOP stations were providing BUFR (Fig. 9), although 5% of these provided BUFR at lower temporal frequency. WMO is now recommending that hourly reports should be exchanged globally. Some countries have taken the move to BUFR as an opportunity to send data from more stations, which is welcome. To encourage exchange of data from additional networks (not operated by NMSs), WMO is promoting a change from 5-digit identifiers to much more complicated WIGOS identifiers (e.g., Camborne, UK, changes from 03808 to 0-20000-0-03808 (see Appendix 2); the last part is allowed to be up to 16 characters long). This is a challenge for NWP centers because all their systems have to change, from data ingest, preprocessing, bias correction, assimilation to monitoring and verification. By comparison, the buoy change a few years ago from 5-digit to 7-digit identifiers was much easier to cope with. Virtually all ships and buoys now report in BUFR, most ships and moored buoys also report in TAC, and only a small number of drifting buoys currently report in TAC.

Many of these BUFR SYNOP reports have been reformatted from TAC (FM-12) SYNOP reports; reformatting in this case is more acceptable than for radiosonde data, since the basic character of the SYNOP data is the same in TAC and BUFR (Table 2). However, the problems related to using an internal station list to provide the latitude, longitude, and station elevation for BUFR reports are the same as for radiosonde data (see Appendix 2 for an example of a large error in BUFR elevation).

Migration to IWXXM: While the WMO has been carrying out the migration from TAC to BUFR formats, ICAO has also embarked on a major format change from their TAC formats (METAR, SPECI, etc.) to IWXXM (ICAO Weather Exchange Model), an xml-based format formally described in ICAO Annex III and WMO Pub 306, I.3 (WMO 2020b). IWXXM is much less compact than METAR and is intended to be a machine-to-machine format like BUFR. The current schedule for the migration requires all countries to produce messages in IWXXM by November 2020 and to discontinue METARs in 2026 (Murphy 2019). At least initially, many countries (including the U.S.) will be generating IWXXM messages from METAR messages. (See Appendix 2 for a discussion of station metadata for METAR/IWXXM reports.)

Crowd sourced observations: The use of citizen or crowd-sourced observations has generated interest in recent years. Nipen et al. (2020) and their references provide an insight into the potential and the problems: quality control is important, and traditional monitoring methods may not work if the data has been anonymized. Private weather stations seem the easiest citizen observations to use. The use of smartphone pressures raises more issues: obtaining accurate height metadata and dealing with biases. The height issue is somewhat easier for vehicles. Hintz et al. (2019) review recent work.

4 Aircraft-Based Observations

Aircraft-based observations (ABO) have been in use by the operational meteorology community since the post-World War I era. In the early days, aircraft soundings were made by fastening an "aerometeorograph" (chart recorder) to the wing struts of a biplane; the chart recorded pressure, temperature, and relative humidity and was used to generate "APOBs" (airplane observations) (Moninger et al. 2003). Starting in 1931, these flights were done each day at a network of Army, Navy, and Weather Bureau stations (Marks 1940), but they were replaced by radiosondes by 1939 (Benjamin et al. 2018). Radiosonde data were initially validated by comparison with APOBs (Diamond et al. 1938).

From the early days of commercial aviation, meteorological information has also been passed along by pilot voice reports to air traffic control (ATC) and from there encoded for dissemination in both the aviation and meteorological communities. AIREPs are routine over-ocean reports that include temperature and winds at cruising altitudes and are the voice reports of most interest to NWP.[6] Although their main purpose is to provide position reports for aircraft out of radar range, they have been used by meteorologists since at least 1965 (Chadsey et al. 1967). Over the past 15 years, voice AIREPs have been increasingly replaced by automated ADS-C reports.

The modern era of automated reporting for ABO began with FGGE (First GARP (Global Atmospheric Research Program) Global Experiment) in 1978–1979. Widebody aircraft of that era used Inertial Navigation Systems (INS), which made wind speed and direction measurements practical (Sparkman et al. 1981). Two different efforts were launched, both of which accessed the aircraft's avionics for position and meteorological data—ASDAR, which transmitted observations via geostationary meteorological satellites, and ACARS, which transmitted observations via VHF or HF radio (and eventually commercial satellites) (WMO 2017). The first generation ASDAR system was fielded on 17 aircraft in time to participate in FGGE and provide real-time observations to the GTS (Sparkman et al. 1981). Although a second-generation ASDAR program was operational during 1991–2007, the ACARS approach soon gained ascendency since airlines were installing that system to gather data from aircraft in flight for their own purposes. Starting with the ACARS installations on 11 B-747 aircraft reported by Sparkman et al. (1981), virtually all commercial aircraft flying today are equipped with ACARS.

Automated meteorological reports transmitted via ACARS and disseminated using WMO data formats (either TAC or BUFR) are referred to generically as AMDAR (Aircraft Meteorological Data Relay) reports. Australia was the first country to establish an AMDAR program with observations beginning in 1986. The U.S. AMDAR program (MDCRS) began operation in 1991 with data initially from

[6] Since the 1930s, PIREPs (Pilot Reports) have provided information about hazardous weather (especially icing and turbulence) encountered in flight over the U.S. and Canada (Petersen 2016); these typically do not contain information usable for global NWP but they can be important for verification of flight hazards.

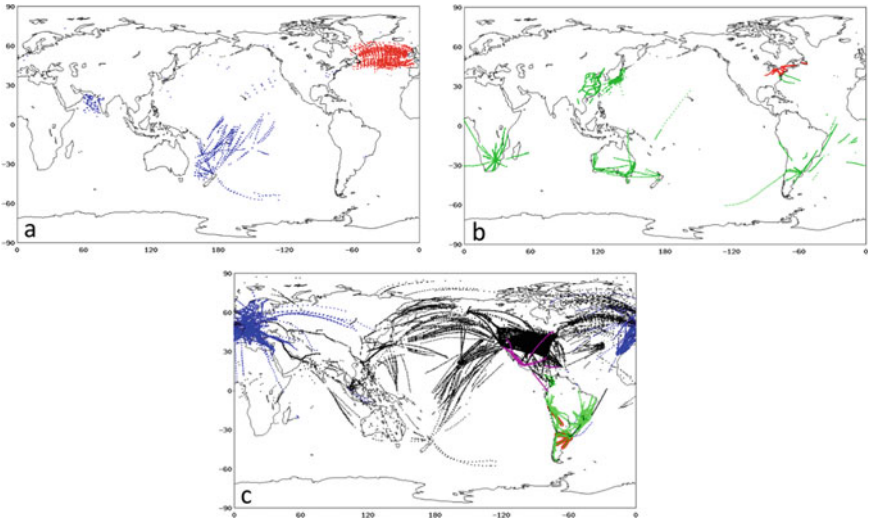

Fig. 14 ABO data coverage for observations received at FNMOC between 0900 and 1500 UTC 7 January 2020. **a** AIREP-format observations with known ADS-C reports (red) and other AIREPs (blue); **b** non-standard BUFR AMDAR (red) and text AMDAR (green); and **c** standard BUFR AMDAR from programs in the U.S. (black), Europe (blue), Mexico (magenta), LATAM (green), and Argentina (red)

three airlines (Homans 2007) expanding to nine by 2016. European programs began with KLM in 1993; currently 14 airlines participate in the EUMETNET E-ABO program (EUMETNET 2020). Other AMDAR programs have been established in Asia (China, Hong Kong, Japan, and South Korea), South America (Argentina and LATAM), North America (Canada and Mexico), New Zealand, and South Africa (Stringer 2020). An example of ABO coverage for data received at FNMOC is shown in Fig. 14. Note that coverage varies not only geographically, with fewer observations over oceans and in the Southern Hemisphere, but also by time of day and day of week (not shown), with fewer observations during the night and on weekends.

Moninger et al. (2003, Fig. 2) shows the increase in aircraft wind observations that occurred leading up to FGGE, with a peak of roughly 6000 per day during FGGE. The number of reports per day increased sharply during the 1990s to nearly 35,000 per day by 1999. Figure 15 shows the ongoing increases since then in terms of observations used in the ECMWF analysis,[7] with more aircraft from more countries/airlines providing data, especially in the past decade. At the end of 2019, roughly 1.9 M data values were used in the ECMWF analysis, which corresponds to over 600,000 reports per day—roughly 100 times as many as in the NCEP reanalysis for the FGGE year.

[7] In this graph, temperature, humidity, u-wind, and v-wind count as individual measurements. Since humidity observations are relatively few in number, the values here can be divided by 3 to give a report count for comparison with the Moninger et al. (2003) values.

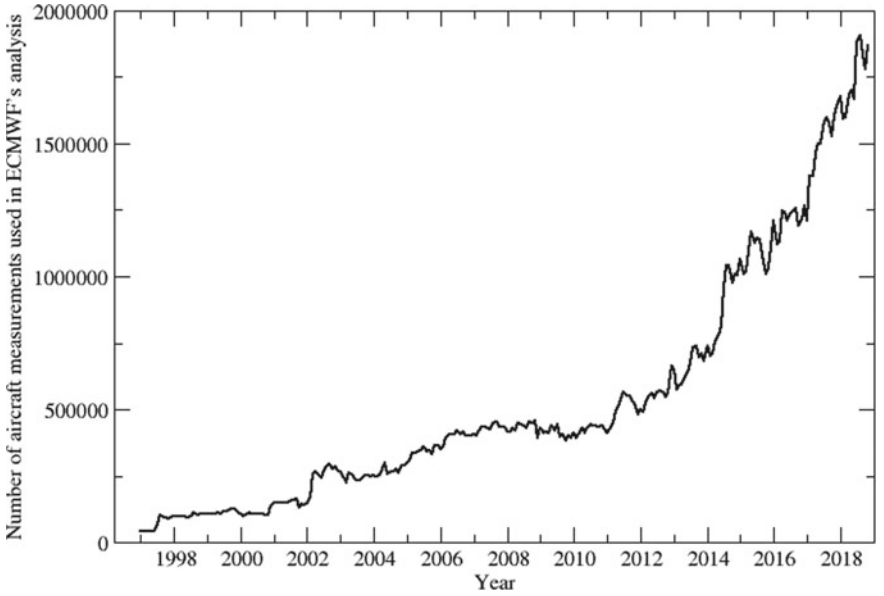

Fig. 15 Average number of aircraft measurements assimilated per day at ECMWF as a function of year. Note that temperature, u-wind, and v-wind count as separate measurements (as does the relatively small number of humidity measurements). ECMWF currently uses approximately 75% of AMDAR and AIREP reports received (Figure courtesy of Lars Isaksen)

Measurement Characteristics

Aircraft observations have a more complex relationship between sensor systems and meteorological variables than radiosonde and surface data. Ignoring humidity for the moment, there are two fundamental sensors in use—the pitot-static probe and the total air temperature (TAT) probe. The pitot-static probe measures total air pressure (static plus dynamic) and static air pressure. Values from the aircraft's navigation system are also required, not only for the location of the observations, but also for the determination of winds. A thorough description of the interplay in these measurements in driving meteorological variables is given in Chapter 3 of WMO (2018b) and summarized here.

Air pressure: Static pressure is converted to pressure altitude in the aircraft's avionics by assuming standard atmosphere conditions. If the aircraft's altimeter was set to the standard atmosphere value of 1013.25 hPa (referred to as "QNE"), then pressure altitude is a function only of pressure. This is the case for most AMDAR reports—WMO (2018b, p. 59) in fact states, "The pressure altitude reported by the AMDAR on-board software should always be with respect to ICAO mean sea-level pressure [QNE] only."

At the same time, most countries require smaller aircraft to use either the altimeter setting "QNH", a form of sea level pressure based on the standard atmosphere, or

"QFE" the observed surface pressure at the airport's field elevation. If QNH is used, the altimeter displays the field elevation while the aircraft is on the ground, while if QFE is used, the altimeter displays an altitude of zero while the aircraft is on the ground. The baseline value QNH or QFE must be known to convert pressure altitude to pressure below the transition level. Above the "transition level" (which varies by country (e.g., 18000 ft in the U.S.)—see https://www.skybrary.aero/index.php/Transition_Altitude/Level), all aircraft set their altimeters to the standard pressure QNE. From an aviation perspective, altitudes above the transition level are referred to as flight levels ("FL") and are generally reported in hundreds of feet—18000 ft becomes FL180. However, the terms "pressure altitude" and "flight level" are often used interchangeably. Regardless of which term is used, the important thing to know when using ABO data is whether data below the transition level use QNE or QNH/QFE as a reference, although nearly all AMDAR observations use QNE. The simplest way to clear up any confusion would be to have data providers convert pressure altitudes to pressure, taking QNH/QFE into account as necessary!

In order to meet aviation requirements, the pressure uncertainty must be 1.5 hPa or less, but the uncertainty in pressure computed from pressure altitude can be larger at low levels if pressure altitudes in the downlinked messages (or the final formatted AMDAR messages) are reported in hundreds of feet. Pressure altitudes are currently reported in the downlinked messages in tens of feet (WMO 2018b).

Temperature: Total air temperature is defined as the temperature that would result from compressional heating if the airflow around the aircraft were completely brought to rest. To mimic this, the most commonly used TAT probe slows down the flow inside the probe to just a few m/s before measuring the temperature (WMO 2018b). The static or ambient temperature is then computed from the probe temperature, the Mach number, and the probe recovery factor, which is typically 0.98 (WMO 2018b). For an aircraft cruising at Mach 0.85, the difference between the probe temperature and the ambient temperature is over 30 K. In principle, the Mach number is defined as the true airspeed divided by the speed of sound in free air and so is itself a function of the static air temperature. In practice, the Mach number is derived from measurements of static pressure and total air pressure from the pitot-static probe (e.g., WMO 2018b, p. 60, Eq. 3.5). The uncertainty of the temperature measurement ranges from 0.3 K at low Mach numbers to 0.4 K at Mach 0.8. Sources of error include evaporative cooling of up to 3 K after the sensor has been wetted in cloud, and a loss of accuracy at very low airspeeds (e.g., prior to take-off). Drüe (2011) also discusses a possible dependency on pitch angle.

Bias is a significant issue with ABO. Dee and Uppala (2009) found a small positive global mean O-A for aircraft data at 200 hPa in the ERA-Interim reanalysis that was accompanied by a small negative O-A for radiosonde data at that level; the latter began in about 1999 when the numbers of aircraft data increased sharply. Ballish and Kumar (2008) similarly found positive monthly average O-B values for AMDAR data in the 300–200 hPa layer, in contrast to negative values for radiosonde data in the same layer. The ABO temperature biases varied by aircraft type, pressure, and flight phase, with aircraft descents tending to be colder than ascents. These results led

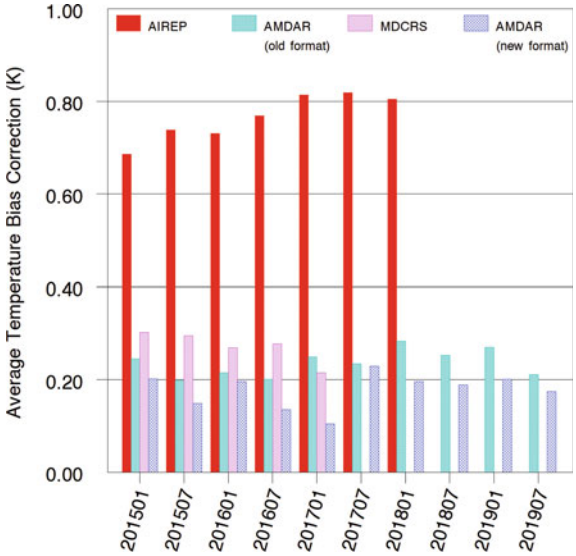

Fig. 16 Average bias correction applied at ECMWF in ERA5. The use of AIREP temperatures was stopped in 2018, and the format for MDCRS data was changed to use the AMDAR BUFR template version 7 ("new format") in 2017. AMDAR averages are larger for pressures less than 300 hPa, but with a lot of variability (SD ~0.4 K) (Courtesy of Lars Isaksen)

Ballish and Kumar (2008) to propose airframe- and pressure-dependent corrections of roughly -0.5 to -1.0 K at cruising altitudes (300–150 hPa) for most airframe types studied. Drüe et al. (2008) also looked at airframe-dependent temperature biases in a special dataset containing high-resolution Lufthansa descents into the Frankfurt's Rhein/Main airport. The systematic deviation of temperature (relative to the mean profile) varied from near zero for A330 and A340 aircraft to roughly +0.3 K for A319 and A321 aircraft and -0.5 K for A320 aircraft. The differences in the latter three are noteworthy since these aircraft types are part of the same family with essentially the same nose section where the sensors are located but a difference in overall length.

Performing a bias correction is therefore an important factor in assimilating aircraft temperature data, especially for the large amount of data at cruising altitudes that can dominate the analysis in regions with heavy aircraft traffic. Isaksen et al. (2012) and Ingleby et al. (2018) describe the application of the ECMWF variational bias correction (VarBC) scheme to correct data from individual aircraft. The bias correction uses a constant that is a function of aircraft registration number (implemented in 2011), with additional terms based on ascent rate and descent rate (implemented in 2018). Isaksen et al. (2012) showed that this scheme led not only to improving the fit of the background and analysis to the aircraft observations, but it also improved the fit to radiosonde temperatures and GPS radio occultation data. A bulk bias correction of AIREP data (which do not generally report registration numbers) was tested at ECMWF and reduced the 200 hPa warm bias in the analysis and forecasts over the North Atlantic. Although it was not used operationally (Ingleby et al. 2018), this overall bias correction scheme is used in the ERA5 reanalysis (Hersbach et al. 2020), with average bias corrections by year and data format shown in Fig. 16. Note that ECMWF decided to stop assimilating AIREP temper-

atures in 2018, given the availability of higher-precision ADS-C data through the MDCRS data stream (Ingleby et al. 2019).

Other centers also apply a temperature bias correction to aircraft data. ECCC uses a static correction as a function of pressure that ranges from 0.1 K for the 700–500 hPa layer to 0.5 K for the 300–100 hPa layer (Buehner et al. 2015). The UK Met Office also uses a static approach, but they base their corrections on monthly O-B statistics in three pressure layers for individual aircraft (Colin Parrett and Adam Maycock, personal communication). In this scheme, corrections are only applied to aircraft with a bias greater than 0.5 °C, which results in corrections to roughly 40% of aircraft temperatures. NCEP followed the ECMWF strategy in using VarBC to determine corrections using registration number and ascent/descent rates, but correcting AIREP-formatted data without reference to registration numbers. The NCEP scheme also uses a tension-spline algorithm to better define the time used in computing the ascent/descent rates (Zhu et al. 2015).

However, the ideal solution for NWP would be to correct the temperature bias at the source in the onboard software, as discussed at the 2020 ECMWF/EUMETNET Workshop on Aircraft Weather Observations and Their Use (https://www.ecmwf.int/en/learning/workshops/workshop-aircraft-weather-observations-and-their-use). This would require addressing any truncation of values in both the onboard software and the downlink format, as well as other sources of bias. For example, ECCC worked with FLYHT (data provider for First Air and Canada North airlines) to correct a configuration-related problem that had led to differences of up to 4 °C between mid-tropospheric ascent and descent temperatures (Zaitseva 2020). A 1-min mismatch between temperature and altitude had resulted from the use of 64-s reporting; switching to 1-s reporting removed this problem (Frédéric Lenormand, personal communication). In addition to timing differences, de Haan et al. (2020) found that biases also result from corrections applied to pressure altitude without the modified pressure altitude being used in turn to modify the temperature. They devised a process using NWP temperatures and indicated airspeed measurements from Mode-S EHS observations (see discussion below) that successfully corrects the bias in AMDAR temperatures. If the temperature bias from these sources could be remediated in the onboard software for at least some aircraft, then temperatures from those aircraft would not only be more valuable to data assimilation, but they could be used as additional anchor observations in the VarBC correction of radiance biases (Eyre 2016). At present, using VarBC to correct aircraft temperature biases is an imperfect solution with a residual bias remaining for some aircraft tracks (Ingleby et al. 2020a).

Wind speed and direction: Fundamentally, the wind vector is computed as the vector difference between the airspeed vector (the motion of the aircraft relative to the air) and the groundspeed vector (the motion of the aircraft relative to the ground). Since these two vectors are much larger than the wind vector, they must be measured to a high degree of accuracy. Lenschow (1986) presents the full equations to compute the three-dimensional wind from aircraft measurements, taking into account the aircraft's

attitude (three-dimensional orientation). In common usage, this is reduced to a simplified two-dimensional solution for the horizontal wind components, assuming that the aircraft is perfectly aligned with the airstream, and in straight and level flight (WMO 2003, 2018b). These assumptions are most often violated when the aircraft executes a turn, so the AMDAR formats include a flag to indicate when the roll angle exceeds a rather conservative 5°. In principle, pitch angle can also come into play, but it is not reported.

The airspeed vector is defined as the motion of the aircraft relative to the air; it is given by the heading and true airspeed of the aircraft. Heading is defined as the angle between the longitudinal axis of the aircraft and true north; it is measured by the aircraft's navigation system, which in most aircraft reporting AMDAR is given by an inertial reference system (IRS), GNSS, or a combination. True airspeed is determined from (dry) static temperature and Mach number, although Lenschow (1986) and Khelif et al. (1999) show that a significant error in true airspeed can occur by neglecting humidity effects in very moist air such as in the tropics. The groundspeed vector is defined as the motion of the aircraft relative to the ground, given by the track angle (with respect to true north) and the speed of the aircraft relative to the earth. These are determined either from a succession of locations of the aircraft from the aircraft's navigation system or by integrating the accelerations measured by the IRS.

The uncertainty of the wind then depends on the uncertainty in the basic measurements of static and dynamic pressure and total air temperature and in the values provided by the navigation system. In the absence of gross temperature errors, the uncertainty in true airspeed is dominated by the uncertainty in Mach number, which itself depends on the uncertainty in the basic pressure measurements (WMO 2018b). An uncertainty in Mach number of 0.2% would yield an uncertainty in airspeed of roughly 0.5 m/s. Uncorrected IRS drift can also lead over time to errors in location and therefore wind that are a function of flight duration. Departures from the assumptions made in deriving the two-dimensional horizontal wind equation include significant side slip, roll angle, or pitch angle, but these are typically counted more as error than uncertainty. Overall, WMO (2018b) cites a typical vector uncertainty of 2–3 m/s.

Drüe et al. (2008) looked at wind errors in an aircraft-relative reference frame, projecting the wind into longitudinal (along-fuselage) and lateral (cross-fuselage) components. Their estimate of the random component of the error was similar between these two components, ranging from 0.6 m/s below 1 km to roughly 1.5 m/s at 4 km. The longitudinal systematic component of the error comes primarily from the measurement of true airspeed and was estimated to be close to zero for A300 and A319 aircraft, but with biases of -0.6 m/s for A320 and +0.3 m/s for A321 aircraft (except at lowest levels). A330 and A340 aircraft had biases of -0.5 m/s. The lateral systematic component of the error depends primarily on roll angle.

Some large non-random errors in aircraft winds have been present at times. In 2011, systematic wind direction errors of 180° were found in ADS-C reports over the North Atlantic from Gulfstream aircraft. A different airframe/airlines-dependent 180° wind error is documented in Pauley (2002); these errors are isolated and occur

only when the actual wind direction is precisely 0° (or 180°) but is reported as 180° (or either 0° or 360°). This rare error is still seen in current data and presumably results from either an arctangent error or an averaging artifact in the onboard software. A more serious wind direction error occurs when the sign of the v-component of the wind is reversed in ADS-C reports from B787 aircraft for southerly flow (Fig. 17). Based on the characterization of the error by Hill (2020), Ingleby and Isaksen (2020) switch the sign of the v-component for AIREP- and BUFR-formatted ADS-C winds when this gives a significantly better fit to the background ("vswitch" check). Detecting aircraft-type dependent errors, reporting them to the source, and devising means of correcting or rejecting erroneous data would be greatly simplified if the aircraft type and airline were provided as metadata, as discussed at the 2020 ECMWF/EUMETNET Workshop on Aircraft Weather Observations and Their Use (see link above).

Humidity: The only humidity sensor currently used on board AMDAR-reporting aircraft is the WVSS-II (Water Vapor Sensing System, second generation), which as

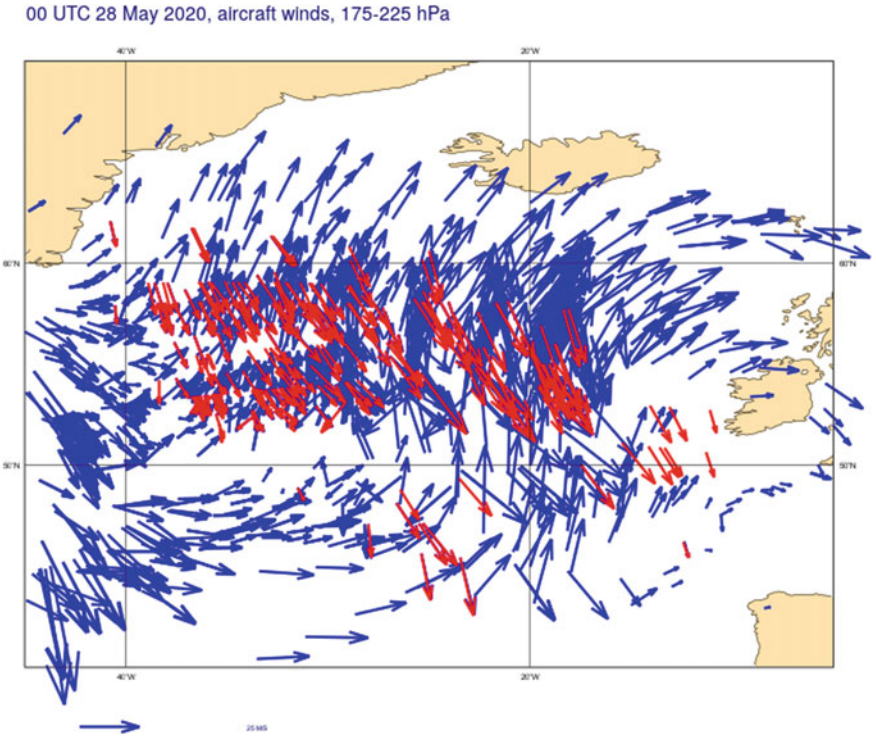

Fig. 17 Aircraft winds (in blue) in the 175–225 hPa over the North Atlantic for layer for 0000 UTC 28 May 2020 with erroneous winds as detected (imperfectly) by the ECMWF "vswitch" check (where the v-component of the wind is flipped under certain conditions) shown in red. The blue arrow at the bottom represents 25 m/s

of 2016 had been fielded on 148 aircraft mostly flying U.S. domestic routes (Petersen et al. 2016). The WVSS-II uses a laser diode system that senses the water vapor absorption in a narrow band of wavelengths centered on 1.37 μm. This measurement is used to derive water vapor concentration (ppmv), which is then combined with WVSS-II-measured pressure and temperature to yield water vapor mixing ratio (WMO 2018b).

A number of comparisons have been made to assess the performance of the WVSS-II. Helms et al. (2010) describe chamber tests by both NWS and Deutscher Wetterdienst (DWD) that demonstrated a relative uncertainty of 5–10% for WVSS-II mixing ratios from 20 to 0.05 g/kg and pressures greater than 200 hPa. These tests suggest a detection limit of 0.02 g/kg. Vance et al. (2015) performed flight tests comparing two WVSS-II sensors with three chilled-mirror sensors and found good agreement, especially with the WVSS-II fed with a Rosemount inlet (such as commonly used for TAT probes) rather than the standard flush-mount inlet ("air sampler") which had a moist bias at the lowest mixing ratios. However, the flush-mount inlet performed better in the presence of liquid water in cloud and had the fastest response time of the five sensors in the test. A comparison was also made between special RS92 radiosonde launches and WVSS-II profiles (Petersen et al. 2016). The overall mean difference between the radiosonde and WVSS-II measurements below 400 hPa was 0.15 g/kg with a standard deviation of 0.62 g/kg; aircraft-to-aircraft differences were even smaller, suggesting that the WVSS-II performs at least as well as RS-92 humidity sensor. Further information on testing is summarized in WMO (2019d).

While these studies demonstrate the overall high quality of the WVSS-II humidity observations, Pauley and Baker (2014) show an example of a pathology that can occur in this system. A particular aircraft gave humidity values that increased with height to 11 g/kg at 200 hPa, which equates to a relative humidity of over 20000%. A supersaturation check would reject all of the observations above roughly 700 hPa in this case, but the values below that level also increased with height, compared to the decrease with height seen in a nearby radiosonde sounding. According to Bryce Ford (SpectraSensors, personal communication), this behavior in the WVSS-II mixing ratios would result from either a blocked air sampler or an incorrectly installed one-way hose between the air sampler and the laser diode chamber. In either case, surface air would be trapped within the chamber leading to the wildly incorrect mixing ratio measurements. Rejecting the entire humidity profile when more than 1/3 of the observations are rejected can handle this pathology.

The WVSS-II humidity is provided with quality flags that indicate potential problems with the data; these should be utilized even though very few values are flagged as bad. Mixing ratios of 0.000 g/kg are present in current data without being flagged as bad—these should not be assimilated. Consideration should be given to excluding values less than the notional detection limit of 0.02 g/kg.

Turbulence: Although aircraft reports of turbulence are not assimilated in operational NWP systems, they are important for verification of turbulence forecasts (e.g., Storer et al. 2019). Two measures are currently in use—DEVG (Derived Equivalent Vertical Gust) and EDR (Eddy Dissipation Rate), both of which are aircraft-independent

turbulence measures. As of February 2020, over 200 aircraft were reporting EDR with an average of roughly 17000 reports per day, and approximately 350 aircraft were reporting DEVG with an average of roughly 20000 reports per day (Taylor 2020).

Related Observing Systems

ADS-C: Over the past approximately 15 years, voice AIREPs have largely been replaced by automated ADS-C (Automatic Dependent Surveillance-Contract) reports for over-ocean position reports required for ATC. These are provided by contract between a given aircraft and an Air Navigation Service Provider (ANSP) that can optionally require temperatures and winds (Williard 2020). Starting in the mid-2000s, ADS-C reports encoded in AIREP format have been disseminated on the GTS, especially over the North Atlantic (Fig. 14a). Although ADS-C reports from a subset of airlines were included in MDCRS starting in 2014, a much larger set of ADS-C reports started being disseminated on the GTS as part of MDCRS in 2017 (e.g., the tracks that stop abruptly south of the main North Atlantic routes in Fig. 14c). As of February 2020, 38 out of the 65 ANSPs required meteorological data in ADS-C position reports (Williard 2020). In a study of KLM ADS-C reports over a 76-day period, de Haan et al. (2013) found that ADS-C reports were comparable in quality to AMDAR reports. However, they are not without problems as seen in the previous discussion of wind errors.

ADS-B: Another type of report downlinked for ATC purposes is ADS-B: Automatic Dependent Surveillance-Broadcast. These reports are broadcast by commercial aircraft and can be used by ATC as a replacement for secondary radar and can also be received by other aircraft for situational awareness. These reports are mandated in some regions/countries; they include location but not meteorological data. However, Stone and Kitchen (2015) developed a technique for using ADS-B-reported pressure altitude and pressure-altitude correction (defined as the difference between the GNSS altitude and the pressure altitude) to define a layer-mean temperature.

AFIRS AMDAR: FLYHT Aerospace Solutions, Ltd. has developed AFIRS (Automated Flight Information Reporting System), a satellite communications system that provides an alternate to ACARS and can also provide ABO (WMO 2017, 2018b). The meteorological data from AFIRS have the same characteristics as AMDAR, and are available in some otherwise data sparse regions. These data are commercially available and have been in use at NCEP by contract, but they had not gained widespread use in NWP prior to FLYHT making the data freely available to the NWP community during the COVID-19 crisis.

TAMDAR: TAMDAR (Tropospheric Airborne Meteorological Data Reporting) had its genesis in a NASA project to develop electronic pilot reporting capability in conjunction with AirDat LLC, who developed a combined sensor that can measure temperature, relative humidity, static and dynamic pressure, and icing and from those compute pressure altitude, indicated/true airspeed, turbulence, and winds (Daniels et al. 2004). The TAMDAR electronics also includes GPS capability for time and

geometric altitude and includes its own satellite communications. TAMDAR installations have been primarily on regional carriers, which make shorter flights often into smaller airports; as a result, TAMDAR complements AMDAR. TAMDAR is currently operated by FLYHT Aerospace Solutions, Ltd., and a description of the TAMDAR measurements is given in WMO (2018b).

While the performance and benefits for TAMDAR have been well studied (e.g., Moninger et al. 2010; Ingleby et al. 2019), these data have not gained widespread use in global NWP (with the exception of NCEP) because of their cost. However, the current owner, FLYHT Aerospace Solutions, Ltd., has made TAMDAR data freely available to the NWP community during the COVID-19 crisis, which has led many centers to implement it to help mitigate the substantial decrease in aircraft data related to COVID-19.

HDOB: In addition to dropsonde data, the NOAA and U.S. Air Force hurricane hunter aircraft also have been providing "high-density high-accuracy" flight-level winds, temperatures, and humidity since the 2007 hurricane season. These so-called HDOB observations use a 30-s time interval and 30 s averages, with the exception of peak values that use 10 s averages. Each report includes time, latitude, longitude, static pressure, geopotential height, extrapolated sea level pressure (or D-value), air temperature, dewpoint temperature, flight-level wind direction and speed, plus peak flight-level wind speed, peak SFMR surface wind speed, SFMR-derived rain rate, and QC flags. The ASCII code form for these observations is described in Appendix G of the National Hurricane Operations Plan (OFCM 2020a). HDOB data have been assimilated in the HWRF system since 2017 (Tong et al. 2018).

Mode-S: An entirely different means of obtaining meteorological information from aircraft is commonly called "Mode-S". These data are derived from information automatically downlinked from an aircraft's transponder in response to being queried by ATC radars operating using Mode-S Enhanced Surveillance (EHS) and are very dense in space and time for regions covered by these radars, not only because the queries are made frequently (every 4–6 s), but also because all aircraft respond to the queries. One advantage these data have over AMDAR data is that they have no communication costs, since Mode-S data are already being downlinked for ATC purposes (EUMETNET 2015). The communication costs for AMDAR and ADS-C data are typically shared in some way with NMSs such as NOAA or organizations such as EUMETNET.

There are several different variants of Mode-S data. Arguably, the simplest to use is Mode-S MRAR (Meteorological Routine Air Report) data, which contain temperature and wind observations, but which are provided by only a small fraction of aircraft. Collocations showed that these data are very similar in quality to AMDAR data (Strajnar 2012), but like AMDAR data need a temperature bias correction (de Haan 2014).

Temperatures and winds can also be derived from the downlinked Mode-S data related to the aircraft sensor data. The process used by EMADDC (European Meteorological Aircraft Derived Data Center) at KNMI is described below (EUMETNET 2015; Sondij 2020). EMADDC processes data from several sources, including ATC

centers, NMS-operated ADS-B receivers (e.g., UK Met Office), and commercially operated ADS-B receivers (e.g., Air Support) (Ingleby and Isaksen 2020). Data coming from ATC radars use the radar-observed slant range and azimuth to the aircraft to determine the latitude and longitude of the observation, while data coming from ADS-B receivers (that receive both ADS-B and Mode-S reports) use the ADS-B locations. The aircraft providing Mode-S data responds to a radar query by downlinking altitude, roll angle, true track angle, groundspeed, either track angle rate or true airspeed, magnetic heading, indicated airspeed and/or Mach number, and climb/descent rate (EUMETNET 2015).

These parameters require careful processing to accurately determine meteorological variables. The groundspeed vector specified by the true track angle and ground speed is quite accurate. The airspeed vector requires the conversion of magnetic heading to true heading (i.e., with respect to true north). The magnetic heading is not measured but rather is determined on board the aircraft from the true heading output by the navigation system and a magnetic declination table, which may be out-of-date and is unknown. EMADDC determines a heading correction for each aircraft based on an estimated date of the magnetic declination table; different estimation methods are used, some of which use NWP winds (Sondij 2020). Temperature was originally derived from the Mach number and true airspeed, but a more accurate method was defined based on the indicated airspeed and static pressure that is in use in the EMADDC test system. After averaging over 20 s, these temperatures have a standard deviation close to that for AMDAR and near-zero bias.

Data from the EMADDC test system was offered to the operational NWP community in April 2020 to help offset the loss in aircraft data associated with COVID-19. ECMWF began testing these data in May 2020 and found that temperatures computed from indicated airspeed and static pressure were essentially unbiased but had standard deviations that were a bit larger than bias-corrected BUFR AMDAR temperatures (Ingleby and Isaksen 2020). One aspect of using Mode-S data in a global model that needs to be addressed is data thinning, since these data are very dense in space and time over Europe. After applying their standard thinning, ECMWF assimilated only 5% of the available Mode-S data (Fig. 18). These data gave improvements mostly in the upper troposphere–a 6% improvement in 12-h wind forecasts and a 3–4% improvement in temperature compared to European radiosondes (Ingleby et al. 2020c). ECMWF began assimilating Mode-S winds in their operational system on 28 July 2020, and now uses more Mode-S than AMDAR reports—remarkable given that Mode-S data are currently only available over Europe.

UAV: Using Unmanned Aerial Vehicles (UAVs) to obtain meteorological observations has been gaining attention in recent years (e.g., WMO 2019e), although a number of problems must be resolved before they can be used operationally, including airspace restrictions, autonomous operations, and data formatting and sharing. Pilot projects have demonstrated the ability of UAVs to provide boundary layer profiling and air quality monitoring among other uses.

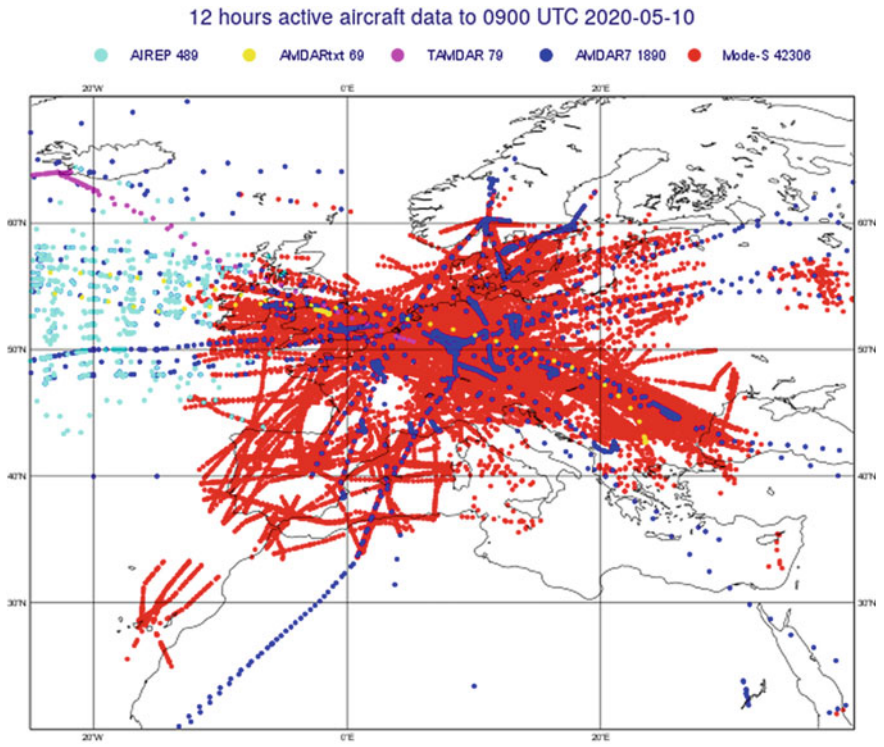

Fig. 18 Distribution of Mode-S aircraft reports (red) after ECMWF thinning for 2100 UTC 9 May to 0900 UTC 10 May 2020. Other aircraft types assimilated at ECMWF are shown for comparison, including AIREP (cyan), TAC AMDAR (yellow), TAMDAR (magenta), and BUFR AMDAR (dark blue). The numbers of reports for each category are shown in the legend

Data Formatting

Downlink formatting considerations: The specification for the current downlink format is the AMDAR Onboard Software Functional Requirements Specification (AOSFRS) (WMO 2014), which is based on the older ARINC 620 format described in Appendix II of the *AMDAR Reference Manual* (WMO 2003). The AOSFRS specification provides for latitude and longitude to be reported in degrees-minutes-seconds, and time in seconds into the month (although not all airlines have the software capability to report seconds (Stewart Taylor, personal communication). The precision of the meteorological variables is listed in Table 3. Most airlines have versions of ARINC 620 that allow for "up-linking" to the aircraft, i.e., to select data in defined geographic boxes, specified flight phases, and reporting resolution at airports. This capability is used by the E-AMDAR program in its E-AMDAR Data Optimization System (E-ADOS) to reduce communication costs by, for example, reducing the number of profiles downlinked at busy airports (WMO 2012b).

Table 3 Reporting precision used in ABO messages

Variable	AIREP precision	AOSFRS precision (downlink)	FM-42 AMDAR precision	BUFR precision AMDAR template v7
Pressure altitude (or flight level)	100 ft	10 ft	100 ft	1 m
Temperature	1.0 °C	0.1 °C	0.1 °C	0.1 °C (after conversion from K in 0.01 K)
Humidity	–	Mixing ratio 0.001 g/kg	–	Mixing ratio 0.001 g/kg
Wind direction	1°	1°	1°	1°
Wind speed	1.0 kt	1.0 kt	1.0 kt	0.1 m/s

AOSFRS uses a single format for all flight phases, but varies the temporal resolution according to the uplinked configuration. Ascent and descent observations can use either a pressure-based reporting scheme (e.g., every 25 hPa) or a time-based reporting scheme (e.g., every 20 s), with the ascents split into a higher resolution "part 1" and a lower resolution "part 2". The default configuration settings in E-ADOS are (1) ascent part 1: 10 s interval for 150 s; (2) ascent part 2: 50 s interval for 1050 s; (3) en-route: 7 min interval; (4) descent: 40 s interval starting at FL180 (Stewart Taylor, personal communication). However, some aircraft use the older ARINC 620 specification and provide a time and location only for the first observation in ascent part 1, and time only for the first observation in ascent part 2. In this case, times are computed using the specified time intervals; locations in part 1 are often interpolated, but sometimes are set to the initial location. These are what Drüe (2011) refers to as "faked positions".

TAC FM-42 formatting considerations: The TAC format in use for AMDAR data is specified as FM-42 (WMO 2019a), with the reporting precision for meteorological variables shown in Table 3. One of the deficiencies of this format is that it does not specify the month or the year, instead only giving the day of the month and the time. There have been cases where month-old data have been retransmitted on the GTS (Brad Ballish, personal communication—see the example in Pauley et al. 2014, slide 19). When this occurs, nothing in the report alerts the user to its age; both current and old observations can be present from the same aircraft, potentially separated by unrealistic distances.

In terms of metadata, aircraft identifiers are registration or "tail" numbers that have typically been anonymized using 5–8 character combinations of letters and numbers. Some AMDAR programs use the first two characters to indicate the program (e.g., AU0171 for an Australian aircraft). Latitudes and longitudes are given in degrees and minutes. The flight phase is indicated as ascent, descent, level, or unsteady, with the latter indicating that the roll angle threshold (usually 5°) has been exceeded and

that the winds are therefore suspect. Unfortunately, there is no means of reporting airframe type in FM42.

FM-42 provides two variables for pressure altitude, which can be confusing since they are both named "pressure altitude". The pressure altitude reported in Sect. 2 of the observation is defined with respect to the standard pressure QNE (WMO 2019a, p. A-194), while the pressure altitude reported in Sect. 3 is defined with respect to QNH (WMO 2019a, p. A-77). The regulation for the latter states that reports at 700 hPa and below use pressure altitude derived from QNH and the field elevation, while heights above 700 hPa use pressure altitude based on QNE. As an example, Japan makes use of this to report pressure altitudes below 1000' (305 m) that reference a local QNH in Sect. 3 (with the pressure altitude in Sect. 2 set to missing), and pressure altitudes at 1000' and above that reference QNE in both Sects. 2 and 3.

BUFR formatting considerations: Roughly 98% of the ABO currently available for NWP (disregarding Mode-S) are made available in BUFR using the AMDAR template version 7 (WMO 2019b), referred to here as "standard" BUFR. A small number of observations are also provided in "non-standard" BUFR, for example, data from the Canadian AMDAR program, AFIRS AMDAR, and TAMDAR. This discussion will focus on the characteristics of standard AMDAR BUFR, with the precision for meteorological variables given in Table 3.

Both flight numbers and registration numbers are available in standard BUFR, but the flight number is usually only populated for MDCRS data. European AMDAR registration numbers are anonymized as "EU" followed by a four-digit number, with some reports repeating the registration number for the flight number, while MDCRS uses 8-character combinations of letters and numbers in their anonymization for both registration number and flight number, with no obvious pattern. Latitude and longitude are given in degrees to five decimal places in BUFR, although the functional precision is often three decimal places. Dates are given as year, month, and day and times are given as hours, minutes, and seconds, but seconds are set to zero for most observations. Only a single pressure altitude variable is used in standard AMDAR BUFR; it is referred to as "flight level" in most of the documentation and is assumed to be pressure altitude based on the standard pressure QNE. This variable should be consistent with the Sect. 2 pressure altitude in FM42.

Over the past few years, increasing numbers of ADS-C reports have been made available through the MDCRS data stream. These can only be distinguished from AMDAR reports by the GTS bulletin header used for their dissemination, although identifying the ADS-C reports inside the BUFR messages was recommended at the 2020 ECMWF/EUMETNET Workshop on Aircraft Weather Observations and Their Use.

TAC AIREP formatting considerations: The AIREP format is governed by ICAO rather than WMO. Most AIREP documentation describes pilot procedures rather than the code format, but a brief description of the format is given in the U.S. FAA publication on Aviation Weather Services (FAA 2019) and a somewhat longer one in the publication on Flight Services (FAA 2020). In principle, ICAO TAC codes are being transitioned to IWXXM, but at present, there is no clear picture on whether

AIREP will be transitioned to IWXXM or whether the ADS-C reports currently being converted to AIREPs will be provided in BUFR via MDCRS instead.

Since AIREPs are provided primarily for position reports for ATC, the aircraft identifier ("call sign") is a flight number rather than a registration number (although some ADS-C reports also include a registration number in the comments at the end of the message). This means that reject listing and bias correction are difficult at best for AIREP data as previously described. Voice reports, which are no longer prevalent, often gave locations with respect to waypoints which required a dictionary to convert them to latitude and longitude; incorrect locations at times resulted from ambiguous waypoint names or out-of-date waypoint dictionaries (Brad Ballish, personal communication). If locations are given as latitudes and longitudes, they are given in degrees and minutes. Prior to roughly 2005, voice reports were made every 10° longitude across the North Pacific and North Atlantic, often with locations and meteorological variables heavily rounded. At present, nearly all observations provided in AIREP format are ADS-C reports, typically cruise-level reports provided at roughly 15 min intervals with precision as shown in Table 3.

Preprocessing and quality control for NWP

There are a number of unique challenges involved in preprocessing ABO data. Duplicate removal is complicated by the anonymization of identifiers. The same aircraft can report in AIREP using a plain-text flight number, in AMDAR using the country/region-specific anonymization, and in Mode-S, using yet another anonymization (Ingleby and Isaksen 2020). Such cross-type duplicates are difficult to detect, especially since units and reporting precisions differ between types. Using plain-text aircraft registration numbers would circumvent this problem, if the airlines and employee representatives would agree. If not, the 2020 ECMWF/EUMETNET Workshop on Aircraft Weather Observations recommended that common anonymization tables be used, at least between AMDAR and Mode-S reports.

Track-checking is recommended to ensure that valid locations, times, and pressure altitudes are assigned to each observation. However, this can be difficult when seconds are not reported for ascent/descent observations and when ascent part 1 observations are given the same location. Additional QC tests can be applied by comparing neighboring observations in a particular flight. For example, sometimes values for time, latitude, longitude, or meteorological variables can become "stuck" partway through a flight (or for a whole flight) and report constant values.

Thinning is also important for ABO data. FNMOC applies along-track thinning especially for en-route reports with small time differences. ECMWF rejects data from the lowest 30 hPa in ascents and descents due to large biases, especially for temperatures. They also apply thinning for each aircraft to 35 km horizontally and 7.5 hPa vertically (Ingleby et al. 2018), which removes roughly 25% of AMDAR/AIREP reports and 95% of Mode-S reports, primarily near airports. The UK Met Office thins ABO to an even greater extent, using only about one-quarter of the AMDAR/AIREP data used by ECMWF. NCEP is testing a process that uses a Hilbert curve projection

to determine areas of high ABO density; the observations in that region are then downweighted rather than thinned (Hill 2020).

5 Summary and Discussion

As described in detail in the previous sections, the primary in-situ observations currently used for NWP originate from radiosondes, surface platforms (including land stations, buoys, and ships), and aircraft. Each of these provides a heterogeneous dataset, with random and systematic errors that vary in the details of instrumentation, degree of automation, and data format. Metadata must also be correct and correctly accounted for—location metadata are more problematic than one might expect (Appendix 2). The data formats make provision for radiosonde type (although it may be missing), but not for aircraft type; both affect the error characteristics for that platform. To summarize, in-situ observations are messy and require care to be used optimally in data assimilation.

Radiosondes: There is a migration to binary codes underway with increased reporting of high vertical resolution data; by August 2020, 44% of stations were providing high resolution profiles. The reporting of position information at each level enables a better comparison of radiosonde and forecast background values at upper levels. In addition, some radiosonde systems allow data to be transmitted while the radiosonde is descending. The data from some radiosonde descents, with the exception of the topmost portion, has been shown to be of similar quality to the data from radiosonde ascents.

Surface: Many countries make only six- or three-hourly SYNOP reports available on the GTS; as supported by the WMO, global NWP centers would like to receive hourly reports. More data, especially surface pressure over the oceans is also highly desirable—this could be achieved by increasing the proportion of buoys with pressure sensors (currently only about 50%). Potentially there are huge numbers of other data sources (‘crowd-sourced’ or ‘internet of things’) but these come with additional quality issues, and may be more useful for limited area forecasting. The move to WIGOS Station Identifiers is intended to allow for such expansion, but it imposes a major software change for NWP centers that will also affect radiosonde observations.

Aircraft: The last two decades have seen a massive increase in the volume of observations from commercial aircraft, at least until the COVID-19 pandemic led to a precipitous drop in aircraft flights in March 2020. The data are very useful, but a wind direction error affecting a small subset of reports from a particular aircraft type has shown that basic metadata not currently provided (such as aircraft type and airline) would be useful for QC purposes. Discussion has also started on the possible correction of temperature biases at the source (rather than in NWP systems). Established aircraft reports come from only a small proportion of flights; early results from using much denser ‘Mode-S’ reports that are generated for Air Traffic Control (currently only available over Europe) are promising.

Not all of these characteristics are fully accounted for in data assimilation systems, so NWP centers have developed processes to counteract many of the problems discussed in this chapter. Observation errors are typically assumed to be Gaussian, unbiased, and uncorrelated in time and space and with the model background. Observations that have large departures from the model background contribute to the tails of the distribution; these observations are deleted (or downweighted) either in the data preprocessors through specific QC checks or in O-B checks that are part of the data assimilation system. Observation biases are addressed either through corrections during preprocessing or through variational bias correction as part of the data assimilation system. Aircraft temperatures are usually subjected to bias correction; radiosonde temperatures and pressures from surface platforms may be as well. (The ECMWF correction of radiosonde temperatures is updated monthly using the best subset of radiosonde data as a reference so that the data can still be considered as anchor measurements. ECMWF is also gradually using more profiles uncorrected.) Surface pressure biases are typically associated with uncorrected errors in station (barometer) height, so correcting station metadata errors can reduce the need for surface pressure bias correction. High-resolution observations that are part of a particular radiosonde sounding or a particular aircraft flight can contain correlated error; data thinning is often employed in mitigation. However, these processes are imperfect at best. For example, using the model background to supply the height-to-pressure conversion for pilot balloon winds is better than using a standard atmosphere conversion, but introduces the possibility of error correlated with the model background. In principle, however, one could directly use height as the vertical coordinate for some or all radiosonde and pilot balloon profiles.

Even so, in-situ observations play an important role in operational NWP. Figure 19 shows the FSOI and counts by data category for January 2020 for both ECMWF and FNMOC. (FSOI is a measure of the error reduction in 24-h forecasts brought about by assimilating observations (Langland and Baker 2004) and is typically computed as part of operational forecast systems that use 4DVAR.) There are differences between centers in how FSOI is computed—for example, ECMWF uses a dry energy error norm, while FNMOC uses a moist energy error norm. There are also differences in data assimilation systems that affect the impact that observations have–for example, ECMWF uses a 12-hr assimilation window beginning at 0900 or 2100 UTC, while FNMOC uses a 6-h assimilation window beginning at 0300, 0900, 1500, or 2100 UTC. These differences both affect the FSOI for radiosonde data, since the *direct* contribution of humidity observations is accounted for with a moist energy error norm (but not with a dry energy error norm), and since most radiosondes are launched roughly an hour before 0000 and 1200 UTC, putting them comparatively early in the ECMWF assimilation window where observations have less FSOI impact than they do late in the assimilation window (e.g., Ingleby et al. 2019).

Note that the volume of data assimilated is dominated by satellite observations for both of the systems depicted in Fig. 19 for January 2020. In the ECMWF system, satellite radiances account for 85.5% of the observations assimilated, with AMVs contributing 2.3%, GPS-RO 1.7% and the Aeolus satellite-based wind lidar another 0.4% for a total of 89.9%. Aeolus winds and all-sky radiances are not used in the

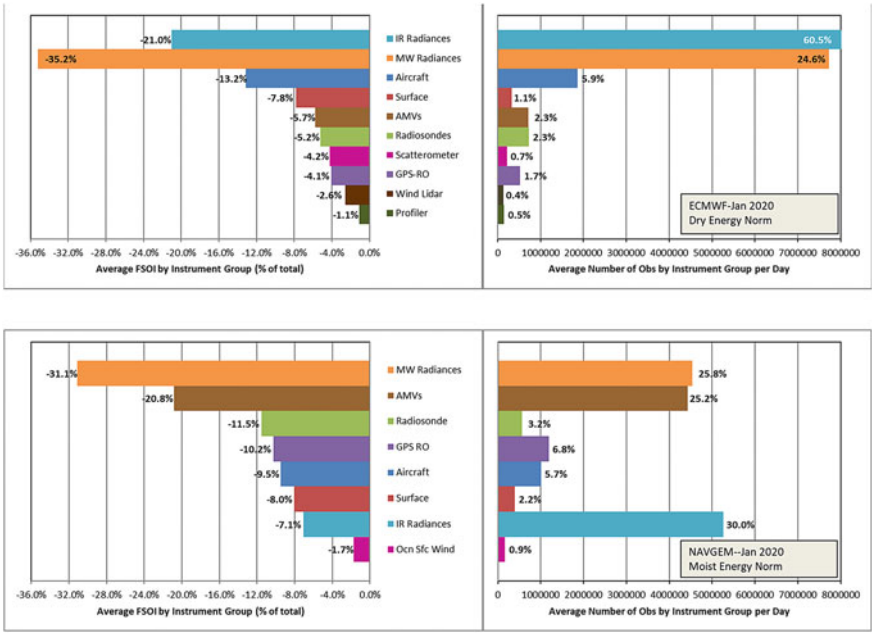

Fig. 19 Forecast Sensitivity to Observation Impact (FSOI) (left) expressed as a percentage of the total and counts of assimilated observations (right) by instrument group for ECMWF's IFS (top) and for the U.S. Navy's NAVGEM (bottom) for January 2020. The instrument groups are labeled separately for the top and bottom figures according to the legend in the center. Note that the counts for IR radiances are truncated in the top plot in order to better depict the counts for the remaining types

FNMOC system and only roughly one-quarter of the number of IR radiances are used. Even so, radiances account for 55.8% of the observations, AMVs 25.2%, and GPS-RO 6.8%, for a total of 87.8%. In contrast, in-situ observations accounted for 9.3% of the observations assimilated in the ECMWF system and 11.1% in the FNMOC system. The percentages of satellite and in-situ observations are remarkably similar between the two systems, given the large differences in types and amounts of data assimilated. However, the impact of the in-situ observations is much greater than the counts would suggest. In-situ observations contributed 26.2% of the error reduction in 24-h forecasts in the ECMWF system and 29.0% in the FNMOC system. Within this group, aircraft data play the largest role in the ECMWF system, while radiosonde data play the largest role in the FNMOC system, in no small part because FNMOC does not currently assimilate flight-level temperatures.

A complementary and arguably more definitive way of evaluating the influence observations have in a data assimilation system is through data denial experiments (OSEs). These allow the evaluation of the impact a group of observations has throughout the forecast period, but require separate data denial and control model runs. They measure the cumulative impact of withholding a group of observations,

whereas FSOI measures the impact of that group of observations on a particular forecast range in the presence of all of the observations, in effect, restarting from an 'all data' forecast each time. A recent set of OSEs performed at ECMWF (Bormann et al. 2019) for June–September 2016 and December 2017–March 2018 included one in which "conventional" data were denied, with conventional data defined as the three groups of in-situ observations described in this chapter plus wind profiler observations. In medium-range (2–10 day) forecasts, the impact from denying conventional observations was greater in the Northern Hemisphere than denying microwave (MW) radiances, infrared (IR) radiances, GPS RO, or Atmospheric Motion Vectors (AMVs). Increases in 3-day forecast error for the summer and winter experiments combined were 10% for 500 hPa geopotential heights, 7% for both 850 hPa and 200 hPa vector wind, 7% for 100 hPa temperature, and 5% for 850 hPa relative humidity. Denying microwave radiances had the dominant impact for the same variables for the Southern Hemisphere, while results were more mixed in the tropics. The impact of the data denials generally decreased with increasing forecast range.

Breaking this down a bit further, the UK Met Office performed a 90-day study starting from 15 August 2019, where separate experiments were performed to deny aircraft data and to deny radiosonde data (Ingleby et al. 2020d, Supplement S2). In the experiment that denied aircraft data, forecast errors increased by the greatest amount for short-range forecasts in the upper-troposphere, with increases for 12-h forecasts at 250 hPa of 9% for winds, 11% for temperatures, and 8% for geopotential heights. The UK Met Office scorecard combines numeric changes in wind, temperature, and geopotential height at various pressure levels and forecast times for the Northern Hemisphere, Southern Hemisphere, and tropics into a score signifying percentage improvement. The overall score for the experiment denying aircraft data was -1.1%, very similar to the score of -1.2% for the experiment denying radiosonde data. An ECMWF study denying aircraft data for 28 January to 30 April 2019 similarly showed the greatest impact at 250 hPa and for short forecast ranges, with most of the impact coming from aircraft winds (Ingleby et al. 2020d).

Although the focus of this chapter is on global NWP, results from data denial experiments using the NOAA Rapid Refresh (RAP) NWP system illuminate the importance of in-situ data for data-dense regions such as the contiguous U.S (James and Benjamin 2017). The RAP uses an hourly update cycle in its data assimilation system, using observations from radiosondes, aircraft, land and marine surface platforms, radar, wind profilers, satellites (AMVs and radiances), as well as GPS-derived IWV and lightning observations. A comprehensive set of data denial experiments was performed for observations from May 2013, July 2014, and January 2015. Their results showed that aircraft data were the most important observation type overall for short-range forecasts of wind, temperature, and RH. Surface and radiosonde data had secondary but similar impacts for winds and temperature and comparable impact for RH.

The current COVID-19 crisis has inadvertently provided a real-time data denial experiment, but one with no corresponding control for comparison. The decrease in aircraft data was very dramatic in the second half of March 2020, with the counts reaching a minimum in late April (Ingleby et al. 2020d; James et al. 2020). After

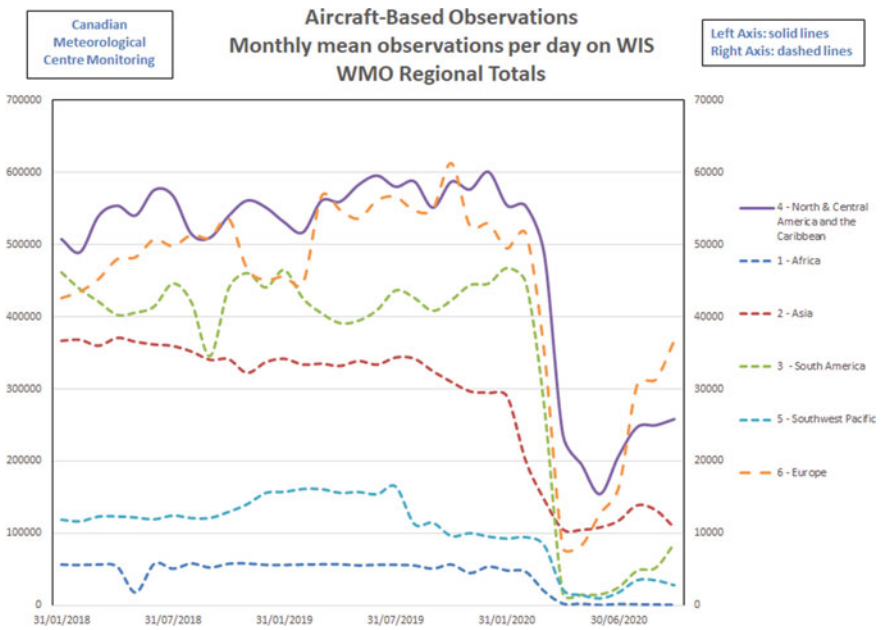

Fig. 20 Time series of ABO counts by WMO region. Note that the Region IV numbers are plotted against the left axis, while the other regions are plotted against the right axis (Figure courtesy of Yulia Zaitseva, Canadian Meteorological Center)

that, counts rebounded somewhat in May through July 2020 to not quite half of the data volume in February 2020 (Ingleby et al. 2020d). However, the decrease in aircraft data and the rebound was not the same in all regions. Figure 20 shows the ABO time series broken down by WMO region and clearly portrays the COVID-19 decrease in all six regions. However, the rebound has been uneven, with Europe and North America showing the greatest recovery and other regions showing only slow increases if that. Given this large decrease in aircraft data and the results from the data denial studies mentioned above, did forecast skill degrade during spring and summer 2020?

As mentioned above, studies that deny aircraft data show increased forecast errors, but they don't answer the question about what happens when aircraft data are reduced rather than eliminated. A comparison of FSOI between January 2020 (Fig. 19) and April 2020 (not shown) from the U.S. Navy's NAVGEM shows that average counts for aircraft data translate to 1.03M observations per day in January but only 371K observations per day in April. However, the FSOI did not drop quite as precipitously, with aircraft data giving 9.5% of the 24-h forecast error reduction in January but 4.2% of the error reduction in April. One of the interesting aspects of FSOI is its ability to capture tradeoffs between observation types. Often, adding a new observation type can show beneficial impact for that observation type while at the same time reducing the beneficial impact from other observation types. The reverse happened for April

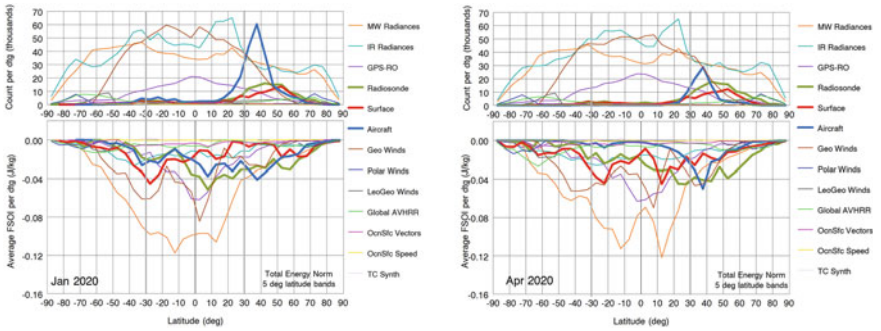

Fig. 21 Average counts (top) and FSOI (bottom) by 5° latitude band for January 2020 (left) and April 2020 (right) for instrument type groups used operationally in the U.S. Navy's NAVGEM. Averages are given per 6-h date-time-group (dtg) time window, and each variable, (temperature, u-wind, v-wind, etc.) counts as a separate observation. Note that negative values for FSOI indicate a reduction in 24-h forecast error and therefore beneficial impact

2020–while the beneficial impact decreased for aircraft data, it increased especially for surface observations (from 8% in January to 10.3% in April). Global AVHRR satellite winds include EUMETSAT single and triplet MetOp winds in polar regions as well as dual MetOp winds with global coverage; their beneficial impact increased from 3.6% in January to 5.3% in April.

Figure 21 shows how the FSOI and counts break down by latitude band for January and April 2020. The sharp decrease in aircraft data counts is shown by comparing the heavy blue lines in the top graphs, with a large peak in Northern Hemisphere mid-latitudes that decreased roughly by half in April. However, the loss of impact of the aircraft observations is felt more in the Southern Hemisphere mid-latitudes and in the tropics, as can be seen by the nearly zero heavy blue line for those latitudes in April compared to the significant impact seen in January. This agrees with the drop to very small numbers of observations seen in Fig. 20 for South America, the Southwest Pacific (which includes Australia and New Zealand), and Africa, and is consistent with the significant regional impact found for aircraft data for South America by Cotton and Eyre (2019). Data from surface platforms are depicted in Fig. 21 as heavy red lines. Although little difference is seen in the distribution of counts, the beneficial impact for these observations is increased in the tropics and Northern Hemisphere mid-latitudes in April compared to January. Increased beneficial impact for Global AVHRR winds and for radiosondes is also present in April on the north side of the aircraft data peak. The ability of observations of different types to compensate for the loss of aircraft data is an example of what Bormann et al. (2019) refer to as the complementarity and resilience of the global observing system.

The impact of an observation in a data dense area is less than that of a similar observation in a data sparse area: for example the FSOI per datum of surface pressure from buoys/ships is about six times that of surface pressure from land stations (ECMWF, January 2020 data). Table 4 summarizes results for different 'radiosonde types' in January 2020. PILOT profiles contain winds only—an average of 20 levels

Table 4 Summary of assimilated radiosonde (RS) numbers and FSOI per profile (arbitrary units) for the ECMWF system, Jan 2020

	PILOT TAC	LAND RS TAC	LAND RS BUFR	SHIP RS BUFR
# of profiles used	14259	19954	21094	386
# data per profile	40	138	889	940
Ratio of FSOI per profile	4	10	15	30

were used (the '# data' row counts u and v separately) they do not extend as high as most radiosonde ascents. For land radiosonde reports many more levels are reported and used from the BUFR reports, and per profile the FSOI is 50% larger than from TAC reports. This is probably an underestimate of the impact of high-resolution reporting for NWP because the high-resolution reports also tend to be in areas with good spatial coverage. The numbers of reports from ships are much smaller (and the error bars on the FSOI correspondingly larger), but the FSOI per profile is about double that from high-resolution land reports. Figure 22 shows a breakdown in the vertical by variable for radiosondes and aircraft—for both the wind FSOI is more than twice the temperature FSOI (humidity FSOI is much smaller, partly due to the use of a dry energy norm). Satellites provide a lot of temperature information but less wind information and this is part of the reason for the dominance of wind impact. For radiosondes there is a maximum of wind impact at 100 hPa, mainly coming from the tropics, because of the lack of other wind data at those levels (although ECMWF started using wind data from the Aeolus satellite on 10 January 2020). In the extratropics background errors are somewhat larger in winter, combined with the disproportionate number of in-situ observations in the northern hemisphere this means that the global FSOI of in situ observations is slightly larger in the northern winter than in the northern summer.

Ingleby et al. (2020d) examined the impact of the loss of aircraft data on model skill in the ECMWF global system by comparing time series of model verification statistics. Since the greatest impact of denying aircraft observations was seen in the upper troposphere at short forecast ranges, the time series of RMS vector wind error for 20° N–90° N in 24-h forecasts were examined at 500, 250, and 100 hPa for late February through July for three years—2018, 2019, and 2020 (Ingleby et al. 2020d, Fig. 3). The 250 hPa RMS vector wind errors show considerable day-to-day, week-to-week, and even year-to-year variability, but no step increase in error corresponding to the large drop in aircraft data is present. A comparison of 500 hPa RMS height error and 500 hPa height anomaly correlation for 72 h forecasts for the Northern Hemisphere extratropics from a number of operational global NWP systems from January 2016 through July 2020 shows that the forecast skill in spring 2020 was not unusual compared to other springs. The paper concluded that any loss of forecast skill was too small to show up in the usual global model metrics and was partly offset by increases in other observation types.

James et al. (2020) performed a partial data denial study in which 80% of the aircraft data were denied over the contiguous U.S. in two 10-day experiments using

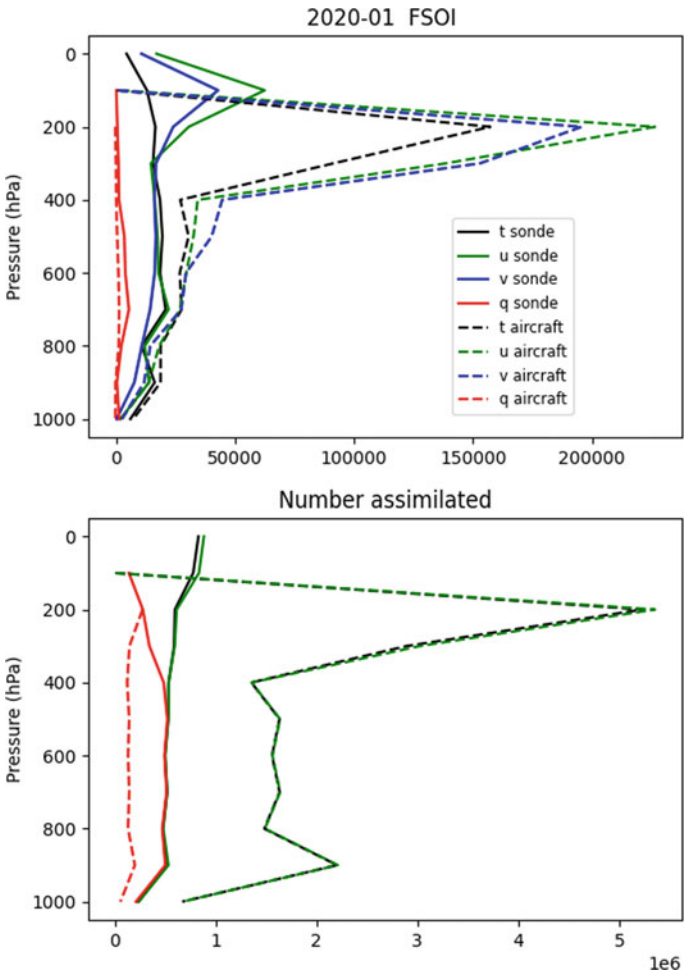

Fig. 22 Total aircraft and radiosonde negative FSOI (top) and data counts (bottom) for January 2020 in the ECMWF system calculated in 100 hPa layers and by variable (u and v wind data counts are identical, t—temperature, q—specific humidity)

a later version of the RAP system than was used in James and Benjamin (2017). Their results showed relatively small increases in RMS forecast error compared to denying all aircraft data. Focusing on their February 2019 experiment, the results for 3, 6, 9, and 12-h forecasts show that denying 80% (all) of the aircraft data led to an increase in RMS vector wind error of 0.1–0.15 m/s (0.15–0.45 m/s) and an increase in RMS temperature error of 0.02–0.04 K (0.05–0.14 K). These results likely underestimate the actual impact of either an 80% cut or total denial of aircraft observations, since aircraft observations were still used in the global data assimilation system providing lateral boundary conditions for these regional data assimilation experiments. Even

so, the result that using only 20% of the aircraft observations still gives significant impact is consistent with the global model results and the FSOI results described above.

In summary, in-situ observations play an important role in global data assimilation despite being heavily outnumbered by satellite observations. Observations from radiosondes, aircraft, and surface platforms provide over one-quarter of the error reduction (FSOI) in 24-h forecasts from roughly 10% of the assimilated observations. These data types complement each other and a degree of redundancy in the observations compensates for both short-term data outages (such as periodically occur in operational data streams) as well as longer-term outages such as the COVID-19 drop in aircraft data. In-situ observations also play a role in providing anchor observations for radiance bias correction schemes. Improvements in these observations, such as the improved precision, vertical resolution, and location information in high-resolution BUFR radiosonde data can then play a role in improving the accuracy of bias correction, as should future improvements such as correcting aircraft temperature biases at the source.

Acknowledgements The authors would like to thank Lars Isaksen (ECMWF), Chris Thomas (UKMO), and Nancy Baker (NRL) for their thoughtful comments on the manuscript. They would also like to thank Yulia Zaitseva (ECCC) for the use of Figure 20 and for helpful answers to questions, along with Frédéric Lenormand (ECCC), Stewart Taylor (E-AMDAR), Holger Vömel (NCAR), Colin Parrett (UKMO), Adam Maycock (UKMO), Judy St. James (ECCC), Chet Schmitt (NOAA), Ming Hu (ESRL), Bryce Ford (SpectraSensors), Brad Ballish (now retired from NCEP), and Dan Tyndall (NRL). Finally, thanks go to Stan Benjamin (ESRL) for providing an early copy of James et al. (2020) and to Melissa Eighmy Brown (U. Minn) and Blake Sorenson (U. ND) for assistance with references not readily available on the internet during these COVID telework times. The first author gratefully acknowledges support from the Naval Research Laboratory under program element 062435N.

Appendix 1: Definitions of Acronyms

4DVAR	4-Dimensional VARiational data assimilation
ABO	Aircraft-Based Observations
ACARS	Aircraft Communications, Addressing, and Reporting System
ADS-B	Automatic Dependent Surveillance—Broadcast
ADS-C	Automatic Dependent Surveillance—Contract
AFIRS	Automated Flight Information Reporting System
AIREP	Aircraft Report
AMDAR	Aircraft Meteorological Data Relay
AMMA	African Monsoon Multidisciplinary Analysis
ANSP	Air Navigation Service Provider
AOSFRS	AMDAR Onboard Software Functional Requirements Specification
APOB	Airplane Observation

ARINC	Aeronautical Radio, Incorporated
ASAP	Automated Shipboard Aerological Program
ASDAR	Aircraft to Satellite Data Relay
ASPEN	Atmospheric Sounding Processing Environment
ATC	Air Traffic Control
AVAPS	Airborne Vertical Atmospheric Profiling System
BSRN	Baseline Surface Radiation Network
BUFR	Binary Universal Format for the Representation of meteorological data
C-MAN	Coastal-Marine Automated Network
CNES	Centre National d'Etudes Spatiales
CQC	Complex Quality Control (of radiosonde heights and temperatures)
DA	Data Assimilation
DEVG	Derived Equivalent Vertical Gust
DMI	Danish Meteorological Institute
DWD	Deutscher Wetterdienst (Germany's National Meteorological Service)
E-ABO	EUMETNET Aircraft-Based Observations
E-ADOS	E-AMDAR Data Optimization System
E-ASAP	EUMETNET Automated Shipboard Aerological Program
ECCC	Environment and Climate Change Canada
ECMWF	European Centre for Medium-Range Weather Forecasts
EDA	Ensemble of Data Assimilations
EDR	Eddy Dissipation Rate
EHS	Enhanced Surveillance, associated with Mode-S on ATC radars
EMADDC	European Meteorological Aircraft Derived Data Center
ERA-Interim	Interim ECMWF ReAnalysis, the predecessor to ERA5
ERA5	ECMWF ReAnalysis, fifth generation
EUMETNET	European Meteorological Network
FGGE	First GARP (Global Atmospheric Research Program) Global Experiment
FL	Flight Level (usually reported in hundreds of feet
FNMOC	(U.S. Navy's) Fleet Numerical Meteorology and Oceanography Center
FSOI	Forecast Sensitivity to Observation Impact
GCOS	Global Climate Observing System
GNSS	Global Navigation Satellite System (which includes GPS)
GPS	Global Positioning System
GPS RO	GPS Radio Occultation observations
GRUAN	GCOS Reference Upper-Air Network
GTS	Global Telecommunications System
GUAN	GCOS Upper-Air Network
HDOB	High Density (high accuracy) Observations (from hurricane hunter aircraft)
HRRR	(NCEP's) High-Resolution Rapid Refresh model

HWRF	(NOAA's) Hurricane Weather Research and Forecasting model
ICAMS	(U.S.) Interagency Council for Advancing Meteorological Services (formerly OFCM)
ICAO	International Civil Aviation Organization
IFS	(ECMWF's) Integrated Forecasting System
IMOP	WMO's Instruments and Methods of Observation Program
INS	Inertial Navigation System
IOM	Instruments and Observing Methods
IRS	Inertial Reference System
IWV	Integrated Water Vapor
IWXXM	ICAO Weather Exchange Model
KNMI	Koninklijk Nederlands Meteorologisch Instituut (Royal Dutch Meteorological Institute)
MDCRS	Meteorological Data Collection and Reporting System
METAR	Meteorological Terminal Air Report or METeorological Aerodrome Report
Mode-S EHS	Mode Selective Enhanced Surveillance
MRAR	Meteorological Routine Air Report
NASA	(U.S.) National Aeronautics and Space Administration
NAVGEM	(U.S.) Navy Global Environmental Model
NCAR	(U.S.) National Center for Atmospheric Research
NCEP	(U.S.) National Centers for Environmental Prediction
NMS	National Meteorological (or Hydrometeorological) Service
NOAA	(U.S.) National Oceanic and Atmospheric Administration
NRL	(U.S.) Naval Research Laboratory
NSF	(U.S.) National Science Foundation
NWP	Numerical Weather Prediction
NWS	(U.S.) National Weather Service
O-A	Observation minus Analysis
O-B	Observation minus Background
OFCM	U.S. Office of the Federal Coordinator for Meteorological Services and Supporting Research, reorganized as the Interagency Council for Advancing Meteorological Services (ICAMS)
OIS	(WMO's) Operational Information Service
ON29	Office Note 29 internal format used at NCEP (predecessor to BUFR)
OSCAR	(WMO's) Observing Systems Capability Analysis and Review tool
OSE	Observing System Experiment (also called data denial experiment)
PILOT	WMO TAC format for wind profiles
PIRATA	Prediction and Research Moored Array in the Tropical Atlantic
PIREP	Pilot Report
Pmsl	Mean sea-level pressure
Psfc	Surface pressure, also sometimes called station pressure
QC	Quality Control
QFE	Aviation "Q" code for surface pressure at the field elevation

QNE	Aviation "Q" code for the standard altimeter setting (1013.25 hPa)
QNH	Aviation "Q"-code for altimeter setting
RAMA	Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction
RAWS	Remote Automated Weather Station
RH	Relative Humidity
RWC	Regional WIGOS Center
SFMR	Stepped Frequency Microwave Radiometer
SPECI	"Special" METAR report, triggered when certain criteria are met
SYNOP	WMO TAC format for surface data
TAC	Traditional Alphanumeric Codes
TAMDAR	Tropospheric Airborne Meteorological Data Reporting
TAT	Total Air Temperature
TAO/TRITON	Tropical Atmosphere Ocean/ TRIangle Trans-Ocean buoy Network
TEMP	WMO TAC format for radiosonde data
UAV	Unmanned Aerial Vehicle
UKV	UK Variable resolution component of the Met Office Unified Model
VarBC	Variational Bias Correction
VOS	Volunteer Observing Ship
WBAN	(United States) Weather Bureau-Army-Navy
WDQMS	WIGOS Data Quality Monitoring System
WIGOS	WMO Integrated Global Observing System
WIS	WMO Information System
WMO	World Meteorological Organization
WSI	WIGOS Station Identifier
WVSS-II	Water Vapor Sensing System, second generation
WWW	World Weather Watch

Appendix 2: Station Metadata Considerations

In order for in-situ observations to be useful for data assimilation, they need to be accompanied by metadata that accurately describe the location of the observation. Historically, the station metadata—latitude, longitude, and elevation—for radiosonde and land surface stations were omitted from TAC-formatted observations in order to conserve space and bandwidth. NWP centers therefore maintain local station lists indexed by station identifiers (5-digit WMO block station numbers or 4-character ICAO identifiers) to supply this information. Even though an advantage of BUFR is its inclusion of station metadata, a few stations have metadata errors, making it advisable to check the BUFR-provided values against a local station list. This is especially true for "reformatted" BUFR, where the station metadata are provided by a station list

that can be difficult to correct. This appendix describes some considerations about station metadata and provides examples of errors.

Station list maintenance: Before the implementation of OSCAR/Surface in July 2016, station metadata indexed by block-station numbers were collected and published by the WMO as Pub. 9, Vol. A. [8] Pub 9 Vol A gave latitudes and longitudes in degrees-minutes-seconds, often with only one-minute accuracy (i.e., 0.0167°). OSCAR/Surface uses decimal degrees, but many of the entries were converted from the Pub 9 Vol A list and so have the same fundamental accuracy. Elevations in both are generally given in meters. The OSCAR/Surface database (https://oscar.wmo.int/surface/#/), however, is only as good as WMO member countries make it. While some countries keep their entries accurate and up-to-date, other countries are slow to make corrections and may not even provide complete metadata. For example, at least two dozen stations that currently provide regular radiosonde reports lack elevations.

Maintaining station lists is a necessary but underappreciated activity at NWP centers. Stations move and may or may not be given a new identifier, equipment is modernized which may change the location or especially the elevation, and errors in metadata are detected, requiring some detective work to determine correct values. Ideally, WMO member countries should work to ensure the accuracy of the OSCAR/Surface entries they are responsible for, but at the same time, NWP centers need to have a mechanism for overriding known errors in OSCAR/Surface.

WIGOS Station Identifiers: Prior to July 2016, the WMO had assigned five-digit "block-station" numbers as identifiers for radiosonde, pilot balloon, and surface land stations, with the first two digits specifying a country or region and the last three specifying a particular station. Radiosonde profiles or surface observations taken from ships or "mobile" stations use alphanumeric identifiers that are typically 4–7 characters in length; these are not included in station lists since the SHIP and MOBIL formats of TEMP and SYNOP include the latitude, longitude, and elevation of the observation. Station identifiers are not used for dropsonde data, although the TEMP DROP format does include an observation number that can be used to identify a particular dropsonde.

The WMO has begun a migration from the legacy "block-station" identifiers to longer WIGOS Station Identifiers (WSIs), as part of the WMO Integrated Global Observing System (WIGOS) initiative. A description of WSIs is given in Attachment 2.2 of the WIGOS Manual (WMO 2019c) and in OFCM (2020b). The motivation for moving to longer identifiers was to both accommodate future high-resolution datasets such as surface mesonets and to provide more identifiers for countries that were running out of block-station numbers. For example, Canada started recycling obsolete block-station numbers when they had used all the numbers in block 71, which caused no small amount of confusion at FNMOC where the obsolete locations

[8] WMO Pub 9 Vol A was originally provided in document form (e.g., the WMO (2012a) document) and in flatfile form through April 2016 (https://www.wmo.int/pages/prog/www/ois/volume-a/vola-home.htm). A "legacy" flatfile similar to the Pub 9 Vol A flatfile has been produced from the OSCAR/Surface metadata from July 2016 through the present and is available at the Vol A website listed here.

were still in their local station list. Even though the motivation is prompted more by surface data considerations, radiosonde stations and even aircraft will also be assigned WSIs.

WSIs are laid out in four parts:

- WIGOS Identifier Series—only series "0" has currently been defined
- Issuer of Identifier—WMO Program or Country/National Identifier (5-digit number)
- Issue Number—typically "0" for WMO programs; 0–65535 for nationally defined schemas
- Local Identifier—legacy identifier for existing stations (e.g., block-station number); up to 16 characters (A–Z, a–z, 0–9) for nationally defined schemas

When OSCAR/Surface was set up in 2016, metadata for existing radiosonde and surface stations with block-station numbers were imported from WMO Pub 9, Vol A and assigned WSIs with the Issuer of Identifier set to either 20000 (or 20001 for the radiosonde site when an identifier has separate entries for the surface and radiosonde sites), the Issue Number set to 0, and the Local Identifier set to the block-station number. In searching for a particular WSI in OSCAR/Surface, the WSI is listed with the four parts separated by dashes (e.g., the WSI 0-20001-0-72662 corresponds to the radiosonde station using block-station number 72662). Station searches in OSCAR/Surface can use block-station numbers or WSIs.

Note that TAC formats such as TEMP, PILOT, and SYNOP cannot use WSIs—these older code forms can only use block station numbers. A four-part BUFR descriptor has been established to allow the use of WSIs in BUFR messages, with the current recommendation that stations with block-station numbers provide both forms of the identifier in the BUFR message. At present, very few countries are using WSIs, although many are making plans for their use. Further details about WSIs can be found in WMO (2019c) for the WMO perspective and in OFCM (2020b) for the U.S. implementation.

From the perspective of NWP, this large change in length for station identifiers is a significant problem. While decoding WSIs from BUFR files is not difficult, accommodating them in databases and using them in existing data assimilation systems is a problem that will take considerable effort at NWP centers. Any code changes will also need to be backward compatible to accommodate reanalyses and retrospective studies in addition to the mix of block-station numbers and WSIs that will be in use operationally for at least the next decade if the TAC to BUFR migration is any indication. ECMWF is using the strategy of adding a new character variable long enough to hold WSIs in addition to its usual eight-character identifier. If a station only reports a WSI, a unique eight-character identifier is then generated for use within the NWP system.

Station elevation: Elevation is a problem both because there are multiple elevations to choose among and because small errors in elevation have a greater impact than small errors in latitude or longitude. To be specific, an error of a few kilometers in the location of a station can be tolerated in all but the highest resolution numerical

models, while an error of a few meters in elevation can make the difference between an observation being used or being rejected. An error of 8 to 9 m in elevation implies a surface pressure error of roughly 1.0 hPa.

Using the terminology from WMO Pub 9 Vol A, there are two elevations that are reported: "Hp" is the barometer height and "Hha" is either "H", the ground elevation in the vicinity of the station, or "HA", the field elevation defined as the official altitude for an airport station (WMO 2012a). (These two elevations are also present in OSCAR/Surface but are neither well labeled nor easily found from the website interface.) For radiosonde data, the desired elevation is the one that sets the baseline with the surface pressure observation—"Hp"; this baseline is used in the hydrostatic calculation of height (or pressure depending on the radiosonde type). However, the U.S. practice is to modify the pressure observation to "correct" it to the launch elevation (or release height) (Table J-1, NWS 2010b). This means that "Hha" is the correct elevation to use for U.S. radiosonde stations. Choosing the wrong elevation can lead to height errors in the entire sounding. Figure 23 shows an example of Hha rather than Hp being used for a station that switched from using RS92 to RS41 radiosondes (Ingleby et al. 2016b, Fig. 5). The switch initially led to height O-B values increasing from 10 m or less near the surface to roughly 40 m. A comparison of the values before (red) and after (blue) a partial correction and after

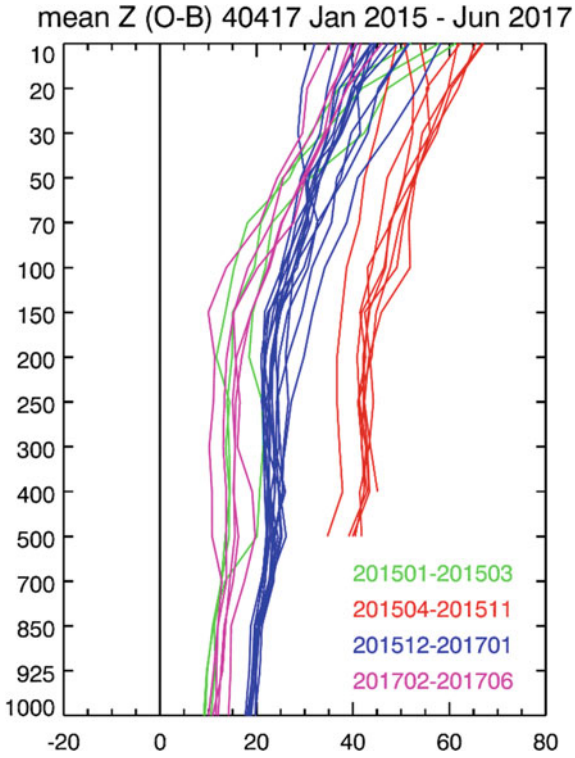

Fig. 23 Mean monthly height O-B departures (in m) as a function of pressure (in hPa) color-coded by time period for station 40417, showing the impact of using an incorrect station elevation. The green lines are for a period prior to the switch from RS92 to RS41 radiosondes, the red lines are for a period after the switch, the blue lines are for a period after a partial correction was made, and the magenta lines after a full correction was made

a full correction (magenta) demonstrates that the elevation error leads to a constant height error.

Choosing the correct elevation is also an issue for native BUFR radiosonde reports. There are several elevations included in the standard radiosonde BUFR template, which can be a source of confusion. In principle, the most important of these is the radiosonde release height; this should be the height corresponding to the surface pressure and so the baseline for the hydrostatic calculation. The release height may or may not be the same as the station height and the geopotential height reported for the surface level in the sounding, although it should be at least close to as the latter. The release height is also often missing, in which case the station height is usually the best choice for the baseline height. The format also allows a barometer height to be reported, since the pressure given for the surface level is from an external barometer and not from the radiosonde itself. However, the barometer height is often missing and sometimes suspiciously set to the same value for a group of stations within a particular country.

GPS elevation: Another issue with elevation data in general is the use of GPS elevations. While GPS provides an easy way to determine the station metadata for a particular site, GPS elevations are not the same as elevations above mean sea level and require conversion as described above in the discussion of geopotential heights. Ingleby (2017) presents an example for radiosonde station 04417—the "Geosummit" station in the heart of Greenland, which is operated by the U.S. National Science Foundation in association with the Danish Meteorological Institute (DMI). A 35 m height bias for this station with respect to the ECMWF background was noted in early 2016 at levels between 700 and 150 hPa. An exchange of emails confirmed that the radiosonde ground station was using an unadjusted GPS height of 3255 m, which was changed in July 2016 to 3216 m and subsequently updated by DMI in OSCAR/Surface. Unfortunately, the radiosonde ground station was moved in August 2019 and a GPS elevation was again entered, with the height set to 3258 m. This was corrected to 3208 m in October 2020. This cautionary tale has two conclusions: (1) GPS is a tempting technology to determine station metadata but provides elevations that require (and might not receive) correction and (2) metadata (and data) problems that get fixed may not stay fixed!

Dual location stations: Some stations use the same station identifier for both a surface station and a radiosonde station. In the U.S., surface stations are almost always associated with airports, reflecting the practice of generating WMO SYNOP reports from ICAO METAR reports. Upper-air stations are almost always associated with NWS offices, many of which were relocated in the 1990s as part of the NWS restructuring and modernization (National Research Council 2012). As a result, there are some stations that use the same station identifier for a surface station and a radiosonde station that are not collocated either in terms of location (latitude and longitude) or elevation; these are referred to as dual location stations and are described in more detail in OFCM (2020b). While OSCAR/Surface and even WMO Pub 9 Vol A are capable of having separate entries for the surface and radiosonde station, at present some U.S. dual location stations only have a single entry in OSCAR/Surface.

The Weather Forecast Office in Tucson, AZ, is an example of a dual location station (72274) that should have two entries but only has one. The radiosonde station is located at the University of Arizona (32.2278° N, 110.9560° W) with radiosondes launched from the roof of a building at an elevation of 751 m, agreeing with the station metadata in OSCAR/Surface. The surface station is located at the airport (32.1313° N, 110.9552° W) roughly 10 km away with a barometer height of 795 m and a field elevation of 805 m.

Metadata considerations for radiosonde data: The station metadata provided in "native" BUFR radiosonde reports come from information entered in the radiosonde ground station and are generally both accurate and precise, with latitude and longitude typically given in decimal degrees to five decimal places and elevation given to the nearest meter. However, problems with launch location can still occur in native BUFR reports. The two most common errors in the early days were omitting the negative sign for west longitude or entering latitudes/longitudes with the fractional degrees past the decimal point set to the minutes from values specified as degrees and minutes. At present, the errors are more subtle. For example, at least a few stations use a low-precision launch latitude or longitude. When drift offsets obtained by GPS are incorporated to give the actual location at individual levels in the sounding, these stations appear to have a jump in location for the first level or two that can be as large as a few km. However, if radar is used rather than GPS, the offsets are calculated relative to the radar and so do not show a jump in location if the launch location either contains error or lacks precision. So, GPS locations are self-correcting to a certain extent, but radar systems will give no indication of an inaccurate location.

However, "reformatted" BUFR is both more susceptible to errors in station metadata and seemingly more resistant to correcting these errors. In the U.S., both native and reformatted BUFR messages are generated for most NWS radiosonde stations, but a subset of stations including military sites uses equipment that does not generate native BUFR and so only have "reformatted" BUFR. For example, Shemya Air Force Base, Alaska (70414), is located far out in the Aleutian Islands past the date line. The station list used in generating reformatted BUFR lists the longitude as 174.11° W rather than the correct value of 174.10° E from OSCAR/Surface.

One final comment should be made in terms of the metadata provided in TEMP DROP messages. Metadata are not only provided for the launch point, but also for the "splash" point, and the first and last wind locations. In addition, an aircraft identifier is provided as well as a storm name/mission name. This extra metadata is encoded in nationally developed code groups that for the U.S. are described in Appendix G in the National Hurricane Operations Plan (OFCM 2020a). Native BUFR dropsonde messages do not encode these extra locations; the availability of balloon drift locations at each level make them unnecessary.

Metadata considerations for surface data: Since surface pressure is the most important variable reported in land or marine surface data messages for data assimilation, it is the one that is most important to match to an elevation. In practice, some NWP centers use "Hha" for the elevation for SYNOP stations, although the height of the surface pressure sensor is preferable (Ingleby 2015). BUFR SYNOP

reports are often reformatted from TAC SYNOP, with station metadata provided by an internal station list with the same issues described for reformatted radiosonde data. For example, the BUFR SYNOP reports for Portland, OR, (72698) give the station elevation as 456 m, rather than the elevation of 7m stated in OSCAR/Surface. The elevation of 456 m is correct for Pendleton, OR, (72688); this appears to be a simple transcription error between stations with similar WMO identifiers and names. As for radiosonde data, the station metadata provided in BUFR SYNOP reports should be checked against a local station list as the data are being processed for data assimilation.

Correctly specifying station metadata is a problem with METARs that will carry forward into the IWXXM era. Like SYNOP, the METAR format does not include latitude, longitude, and station (field) elevation, so these must be specified from a local station list. IWXXM can optionally include the latitude and longitude of the "Aerodrome Reference Point", defined as the "approximate geometric center of all usable runways" (https://www.ngs.noaa.gov/AERO/arpcomp/arpframe.html). Elevation will likely not be provided at least in the short term, so the need for a local station list for METAR/IWXXM stations will continue, as will the need to check the metadata that are provided. Determining correct values to include in a local station list is a greater problem for METARs than SYNOPs because of the lack of an official source for such metadata. There are online sources that give the airport name that corresponds to a given four-letter ICAO identifier and sometimes even a latitude and longitude for the airport, but elevations are more difficult to locate. Although OSCAR/Surface does not index entries by ICAO identifiers, some information is available there, indexed by city or airport name. Ingleby (2015) found that in March 2013 nearly 3% of SYNOPs were unusable because their locations were unknown, but that number jumps to 11% for METARs.

As discussed previously, U.S. SYNOP reports are reformatted from METAR reports using an unknown station list that can contain errors in elevation and in matching ICAO and WMO station identifiers. An extreme example of such an error was found in a search for duplicate reports between METARs and SYNOPs by only using meteorological variables (Brad Ballish, personal communication). This search revealed that the WMO identifier for Flagstaff, AZ, (72376) is being matched to the ICAO identifier for Farmington, NM, (KFMN), over 350 km away with a station elevation difference of roughly 500 m. While ideally this error should be corrected at the source, it could also be corrected locally by assigning the latitude, longitude, and elevation for KFMN (36.750°N, -108.229°E, 1678 m) to 72376 for SYNOP data. Note that 72376 has a single entry in OSCAR/Surface that gives the latitude, longitude, and field elevation as 35.230°, -111.822°, 2181 m, which are the values included in BUFR SYNOP reports for this station; using the OSCAR/Surface metadata in this case would lead to significant errors not only for surface pressure but for temperature, humidity, and winds as well. Putting 72376 on the SYNOP reject lists for each variable would certainly be warranted!

References

Agustí-Panareda A, Vasiljevic D, Beljaars A, Bock O, Guichard F, Nuret M, Garcia Mendez A, Andersson E, Bechtold P, Fink A, Hersbach H, Lafore J-P, Ngamini J-B, Parker DJ, Redelsperger J-L, Tompkins AM (2009) Radiosonde humidity bias correction over the West African region for the special AMMA reanalysis at ECMWF. Q J R Meteorol Soc 135:595–617. https://doi.org/10.1002/qj.396

Andersson E, Järvinen H (1999) Variational quality control. Q J R Meteorol Soc 125:697–722. https://doi.org/10.1002/qj.49712555416

Ballish BA, Kumar VK (2008) Systematic differences in aircraft and radiosonde temperatures: implications for NWP and climate studies. Bull Am Meteorol Soc 89:1689–1708. https://doi.org/10.1175/2008BAMS2332.1

Bédard J, Laroche S, Gauthier P (2015) A geo-statistical observation operator for the assimilation of near-surface wind data. Q J R Meteorol Soc 141:2857–2868. https://doi.org/10.1002/qj.2569

Benjamin SG et al (2016) A North American hourly assimilation and model forecast cycle: the rapid refresh. Mon Weather Rev 144:1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Benjamin SG, Brown JM, Brunet G, Lynch P, Saito K, Schlatter TW (2018) 100 years of progress in forecasting and NWP applications. Meteorol Monogr 59:13.1–13.67. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1

Bonavita M, Dahoui M, Lopez P, Prates F, Hólm E, De Chiara G, Geer A, Isaksen L, Ingleby B (2017) On the initialization of Tropical Cyclones. ECMWF Technical Memorandum 810, 40 pp. https://doi.org/10.21957/insgrw8u6

Bormann N, Lawrence H, Farnan J (2019) Global observing system experiments in the ECMWF assimilation system. ECMWF Technical Memorandum 839, 24 pp. https://doi.org/10.21957/sr184iyz

Brock FV, Crawford KC, Elliott RL, Cuperus GW, Stadler SJ, Johnson HL, Eilts MD (1995) The Oklahoma mesonet: a technical overview. J Atmos Ocean Technol 12:5–19. https://doi.org/10.1175/1520-0426(1995)012%3c0005:TOMATO%3e2.0.CO;2

Brohan P, Allan R, Freeman JE, Waple AM, Wheeler D, Wilkinson C, Woodruff S (2009) Marine observations of old weather. Bull Am Meteorol Soc 90:219–230. https://doi.org/10.1175/2008BAMS2522.1

Buehner M et al (2015) Implementation of deterministic weather forecasting systems based on ensemble-variational data assimilation at environment Canada. Part I: the global system. Mon Weather Rev 143:2532–2559. https://doi.org/10.1175/MWR-D-14-00354.1

Cardinali C (2009) Monitoring the observation impact on the short-range forecast. Q J R Meteorol Soc 135:239–250. https://doi.org/10.1002/qj.366

Chadsey EM, Moore PR, Rush RE, Smith JE, Vederman J (1967) Utilization of aircraft meteorological reports at WBFC Honolulu. Weather Bureau Technical Memorandum PR-3, 16 pp. ftp://ftp.library.noaa.gov/noaa_documents.lib/NWS/NWS_PR/ESSA_WBTM_PR/ESSA_WBTM_PR-3.pdf.

Cohn SA et al (2013) Driftsondes: providing in situ long-duration dropsonde observations over remote regions. Bull Am Meteorol Soc 94:1661–1674. https://doi.org/10.1175/BAMS-D-12-00075.1

Collins WG (2001) The operational complex quality control of radiosonde heights and temperatures at the National Centers for Environmental Prediction. Part I: description of the method. J Appl Meteorol 40:137–151. https://doi.org/10.1175/1520-0450(2001)040%3c0137:TOCQCO%3e2.0.CO;2

Compo GP et al (2011) The twentieth century reanalysis project. Q J R Meteorol Soc 137:1–28. https://doi.org/10.1002/qj.776

Cotton J, Eyre J (2019) Forecast Sensitivity to Observations Impact (FSOI) by country or region. UK Met Office Forecasting Research Technical Report No: 636, 18 pp. https://digital.nmla.metoffice.gov.uk/IO_3ded5555-2a07-4529-9b7f-1c2bad07096e/

Cucurull L, Anthes RA, Tsao L (2014) Radio occultation observations as anchor observations in numerical weather prediction models and associated reduction of bias corrections in microwave and infrared satellite observations. J Atmos Ocean Technol 31:20–32. https://doi.org/10.1175/JTECH-D-13-00059.1

Daniels TS, Tsoucalas G, Anderson M, Mulally D, Moninger W, Mamrosh R (2004). Tropospheric Airborne Meteorological Data Reporting (TAMDAR) sensor development. In: Preprints, 11th conference on aviation, range, and aerospace meteorology, Hyannis, MA. American Meteorological Society. https://ams.confex.com/ams/pdfpapers/81841.pdf.

Dee DP, Uppala S (2009) Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. Q J R Meteorol Soc 135:1830–1841. https://doi.org/10.1002/qj.493

de Haan S (2014) Availability and quality of Mode-S MRAR (BDS4.4) in the MUAC area: a first study. KNMI Internal Report IR 2014-01, 24 pp. http://mode-s.knmi.nl/documents/IR-2014-01.pdf

de Haan S, Bailey LJ, Können JE (2013) Quality assessment of Automatic Dependent Surveillance Contract (ADS-C) wind and temperature observation from commercial aircraft. Atmos Meas Tech 6:199–206. https://doi.org/10.5194/amt-6-199-2013

de Haan S, van der Meulen J, de Jong P (2020) Characterization of AMDAR temperature bias. Presented at the EUFAR (virtual) workshop on atmospheric temperature measurement from research and operational aircraft, 11–12 Nov 2020. https://www.eufar.net/shared_subjects/d/JH5VVONXTDMEYCJOAYIAH6WKPRO4UL4GR756RFMFB5QBKNX6YVHQ/

Derber JC, Wu W (1998) The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. Mon Weather Rev 126:2287–2299. https://doi.org/10.1175/1520-0493(1998)126%3c2287:TUOTCC%3e2.0.CO;2

de Rosnay P, Balsamo G, Albergel C, Muñoz-Sabater J, Isaksen L (2014) Initialisation of land surface variables for Numerical Weather Prediction. Surv Geophys 35(3):607–621. https://doi.org/10.1007/s10712-012-9207-x

de Rosnay P, Isaksen L, Dahoui M (2015) Snow data assimilation at ECMWF. ECMWF Newsl 143 (Spring 2015). https://doi.org/10.21957/lkpxq6x5

Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background, and analysis-error statistics in observation space. Q J R Meteorol Soc 131:3385–3396. https://doi.org/10.1256/qj.05.108

Diamond H, Hinman WS, Lapham EG (1938) Comparisons of soundings with radio-meteorographs, aerographs and meteorographs. Bull Am Meteorol Soc 19:129–141. https://doi.org/10.1175/1520-0477-19.4.129

Dirksen RJ, Sommer M, Immler FJ, Hurst DF, Kivi R, Vömel H (2014) Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde. Atmos Meas Tech 7:4463–4490. https://doi.org/10.5194/amt-7-4463-2014

Dow G, Macpherson B (2013) Benefit of convective-scale data assimilation and observing systems in the UK Models. Forecasting Research Technical Report 585, Met Office. https://digital.nmla.metoffice.gov.uk/file/sdb%3AdigitalFile%7Cffea7436-b297-43b1-b38a-cb8bf8677601/

Douglas MW, Murillo JM, Orozco RK, Mejia JF (2008) Underutilized observations for studying tropical climate variations: the historical pilot balloon database. In: Preprints, 24th conference on interactive information processing systems, New Orleans, LA. American Meteorological Society, 20–24 Jan 2008. https://ams.confex.com/ams/pdfpapers/135253.pdf

Drüe C, Frey W, Hoff A, Hauf Th (2008) Aircraft type-specific errors in AMDAR weather reports from commercial aircraft. Q J R Meteorol Soc 134:229–239. https://doi.org/10.1002/qj.205

Drüe C (2011) Reconstruction of aircraft trajectories from AMDAR weather reports. J Atmos Ocean Technol 28:921–932. https://doi.org/10.1175/2011JTECHA1499.1

Dutra E, Muñoz-Sabater J, Boussetta S, Komori T, Hirahara S, Balsamo G (2020) Environmental lapse rate for high-resolution land surface downscaling: an application to ERA5. Earth Space Sci 7. https://doi.org/10.1029/2019EA000984

Eden P (2009) Traditional weather observing in the UK: an historical overview. Weather 64:239–245. https://doi.org/10.1002/wea.469

Edwards D, Anderson G, Oakley T, Gault P (2014) Met Office intercomparison of Vaisala RS92 and RS41 Radiosondes. Met Office, 89 pp. https://www.vaisala.com/sites/default/files/documents/Met_Office_Intercomparison_of_Vaisala_RS41_and_RS92_Radiosondes.pdf

Eskridge RE, Alduchov OA, Chernykh IV, Panmao Z, Polansky AC, Doty SR (1995) A Comprehensive Aerological Reference Data Set (CARDS): rough and systematic errors. Bull Am Meteorol Soc 76:1759–1776. https://doi.org/10.1175/1520-0477(1995)076%3c1759:ACARDS%3e2.0.CO;2

EUMETNET (2015) Final report of the EUMETNET aircraft derived data feasibility study expert team, EUMETNET, Brussels, Belgium, 99 pp. http://mode-s.knmi.nl/documents/EUMETNET_ADD_Report_FINAL_v1.0_03102015.pdf

EUMETNET (2020) Observations—E-ABO. https://www.eumetnet.eu/activities/observations-programme/current-activities/e-amdar/. Accessed 4 Sept 2020

Eyre JR (2016) Observation bias correction schemes in data assimilation systems: a theoretical study of some of their properties. Q J R Meteorol Soc 142:2284–2291. https://doi.org/10.1002/qj.2819

FAA (2019). Aviation weather services. Advisory Circular AC 00–45H, 348 pp. https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_00-45H_CHG_2.pdf

FAA (2020) Flight services: Chapter 6—International operations, Section 1—Messages and formats. FAA Order JO 7110.10AA. https://www.faa.gov/air_traffic/publications/atpubs/fs_html/chap6_section_1.html. Accessed 3 Jan 2021

Fiebrich CA (2009) History of surface weather observations in the United States. Earth Sci Rev 93:77–84. https://doi.org/10.1016/j.earscirev.2009.01.001

Haiden T, Dahoui M, Ingleby B, de Rosnay P, Prates C, Kuscu E, Hewson T, Isaksen L, Richardson D, Zuo H, Jones L (2018) Use of in situ surface observations at ECMWF. ECMWF Technical Memorandum 834, 26 pp. https://doi.org/10.21957/dj9lpy4wa

Helms D, Hoff A, Smit HGJ, Taylor S, Carlberg S, Berechree M (2010) Advancements in the AMDAR humidity sensing. Presented at technical conference on meteorological and environmental instruments and methods of observation, Helsinki, Finland, WMO, 30 Aug–1 Sept 2010. https://library.wmo.int/pmb_ged/wmo-td_1546_en/2_1_Helms_USA.doc

Hersbach H et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146:1999–2049. https://doi.org/10.1002/qj.3803

Hill C (2020) The U.S. perspective in the monitoring and assimilation of aircraft-based observations. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1205/attachments/446/801/ABO-WS-Hill.pdf

Hintz KS, O'Boyle K, Dance SL et al (2019) Collecting and utilising crowdsourced data for numerical weather prediction: propositions from the meeting held in Copenhagen, 4–5 December 2018. Atmos Sci Lett 20.https://doi.org/10.1002/asl.921

Hock TF, Franklin JL (1999) The NCAR GPS dropwindsonde. Bull Am Meteorol Soc 80:407–420. https://doi.org/10.1175/1520-0477(1999)080%3c0407:TNGD%3e2.0.CO;2

Hodge MW (1956) Superadiabatic lapse rates of temperature in radiosonde observations. Mon Weather Rev 84:103–106. https://doi.org/10.1175/1520-0493(1956)084%3c0103:SLROTI%3e2.0.CO;2

Hollingsworth A, Lönnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I. The wind field. Tellus 38A:111–136. https://doi.org/10.3402/tellusa.v38i2.11707

Hollingsworth A, Shaw DB, Lönnberg P, Illari L, Arpe K, Simmons AJ (1986) Monitoring of observation and analysis quality by a data assimilation system. Mon Weather Rev 114:861–879. https://doi.org/10.1175/1520-0493(1986)114%3c0861:MOOAAQ%3e2.0.CO;2

Homans A (2007) ARINC's role in AMDAR: Meteorological Data Collection and Reporting System (MDCRS). Presented at the short course on meteorological applications of aircraft weather data, San Antonio, TX, 14 Jan 2007. https://amdar.noaa.gov/2007course/Homans.ppt

Houchi K, Stoffelen A, Marseille G, De Kloe J (2015) Statistical quality control of high-resolution winds of different radiosonde types for climatology analysis. J Atmos Ocean Technol 32:1796–1812. https://doi.org/10.1175/JTECH-D-14-00160.1

Hyland RW, Wexler A (1983) Formulations for the thermodynamic properties of the saturated phases of H_2O from 173.15 K to 473.15 K. ASHRAE Trans 89(2A):500–519

Ide K, Courtier P, Ghil M, Lorenc AC (1997) Unified notation for data assimilation: operational, sequential and variational. J Meteorol Soc Jpn 75:181–189. https://doi.org/10.2151/jmsj1965.75.1B_181

Ikonen I, Demetriades NWS, Holle RL (2010) Vaisala dropsondes: history, status, and applications. In: Proceedings of the 29th conference on hurricanes and tropical meteorology, Tucson, AZ, 10–14 May 2010. https://ams.confex.com/ams/pdfpapers/168031.pdf.

Ingleby NB (1995) Assimilation of station level pressure and errors in station height. Weather Forecast 10:172–182. https://doi.org/10.1175/1520-0434(1995)010%3C0172:AOSLPA%3E2.0.CO;2

Ingleby B (2010) Factors affecting ship and buoy data quality: a data assimilation perspective. J Atmos Ocean Technol 27:1476–1489. https://doi.org/10.1175/2010JTECHA1421.1

Ingleby B (2015) Global assimilation of air temperature, humidity, wind and pressure from surface stations. Q J R Meteorol Soc 141:504–517. https://doi.org/10.1002/qj.2372

Ingleby B (2016) On the accuracy of different radiosonde types. In: WMO technical conference on meteorological and environmental instruments and methods of observation (CIMO TECO 2016), Madrid, Spain, 27–30 Sept 2016. https://www.wmo.int/pages/prog/www/IMOP/publications/IOM-125_TECO_2016/Session_4/O4(8)_Ingleby_RadiosondeTypes.pdf

Ingleby B (2017) An assessment of different radiosonde types 2015/2016. ECMWF Technical Memorandum 807, 69 pp. https://doi.org/10.21957/0nje0wpsa

Ingleby B et al (2016a) Progress toward high-resolution, real-time radiosonde reports. Bull Am Meteorol Soc 97:2149–2161. https://doi.org/10.1175/BAMS-D-15-00169.1

Ingleby B, Dahoui M, Lehmuskero M (2016b) Improved radiosonde height reports: reporting precision, GPS reference height and other issues. In: WMO technical conference on meteorological and environmental instruments and methods of observation (CIMO TECO 2016), Madrid, Spain, 27–30 Sept 2016. https://www.wmo.int/pages/prog/www/IMOP/publications/IOM-125_TECO_2016/Session_4/P4(18)_Ingleby_Improved_radiosonde_height.pdf

Ingleby B, Edwards D (2015) Changes to radiosonde reports and their processing for numerical weather prediction. Atmos Sci Lett 16:44–49. https://doi.org/10.1002/asl2.518

Ingleby B, Edwards D (2018) Radiosonde descent reports: encouraging preliminary results. In: WMO technical conference on meteorological and environmental instruments and methods of observation (CIMO TECO 2018), Amsterdam, Netherlands, 8–11 Oct 2018. https://www.wmo.int/pages/prog/www/IMOP/documents/O3_5_Ingleby_ExtendedAbstract.pdf

Ingleby B, Huddleston M (2007) Quality control of ocean temperature and salinity profiles—historical and real-time data. J Mar Syst 65:158–175. https://doi.org/10.1016/j.jmarsys.2005.11.019

Ingleby B, Isaksen L (2018) Drifting buoy pressures: impact on NWP. Atmos Sci Lett 19.https://doi.org/10.1002/asl.822

Ingleby B, Isaksen L (2020) Use of Mode-S data and other aircraft issues, available on request

Ingleby B, Isaksen L, Kral T (2019) Evaluation and impact of aircraft humidity data in ECMWF's NWP system. ECMWF Technical Memorandum No. 855, https://doi.org/10.21957/4e825dtiy

Ingleby B, Isaksen L, Kral T, Haiden T, Dahoui M (2018) Improved use of atmospheric in situ data. ECMWF Newsl 155:20–25 (Spring 2018). https://doi.org/10.21957/cf724bi05s

Ingleby NB, Lorenc AC (1993) Bayesian quality control using multivariate normal distributions. Q J R Meteorol Soc 119:1195–1225. https://doi.org/10.1002/qj.49711951316

Ingleby B, Moore D, Sloan C, Dunn R (2013) Evolution and accuracy of surface humidity reports. J Atmos Ocean Technol 30:2025–2043. https://doi.org/10.1175/JTECH-D-12-00232.1

Ingleby B, Dahoui M, Isaksen L (2020a) ECMWF: aircraft assimilation and data monitoring. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use,

ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1198/attachments/442/797/ABO-WS-Ingleby.pdf

Ingleby B, Prates F, Isaksen L, Bonavita M (2020b) Recent BUFR dropsonde data improved forecasts. ECMWF Newsl 162:9 (ECMWF, Reading, UK). https://www.ecmwf.int/en/newsletter/162/news/recent-bufr-dropsonde-data-improved-forecasts

Ingleby B, Sondij J, de Haan S (2020c) EMADDC Mode-S: a new source of aircraft data over Europe. ECMWF Newsl 164:12. https://www.ecmwf.int/en/newsletter/164/news/emaddc-mode-s-new-source-aircraft-data-over-europe

Ingleby B, Candy B, Eyre J, Haiden T, Hill C, Isaksen L, Kleist D, Smith F, Steinle P, Taylor S, Tennant W, Tingwell C (2020d) The impact of COVID-19 on weather forecasts: a balanced view. Accepted by Geophys Res Lett. https://doi.org/10.1029/2020GL090699

Isaksen L, Bonavita M, Buizza R, Fisher M, Haseler J, Leutbecher M, Raynaud L (2010) Ensemble of data assimilations at ECMWF. ECMWF Technical Memorandum No. 636, 48 pp. https://doi.org/10.21957/obke4k60

Isaksen L, Vasiljevic D, Dee D, Healy S (2012) Bias correction of aircraft data implemented in November 2011. ECMWF Newsl 131:6 (ECMWF, Reading, UK). https://www.ecmwf.int/en/elibrary/14591-newsletter-no-131-spring-2012

Ivanov A, Kats A, Kurnosenko S, Nash J, Zaitseva N (1991) WMO international radiosonde comparison: phase III (Dzhambul, USSR, 1989), Final Report, WMO/TD-No. 451, Instruments and Observing Methods Report No. 40, WMO, 160 pp. https://library.wmo.int/doc_num.php?explnum_id=9547

Jaatinen J, Pälä E (1997) Windfinding accuracy of terrestrial navaids. Vaisala News 146:35–38. https://www.vaisala.com/sites/default/files/documents/VN146_Windfinding_Accuracy_of_Terrestrial_Navaids.pdf

James EP, Benjamin SG (2017) Observation system experiments with the hourly updating rapid refresh model using GSI hybrid ensemble-variational data assimilation. Mon Weather Rev 145:2897–2918. https://doi.org/10.1175/MWR-D-16-0398.1

James EP, Benjamin SG, Jamison BD (2020) Commercial aircraft-based observations for NWP: global coverage, data impacts, and COVID-19. J Appl Meteorol Climatol. https://doi.org/10.1175/JAMC-D-20-0010.1

Janjic T, Bohrmann N, Bocquet M, Carton JA, Cohn SE, Dance SL, Losa SN, Nichols NK, Potthast R, Waller JA, Weston P (2018) On the representation error in data assimilation. Q J R Meteorol Soc 144:1257–1278. https://doi.org/10.1002/qj.3130

Jauhiainen H, Survo P, Lehtinen R, Lentonen J (2014) Radiosonde RS41 and RS92 key differences and comparison test results in different locations and climates. In: TECO-2014, WMO technical conference on meteorological and environmental instruments and methods of observation, Saint Petersberg, Russian Federation, 7–9 July 2014. https://library.wmo.int/pmb_ged/iom_116_en/Session%203/P3_16_Juhiainen_Radiosonde_RS41_RS92_Key_Differences_Comparison_Test Results.pdf

Kaisti K, Saarnimo T, Karttunen C (1994) Windfinding in radiosonde using GPS. In: TECO-94, WMO technical conference on instruments and methods of observation, Geneva, Switzerland, 29 Feb–2 Mar 1994. https://library.wmo.int/doc_num.php?explnum_id=9607

Kalnay E (2003) Atmospheric modeling, data assimilation, and predictability. Cambridge University Press, 341 pp

Kent EC, Woodruff SD, Berry DI (2007) Metadata from WMO Publication No. 47 and an assessment of voluntary observing ship observation heights in ICOADS. J Atmos Ocean Technol 24:214–234. https://doi.org/10.1175/JTECH1949.1

Keyser D (2000) RAOB/PIBAL balloon drift latitude, longitude, and time calculation in PREPBUFR. https://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/balloon_drift_for_TPB.htm.

Khelif D, Burns SP, Friehe CA (1999) Improved wind measurements on research aircraft. J Atmos Ocean Technol 16:860–875. https://doi.org/10.1175/1520-0426(1999)016%3c0860:IWMORA%3e2.0.CO;2

Kistler R et al (2001) The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation. Bull Am Meteorol Soc 82:247–268. https://doi.org/10.1175/1520-0477(2001)082%3c0247:TNNYRM%3e2.3.CO;2

Krockauer R, Prates C (2018) Ship-launched radiosondes plug weather data gap. ECMWF Newsl 157 (Autumn 2018). https://ecmwf.net/en/newsletter/157/news/ship-launched-radiosondes-plug-weather-data-gap

Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201. https://doi.org/10.3402/tellusa.V56I3.14413

Larouche S, Sarrazin R (2013) Impact of radiosonde balloon drift on numerical weather prediction and verification. Weather Forecast 28:772–782. https://doi.org/10.1175/WAF-D-12-00114.1

Lawrence H, Bormann N, Sandu I, Day J, Farnan J, Bauer P (2019) Use and impact of Arctic observations in the ECMWF Numerical Weather Prediction system. Q J R Meteorol Soc 145:3432–3454. https://doi.org/10.1002/qj.3628

Lehtinen R, Survo P, Jauhiainen H (2016) Comparison of two radiosonde pressure measurements: pressure sensor vs. GPS-derived pressure. In: 18th symposium on meteorological observation and instrumentation, New Orleans, Louisiana, 10–14 Jan 2016. https://ams.confex.com/ams/96Annual/webprogram/Manuscript/Paper285261/11_1_Comparison_of_Two_Radiosonde_Pressure_Measurements.pdf

Lenschow DH (1986) Aircraft measurements in the boundary layer. In: Lenschow DH (ed) Probing the atmospheric boundary layer. American Meteorological Society, Boston, pp 39–55

Liu Y, Tang N (2014) Humidity sensor failure: a problem that should not be neglected. Atmos Meas Tech 7:3909–3916. https://doi.org/10.5194/amt-7-3909-2014

Lopez P (2011) Direct 4D-Var assimilation of NCEP stage IV radar and gauge precipitation data at ECMWF. Mon Weather Rev 139:2098–2116. https://doi.org/10.1175/2010MWR3565.1

Madonna F, Kivi R, Dupont J-C, Ingleby B, Fujiwara M, Romanens G, Hernandez M, Calbet X, Rosoldi M, Giunta A, Karppinen T, Iwabuchi M, Hoshino S, von Rohden C, Thorne PW (2020) Use of automatic radiosonde launchers to measure temperature and humidity profiles from the GRUAN perspective. Atmos Meas Tech 13:3621–3649. https://doi.org/10.5194/amt-13-3621-2020

Marks AM (1940) Reminiscences on "A P O B" flights and fliers. Bull Am Meteorol Soc 21:18–22. https://doi.org/10.1175/1520-0477-21.1.18

Milan M, Macpherson B, Tubbs R, Dow G, Inverarity G, Mittermaier M, Halloran G, Kelly G, Li D, Maycock A, Payne T, Piccolo C, Stewart L, Wlasak M (2019) Hourly 4D-Var in the Met Office UKV operational forecast model. Q J R Meteorol Soc https://doi.org/10.1002/qj.3737

Miloshevich LM, Paukkunen A, Vömel H, Oltmans SJ (2004) Development and validation of a time-lag correction for Vaisala radiosonde humidity measurements. J Atmos Ocean Technol 21:1305–1327. https://doi.org/10.1175/1520-0426(2004)021%3c1305:DAVOAT%3e2.0.CO;2

Mitchell CL, Wexler H (1941). How the daily forecast is made. In: Climate and man: yearbook of agriculture. U.S. Department of Agriculture, pp 579–598. https://naldc.nal.usda.gov/download/IND43893790/PDF

Moninger WR, Benjamin SG, Jamison BD, Schlatter TW, Smith TL, Szoke EJ (2010) Evaluation of regional aircraft observations using TAMDAR. Weather Forecast 25:627–645. https://doi.org/10.1175/2009WAF2222321.1

Moninger WR, Mamrosh RD, Pauley PM (2003) Automated meteorological reports from commercial aircraft. Bull Am Meteorol Soc 84:203–216. https://doi.org/10.1175/BAMS-84-2-203

Murphy P (2019) IWXXM status and operational implementation. Presented at the Air Transportation Information Exchange Conference (ATIEC) 2019, Tysons Corner, VA, 23–24 Sept 2019. Presentation available at https://www.faa.gov/air_traffic/flight_info/aeronav/atiec/media/Presentations/Day%201%20AM%20003%20Pat%20Murphy%20IWXXM.pdf

Nash J (2015) Measurement of upper-air pressure, temperature and humidity. Instruments and Observing Methods Report No. 121, WMO, 87 pp. https://library.wmo.int/doc_num.php?explnum_id=7366

Nash J, Oakley T, Vömel H, Wei L (2011) WMO intercomparison of high quality radiosonde systems (Yangjiang, China, 2010). Instruments and Observing Methods Report No. 107, WMO, 238 pp. https://library.wmo.int/doc_num.php?explnum_id=9467

National Research Council (2012) The national weather service modernization and associated restructuring: a retrospective assessment. The National Academies Press, Washington, DC, 120 pp. https://doi.org/10.17226/13216

Nipen TN, Seierstad IA, Lussana C, Kristiansen J, Hov Ø (2020) Adopting citizen observations in operational weather prediction. Bull Am Meteorol Soc 101:E43–E57. https://doi.org/10.1175/BAMS-D-18-0237.1

NWS (2010a) National weather service observing handbook no. 1: marine surface weather observations. U.S. Department of Commerce, 150 pp. https://www.vos.noaa.gov/ObsHB-508/ObservingHandbook1_2010_508_compliant.pdf.

NWS (2010b) Rawinsonde observations. National Weather Service Manual 10-1401, 208 pp. https://www.nws.noaa.gov/directives/sym/pd01014001curr.pdf

NWS (2013) Snow measurement guidelines for national weather service surface observing programs. U.S. Department of Commerce, Silver Spring, MD, 14 pp. https://www.weather.gov/media/coop/Snow_Measurement_Guidelines-2014.pdf

OFCM (1997) Federal meteorological handbook no. 3. Office of the Federal Coordinator for Meteorological Services and Supporting Research, FCM-H3-1997, 191 pp. https://www.icams-portal.gov/resources/ofcm/fmh/FMH3/00-entire-FMH3.pdf

OFCM (2020a) National hurricane operations plan. Office of the Federal Coordinator for Meteorological Services and Supporting Research, FCM-P12-2020, 178 pp. https://www.icams-portal.gov/resources/ofcm/nhop/2020_nhop.pdf

OFCM (2020b) U.S. federal meteorological data management handbook: a guide to standards and best practices. Office of the Federal Coordinator for Meteorological Servides and Supporting Research, FCM-H13-2020, 53 pp. https://www.icams-portal.gov/resources/ofcm/fmh/fmdmp/fmdmp.pdf

Pauley PM (1998) An example of uncertainty in sea level pressure reduction. Weather Forecast 13:833–850. https://doi.org/10.1175/1520-0434(1998)013%3c0833:AEOUIS%3e2.0.CO;2

Pauley PM (2002) Telling North from South: an example of an error in automated aircraft data. Weather Forecast 17:334–336. https://doi.org/10.1175/1520-0434(2002)017%3c0334:TNFSAE%3e2.0.CO;2

Pauley P, Anderson A, Ballish B, Benichou H, Huang X-Y, Ingleby B, Isaksen L, Jacobs N, Kleinert T, Moninger W, Parrett C, Taylor S, Verner G, Zaitseva Y (2014) Quality control for aircraft data in numerical weather prediction systems. Invited oral presentation, 2014 NOAA aircraft data workshop, Annapolis, MD, 29–30 Apr 2014. https://doi.org/10.13140/RG.2.2.35860.73608

Pauley P, Baker N (2014) Assimilating MDCRS humidity in NAVGEM. Poster presented at the 2014 NOAA aircraft data workshop, Annapolis, MD, 29–30 Apr 2014. https://doi.org/10.13140/RG.2.2.19083.52008

Petersen RA (2016) On the impact and benefits of AMDAR observations in operational forecasting—part I: a review of the impact of automated aircraft wind and temperature reports. Bull Am Meteorol Soc 97:585–602. https://doi.org/10.1175/BAMS-D-14-00055.1

Petersen RA, Cronce L, Mamrosh R, Baker R, Pauley P (2016) On the impact and future benefits of AMDAR observations in operational forecasting: part II: water vapor observations. Bull Am Meteorol Soc 97:2117–2133. https://doi.org/10.1175/BAMS-D-14-00211.1

Pettifer R (2009) From observations to forecasts—part 2. The development of *in situ* upper air measurements. Weather 64:302–308. https://doi.org/10.1002/wea.484

Poli P, Moll P, Rabier F, Desroziers G, Chapnik B, Berre L, Healy SB, Andersson E, El Guelai F-Z (2007) Forecast impact studies of zenith total delay data from European near real-time GPS stations in Météo France 4DVAR. J Geophys Res 112:D06114. https://doi.org/10.1029/2006JD007430

Poli P et al (2016) ERA-20C: an atmospheric reanalysis of the twentieth century. J Clim 29:4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1

Rabier F et al (2010) The Concordiasi project in Antarctica. Bull Am Meteorol Soc 91:69–86. https://doi.org/10.1175/2009BAMS2764.1

Ralph FM et al (2017) Dropsonde observations of total integrated water vapor transport within North Pacific atmospheric rivers. J Hydrometeorol 18:2577–2596. https://doi.org/10.1175/JHM-D-17-0036.1

Redelsperger J, Thorncroft CD, Diedhiou A, Lebel T, Parker DJ, Polcher J (2006) African monsoon multidisciplinary analysis: an international research project and field campaign. Bull Am Meteorol Soc 87:1739–1746. https://doi.org/10.1175/BAMS-87-12-1739

Ryan WA, Doesken NJ, Fassnacht SR (2008) Evaluation of ultrasonic snow depth sensors for U.S. snow measurements. J Atmos Ocean Technol 25:667–684. https://doi.org/10.1175/2007JTECHA947.1

Seidel DJ, Sun B, Pettey M, Reale A (2011) Global radiosonde balloon drift statistics. J Geophys Res Atmos 116:D07102. https://doi.org/10.1029/2010JD014891

Shimizu K, Hasebe F (2010) Fast-response high-resolution temperature sonde aimed at contamination-free profile observations. Atmos Meas Tech 3:1673–1681. https://doi.org/10.5194/amt-3-1673-2010

Sondij J (2020) Status of the European Meteorological Aircraft Derived Data Center (EMADDC). Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1194/attachments/440/795/ABO-WS-Sondij.pdf

Sonntag D (1994) Advancements in the field of hygrometry. Z Meteorol 3(2):51–66

Sparkman JK, Giraytys J, Smidt GJ (1981) GARP topics: ASDAR: a FGGE real-time data collection system. Bull Am Meteorol Soc 62:394–400. https://doi.org/10.1175/1520-0477-62.3.381

Stone EK, Kitchen M (2015) Introducing an approach for extracting temperature from aircraft GNSS and pressure altitude reports in ADS-B messages. J Atmos Ocean Technol 32:736–743. https://doi.org/10.1175/JTECH-D-14-00192.1

Storer LN, Gill PG, Williams PD (2019) Multi-model ensemble predictions of aviation turbulence. Meteorol Appl 26:416–428. https://doi.org/10.1002/met.1772

Strajnar B (2012) Validation of Mode-S Meteorological Routine Air Report aircraft observations. J Geophys Res 117:D23110. https://doi.org/10.1029/2012JD018315

Stringer S (2020) Introduction to global aircraft based observations. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1193/attachments/439/794/ABO-WS-Stringer.pdf

Tavolato C, Isaksen L (2015) On the use of a Huber norm for observation quality control in the ECMWF 4D-Var. Q J R Meteorol Soc 141:1514–1527. https://doi.org/10.1002/qj.2440

Taylor S (2020) Turbulence monitoring. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1213/attachments/452/807/ABO-WS-Taylor.pdf

Thorne PW et al (2017) Toward an integrated set of surface meteorological observations for climate science and applications. Bull Am Meteorol Soc 98:2689–2702. https://doi.org/10.1175/BAMS-D-16-0165.1

Tong M et al (2018) Impact of assimilating aircraft reconnaissance observations on tropical cyclone initialization and prediction using operational HWRF and GSI ensemble-variational hybrid data assimilation. Mon Weather Rev 146:4155–4177. https://doi.org/10.1175/MWR-D-17-0380.1

U.S. Weather Bureau (1949) Radiosonde and rawinsonde code (in international form as adopted by the International Meteorological Organization at Paris in 1946 and revised at Washington in 1947). USWB

Vaisala (2017) Vaisala radiosonde RS41 measurement performance. White Paper, 28 pp. https://www.vaisala.com/sites/default/files/documents/WEA-MET-RS41-Performance-White-paper-B211356EN-B-LOW-v3.pdf

Vance AK, Abel SJ, Cotton RJ, Woolley AM (2015) Performance of WVSS-II hygrometers on the FAAM research aircraft. Atmos Meas Tech 8:1617–1625. https://doi.org/10.5194/amt-8-1617-2015

Vasiljevic D (2006) Handling biases in surface pressure (Ps) observations in data assimilation. In: ECMWF/EUMETSAT NWP-SAF workshop on bias estimation in data assimilation, 11 pp, 8–11 Nov 2005. https://www.ecmwf.int/node/12904

Vömel H et al (2007) Radiation dry bias of the Vaisala RS92 humidity sensor. J Atmos Ocean Technol 24:953–963. https://doi.org/10.1175/JTECH2019.1

Waller JA, Dance SL, Nichols NK (2016) Theoretical insight into diagnosing observation-error correlations using observation-minus-background and observation-minus-analysis statistics. Q J R Meteorol Soc 142:418–431. https://doi.org/10.1002/qj.2661

Wang J, Carlson DJ, Parsons DB, Hock TF, Lauritsen D, Cole HL, Beierle K, Chamberlain E (2003) Performance of operational radiosonde humidity sensors in direct comparison with a chilled mirror dew-point hygrometer and its climate implication. Geophys Res Lett 30:1860. https://doi.org/10.1029/2003GL016985

Wang J, Zhang L, Dai A, Immler F, Sommer M, Vömel H (2013) Radiation dry bias correction of Vaisala RS92 humidity data and its impacts on historical radiosonde data. J Atmos Ocean Technol 30:197–214. https://doi.org/10.1175/JTECH-D-12-00113.1

WBAN (1957) Manual of radiosonde observations, circular P. U.S. Government Printing Office. https://www.google.com/books/edition/_/Jc5RuQEACAAJ?hl=en&gbpv=0

WBAN (1963) Manual for radiosonde code, 2nd edn. U.S. Government Printing Office. https://www.google.com/books/edition/_/LeWEAAAAIAAJ?hl=en&gbpv=0

Wexler A (1976) Vapor pressure formulation for water in range 0 to 100 °C. A revision. J Res Natl Bur Stand A Phys Chem 80A(5 and 6):775–785

Wexler A (1977) Vapor pressure formulation for ice. J Res Natl Bur Stand A Phys Chem 81A(1):5–20

Wick GA, Hock TF, Neiman PJ, Vömel H, Black ML, Spackman JR (2018) The NCAR–NOAA global hawk dropsonde system. J Atmos Ocean Technol 35:1585–1604. https://doi.org/10.1175/JTECH-D-17-0225.1

Williard A (2020) ADS-C and AIREPs. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1204/attachments/445/800/ABO-WS-Williard.pdf

WMO (2003) Aircraft Meteorological Data Relay (AMDAR) reference manual. WMO-No. 958, 80 pp. https://library.wmo.int/doc_num.php?explnum_id=9026

WMO (2012a) Observation stations. WMO-No. 9, 593 pp. https://library.wmo.int/doc_num.php?explnum_id=9896

WMO (2012b) The EUMETNET AMDAR optimization program (E-ADOS)—a component of the E-AMDAR programme. WMO AMDAR Panel Newsletter, vol 4, Oct 2012. https://sites.google.com/a/wmo.int/amdar-news-and-events/volume-4-october-2012/theeumetnetamdaroptimizationprograme-ados

WMO (2014) AMDAR onboard software functional requirements specification. WMO-No. 958, 97 pp. https://library.wmo.int/doc_num.php?explnum_id=7360

WMO (2017) Guide to aircraft-based observations. WMO-No. 1200, 141 pp. https://library.wmo.int/doc_num.php?explnum_id=4120

WMO (2018a) Guide to meteorological instruments and methods of observation, vol I—measurement of meteorological variables. WMO-No. 8, 548 pp. https://library.wmo.int/doc_num.php?explnum_id=10179

WMO (2018b) Guide to instruments and methods of observation, vol III—observing systems. WMO-No. 8, 426 pp. https://library.wmo.int/doc_num.php?explnum_id=9872

WMO (2019a) Manual on codes, vol. I.1, part A—alphanumeric codes. WMO-No. 306, 480 pp. https://library.wmo.int/doc_num.php?explnum_id=10235

WMO (2019b) Manual on codes, vol. I.2, part B—binary codes and part C—common features to binary and alphanumeric codes. WMO-No. 306, 1180 pp. https://library.wmo.int/doc_num.php?explnum_id=10310

WMO (2019c) Manual on the WMO integrated global observing system, annex VII to the WMO technical regulations. WMO-No. 1160, 152 pp. https://library.wmo.int/doc_num.php?explnum_id=10145

WMO (2019d) Tests, Comparisons and operational performance of the Water Vapor Sensing Systems (WVSS-II). Instruments and Observing Methods Report No. 133, 45 pp. https://library.wmo.int/index.php?lvl=notice_display&id=21509#.X29_63lKguU

WMO (2019e) Report on the workshop on use of Unmanned Aerial Vehicles (UAV) for operational meteorology, 45 pp. https://library.wmo.int/doc_num.php?explnum_id=9951

WMO (2020a) Global Telecommunication System (GTS). https://community.wmo.int/activity-areas/global-telecommunication-system-gts

WMO (2020b) Manual on codes, vol. I.3, part D—representations derived from data models. WMO No. 306, 272 pp. https://library.wmo.int/doc_num.php?explnum_id=10285

Zaitseva NA (1993) Historical developments in radiosonde systems in the Former Soviet Union. Bull Am Meteorol Soc 74:1893–1900. https://doi.org/10.1175/1520-0477(1993)074%3C1893:HDIRSI%3E2.0.CO;2

Zaitseva Y (2020) Global Aircraft Based Observation (ABO) data monitoring. Presented at the ECMWF/EUMETNET workshop on aircraft weather observations and their use, ECMWF, Reading, UK, 12–13 Feb 2020. https://events.ecmwf.int/event/168/contributions/1202/attachments/444/799/ABO-WS-Zaitseva.pdf

Zhao Q, Cook J, Xu Q, Harasti PR (2008) Improving short-term storm predictions by assimilating both radar radial-wind and reflectivity observations. Weather Forecast 23:373–391. https://doi.org/10.1175/2007WAF2007038.1

Zhu Y, Derber JC, Purser RJ, Ballish BA, Whiting J (2015) Variational correction of aircraft temperature bias in the NCEP's GSI analysis system. Mon Weather Rev 143:3774–3803. https://doi.org/10.1175/MWR-D-14-00235.1

GNSS-RO Sounding in the Troposphere and Stratosphere

Benjamin Ruston, Neill Bowler, Sean Healy, and Hui Shao

Abstract The use of Global Navigation Satellite System Radio Occultations (GNSS-RO) for vertical sounding of temperature and moisture in the atmospheric column has become a standard practice of many numerical weather prediction (NWP) centers. The introduction of this observation has seen broad positive impact on analyses and forecasts. On longer timescales the impact of the introduction of this data type in re-analyses can be clearly seen. Further, the observations can be used without bias correction and the consistency between sensors is very good allowing these observation to serve as anchoring observations. This is particularly helpful to constrain the bias-correction applied to satellite radiances. In the following chapter we explore the fundamentals of the measurement, the derivation of the typical observation which is used in NWP, the assimilation methods and error assumptions which are used, and finally some conjecture on the direction to improve the use of the observations and what future measurement systems may look like.

B. Ruston (✉)
Naval Research Laboratory (NRL), Monterey, CA, USA
e-mail: ben.ruston@nrlmry.navy.mil

N. Bowler
Met Office, FitzRoy Road, Exeter EX1 3PB, UK
e-mail: neill.bowler@metoffice.gov.uk

S. Healy
European Centre for Medium Range Weather Forecasts (ECMWF), Reading, UK
e-mail: sean.healy@ecmwf.int

H. Shao
Joint Center for Satellite Data Assimilation (JCSDA)/University Corporation for Atmospheric Research (UCAR), College Park, MD, USA
e-mail: huishao@ucar.edu

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_13

1 Fundamentals of the Radio Occultation Measurement

The concept radio occultation measurements was pioneered in the planetary science community in 1960s and 1970s Kliore et al. (1965), Fjeldbo et al. (1971). The use of radio occultation techniques to measure the Earth's atmosphere was also discussed during that period, but the potential costs appeared prohibitive at the time Yunck et al. (2000). In the 1980s a team at the Jet Propulsion Laboratory (JPL) suggested making radio occultation measurements with the Global Positioning System (GPS). This led to the concept of Global Navigation Satellite Systems radio occultation (GNSS-RO), and the "proof of concept" GPS-MET mission in the mid 1990s Ware et al. (1996). A number research and operational missions followed, including CHAMP Wickert et al. (2001), COSMIC/FORMOSAT-3 Anthes et al. (2008), Metop GRAS Luntama (2008) and, more recently, COSMIC-2/FORMOSAT-7 Schreiner (2020). The earlier missions used the American GPS system, but more recent missions—like COSMIC-2/FORMOSAT-7—are also exploiting other signals such as the Russian GLONASS system.

The GNSS-RO technique at its core is based on the simple physics of refraction. It requires the measurement of a Doppler frequency shift of a transmitted signal by a GNSS receiver Melbourne et al. (1994); Kursinski et al. (1997). The measurement geometry is shown in Fig. 1. The transmitters are typically in a medium earth orbit (with orbital periods around 12 h) and the receivers typically in a low Earth orbit (LEO). The signals propagate from the GNSS satellite to the LEO, but the path is slightly curved as a result of refractive index gradients in both the ionosphere and neutral atmosphere. This curvature or bending of the ray path changes the Doppler shift of the signal, when compared with the Doppler shift that would have been measured for a straight line path between the satellites. Most of the bending occurs over a few hundred kilometers of the ray-path between the satellites, where the signal is closest to the Earth's surface. Because of the availability of precise orbital determination (POD) and stable clocks to high precision, down to femtoseconds, this Doppler shift is well measured and is at the core of the stability of the GNSS-RO observation. The relative motion between the transmitter and receiver provide the sounding through the atmosphere, producing either a rising or setting occultation as it views the transmitter. The transmitted signal is in the microwave spectrum, with commonly used GPS frequencies at $f_1 = 1.57542$ GHz and $f_2 = 1.2276$ GHz, referred to as the GPS L1 and L2 signals. The use of relatively long wavelengths (19.1 cm for L1; 24.4 cm for L2) allows the signals to pass through Earth's atmosphere with little interference from particles such as aerosols or clouds.

A typical occultation sounding of the neutral atmosphere will last about one or two minutes. Due to the satellite motions, a slice of the atmosphere is scanned, and the LEO satellite receives signals where the ray paths have different minimum distances to the surface, from zero up to approximately 100 km. The points where the ray paths have the minimum distance to the surface are commonly referred to as the tangent points. The resulting occultation profile has a relatively higher vertical resolution (about a few hundred meters, varying with heights), compared with other satellite

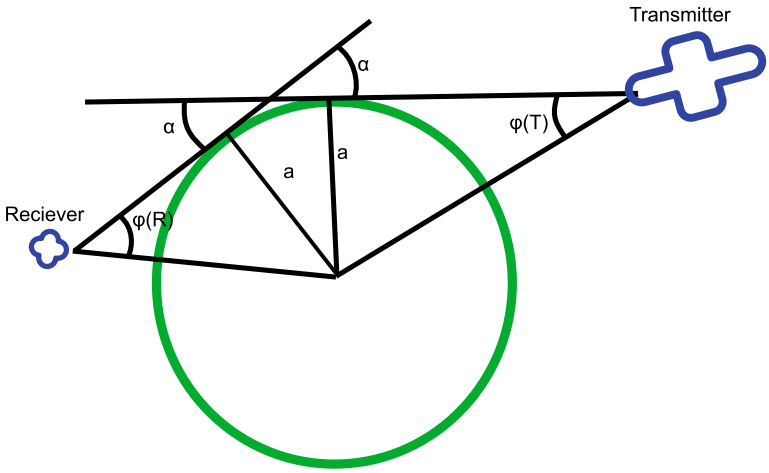

Fig. 1 Schematic of GNSS geometry for a medium Earth orbit transmitter and Low-Earth Orbit (LEO) receiver

measurements. Given the bending measurement reflects the integrated effects along the ray-path in the atmosphere, the horizontal resolution along the ray-path is rather coarse, typically a few hundred kilometers. However, the resolution perpendicular to the "occultation plane" (a plane defined by the positions of a LEO satellite and its occulted GPS satellite, and the center of the local curvature of the occultation) remains fairly high. The measurement from each occultation contains a slanted profile of atmospheric states at each of the tangent points. Due to the satellite motions, the tangent points for one occultation have horizontal shifts, these are as large as 1 degree from the surface to about 40 km (Fig. 2).

The proof of concept GPS-MET mission, led by the Universities Corporation for Atmospheric Research (UCAR), was the first spacebourne sensor to demonstrate the GNSS-RO technique Ware et al. (1996). A series of studies examined data from this early mission and helped to establish the ability of radio occultation to be used for retrievals of temperature Kursinski et al. (1996); Rocken et al. (1997) and geopotential height Leroy (1997) using retrieval techniques developed by planetary scientists. Subsequently information content studies were run, based on variational retieval techniques more closely related to how the measurements would be assimilated into NWP systems. These showed that the GNSS-RO measurements complemented the information provided by high resolution interferometers Collard and Healy (2003), suggesting that these measurements would provide useful information for NWP applications, particularly in the upper troposphere and stratosphere. The key characterisistics that make GNSS-RO measurement an important component of the global observing system are that they can be used without bias correction Eyre (2016), and they have excellent vertical resolution as a result of the limb geometry.

Routine operational assimilation of GNSS-RO measurements into NWP systems began in 2006, and most centres currently assimilate either refractivity

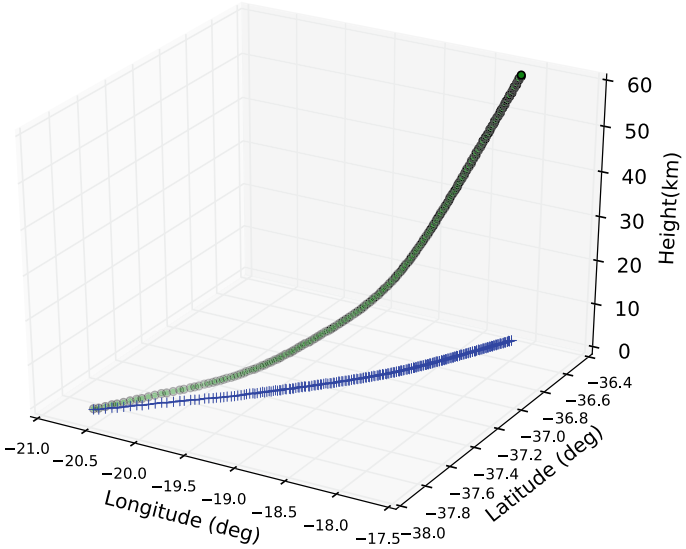

Fig. 2 Tangent point positions (black) as function of impact height (km) and their horizontal projections (blue) within a latitude-longitude plane. The example data are from the COSMIC-2 mission, measured at 11:13 UTC 11 October 2019

or—increasingly—bending angles. The largest impact has generally been seen for upper-tropospheric and stratospheric temperatures, but there are some indications that the latest missions are also having a significant impact lower in the troposphere.

2 Typical Use of GNSS-RO in NWP

2.1 GNSS-RO Processing

It is important to understand the processing of GNSS-RO observations and their information content prior to developing an assimilation strategy. In particular, it should be recognised that GNSS-RO observations are not direct measurements of geophysical quantities such as temperature, pressure and water vapor, even if these quantities can be retrieved from GNSS-RO measurements with a suitable processing system Kursinski et al. (1997), Hajj et al. (2002).

Figure 3 shows a schematic flowchart of the key steps in a GNSS-RO geometrical optics (GO) processing system. The geometrical optics processing assumes that only a single ray arrives at the receiver at a given time, but this is often not the case for ray paths in the troposphere. More than one ray arriving at the receiver is known as "atmospheric multipath". This problem can be mitigated with "wave optics" processing

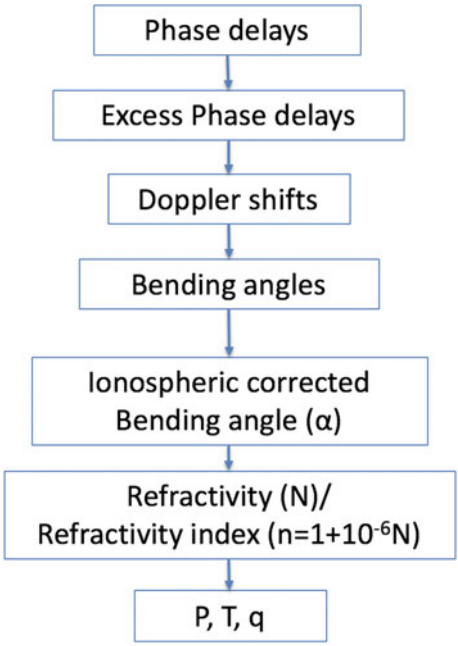

Fig. 3 Schematic flowchart of GNSS-RO measurements and retrievals. P, T and q refer to the air pressure, temperature and specific humidity respectively

techniques, which are essentially a coordinate transforms Gorbunov and Lauritsen (2004); Jensen et al. (2003, 2004) designed to recover the single ray path picture.

Briefly, the GO processing starts with "raw" measurements of the phase delay of radio signals received at two GNSS frequencies during an occultation. Following various calibration and correction procedures Hajj et al. (2002), the "excess" phase delays are computed by subtracting the phase delays expected for a straightline path in a vacuum. The time derivative of the excess phase delays the provide a timeseries of Doppler shift values at both transmitted freqiencies Cucurull et al. (2015).

The total bending angle for each GNSS signal, i, α_i, as a function of the impact parameter, a, can then be derived from the Doppler shift values, by *assuming* that the impact parameter, a, is a constant along the ray-path. This assumption, known as spherical symmetry, implies that horizontal refractive index gradients are zero in the plane of the ray-path, meaning that the refractive index, n, is assumed to be only a function of a height variable, $n(r)$.

The ionospheric contribution to the ray bending can be removed—or corrected—by taking a linear combination of the bending angles at the two GNSS frequencies Vorobev and Krasilnikova (1994). For the the GPS L1 and L2 signals this correction can be writtens as,

$$\alpha(a) = \alpha_1(a) + \frac{f_2^2}{f_1^2 - f_2^2}\left(\alpha_1(a) - \alpha_2(a)\right) \tag{1}$$

where α_1 and α_2 are the bending angles for L1 and L2 signals, interpolated to a common impact parameter value, a, and f_1 and f_2 are the signal frequencies.

The raw observations, of Doppler shift and time delay, are typically measured at high temporal resolution. This corresponds with a very high vertical sampling, much higher than the vertical resolution of NWP models. Therefore, some vertical smoothing is applied to the bending angles and these are sub-sampled to an appropriate resolution in the vertical. This smoothing reduces the noise present in the measurements at the expense of introducing vertical correlations in the errors of the observations. The amount of smoothing applied varies between processing centres, and there is not currently an established procedure for determining the optimal level of smoothing in order to maximise NWP performance.

A key approximation which is often employed to use radio occultation profiles of $\alpha(a)$ is that of spherical symmetry. This assumption can break down, particularly near the surface in the presence of horizontal gradients of humidity. However, making the assumption of spherical symmetry enables the application of an Abel transform pair, relating $\alpha(a)$ to the refractive index, n, as function of a height variable Kursinski et al. (1997). More specifically, the bending angle integral can be written as:

$$\alpha(a) = -2a \int_a^\infty \frac{\frac{d(\ln n)}{dx}}{(x^2 - a^2)^{\frac{1}{2}}} dx \tag{2}$$

where $x = nr$, with r being the radius of a point on the ray path. Conversely, the profile of refractive index is then written as function of $\alpha(a)$,

$$n(x) = \exp\left[\frac{1}{\pi} \int_x^\infty \frac{\alpha(a)}{(a^2 - x^2)^{\frac{1}{2}}} da\right] \tag{3}$$

noting that the upper limit of this integral is ∞, implying some extrapolation of the observed bending angle profile, since it usually stops around 80 km. The extrapolation and smoothing of the bending angles prior to the Abel transform is combined in a processing step known as statistical optimization Healy (2001).

In general, the atmospheric refractive index can be written as function of geophysical quantities via Bean and Dutton (1968); Hajj et al. (2002):

$$N = (n-1)10^6 = a_1 \frac{P}{T} + a_2 \frac{P_w}{T^2} + a_e \frac{n_e}{f^2} + a_w W_w + a_i W_i + O(f^{-3}) \tag{4}$$

where N is known as the refractivity. The quantities on the right hand side are: P_w=$P/(0.622 + 0.378q)$ is the water vapor partial pressure, where P is the air pressure, T is the temperature, q is the specific humidity; n_e is the electron density and f is the signal frequency; W_w and W_i are the liquid water and ice contents; a_1, a_2, a_e, a_w, and a_i are empirical coefficients for each term, respectively. Therefore, refractivity has contributions from four main sources Kursinski et al. (1997), the dry

neutral atmosphere, water vapor, free electrons in the ionosphere and scattering by liquid water and ice particulates.

The ionospheric contribution to N can usually be ignored for NWP applications, because it should be removed with Eq. (1) Vorobev and Krasilnikova (1994). In addition, one can also neglect the scattering by particulates, as for the transmitter frequencies used and for the majority of atmospheric suspensions of water and ice, because their contribution is small compared with other terms in the equation Kursinski et al. (1997); Solheim et al. (1999). This then leaves the dry neutral atmosphere and water vapor terms for the atmosphere typically below about 60 km. This simplifies to an equation using empirically derived constants for a_1 and a_2 Bean and Dutton (1968):

$$N = (n-1)10^6 = 77.6\frac{P}{T} + 3.73 \cdot 10^5 \frac{P_w}{T^2} \tag{5}$$

We note that more sophisticated expressions for refractivity are also used in GNSS-RO data assimilation applications Aparicio and Laroche (2011).

In the stratosphere, and other regions where the contribution of water vapor to the refractivity is small, the refractivity is proportional to density ($N \simeq 77.6\frac{P}{T} \propto \rho$). Given refractivity (or density) as a function of height, the hydrostatic equation can be integrated downwards to give pressure as function of height. The temperature can then be derived by applying the ideal gas law. The temperature retrieved by assuming the humidity can be neglected is often called the "dry temperature" in GNSS-RO.

More generally, when the moisture cannot be neglected – for example in the lower troposphere—the retrieval of T, P and q from N is an under-determined problem and therefore *a priori* information (e.g., an NWP forecast state) is required to solve the geophysical retrieval.

From a data assimilation perspective, the "dry temperature retrievals" outlined here provide a good framework for understanding the measurement technique and understanding assimilation options, but these retrievals are not suitable for direct assimilation into an NWP system.

3 Assimilation Methods and Error Statistic Assumptions

The aim of data assimilation is to combine a forecast of the atmospheric state, $\mathbf{x_b}$, with new observations, $\mathbf{y}$ in a statistically optimal way Lorenc (1986). In variational assimilation methods this involves minimizing a cost function of the form,

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x_b})^{\mathbf{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x_b}) + (\mathbf{y} - H(\mathbf{x}))^{\mathbf{T}}\mathbf{R}^{-1}(\mathbf{y} - H(\mathbf{x})) \tag{6}$$

with respect to $\mathbf{x}$, where $\mathbf{B}$ is the background-error covariance matrix, $\mathbf{R}$ is the observation-error covariance matrix and H is the forward operator, mapping the meteorological state information, $\mathbf{x}$, to observation space. The analysis, $\mathbf{x_a}$, is the

state which minimizes the cost function, and it should be consistent with both $\mathbf{x_b}$ and $\mathbf{y}$, to within their expected error statistics.

Given this general framework, it is clear that the assimilation requires both a good estimate of the GNSS-RO error statistics ($\mathbf{R}$), and an accurate forward operator (H) to map the meteorological state to observation space. These two requirements are not completely independent because the $\mathbf{R}$ matrix is usually assumed to include contributions from both measurements errors, $\mathbf{E}$, and forward model errors, $\mathbf{F}$, so $\mathbf{R} = \mathbf{E} + \mathbf{F}$. The choice of forward model, H, usually represents a trade-off between making $\mathbf{F}$ as small as possible, by including as much physics as possible to retreive all of the available information, versus the computational cost and the timeliness requirements of an operational NWP system.

In addition, it is now generally recognised in NWP that additional processing steps tend to complicate the observation error statistics and correlations. For example, in the context of GNSS-RO, error simulation studies show that vertical error correlations for refractivity tend to be broader than bending angles.

The additional processing steps can also introduce *a priori* information less accurate than the NWP forecast. For example, we do not want to assimilate information provided by a mean state climatology into NWP system. For this reason, it is usually preferable to assimilate variables close to the raw measurement for all observation types. As far as we are aware, no NWP centres assimilate GNSS-RO retrievals of temperature and humidity.

3.1 Forward Operators: H (x)

Eyre (1994) was the first to discuss the GNSS-RO assimilation options in detail. The various processing levels discussed by Eyre are the same as those shown in Fig. 3. At present, most of the operational centres use either the ionospherically-corrected bending angle (Eq. 1) or refractivity (Eq. 3) for data assimilation.

Refractivity Assimilation

When considering assimilation of refractivity, one can use the formulation in Eq (5) as the forward operator to compute the atmospheric refractivity from the model state (T, P, and q) as a function of geopotential height. This is a common approach which was an option taken by many data assimilation systems, e.g., the Gridpoint Statistical Interpolation (GSI) system Cucurull et al. (2007) and the Weather Research and Forecast model Data Assimilation (WRFDA) system. It is relatively straightforward to account for "tangent point drift" in the horizontal plane when assimilating refractivity Cucurull et al. (2007). However, the forward model simulates the refractivity at the horizontal location of the tangent point, while the retrieved refractivity will be related to a quantity which is horizontally averaged in the occultation plane. There are "non-local"(two-dimensional) refractivity and phase operators Syndergaard et al. (2005); Sokolovskiy et al. (2005); Shao et al. (2009) that try account for this by simulating horizontally averaged refractivity values in the two-dimensional occultation plane,

but these are not currently adopted by global operational NWP centres. However, they have been tested in limited-area models Chen et al. (2009).

When considering refractivity assimilation, it is also worth noting the "statistical opimization" step to smooth and extrapolate the bending angles prior performing the Abel transform, Eq.(3). This introduces *a priori* information and for that reason it is advisable not to assimilate the refractivity values above 40 km.

Bending Angle Assimilation

There are several ways to assimilate the GNSS-RO bending angle profiles with varying degrees of complexity. The simplest is a one-dimensional (1D) approach that computes the bending angle integral, Eq (2), at a single representative location Healy (2006), ignoring the actual two-dimensional geometry. Similar to refractivity assimilation, it is straightforward to introduce tangent point drift Poli et al. (2009); Cucurull (2012). The 1D bending angle assimilation technique is an approach broadly adopted by the operational and research communities, given its simplicity and computational efficiency to implement. Rennie (2010) and Cucurull et al. (2013) have both presented experiments comparing such an operator with a refractivity operator in their systems, before switching to the bending angle approach for operational use.

The most complex and accurate GO approach for bending angle assimilation is to simulate bending angle via a ray-tracing method using three-dimensional (3D) refractivity information provided by the NWP forecast. It solves a ray-trajectory equation, which governs the behavior of the radio signal wave under the influence of a refractivity field. The bending angle can be computed by following the ray path. When expressed in a Cartesian coordinate, the general ray-path equation is written as Kravtsov and Orlov (1990):

$$\frac{d^2\mathbf{r}}{ds^2} = n\nabla n \tag{7}$$

where $\mathbf{r}$ is the position vector pointing from the Earth's center to the ray trajectory in the Cartesian coordinate, s is defined by $ds = dl/n$, where l is the length of the ray path and ds is the differential displacement along the ray path. A commonly used form of the ray equation is a set of first-order differential equations:

$$\frac{d\mathbf{r}}{ds} = \mathbf{t} \tag{8}$$

$$\frac{d\mathbf{t}}{ds} = n\nabla n \tag{9}$$

where $\mathbf{t}$ defines the direction of the ray. The ray-trajectory equation can be numerically solved for any given 3D field of n, once either initial conditions (initial position and direction) or boundary conditions (two end point positions) of the ray are prescribed. The boundary problem may require a ray-shooting method, which is expensive computationally and is subject to multiple solutions due to multi-path propagations Zou et al. (1999). Therefore, it is typically solved as an initial value problem. Over the past 20 years, variants of bending angle ray-tracing operators

have been proposed (e.g. Zou et al. (2002); Wee et al. (2010)) but they are not used operationally at the moment.

The 3D ray-tracing method can be simplified by solving the equation in multiple 2D "occultation planes", defined geometrically by the positions of the GNSS and LEO satellites and the local curvature center. Such a bending angle operator is often denoted as a 2D bending angle operator. For a two-dimensional approximation in polar coordinates, ignoring refractive index gradients perpendicular to the GNSS-RO occultation plane, the ray-path equations can be written as Rodgers (2000):

$$\frac{dr}{ds} = cos\phi \tag{10}$$

$$\frac{d\theta}{ds} = \frac{sin\phi}{r} \tag{11}$$

$$\frac{d(\theta + \phi)}{ds} = \frac{-sin\phi}{n}\left(\frac{\partial n}{\partial r}\right)_{\theta} + \frac{cos\phi}{nr}\left(\frac{\partial n}{\partial \theta}\right)_{r} \tag{12}$$

where r and ϕ are the radius and the polar angle at an arbitrary point on the ray path, respectively, θ is the local zenith angle of the ray path.

Implementation of a 2D bending angle operator in an NWP system is more challenging than the 1D operators, since it requires information from multiple horizontal locations of the NWP model state along the specific ray path for each bending angle computation. One such 2D bending angle operator is described by Healy et al. (2007). It is also critical for operational implementation to develop an efficient parallel computing scheme for computational efficiency Healy (2014). Currently, the use of 2D bending angle operators by operational NWP centres is limited Healy (2014). With the advance of computational resources and techniques, 2D or even 3D bending angle assimilation—or other advanced approaches – will be more feasible for operational implementation and research studies.

3.2 Error Statistic Assumptions

As noted above, understanding the measurement errors statistics and using a realistic **R** matrix when assimilating the GNSS-RO measurements is also a key requirement for successfully exploiting these data.

Due to the high precision of the clocks used within the GNSS satellites it is possible for the raw measurements within GNSS-RO to be very precise. Therefore, many of the errors and uncertainties associated with GNSS-RO observations are related to the processing of the measurements and their forward modelling in NWP. These were dealt with in considerable detail in Kursinski et al. (1997) so a brief description will be given here.

At the very highest levels in the neutral atmosphere, the bending induced by the ionosphere can play an important role. The techniques for removing these effects, described above, are not perfect and the residual errors will affect the measured bending angle at upper levels. In the troposphere many of the errors and uncertainties are connected with the variation of water vapour. The calculation of a bending angle from the raw measurement made by the satellite entail an assumption of spherical symmetry (Sect. 2.1). This situation is compounded by many observation operators being one-dimensional which also requires the assumption of spherical symmetry. Although this assumption is reasonable in the stratosphere, it is a poor assumption in the troposphere, since water vapour often varies strongly on short horizontal scales.

These are just two of the main issues which affect GNSS-RO data. Since they affect data at high- and low-levels, GNSS-RO data generally have most impact in the low- to mid-stratosphere. That being said, the observations still play an important role at the other levels.

A commonly used method to estimate the uncertainties in observations is the method of Desroziers et al. (2005). With this method one calculates the cross-product between the innovations and the residuals as

$$E(\mathbf{d}_{\mathrm{o}}^{\mathrm{a}}(\mathbf{d}_{\mathrm{o}}^{\mathrm{b}})^{\mathrm{T}}) = \mathbf{R}_{\mathrm{est}} \tag{13}$$

where $\mathbf{d}_{\mathrm{o}}^{\mathrm{b}}$ is a vector of the differences between the observations and the NWP background forecast (known as the innovations) and $\mathbf{d}_{\mathrm{o}}^{\mathrm{a}}$ is a vector of the differences between the observations and the analysis (known as the residuals). This cross-product provides an estimate of the observation-error covariance matrix. If the data assimilation method used in the analysis is perfect (i.e. is provided with the correct error-covariance matrices and finds the globally optimal solution) then this estimate will correspond to the true observation-error covariance matrix. Thus this method was developed as a consistency check on the inputs provided to the data assimilation. There are other methods to estimate the observation uncertainties Hollingsworth and Lonnberg (1986); Scherllin-Pirscher (2011); Anthes and Rieckh (2018) which may be preferable in certain situations.

Examination of estimated observation uncertainties (the square root of the diagonal elements of $\mathbf{R}_{\mathrm{est}}$) diagnosed using the method of Desroziers indicates that they principally vary with a small number of different quantities Bowler (2020). The key quantities with which the uncertainties vary are the height of the observation, the receiving satellite and the latitude of the tangent point.

The variation of the estimated uncertainty with latitude for Metop-B is shown in Fig. 4. In this (and following figures) the estimated uncertainty has been normalised by the background forecast of the bending angle. This normalisation is chosen because the bending angle can vary by many orders of magnitude with height. In the troposphere the estimated uncertainties are much larger than elsewhere. The reason for this variation is well understood. The tropical troposphere has a high specific humidity. Since water vapour is often not well modelled by NWP systems, in part due to the small-scale variations that it displays, then there are large differences between the observation and the background forecast. In addition the one-dimensional for-

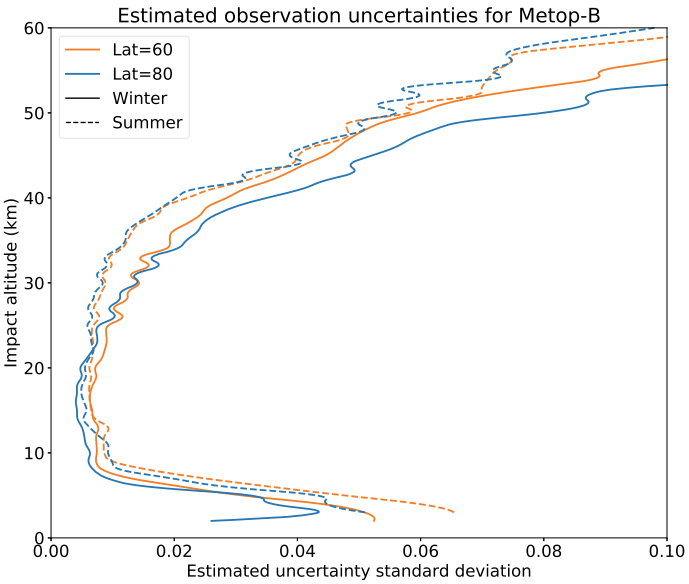

Fig. 4 Diagnosed uncertainties, normalised by the background bending angle, with results binned using a range of latitudes for Metop-B. Each latitude gives the centre of a 20 degree bin, and if the tangent-point of the observation falls within that bin, then the observation is assigned to the given latitude. The statistics for Figs. 4, 5 and 6 are calculated using assimilation statistics from the Met Office's NWP system from the month of January 2020

ward models that are often used (including in the calculation of Fig. 4) make strong assumptions about the spherical symmetry of the atmosphere. These background and forward-modelling errors lead to the observation being ascribed as having large uncertainties.

The variation with receiving satellite is illustrated in Fig. 5. These statistics are calculated from observations within 10 degrees of the equator, so that different latitudinal sampling does not complicate the graph. The statistics from the different satellites are often very similar, despite large differences in the hardware and processing software used. At high altitudes the observation uncertainty estimated for the Metop satellites is much smaller than for the other satellites. This is due to the receiver exhibiting a very low thermal noise and the use of an ultra-stable oscillator in the instrument (C Marquardt, Radio occultation team leader, EUMETSAT, personal communication, 2019).

It is also noticeable that the standard deviation estimated for FY-3C/D is large between around 20–25 altitude. 25km is the height at which the data processing switches from geometric optics to wave optics. For all GNSS-RO data a smoothing is applied to the data within a vertical profile in order to reduce the noise in the data (see Sect. 2.1). The amount of smoothing that is required differs between geometric and wave optics processing. This increase in the standard deviations occurs because the level of smoothing has not been well-matched between the types of processing.

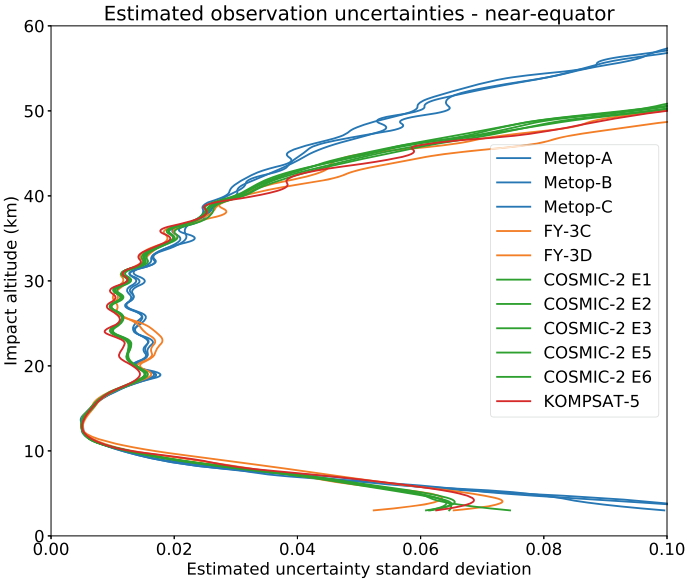

Fig. 5 Diagnosed observation uncertainties, normalised by the background bending angle, with results binned for different satellites. Statistics calculated observations within 10 degrees of the equator

The Metop satellites have the largest standard deviations between around 25–35. This is due to less vertical smoothing being used generally for these satellites.

From the above examples it can be noted that both the hardware used to receive the signals, as well as the software used to process those signals, plays a role in the quality of the observations.

One can also note that the estimated observation uncertainties vary with the season. Figure 6 shows a comparison between the estimated uncertainties in the Northern Hemisphere summer and winter for data from Metop-B. Larger uncertainties are diagnosed in the summer troposphere than in the winter troposphere. This is likely due to the warmer atmosphere containing more water vapour. However, above 25 km altitude the winter atmosphere has larger diagnosed uncertainties. Near the winter pole the average bending angle at high altitudes can become very small. Since the uncertainties are normalised by the background bending angle, this can lead to a larger relative error.

4 GNSS-RO Impact in NWP Systems

In this section we briefly summarise the current impact of GNSS-RO measurements in the ECMWF NWP system. These should be representative of the impact at other NWP centres. The experiments span a 3 month period from January 1 to March

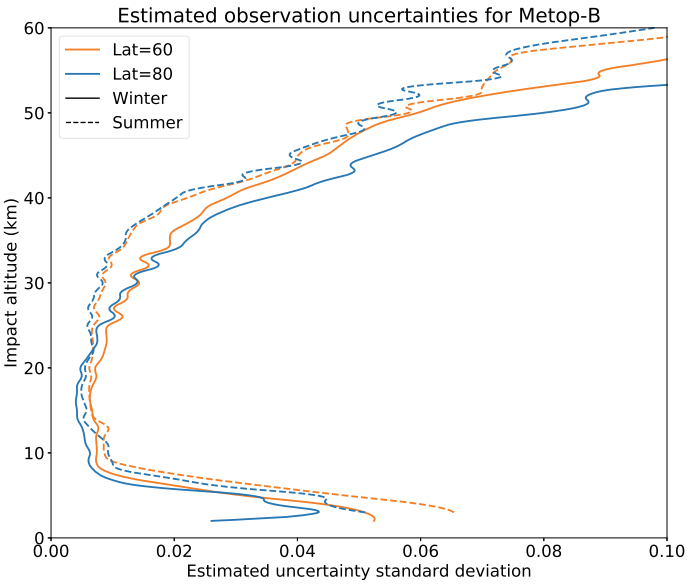

Fig. 6 Diagnosed observation uncertainties, normalised by the background bending angle for Metop-B. Results are shown for different seasons and latitudes in the Northern Hemisphere Extra-Tropics summer (16 June 2018 to 16 July 2018) and winter (January 2020)

31, 2020. They use the ECMWF Integrated Forecast System (IFS) cycle 47R1 in incremental 4D-Var mode, with a 12 h assimilation window. The forecast model is run at Tco399, which is an effective grid spacing of 25 km.

The GNSS-RO data are assimilated with a 2D bending operator. A global bending angle uncertainty model is used, and the assumed uncertainty is only a function of impact height (impact parameter minus radius of curvature). The percentage uncertainty is assumed to fall linearly in impact height from 20 % at 0 km to 1.25 % at 10 km. Above 10 km we assume a constant value of 1.25 % until this reaches a lower limit of 3×10^{-6} rad.

The three experiments shown here are:

1. CONTROL (CTL): Uses the GNSS-RO available operationally at the beginning of 2020, including Metop (A+B+C) GRAS, FY-3C GNOS, TerraSAR-X, Tandem-X and KOMPSAT-5.
2. COSMIC-2 (C2): CTL experiment plus the COSMIC-2 measurements
3. NoRO: CTL minus ALL GNSS-RO measurements

All other observation types used operationally during this period are assimilated in these experiments.

Figure 7 shows the standard deviation of differences between short-range forecasts and globally-distributed radiosonde temperature profiles when the GNSS-RO measurements are assimilated, divided by the standard deviation from the NoRO

experiment. Values less than 100 % indicate that the GNSS-RO are improving the short-range forecasts. The GNSS-RO measurements have a clear positive impact throughout the vertical column, but the largest impact is above 200 hPa, where the percentage improvements exceed 2 % when the COSMIC-2 data are assimilated.

Results qualitatively similar to this have been produced at many NWP centres (e.g., Healy (2006); Cucurull et al. (2007); Aparicio and Deblonde (2008); Rennie (2010)), and they are broadly consistent with earlier GNSS-RO information content studies Collard and Healy (2003), so this is a well established result. In contrast, it has been more difficult to demonstrate a clear impact of GNSS-RO on the humidity fields. However, recent results assimilating COSMIC-2 measurements have now suggested that GNSS-RO are also improving the short-range humidity forecasts. Figure 8 shows the globally averaged short-range forecast departure statistics for ATMS radiances. Channels 6–15 are sensitive to temperature, but channels 18–22 are sensitive to tropospheric humidity. There is a clear improvement in the humidity sensitive channels, of around 1 % when the COSMIC-2 measurements are assimilated. This signal is also reproduced in most other satellite and in situ data types sensitive to tropospheric humidity, and it is an important recent result for GNSS-RO.

Figure 9 shows the combined impact of the GNSS-RO measurements on the medium range forecast error statistics in the COSMIC-2 experiment. This is the fractional change in the zonally averaged standard deviation of the geopotential forecast errors. The forecast errors are calculated using the difference between the forecast and the analysis for each of the experiments (with and without GNSS-RO observations). Values less than 0 (blue) indicate that the forecast errors are being reduced by the GNSS-RO measurements, and hatching indicates statistical significance at the 95 % level. At forecast ranges between T+24h and T+120h there are reductions in the forecast error at most heights and latitudes, demonstrating the broad positive impact of these observations.

5 Future Directions for the Observation and Methods

In the previous sections we outlined some of the general principles applied for use of GNSS-RO in numerical weather prediction in the troposphere and stratosphere. The application described uses transmitters in medium earth orbit and receivers in LEO orbit. However, there are future directions which could expand this network to enhance sampling in areas deemed undersampled, or where it is assumed additional coverage could provide the greatest benefit.

There are two concepts which could help to expand the current GNSS-RO observing network. One of these would be to use a constellation of LEO satellites where one would act as a transmitter and another as a receiver. One such concept uses a wave optics-based retrieval chain Benzon and Hoeg (2016), along with XK and KM bands to probe the atmosphere, allowing for correction of water vapor content, and the potential to explore ozone content Benzon and Hoeg (2016). A network of LEO-LEO satellites could significantly enhance the GNSS-RO network supplying

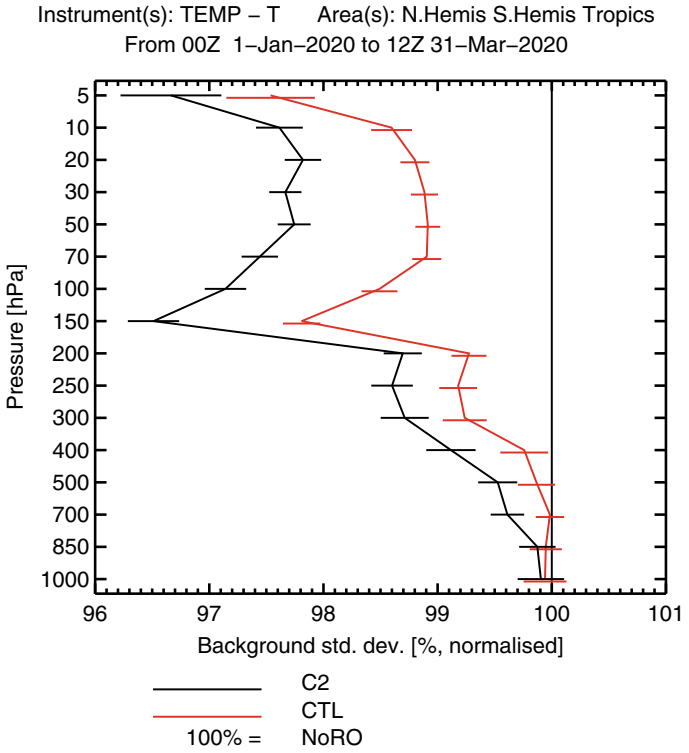

Fig. 7 The normalised standard deviation of short-range forecast departures from radiosonde temperatures on pressure levels, for the C2 (black line) and CTL (red line) experiments. These have been divided by the NoRO standard deviation values, and so values less than 100 % indicate that the GNSS-RO measurements are improving the short-range forecasts

some unique information in addition to current techniques at the L band frequency. A second way to augment the current GNSS-RO network is already being employed which is the implementation on aircraft. This greatly increases the GNSS-RO observation density and has been used for field campaigns. Similar to space-bourne RO, airborne radio occultation (ARO) airborne radio occultation (ARO) measures signal propagation delay from rising and setting GNSS satellites below the local horizon that sample the atmosphere in a region to either side of the aircraft Xie et al. (2008). The application of these ARO could be expanded in the future, a fleet of drones could be deployed to supplement the current GNSS-RO observing network or could be directed for targeted work such as severe weather outbreaks, or extreme weather events.

Another further enhancement of the GNSS-RO observing network could be the use of small and even cube satellites. Since a cube-sat is much cheaper to develop and launch than larger satellites a constellation of cube-sats has the prospect of providing many more observations for a given cost. Various private companies are developing

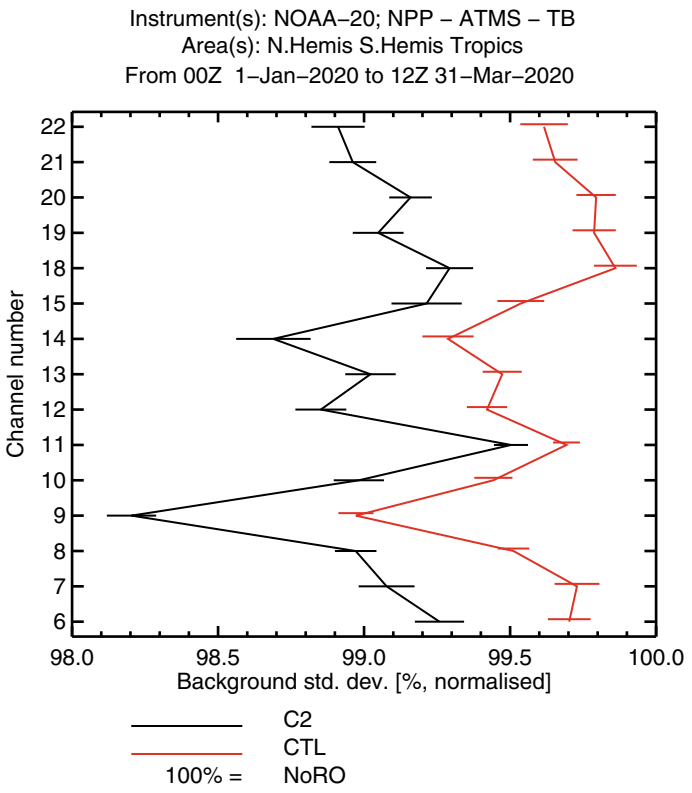

Fig. 8 The normalised standard deviation of short-range forecast departures from ATMS radiances as a function of channel number, for the C2 (black line) and CTL (red line) experiments. These have been divided by the NoRO standard deviation values, and so values less than 100 % indicate that the GNSS-RO measurements are improving the short-range forecasts

such constellations with the hope of selling the observations to government agencies. Preliminary results Bowler (2020); NOAA / NESDIS (2020) indicate that the observations are of sufficient quality for use in NWP. Further, the Met Office and ECMWF assimilated Spire observations operationally in 2020 to help mitigate the loss of aircraft measurements as a result of the COVID-19 pandemic. Figure 10 is the time series of the adjoint based forecast sensitivity to observation impact (FSOI) diagnostic Langland and Baker (2004); Cardinali (2009) in 2020 for the ECMWF operational system. There are two clear jumps in the GNSS-RO (GPSRO) contribution, when COSMIC-2 was assimilated on March 25, 2020 and when Spire data became operational on May 13, 2020. In late March 2020 the contribution from aircraft rapidly decreased as restrictions on air travel greatly reduced the number of observations available.

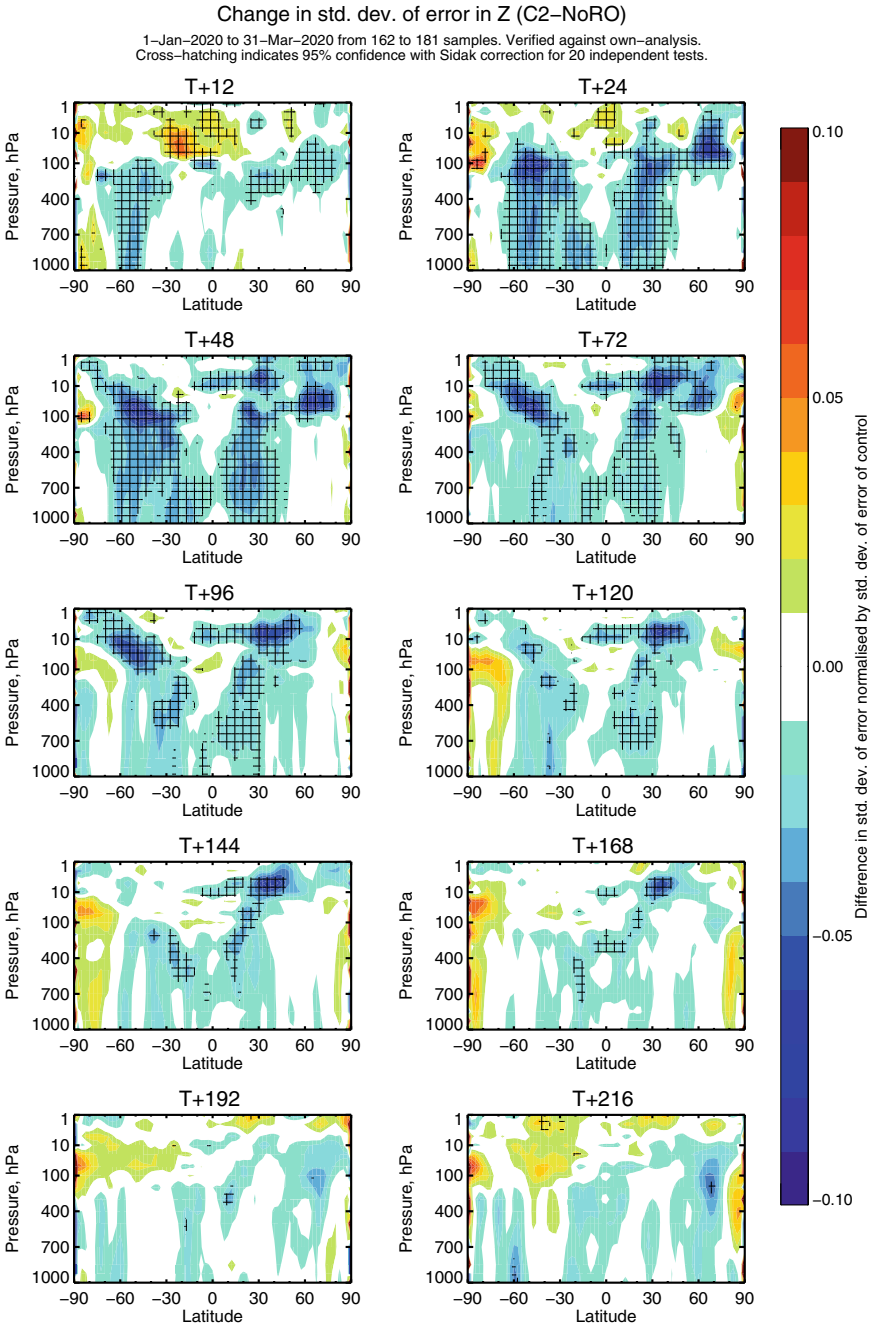

Fig. 9 The zonally averaged fractional change in the standard deviation of the geopotential forecast errors shown as a function of forecast range. The results are comparing the C2 and NoRO experiments, and the verification is against own analysis. The hatching indicates statistical significance at the 95 % level. T+nnn in the figure titles refers to the forecast lead time in hours

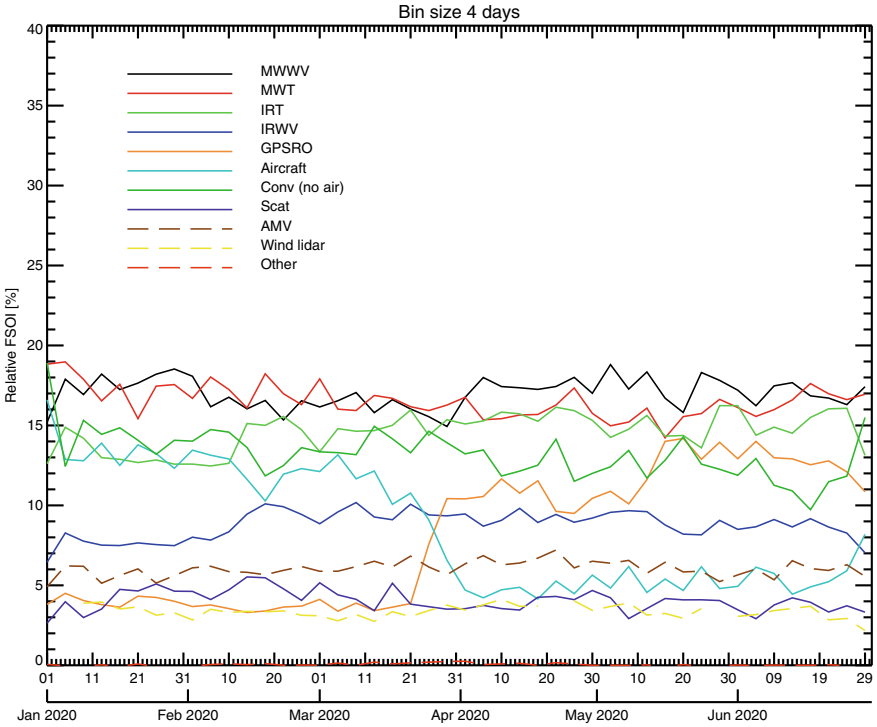

Fig. 10 The timeseries of relative FSOI values for the major observing systems in the ECMWF NWP system. The results are smoothed over 4 d. COSMIC-2 was assimilated from March 25, 2020, and Spire data were used from May 13, 2020

Overall, with the possible exception of 2020, there has been a relatively slow increase in the availability of GNSS-RO measurements since 2006. If some of the concepts described above are adopted routinely, this could increase the density of the observing network. However, an understanding of what constitutes the optimal spatio-temporal sampling is still lacking, and should be devised. An attempt to do just this was done by Harnisch et al. (2013). They examined the reduction in spread of an ensemble of data assimilations when assimilating a number of simulated GNSS-RO observations. They found that the impact did not saturate even when using 128,000 per day. As of 2020, the number of observations routinely available for all operational weather centres is approximately 8000 occultations per day. In the Harnisch et al. (2013) study, it was found that using approxmiately 16 000 profiles accounted for roughly 50% of the benefit of 128 000 occultations, which led the authors to recommend 16,000 - 20,000 occultation profiles per day as a minimum target for future observing networks.

Another new direction being taken in GNSS-RO measurement technique is the use of space-bourne polarametric missions such as the Radio-Occultation and Heavy Precipitation aboard PAZ (ROHP-PAZ) Cardellach et al. (2018). The PAZ spacecraft

has the GNSS payload enhanced to include a dual-polarization RO antenna. The dual-polarization used by the ROHP-PAZ receiver is sensitive to heavy precipitation events and other depolarizing atmospheric effects (e.g. cloud ice). A mission such as this can help to quantify the intense precipitation events, which are known to account for much of the global precipitation totals, but undersampled by the current Earth observing system. Launched on 22 February 2018, the ROHP-PAZ data are being evaluated and disseminated, with standard GNSS-RO profiles being used at many operational weather centers.

Lastly, the enhanced use of the GNSS-RO, whether they are from traditional LEO or LEO-LEO space-based, airborne, and include polarametric information need proper characterization of the uncertainties of the observation themselves. In contrast to radiometers, the degradation of the sensor over time due to changes in the antenna is not present, but stability of the clocks and transmission source of the occulting signal are key parameters which can be used to characterize the differences between the occultations. The satellite systems of the future are likely to be more numerous, but potentially more intermittent and with shorter lifetimes. This will require accurate and voluminous meta-data on the measurements being taken. Having such information can be used for improved analysis and understanding of the different behavoirs and perceived accuracies. To make better use of the GNSS-RO measurements, a more dynamic quality control procedure and observation error assignment could go a long way to delivering the greater potential of these systems.

Acknowledgements This work was supported by funding from the Public Weather Service (PWS) at the Met Office, UK; and by the Office of Naval Reserach. Sean Healy thanks Dr Katrin Lonitz (ECMWF) for help with Sect. 4.

References

Anthes R, Rieckh T (2018) Estimating observation and model error variances using multiple data sets. Atmos Meas Techniq 11:4239–4260. https://doi.org/10.5194/amt-11-4239-2018

Anthes RA, Bernhardt PA, Chen Y, Cucurull L, Dymond KF, Ector D, Healy SB, Ho SP, Hunt DC, Kuo YH, Liu H, Manning K, Mccormick C, Meehan TK, Randel WJ, Rocken C, Schreiner WS, Sokolovskiy SV, Syndergaard S, Thompson DC, Trenberth KE, Wee TK, Yen NL, Zeng Z (2008) The cosmoc/formosat-3—mission early results. Bull Am Meteorol Soc 89(3):313–333. https://doi.org/10.1175/BAMS-89-3-313

Aparicio JM, Deblonde G (2008) Impact of the assimilation of CHAMP refractivity profiles in Environment Canada Global Forecasts. Mon Wea Rev 136:257–275. https://doi.org/10.1175/2007MWR1951.1

Aparicio JM, Laroche S (2011) An evaluation of the expression of the atmospheric refractivity for GPS signals. J Geophys Res 116(D11). https://doi.org/10.1029/2010JD015214

Bean BR, Dutton EJ (1968) Radio meteorology. Dover Publications, New York

Benzon HH, Hoeg P (2016) Wave optics-based leo-leo radio occultation retrieval. Radio Sci 51(6):589–602. https://doi.org/10.1002/2015RS005852

Bowler NE (2020) An assessment of GNSS radio occultation data produced by spire. Q J R Meteorol Soc (2020). https://doi.org/10.1002/qj.3872

Cardellach E, Tomas S, Rius A, Ao CO, de la Torre-Juarez M, Padulles R, Turk FJ, Schreiner B (2018) Polarimetric gnss radio-occultations aboard paz: Commissioning phase and preliminary results. In: IGARSS 2018 - 2018 IEEE International symposium on geoscience and remote sensing IGARSS, pp. 935–937 (2018). 38th IEEE International geoscience and remote sensing symposium (IGARSS), Valencia, SPAIN, 22–27, 2018

Cardinali C (2009) Monitoring the observation impact on the short-range forecast. Q J R Meteorol Soc 135(638):239–250. https://doi.org/10.1002/qj.366

Chen SY, Huang CY, Kuo YH, Guo YR, Sokolovskiy S (2009) Assimilation of gps refractivity from formosat-3/cosmic using a nonlocal operator with wrf 3dvar and its impact on the prediction of a typhoon event. Terrest Atmos Ocean Sci 20(1):133–154. https://doi.org/10.3319/TAO.2007.11.29.01(F3C). 6th FORMOSAT-3/COSMIC Mission Early Results Workshop, Vandenberg, CA, APR 15, 2006

Collard A, Healy S (2003) The combined impact of future space-based atmospheric sounding instruments on numerical weather-prediction analysis fields: A simulation study. Q J R Meteorol Soc 129(593, B):2741–2760 (2003). https://doi.org/10.1256/qj.02.124

Cucurull L (2012) Sensitivity of nwp model skill to the obliquity of the gps radio occultation soundings. Atmos Sci Lett 13(1):55–60. https://doi.org/10.1002/asl.363

Cucurull L, Derber JC, Purser RJ (2013) A bending angle forward operator for global positioning system radio occultation measurements. J Geophys Res-Atmos 118(1):14–28. https://doi.org/10.1029/2012JD017782

Cucurull L, Derber JC, Treadon R, Purser RJ (2007) Assimilation of global positioning system radio occultation observations into ncep's global data assimilation system. Mon Weather Rev 135:3174–3193. https://doi.org/10.1175/MWR3461.1

Culverwell ID, Lewis HW, Offiler D, Marquardt C, Burrows CP (2015) The radio occultation processing package. ROPP Atmos Meas Tech 8:1887–1899. https://doi.org/10.5194/amt-8-1887-2015

Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background and analysis-error statistics in observation space. Q J R Meteorol Soc 131(613, C):3385–3396 (2005). https://doi.org/10.1256/qj.05.108. 4th WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography, Prague, CZECH REPUBLIC, APR 18-22, 2005

Eyre J (1994) Assimilation of radio occultation measurements inot a numerical weather prediction system. Technical Report 199, European Centre for Medium-Range Weather Forecasts (1994)

Eyre JR (2016) Observation bias correction schemes in data assimilation systems: a theoretical study of some of their properties. Q J R Meteorol Soc 142:2284–2291. https://doi.org/10.1002/qj.2819

Fjeldbo G, Kliore A, Eshlemen V (1971) Neutral atmosphere of venus as studied with mariner-v radio occultation experiments. Astronom J 76(2):123–124. https://doi.org/10.1086/111096

Gorbunov ME, Lauritsen KB (2004) Analysis of wave fields by Fourier Integral Operators and their application for radio occultations. Radio Sci 39. https://doi.org/10.1029/2003RS002971

Hajj G, Kursinski E, Romans L, Bertiger W, Leroy S (2002) A technical description of atmospheric sounding by gps occultation. J Atmos Solar-Terrest Phys 64(4):451–469. https://doi.org/10.1016/S1364-6826(01)00114-6

Harnisch F, Healy SB, Bauer P, English SJ (2013) Scaling of GNSS radio occultation impact with observation number using an ensemble of data assimilations. Mon Weather Rev 141:4395–4413

Healy S (2001) Radio occultation bending angle and impact parameter errors caused by horizontal refractive index gradients in the troposphere: A simulation study (vol 106, pg 11,875, 2001). J Geophys Res-Atmos 106(D20):24087. https://doi.org/10.1029/2001JD001201

Healy S, Thepaut J (2006) Assimilation experiments with champ gps radio occultation measurements. Q J R Meteorol Soc 132(615, B):605–623. https://doi.org/10.1256/qj.04.182

Healy SB (2014) Implementation of the ropp two-dimensional bending angle observation operator in an nwp system. Technical report 19, EUMETSAT ROM SAF

Healy SB, Eyre JR, Hamrud M, Thepaut JN (2007) Assimilating gps radio occultation measurements with two-dimensional bending angle observation operators. Q J R Meteorol Soc 133(626, A):1213–1227 (2007). https://doi.org/10.1002/qj.63

Hollingsworth A, Lonnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data 1. the wind field. Tellus A - Dyn Meteorol Oceanography 38:111–136

Jensen A, Lohmann M. Benzon HH, Nielsen A (2003) Full spectrum inversion of radio occultation signals. Radio Sci 38:1040. https://doi.org/10.1029/2002RS002763

Jensen A, Lohmann M, Nielsen A, Benzon HH (2004) Geometrical optics phase matching of radio occultation signals. Radio Sci 39. https://doi.org/10.1029/2003RS002899

Kliore A, Cain D, Levy G, Eshleman V, Drake F, Fjeldbo G (1965) Mariner 4 occultation experiment. Astronaut Aeronaut 3(7):72–73

Kravtsov Y, Orlov Y (1990) Radio meteorology. Springer Series on Wave Phenomena. Springer, Berlin Heidelberg

Kursinski E, Hajj G, Bertiger W, Leroy S, Meehan T, Romans L, Schofield J, McCleese D, Melbourne W, Thornton C, Yunck T, Eyre J, Nagatani R (1996) Initial results of radio occultation observations of earth's atmosphere using the global positioning system. Science 271(5252):1107–1110. https://doi.org/10.1126/science.271.5252.1107

Kursinski E, Hajj G, Schofield J, Linfield R, Hardy K (1997) Observing earth's atmosphere with radio occultation measurements using the global positioning system. J Geophys Res Atmos 102(D19):23429–23465. https://doi.org/10.1029/97JD01569

Langland R, Baker N (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus Ser A-Dyn Meteorol Oceanogr 56(3):189–201. https://doi.org/10.1111/j.1600-0870.2004.00056.x

Leroy S (1997) Measurement of geopotential heights by gps radio occultation. J Geophys Res Atmos 102(D6):6971–6986. https://doi.org/10.1029/96JD03083

Lorenc A (1986) Analysis-methods for numerical weather prediction. Q J R Meteorol Soc 112(474):1177–1194. https://doi.org/10.1002/qj.49711247414

Luntama JP, Kirchengast G, Borsche M, Foelsche U, Steiner A, Healy S, von Engeln A, O'Clerigh E, Marquardt C (2008) Prospects of the eps gras mission for operational atmospheric applications. Bull Am Meteorol Soc 89(12):1863+. https://doi.org/10.1175/2008BAMS2399.1

Melbourne W, Davis E, Duncan C, Hajj G, Hardy K, Kursinski E, Meehan T, Young L (1994) The application of spaceborne gps to atmospheric limb sounding and global change monitoring. Technical Report 94-18, National Aeronautics and Space Administration

NOAA / NESDIS: Commercial Weather Data Pilot (CWDP) Round 2 Summary. https://www.space.commerce.gov/wp-content/uploads/2020-06-cwdp-round-2-summary.pdf (2020)

Poli P, Moll P, Puech D, Rabier F, Healy SB (2009) Quality control, error analysis, and impact assessment of formosat-3/cosmic in numerical weather prediction. Terrest Atmos Ocean Sci 20(1):101–113. https://doi.org/10.3319/TAO.2008.01.21.02(F3C). 6th FORMOSAT-3/COSMIC Mission Early Results Workshop, Vandenberg, CA, APR 15, 2006

Rennie MP (2010) The impact of gps radio occultation assimilation at the met office. Q J R Meteorol Soc 136(646, A):116–131. https://doi.org/10.1002/qj.521

Rocken C, Anthes R, Exner M, Hunt D, Sokolovskiy S, Ware R, Gorbunov M, Schreiner W, Feng D, Herman B, Kuo Y, Zou X (1997) Analysis and validation of gps/met data in the neutral atmosphere. J Geophys Res Atmos 102(D25):29849–29866. https://doi.org/10.1029/97JD02400

Rodgers CD (2000) Inverse methods for atmospheric sounding: theory and practice. Ser Atmos Ocean Planetery Phys. World Scientific (2000). https://doi.org/10.1142/3171

Scherllin-Pirscher B, Steiner AK, Kirchengast G, Kuo YH, Foelsche U (2011) Empirical analysis and modeling of errors of atmospheric profiles from gps radio occultation. Atmos Meas Techniq 4:1875–1890. https://doi.org/10.5194/amt-4-1875-2011

Schreiner WS, Weiss JP, Anthes RA, Braun J, Chu V, Fong J, Hunt D, Kuo YH, Meehan T, Serafino W, Sjoberg J, Sokolovskiy S, Talaat E, Wee TK, Zeng Z (2020) Cosmic-2 radio occultation constellation: First results. Geophys Res Lett 47(4). https://doi.org/10.1029/2019GL086841

Shao H, Zou X, Hajj GA (2009) Test of a non-local excess phase delay operator for gps radio occultation data assimilation. J Appl Remote Sens 3

Sokolovskiy S, Kuo Y, Wang W (2005) Assessing the accuracy of a linearized observation operator for assimilation of radio occultation data: Case simulations with a high-resolution weather model. Mon Weather Rev 133(8)

Solheim F, Vivekanandan J, Ware R, Rocken C (1999) Propagation delays induced in gps signals by dry air, water vapor, hydrometeors, and other particulates. J Geophys Res Atmos 104(D8):9663–9670. https://doi.org/10.1029/1999JD900095

Syndergaard S, Kursinsi E, Herman B, Lane E, Flittnerm D (2005) Refractive index mapping operator for assimilation of occultation data. Mon Weather Rev 133(11):2650

Vorobev V, Krasilnikova T (1994) USSR Phys Atmos Ocean 29

Ware R, Exner M, Feng D, Gorbunov M, Hardy K, Herman B, Kuo Y, Meehan T, Melbourne W, Rocken C, Schreiner W, Sokolovskiy S, Solheim F, Zou X, Anthes R, Businger S, Trenberth K (1996) Gps sounding of the atmosphere from low earth orbit: Preliminary results. Bull Am Meteorol Soc 77(1):19–40. https://doi.org/10.1175/1520-0477(1996)077<0019:GSOTAF>2.0.CO;2

Wee TK, Kuo YH, Lee DK (2010) Development of a curved ray tracing method for modeling of phase paths from gps radio occultation: a two-dimensional study. J Geophys Res Atmos 115. https://doi.org/10.1029/2010JD014419

Wickert J, Reigber C, Beyerle G, Konig R, Marquardt C, Schmidt T, Grunwaldt L, Galas R, Meehan T, Melbourne W, Hocke K (2001) Atmosphere sounding by gps radio occultation: first results from champ. Geophys Res Lett 28(17):3263–3266. https://doi.org/10.1029/2001GL013117

Xie F, Haase JS, Syndergaard S (2008) Profiling the atmosphere using the airborne gps radio occultation technique: a sensitivity study. IEEE Trans Geosci Remote Sens 46(11, 1):3424–3435. https://doi.org/10.1109/TGRS.2008.2004713

Yunck T, Liu C, Ware R (2000) A history of gps sounding. Terrest Atmos Ocean Sci 11(1):1–20. https://doi.org/10.3319/TAO.2000.11.1.1(COSMIC)

Zou X, Liu H, Anthes R (2002) A statistical estimate of errors in the calculation of radio-occultation bending angles caused by a 2d approximation of ray tracing and the assumption of spherical symmetry of the atmosphere. J Atmos Ocean Technol 19(1):51–64. https://doi.org/10.1175/1520-0426

Zou X, Vandenberghe F, Wang B, Gorbunov M, Kuo Y, Sokolovskiy S, Chang J, Sela J, Anthes R (1999) A ray-tracing operator and its adjoint for the use of gps/met refraction angle measurements. J Geophys Res Atmos 104(D18):22301–22318. https://doi.org/10.1029/1999JD900450

Impact of Assimilating the Special Radiosonde Observations on COAMPS Arctic Forecasts During the Year of Polar Prediction

Xiaodong Hong, James D. Doyle, and Daniel P. Tyndall

Abstract We quantify the impact of assimilating the extra radiosonde observations launched in August 2018 during the Year of Polar Prediction (YOPP) on numerical weather prediction forecasts over the Arctic using the U.S. Navy's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS®). The Naval Research Laboratory Atmospheric Variational Data Assimilation System (NAVDAS) is used to assimilate the observations and provide the initial conditions for the COAMPS forecasts. Four data assimilation experiments are conducted as part of an observing system experiment (OSE) to investigate the impact of assimilating the extra radiosonde observations on COAMPS short-term forecasts (up to 48 h). The four experiments explore the impacts of assimilating the additional radiosonde observations in NAVDAS in cold-start or warm-start configuration. Based on the results from the four-month-long data assimilation tests, we find that assimilating the extra YOPP radiosonde observations improves the overall accuracy of the COAMPS short-term atmospheric forecasts over the Arctic for both cold-start and warm-start initializations. The warm-start configuration makes the most significant improvement to the COAMPS Arctic forecasts during the test period. The domain-averaged mean RMSE of the 48 h forecast from the warm-start assimilation is about 17% less than that from the cold-start initialization.

1 Introduction

Accurate atmospheric forecasts over the Arctic are very challenging and are especially important during the critical ice-melting period that features substantial subseasonal variabilities. The numerical weather prediction (NWP) challenges in the Arctic region are mainly due to (1) the inability of models to represent the complex processes accurately and (2) a paucity of observations used to generate initial conditions for NWP models through data assimilation. Radiosonde observations, which

X. Hong (✉) · J. D. Doyle · D. P. Tyndall
Naval Research Laboratory, Monterey, CA 93943, USA
e-mail: xiaodong.hong@nrlmry.navy.mil

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_14

provide highly accurate vertical profiles of temperature, humidity, and wind components, are among some of the most important observations to data assimilation systems, but have sparse coverage across the Arctic. Due to limited financial and human resources, conventional observations are sparse north of 70° N (Lawrence et al. 2019). There are only 76 sounding stations on continents and islands north of 60° N with a normal launch frequency of twice a day (Naakka et al. 2019). There are no regular radiosondes launched over the Arctic Ocean (Naakka et al. 2019).

Previous studies have shown the utility of radiosonde observations over the Arctic Ocean. For example, Naakka et al. (2019) found that high-quality radiosonde observations substantially improved analyses and provided observation information to the data assimilation system that satellite observations could not provide. Other studies suggest that the large uncertainties in Arctic weather forecasts arise partially from the sparse nature of the observational network over the Arctic Ocean. Inoue et al. (2020) showed that additional assimilated Arctic observations in atmospheric ensemble forecasts enhanced the forecast skill of weather and sea-ice forecasts during a strong wind event. Yamazaki et al. (2015) showed that the use of additional radiosonde observations that were launched from the German icebreaker Polarstern from mid-July to early August 2012 intensified a tropopause polar vortex in an atmospheric ensemble reanalysis. Additionally, the ensemble prediction reproduced the formation of the strong Arctic cyclone (named "AC12", the strongest Arctic cyclone observed on record) when the reanalysis was used as the initial conditions due to the improved upper-tropospheric circulation in the Arctic region provided by assimilating the additional radiosonde observations. When the additional radiosondes were excluded from the assimilation, the reanalysis produced a significantly weaker cyclone, showing that the additional radiosonde observations were indispensable for the improved prediction of AC12.

The World Meteorological Organization (WMO) has established a ten-year Polar Prediction Project (PPP) to significantly improve weather and climate prediction over the Polar Regions. One of the critical activities of the PPP is the Year of Polar Prediction (YOPP), which provided an opportunity to address the lack of observation coverage in the Arctic (Jung et al. 2016; Bauer et al. 2016). Three Special Observing Periods (SOPs) of YOPP involve the intensive observation and modeling campaigns in both the Arctic and Antarctic from mid-2017 to mid-2019. Additional YOPP observations collected during the SOPs will be used to develop improved data assimilation systems that mitigate the challenge of observational sparseness over the Polar Regions. The enhanced observations also are extremely valuable for verification to quantify the accuracy of current numerical models and identify systematic model errors over the Polar Regions (Casati et al. 2017).

One of the three YOPP SOPs took place in the Arctic from July to September 2018, with extra radiosondes launched from the locations shown in Fig. 1a (marked with yellow dots). The majority of the extra radiosondes were released at 06Z and 18Z; fewer were released at 00Z and 12Z (Fig. 1b). The extra radiosondes launched during August 2018 provide about 25% more observations than the routine radiosonde

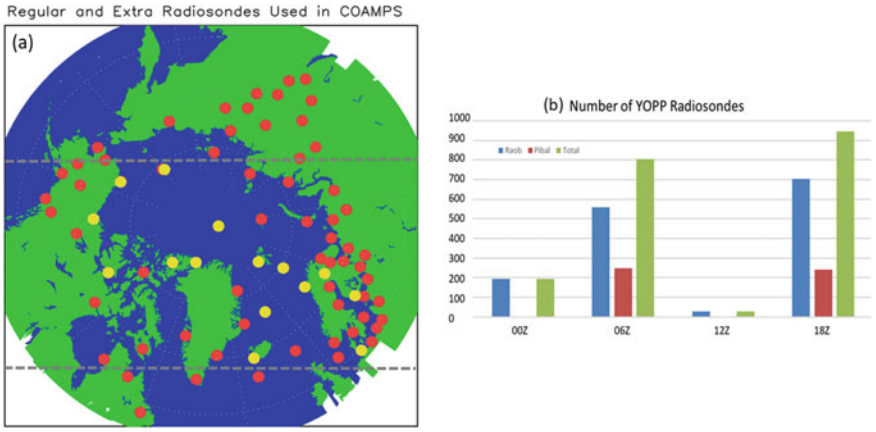

Fig. 1 **a** COAMPS domain and locations of routine (red dots) and YOPP extra radiosonde locations (yellow dots) at 00Z on August 10, 2018. The area, enclosed by the two gray-dashed lines and the left-and-right lateral model boundaries, is used for calculating the average impact for each of the data assimilations in Sect. 4. **b** The total number of extra YOPP radiosondes as a function of instrument type (radiosonde [raob] or pilot balloon [pibal]) and initialization during August 2018

launches. These special YOPP radiosonde observations were disseminated in real-time from the radiosonde ground station through the Global Telecommunication System (GTS) to any operational center subscribing to the observation bulletins. Because the data was transmitted through the GTS in real-time, the Navy's Fleet Numerical Meteorology and Oceanography Center (FNMOC) used the additional radiosondes as part of their routine operational forecasts.

During the July to September 2018 SOP, a typical atmospheric circulation with a substantial variation evolved over the Arctic region through August. As noted earlier, previous studies have examined the impacts of assimilating additional observations on Arctic forecasts during extreme events such as strong winds or polar vortexes. Still, the impact on forecasts during the summer months has been investigated less frequently. In this study, we focus on quantifying the impact of assimilating the extra radiosonde observations during YOPP on COAMPS (Hodur 1997) Arctic forecasts using NAVDAS (Daley and Barker 2001) with an OSE methodology.

We organize the remainder of this paper in the following way. First, we summarize the synoptic features during the transition phase of the atmospheric circulations during August 2018. We then provide a summary of the numerical experiment design in Sect. 3. The discussion of the results is presented in Sect. 4. Finally, we provide a summary and conclusions in Sect. 5.

2 Synoptic Features

The Arctic sea ice extent rapidly decreased during the first two weeks of August 2018, according to the data from the National Snow and Ice Data Center (NSIDC). Fig. 2a, b depict the monthly sea ice extent for July and August 2018, respectively. Sea ice extent declined at approximately 25,100 square miles per day through the first two weeks of August, which is slightly faster than the average rate of 22,000 square miles per day, according to NSIDC. The more rapid retreat in sea ice extent has a strong feedback on the surface energy budget, impacts the atmospheric circulation, and presents a significant challenge in forecasting the Arctic weather accurately during the summer of 2018 (http://nsidc.org/arcticseaicenews/2018/08/).

The COAMPS and the European Centre for Medium Range Weather Forecasts Reanalysis-Interim (ERAI, Dee et al. 2011) mean sea-level pressure analyses both clearly show the substantial variation of atmospheric circulation over the Arctic during August 2018 (Fig. 3). Both analyses contain similar features that captured the significant weekly variability during August. During the first week of August, a strong high-pressure center is located over the central Arctic Ocean, and low-pressure centers are situated over the Kara Sea and northern Canada (Fig. 3a, e). Air temperatures at 850 hPa are warmer over the central Arctic Ocean and colder over the Beaufort Sea, corresponding to the sea-level pressure (SLP) patterns (not shown). The warmer temperatures cause additional sea ice retreat; this would also be seen in NWP forecasts if the atmospheric model were coupled to a sea ice model during the forecast. In the current COAMPS configuration, the sea ice coverage is updated at every initialization through the ocean analysis using the Navy Coupled Ocean Data Assimilation (NCODA, Cummings 2005). During the second week of August, the

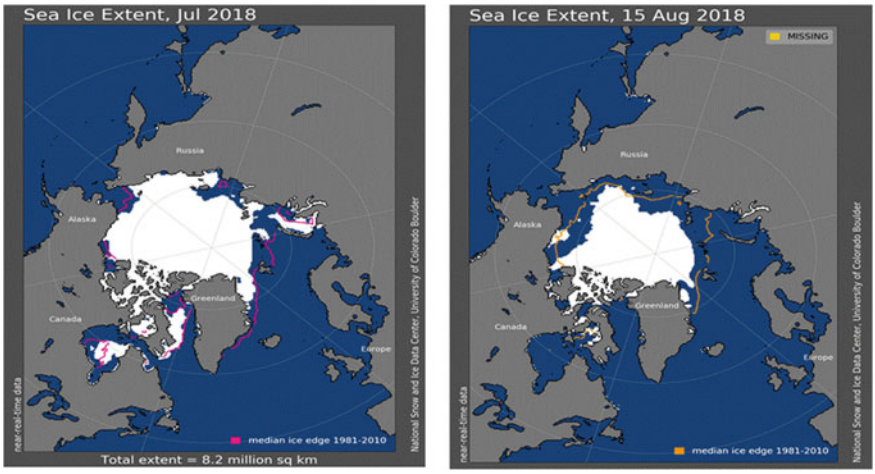

Fig. 2 **a** July 2018 sea ice extent (left panel). **b** August 2018 sea ice extent (right panel). Courtesy from http://nsidc.org/arcticseaicenews/2018/08/

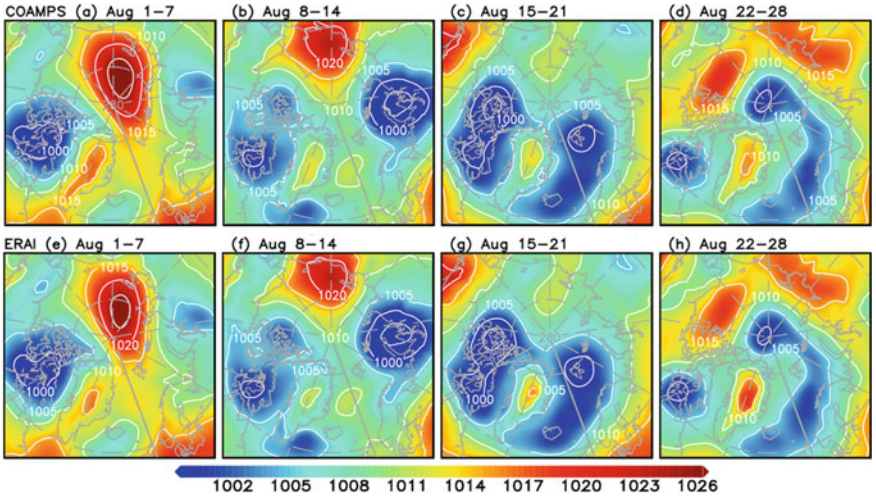

Fig. 3 Mean sea-level pressure from COAMPS (upper panel) analyses and ERAI (lower panel) for **a** and **d** August 1–14, **b** and **e** August 15–21, and **c** and **f** August 22–28, 2018

cyclone center over the Kara Sea intensifies and begins to merge with another cyclone center over northern Canada (Fig. 3b, f). This process pushes the ridge axis off the central Arctic Ocean and results in two cyclones occupying the areas during the third week (Fig. 3c, g). Consequently, the air temperature near the surface becomes colder than during the first week (not shown). The cyclones continue to occupy the central Arctic Ocean and split into three weaker centers during the fourth week (Fig. 3d, h).

3 Experimental Design

In this study, we use the limited area atmospheric component of COAMPS to produce the 48 h atmospheric numerical forecasts every 6 h throughout August over the Arctic during the YOPP. NAVDAS, a three-dimensional variational (3DVAR) data assimilation system, assimilates observations to produce the analysis, which is used as the initial conditions for COAMPS forecasts. The global atmospheric forecast fields from NAVGEM are used as the lateral boundary conditions for COAMPS.

An OSE is conducted using four experiments designed to investigate the impact of assimilating the extra radiosonde observations obtained during August of the YOPP on the COAMPS Arctic forecasts. Two sets of experiments are configured to determine the impact of assimilating the additional radiosondes using NAVDAS. All experiments use identical observation data as inputs for NAVDAS, except two experiments (one in each set) withhold the additional YOPP radiosondes from assimilation. The two sets of experiments are configured to utilize different background fields as

Table 1 Design of four cases

	YC	NYC	YW	NYW
YOPP/NoYOPP	YOPP	NoYOPP	YOPP	NoYOPP
Warm/cold-start	Cold	Cold	Warm	Warm

inputs for NAVDAS. In two of the four experiments, the standard COAMPS initialization is used, which utilizes the previous cycle's 6 h forecast as the background field for NAVDAS (known as a warm-start). In the other two experiments, COAMPS is configured in a non-standard mode to cold-start the system at every forecast cycle. When running in this mode, the NAVGEM analysis is used as the background field for NAVDAS. The NAVGEM analysis is produced by the Navy's operational four-dimensional variational (4DVAR) data assimilation system, NAVDAS-AR (Xu et al. 2005). Configuration of COAMPS to cold-start every cycle is controversial, as using the NAVGEM analysis as the background for NAVDAS introduces background errors that have been correlated with the observation errors. Additionally, the impact of the YOPP radiosondes in experiments that have been cold-started is further obscured, as the NAVGEM analysis used in all experiments assimilated the YOPP radiosondes. While the configuration of the cold-start set of experiments is less than ideal, we believe that the differences between the two sets of experiments provide an interesting contrast between the NAVGEM and COAMPS systems.

For simplicity, we denote the four data assimilation experiments using the nomenclature described in Table 1. YC (YOPP observations with a cold start) and YW (YOPP observations with a warm start) include the routine and the extra observations. In contrast NYC (no YOPP observations and a cold start) and NYW (no YOPP observations and a warm start) only include the routine observations.

The COAMPS Arctic model domain includes all the extra YOPP radiosonde locations as indicated in Fig. 1a. The horizontal model resolution is 15 km with a grid dimension of 402×402 points. There are 60 sigma levels in the vertical with 14 levels in the lowest 1000 m. We use a COAMPS model configuration that is typical in routine operational applications to produce all of the forecasts in this study. Analyses from the four cases are compared to ERAI reanalysis as shown in Fig. 3, and forecasts are verified using all available surface and upper-air observations to examine the impact of the extra observations on the forecasts.

4 Discussion of Results

This section focuses our discussions on the following diagnostics: the analysis increments, the analysis RMSE, and the COAMPS Arctic forecasts. To examine the overall impact of assimilating the extra radiosonde observations of the YOPP on the COAMPS Arctic forecasts, we employ various spatial and temporal averages of

the diagnostics mentioned above to analyze the results of the four data assimilation experiments.

(1) *Impact of assimilating the extra YOPP observation on mean analysis increments*

The analysis increment is defined as the difference between the analysis and the background field. It is the data assimilation system's correction to the background field as informed by the observation information assimilated. NAVDAS creates the analysis increments using a 3DVAR methodology, which utilizes a background error covariance based on geostrophic and hydrostatic balance, specified observation errors, and observation innovations, which are the differences between the observations and the background field. Mean analysis increments can be used to help identify systematic biases in the background field. When an OSE methodology is used, the difference of the mean analysis increments between experiments can show the impact of the observing system under investigation (the additional YOPP radiosondes in the case of this study). Fig. 4 shows time and space averaged analysis increments from fore-

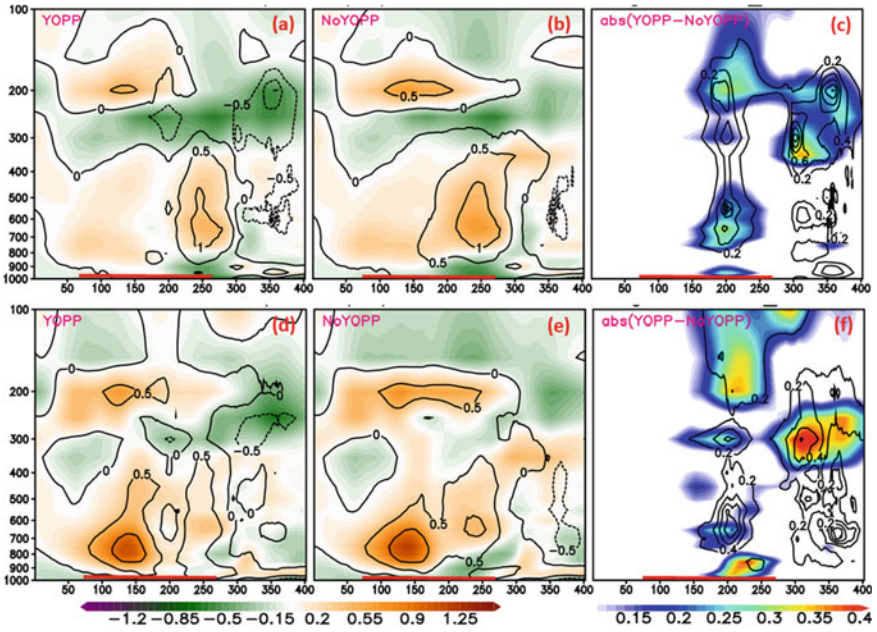

Fig. 4 Monthly mean analysis increments of air temperature (color shading) and dew-point depression (contours) for all model runs initialized at 06Z during August 2018 for cold-start (top panels) and warm-start (bottom panels) experiments using the YOPP radiosondes (left panels) and withholding the YOPP radiosondes (middle panels), and the absolute difference between YOPP and NoYOPP analyses (right panels). The vertical cross-section is the mean of the areas bounded by the two gray-dashed lines in Fig. 1a. The labels of x-axes are horizontal grid points with the red solid-line corresponding to the Arctic area

casts initialized at 06Z over the COAMPS domain highlighted in Fig. 1a; increments are spatially averaged along the y-direction and temporally averaged through the entire month. Fig. 4 also shows the absolute difference in the mean analyses in each set of experiments between experiments assimilating (YOPP) and experiments withholding (NoYOPP) the additional radiosondes; the absolute difference is the impact the additional radiosondes had on the NAVDAS analysis.

In the cold-start case, the assimilation of the additional YOPP observations results in more negative temperature mean analysis increments (Fig. 4a) and more negative dew-point depression mean analysis increments (dashed contours in Fig. 4a) around the 200–400 hPa layer as compared to the mean analysis increments generated from withholding the YOPP observations. The more negative temperature and dew-point depression increments indicate that the NAVGEM analysis was too warm and too humid than the additional YOPP observations. Note that because we are cold-starting in both of these experiments, the background fields used here are identical. The absolute differences of the analyses between the YC (Fig. 4a) and no NYC (Fig. 4b) analyses displayed in Fig. 4c further demonstrate the impact of the extra data on the analysis increments with the cold-start data assimilation option.

For the warm-start cases (YW and NYW), the mean analysis increment differences between the two cases are due to the additional YOPP radiosonde observations. For the warm-start cases, there is both a direct and indirect contribution from the YOPP radiosondes. The direct contribution comes impacts associated with the assimilation of the radiosondes during each forecast cycle. The indirect contribution is from the background field, which was impacted by the previous cycle's assimilation of the additional YOPP observations. The temperature and dew-point depression mean analysis increments from the YW and NYW cases are depicted in Fig. 4d, e; the absolute difference in the mean analyses between the two cases is depicted in Fig. 4f. The experiments utilizing the warm-starts require slightly different adjustments to the background field than the cold-start cases: the background over most of the domain between 400 and 300 hPa needs to be warmed and made more humid (Fig. 4d, e). Both the warm start and cold start cases also require positive adjustments to the background temperatures between 900 and 600 hPa; however, the increments are more substantial for the warm-start experiments. The differences between warm-start configured YOPP and NoYOPP cases are larger than the differences in the cold-start case; this is likely due to the presence of YOPP observations in the NAVGEM fields used as the background. The colocation of the significant differences between the YOPP and NoYOPP cases in both COAMPS initializations further supports that the additional radiosondes impact the analyses.

(2) *Impact of assimilating the extra YOPP observations on the reduction in the mean analysis RMSE*

The month-long temporal and horizontal mean difference of analysis RMSE profiles between the cases with and without assimilated YOPP radiosonde observations ($DRMSE_{TM}$) is defined as below:

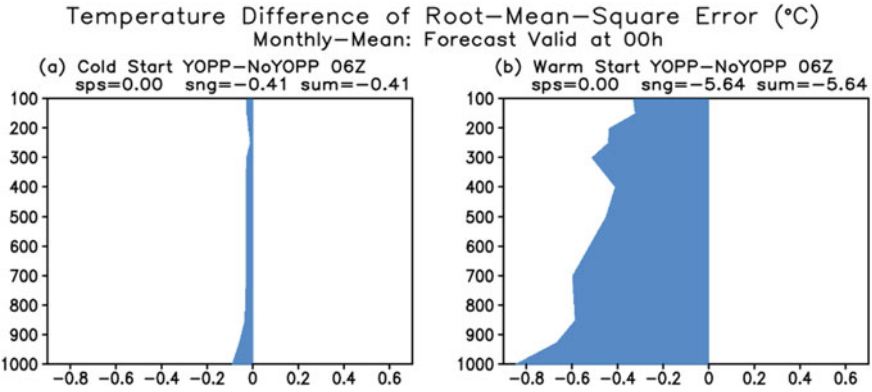

Fig. 5 Mean difference of the air temperature analysis RMSE profiles between YOPP and NoYOPP from all the cold and warm-start cases at 06Z over the August period. The RMSEs for the sum of positive (sps), negative (sng), and the total sum of positive and negative (sum) over the whole domain are listed on the top of each plot

$$DRMSE_{TM} = \frac{1}{T}\sum_{t=1}^{t=T}\left\{\frac{1}{M}\sum_{i=1}^{i=M} RMSE(YOPP)_i - \frac{1}{M}\sum_{i=1}^{i=M} RMSE(NoYOPP)_i\right\}_t \quad (1)$$

where the subscript *TM* is for temporal and horizontal mean difference, the outer summation upper limit of the cycle summation T is the total number of the cycles with an initialization time of 06Z (31), the inner summation is for the horizontal mean, and M is the total number of observation and analysis pairs (28). $RMSE(YOPP)$ and $RMSE(NoYOPP)$represent the temperature RMSEs for YOPP and NoYOPP assimilations validated against radiosondes, respectively. Equation (1) is applied to both the cold-start cases (YC and NYC) and warm-start cases (YW and NYW). The DRMSE for each set of initialization experiments are depicted in Fig. 5a, b. The negative values (blue color) indicates the RMSE is smaller for YOPP than NoYOPP. The sum of positive (sps), negative (sng), and total sum of positive and negative (sum) listed on the top of each plot are summed vertically through the depth of the atmosphere.

The RMSEs of the analyses are calculated using only the routine observations to avoid having the additional YOPP observations used in the verification. The RMSE with the YOPP radiosonde observations assimilation is significantly smaller than the NoYOPP case for both cold and warm-starts. This indicates that the extra YOPP radiosonde observations improve the analyses used as the initial conditions for the forecasts. The RMSE difference between the two warm-start cases (Fig. 5b) is substantially larger than between the cold-start cases (Fig. 5a). The total sum of RMSE from the warm-start cases is −5.64, an RMSE reduction of than more than 13 times of that (−0.41) of the cold-start cases. This result is a consequence of the higher quality background fields used during every 6 h update cycle in the warm-start configuration, which are partially impacted by the assimilation of YOPP observations

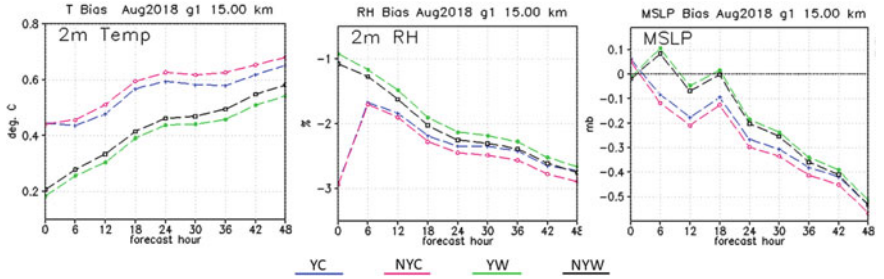

Fig. 6 Domain mean bias as a function of forecast lead time over August for **a** 2 m air temperature, **b** 2 m relative humidity, and **c** sea-level pressure

from the previous cycle. The smaller reduction between the YOPP and NoYOPP case using the cold-start initialization is likely a consequence of the assimilation of the YOPP radiosonde observations as part of the NAVGEM data assimilation process as well as the significantly coarser resolution of the NAVGEM analysis compared to the COAMPS 6 h forecast. The COAMPS warm-start experiment demonstrates that assimilating the extra radiosondes in the data-sparse Arctic region is very beneficial and significantly reduces the RMSE for the analyses.

(3) *Impact of assimilating the extra YOPP observations on the COAMPS Arctic forecasts*

Figure 6 compares the mean biases of surface variables as a function of forecast lead time for each of the four cases. When assimilating the YOPP radiosonde observations for both cold- and warm-start configurations, the mean biases are consistently smaller than their data withheld counterparts. The additional YOPP radiosonde observations' assimilation reduces warm biases by about 5% for both cold and warm starts (Fig. 6a). The warm-start cases, in general, have smaller biases than the cold-start cases, as expected. The reduction in bias from the warm-start cases is most significant for the 2 m air temperature, reducing the bias by approximately 20% of the cold-start cases. Other forecast variables shown in Fig. 6 b and c have similar bias tendencies and reductions as the 2 m air temperature depicted in Fig. 6a. In general, the initial biases are smaller in the warm-start cases than the cold-start cases, consistent with the analysis errors presented in Fig. 5. The initial biases of 2 m relative humidity for the cold-start cases are substantially larger than the warm-start cases (Fig. 6b); this is likely due to larger biases present in the NAVGEM analyses. The COAMPS model quickly reduces these errors within the first 6 h of the forecast as the model spins up. Initial fields using the NAVGEM analyses as the background in the cold-start cases result in warmer (Fig. 6a) and drier (Fig. 6b) forecast biases than using the COAMPS 6 h forecast as the background in the warm-start cases. This result is consistent with the mean analysis increments depicted in the cold-start panels of Fig. 4.

The impact of assimilating YOPP extra radiosondes on COAMPS 48 h forecast are further quantified by examining the time series of the domain averaged differences

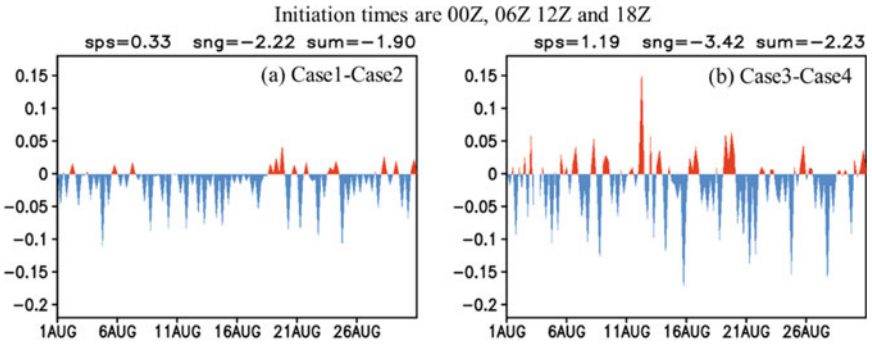

Fig. 7 Time series of the whole model domain averaged differences of RMSE for 48 h forecasts between YOPP and NoYOPP assimilations for air temperature for **a** cold-start and **b** warm-start for all data assimilation cycles. The RMSEs for the sum of positive (sps), negative (sng), and the total sum of positive and negative (sum) are listed on the top of each plot

of RMSE between the YOPP and NoYOPP assimilation experiments. The whole model domain averaged differences of RMSE profiles ($DRMSE_{HM}$) are defined as

$$DRMSE_{HM} = \frac{1}{H}\sum_{z=1}^{z=H}\left\{\frac{1}{M}\sum_{i=1}^{i=M} RMSE(YOPP)_i - \frac{1}{M}\sum_{i=1}^{i=M} RMSE(NoYOPP)_i\right\}_z \quad (2)$$

where the subscript HM denotes the domain mean (i.e. vertical and horizontal mean). The upper limit of the vertical summation H is the total number of vertical levels from 1000 to 10 hPa, and M is the total number of observation and forecast pairs across the domain. The $RMSE(YOPP)$ and $RMSE(NoYOPP)$ follow the convention from Eq. (1). The sum of positive (sps), negative (sng) and total sum of positive and negative (sum) over the whole domain are also listed on the top of each plot. The time series of $DRMSE_{HM}$ is depicted for cold-start (Fig. 7a) and warm-start (Fig. 7b) cases for all data assimilation cycles initialized at 00Z, 06Z, 12Z, and 18Z. Negative values (blue color) are present for the majority of the time series of the cold-start assimilation experiments (Fig. 7a), indicating that 48 h forecast RMSE is smaller when the YOPP radiosondes are assimilated compared to when they are withheld. The absolute value of the total negative $DRMSE_{HM}$ (−1.9) is over 6 times greater than the total positive $DRMSE_{HM}$ (0.33) (Fig. 7a), demonstrating a pronounced reduction of RMSE when the additional YOPP radiosondes are assimilated. During the period between August 13 and 16, the higher pressures over the central Arctic Ocean transitioned to lower pressures, which was a major forecasting challenge during the month. Assimilating the additional YOPP radiosondes provided consistent RMSE reduction, showing the value of the additional observation on reducing forecast uncertainty. Improvement in the RMSE was also seen when the additional radiosonde observations were assimilated using the warm-start configuration, as shown in Fig. 7b. The absolute values of the total negative $DRMSE_{HM}$ (−2.23) is nearly 3 times greater than the total positive

$DRMSE_{HM}$ (1.19). The total sum of positive and negative $DRMSE_{HM}$ (−2.23) from the warm start cases is 17% more than from the cold start case (−1.9), indicated that assimilating YOPP radiosonde observations in a warm start configuration produces the best 48 h forecast error reduction out of all the configurations explored by this study.

5 Summary and Conclusion

Substantial sub-seasonal variabilities during the critical ice-melting period make atmospheric forecasts over the Arctic challenge. These challenges stem from the inability of forecast models to accurately represent the complex processes in the Arctic and the lack of adequate observation coverage. Forecasts during August 2018 were especially challenging due to the evolution of atmospheric circulation with substantial variation over the month-long period. This paper quantified the impact of additional radiosonde launches as part of YOPP on reducing COAMPS forecast errors over the Arctic.

An OSE study was conducted to evaluate the impacts of the additional radiosonde launches. Comparing the analysis increments between the cases with and without the YOPP radiosonde observations in a cold-start configuration demonstrates that the extra YOPP observations make significant adjustments to the background field. Assimilation of the YOPP radiosonde observations results in more negative mean temperature and dew-point depression increments around the 200–400 hPa layer compared to the mean analysis increments generated when withholding the additional observations. The differences in the analysis increments in the warm-start cases with and without the YOPP radiosonde observations show that they also significantly modify the COAMPS 6 h forecast background fields to generate the analysis. Overall, using the COAMPS 6 h forecasts as the background for data assimilation tends to result in positive temperature increments and negative dewpoint depression increments in the boundary layer and lower portion of the atmosphere than when the NAVGEM analysis is used as the background.

The monthly mean difference of the RMSE profiles between the cases with and without assimilation of the YOPP radiosonde observations shows smaller RMSEs for the analyses that assimilated the extra observations for both the cold-start and warm-start configurations. A more significant reduction of the RMSEs due to the assimilation of the additional YOPP radiosonde observations is apparent for the warm-start cases. This result is due to the direct and indirect impacts of the additional radiosondes: the direct impacts are the result of the current cycle's assimilation of the additional data, while the indirect impacts are the result of an improved background, which is partially a consequence of the previous cycle's assimilation of the data. The improved background was also aided by its higher resolution. The smaller difference of the RMSE profiles for the cold-start configuration is also a result of the assimilation of the additional YOPP radiosonde observations in the NAVGEM analysis, used as the background in both the NYC and YC experiments.

Verification of the near-surface variables shows consistently smaller biases throughout the 48 h forecasts when assimilating the extra YOPP radiosonde observations in both cold and warm-start experiments. The bias is reduced by approximately 5% through the assimilation of the additional YOPP observations. The warm-start cases, in general, have smaller biases than the cold-start cases, and the reduction of biases can be up to 20% during the 48 h forecasts. The initial forecast biases from using the NAVGEM analyses as the background field in the cold-start configuration are larger than using the COAMPS 6 h forecasts as the background field in the warm-start experiments. The COAMPS model can correct the high initial errors quickly as the model spins up.

Based on the results from the four data assimilation tests carried out over August 2018, we find that assimilating the additional YOPP radiosonde observations improve the overall accuracy of the COAMPS short-term atmospheric forecasts over the Arctic for both cold-start and warm-start configurations. The warm-start data assimilation results in the most significant overall improvements in the COAMPS Arctic forecasts during the evaluation period, with a domain-averaged mean RMSE reduction of the 48 h forecast of about 17% of the cold-start assimilation. This research shows the beneficial impact of the extra radiosonde observations in the data-sparse Arctic region; making these additional launches routine would benefit COAMPS operational forecasts.

References

Bauer P, Bradley A, Bromwich D et al (2016) Implementation plan for the year of polar prediction (YOPP). WWRP/PPP no. 4-2016, WMO, 68 pp

Casati B, Haiden T, Brown B, Nurmi P, Lemieux JF (2017) Verification of environmental prediction in polar regions: recommendations for the year of polar prediction. WWRP 2017-1, WMO, 44 pp

Cummings JA (2005) Operational multivariate ocean data assimilation. Q J R Meteorol Soc 131:3583–3604. https://doi.org/10.1256/qj.05.105

Daley R, Barker E (2001) NAVDAS source book 2001: the NRL atmospheric variational data assimilation system. NRL Publication NRL/PU/7530-00-418, 160 pp (available from Marine Meteorology Division, NRL, Monterey, CA 93943-5502)

Dee D, Uppala SM, Simmons AJ (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597. https://doi.org/10.1002/qj.828

Goessling HF, Jung T, Klebe S et al (2016) Paving the way for the year of polar prediction. Bull Am Meteorol Soc 97:4. https://doi.org/10.1175/BAMS-D-15-00270.1

Hodur RM (1997) The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). Mon Weather Rev 125:1414–1430

Inoue J (2020) Review of forecast skills for weather and sea ice in supporting Arctic navigation. Polar Sci. https://doi.org/10.1016/j.polar.2020.100523

Jung T, Gordon ND, Bauer P et al (2016) Advancing polar prediction capabilities on daily to seasonal time scales. Bull Am Meteorol Soc 14:1631–1647

Lawrence H, Bormann N, Sandu I, Day J, Farnan J, Bauer P (2019) Use and impact of Arctic observations in the ECMWF Numerical Weather Prediction system. Q J R Meteorol Soc 145:3432–3454. https://doi.org/10.1002/qj.3628

Naakka T, Nygård T, Tjernström M, Vihma T, Pirazzini R, Brooks IM (2019) The impact of radiosounding observations on numerical weather prediction analyses in the arctic. Geophys Res Lett 46:8527–8535. https://doi.org/10.1029/2019gl083332

Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR. Formulation and initial tests of the linear problem. Tellus 57A:546–559

Yamazaki A, Inoue J, Dethloff K, Maturilli M, König-Langlo G (2015) Impact of radiosonde observations on forecasting summertime arctic cyclone formation. J Geophys Res Atmos 120:3249–3273. https://doi.org/10.1002/2014JD022925

Images Assimilation: An Ocean Perspective

François-Xavier Le Dimet, Hans E. Ngodock, and Innocent Souopgui

Abstract This paper discusses the potential benefits of images assimilation in the context of operational oceanography, with the goal of eventually exploiting the dynamical information contained in sequences of ocean images to improve ocean model predictions. Successful assimilation of ocean images will provide a positive answer to the question whether meaningful dynamical information can be extracted from sequences of satellite ocean color images for the improvement of analyses and forecasts of the ocean circulation. Because in situ observational campaigns are costly and usually very limited in space and time, and satellites with visible bands are increasing in number, coverage and providing images with very high temporal frequency. Operational centers should consider making image assimilation an integral part of their future assimilation systems. Beyond the motivation, we also discuss whether images should be assimilated directly or indirectly, the latter consisting of assimilating information derived from images.

1 Introduction

Velocity is a fundamental and necessary quantity in the dynamics of any fluid. Ocean currents are responsible for the transport of heat, salt, nutrients, and they also impact the movement of ships, gliders, drifting buoys, waves and ice. Currents play a significant role in the variability of ocean conditions at both regional and global scales. Accurate knowledge of ocean currents is critical for navigation, search and rescue.

The potential of ocean surface currents observations to drastically improve ocean circulation analyses and forecasts was demonstrated in recent experiments (Carrier

F.-X. Le Dimet
Laboratoire Jean-Kuntzman, Université Grenoble-Alpes, St. Martin-d'Hères, Grenoble, France

H. E. Ngodock (✉)
The US Naval Research Laboratory, Stennis Space Center, MS, USA
e-mail: Hans.Ngodock@nrlssc.navy.mil

I. Souopgui
Department of Physics, The University of New Orleans, New Orleans, LA, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_15

et al. 2016; Ngodock et al. 2015; Muscarella et al. 2015). Other studies have shown improvement of ocean predictability by assimilating ocean observed velocity data (Mariano et al. 2002; Fan et al. 2004; Taillandier et al. 2006; Nilsson et al. 2012). However, apart from coastal high frequency (HF) radar and sparse moored buoys, ocean currents are hardly observed.

The majority of observations used in correcting the ocean circulation consist of sea surface temperature (SST), sea surface height (SSH), and subsurface temperature and salinity (T/S) profiles. Assimilation of these observations provides some correction to the velocity field. For example, SSH assimilation provides geostrophic correction of the velocity field for the mesoscale circulation, and T/S profiles assimilation provides correction of the velocity field through the pressure gradient. However, the spatial distribution of these observations does not allow a reliable reconstruction of the velocity field, and accurate forecasts may be needed in locations where T/S profiles cannot be sampled or in coastal waters shallower than 200 m where SSH is not available. Thus, other types of remotely sensed observations that enable the correction of the circulation need to be exploited.

Sequences of ocean color images from satellites can capture the dynamics of the ocean, as they depict optical evolution of physical quantities and properties in the ocean. For example Fig. 1 shows a couple of Gulf Stream eddies from the infrared channel (Sea Surface Temperature) and the visible channel (Chlorophyll concentration) onboard MODIS. The similarity of the structures from those two channels shows that information from model variables (e.g. Temperature) can be obtained from other quantities (Chlorophyll concentration). Lateral displacements of these

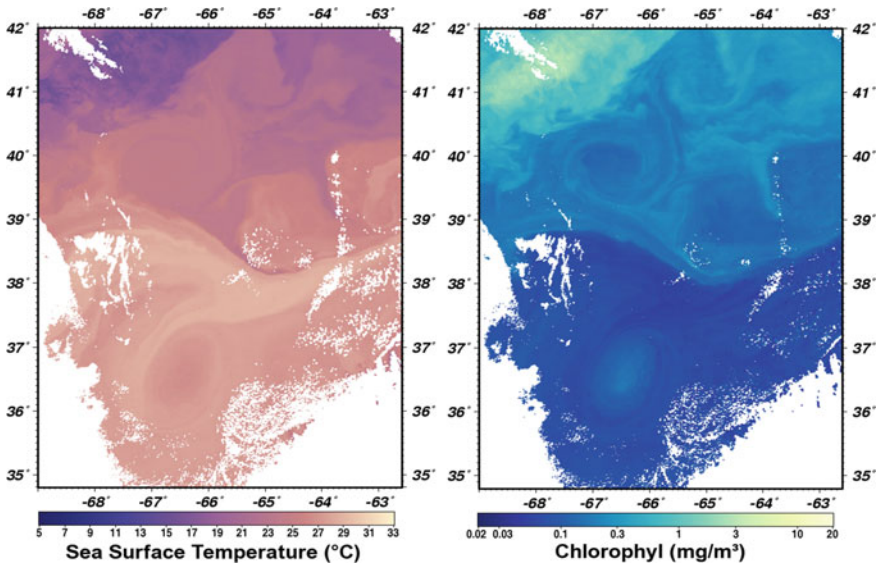

Fig. 1 Image of Sea Surface Temperature and Chlorophyl (Courtesy of NASA for research and educational use, oceancolor.gsfc.nasa.gov)

quantities are mainly due to advection from ocean currents. The evolution of images thus contains information about ocean currents, fronts and eddies. This information can be either extracted or exploited through the assimilation of image sequences, in order to provide more accurate analyses and forecasts of the circulation. There are areas where in-situ observations cannot be collected, and only satellite observations are available. Sometimes traditional SST and SSH are not enough to constrain the circulation, and velocity observations are not available either. Satellite images in this case will contain information that can be used to correct model forecasts via data assimilation.

Ocean images from satellites are abundant and not exploited for dynamical correction of the ocean circulation. Several satellites exist that provide frequent imagery of the ocean, in different locations around the globe. They are available up to the coast, where SSH is not reliable and in-situ observations cannot be sampled. The assimilation of sequences of images will improve the analysis and forecast of ocean currents, fronts and eddies, a gap that has been identified in the US Navy operations.

Prior studies show that velocity fields can be extracted from a sequence of images. Methods of extraction ranges from the particle image velocimetry (Adrian 1991) to the assimilation of images into simple models describing the image motion (Herlin et al. 2006). Velocity derived by such methods can be used as observation of the velocity; that is the case of cloud motion vector used in data assimilation for atmospheric models (Schmetz et al. 1993). The drawback of such approach is that the model considered for the image evolution is totally decoupled from the underlying physical process, thus yield unrealistic velocities. The inferred velocity is usually not accurate because the process of inferring the velocity is not constrained, i.e., is detached from the dynamics of the modeled fluid.

Other studies show that images can be successfully assimilated directly into models like the shallow water coupled with the image evolution model (Titaud 2009; Souopgui 2010). Because of the density of velocity generated by such method, the combination of images or velocity extracted from images, directly with observations of other variables of the model (temperature, surface elevation, etc.) requires the construction of cross-covariance between variables, or a data assimilation method that can inherently handle the cross-covariances.

2 Images Source and Processing: The Ocean Example

At the present time the earth is permanently observed by a large number of satellites in several wavelengths. For the ocean, quantities of interest captured by satellites are sea surface temperature (SST), sea surface height (SSH), ocean color and other quantities in the satellite visible bands, e.g. Fig. 1.

Note that images of the ocean are two dimensional signatures of three dimensional phenomena. These images are basically a set of pixels. However, they contain visible structures such as fronts, vortex and singularities. The information in the image is transported by these structures. Therefore it can be considered from two viewpoints:

the Eulerian viewpoint in which the evolution of the flow is described from a fixed frame, and the Lagrangian viewpoint in which the description is follows the evolution of the flow). A difficulty arises when a flow may have both an Eulerian and a Lagrangian character. Such is the case for Lenticularis clouds in the atmosphere: they look almost steady state, but in reality they are the signature of a strong wind. If the winds where estimated from the displacement of the lenticularis cloud then the result will be far away from the truth. Cloud are particles of water; they are gaseous phase near the ground, in liquid phase when they lift up and become visible, then turn back to vapor/gaseous state again during the descent. As a consequence, it is necessary to take into account the physics of "what is seen". This means that images should be assimilated directly, not the effects seen in the images. It follows that sequences of images should be considered for assimilation rather than individual images, as the latter contain very limited information about the underlying dynamics. Another reason to consider sequences of images is that they contain information about the dynamics of discontinuities such as fronts and singularities. Single or flat images do not provide such information. It is thus important to isolate or extract the discontinuities and define them in a functional space with adequate topology that allows for variational calculus in conjunction with the dynamical model at hand. Care should be taken so that the topology is not too regularizing, otherwise the information in the discontinuities will be lost.

For The particular case of the ocean, images can be obtained from visible-band imagery at high temporal frequency or geosynchronous, and at high horizontal resolution from meteorological satellites. Examples of existing such satellites are listed in Table 1.

The GOES-R series satellites are geostationary meteorological observation platforms and the Advanced Baseline Imager (ABI) is the primary instrument on board. ABI is a passive imaging radiometer with spectral bands from the visible through the infrared. The horizontal spatial resolution is 500 m to 2 km and the observation frequency is 5 min in conus mode (GOES-East) and as high as 30 seconds in mesoscale mode. Despite this very high temporal frequency, the obstacle for utilization of these data for oceanographic applications has been that the ABI contains only

Table 1 examples of existing meteorological satellites with visible bands that provide ocean images at very high spatial and temporal resolution

	GOES	MeteoSat	Himawari	Geo-Kompsat
Visible bands (nm)	470, 640	600, 800	470, 510, 640, 860	470, 509, 639, 863
Frequency (mins)	5	15	2.5–10	<10
Resolution (km)	0.5 (conus mode)	3	0.5–1	0.5–1
Imager	Advanced baseline imager	SEVIRI	Advanced Himawari imager	Advanced meteorological imager
Owner	US	European	Japanese	South Korean

two very broad bands in the visible (470 and 640 nm); the visible bands are not designed for detecting the comparatively weak radiant signal emanating from the surface ocean, also known as the water-leaving radiance (Lw).

For many coastal scenes, however, these limitations may be overcome by new techniques that convolve the GOES-ABI visible band data with information from coincident and dedicated ocean color radiometers (Visible and Infrared Imaging Suite [VIIRS], and the Ocean and Land Colour Imager [OLCI], Jolliff et al. 2019). These sensors are on board polar-orbiting satellites (NPP, NOAA-20, Sentinel 3A/B) and thus provide only 1 image per day (barring any cloud cover), yet, the calibration/validation activities that support these sensors enable very precise determination of Lw across the visible. When this information is used to post-process GOES-ABI data for coastal scenes, unprecedented details on coastal circulation are immediately evident in the true color image sequences. Particularly conspicuous in the high temporal frequency images (every 5 min) is the movement of turbidity plumes emanating from rivers and estuaries as well as the frontal boundaries between turbid shelf waters and the open ocean. Previous studies have shown that it is feasible to extract ocean surface velocity estimates from ocean color image sequences (Yang et al. 2015), but the major obstacle to any pragmatic application has been that without very high temporal frequency (O ~minutes) these estimates are prone to significant errors. Ideally, the frequency of color-enhanced GOES-ABI image sequences is more than sufficient to overcome this obstacle. Yet, aerosol correction, ABI signal noise, and other issues remain to be addressed and require a dedicated research effort in order to exploit the full oceanographic potential of GOES-R datasets.

3 Methods for Image Assimilation and Their Limitations

Data assimilation is the process that minimizes any discrepancies between the observed and modeled phenomena. It requires a direct relationship between the observed and modeled: the model must have variables that relate to the observed. The assimilation of images can then be classified as indirect or direct. In the indirect assimilation, observations are transformed into model variables counterparts, e.g., radiances to temperature or images to atmospheric motion vectors (AMV). In the direct assimilation, the model variables are transformed into observations or a common transformation (into the same metric space) is applied to both the model variables and the images so that they can be compared.

3.1 Indirect Assimilation of Image

Velocities are first estimated from the evolution of images, then they are assimilated as regular observations. As stated above in the introduction, the drawbacks of this approach are that the model considered for the image evolution is totally decoupled

from the underlying physical process, thus yields unrealistic velocities with large observation errors in addition to being correlated. Extracting velocities assume linear dynamics from frame to frame, different from the modeled dynamics. The inferred velocity is usually not accurate because the process of inferring the velocity is not constrained, i.e., is detached from the dynamics of the modeled fluid. In the general case, transforming observations into model counterparts is an ill-posed problem (image sequence to velocity, radiances to temperature). It should thus be avoided, especially in the case of images, since they are two dimensional signatures of three dimensional phenomena, and the underlying physics and dynamics are unknown.

3.2 Direct Assimilation of Images

In the direct image assimilation, no attempt is made to extract the equivalent of model variable; rather a well-suited mathematical space of image is chosen or defined and the calculus of variation is carried out in that space. The question here is how to define such a space. There are three main difficulties in this process. The first is the definition of the space of images, i.e. what is it that is really "seen" in the images: discontinuities, fronts, vortex or singularities? The second difficulty is that images have to be defined in a metric space so that the usual optimization procedures applied to the assimilation or regular observations can also be carried out for the images. The third difficulty is that the observed-modeled relationship that is fundamental to data assimilation requires pseudo or modeled images from which the discrepancies to the observed images are computed. The latter is an expansion of the dynamical model that now includes a component simulating the image evolution.

3.2.1 Mathematical Spaces for Images

Images are a two-dimensional array of pixels. Dynamic information seen in a sequence of images are located in discontinuities and their evolution. For that reason the consideration of an image as an array of pixels is not appropriate for image assimilation; this is confirmed by prior studies (Titaud et al. 2010; Souopgui 2010) and illustrated by Fig. 2, which compares the image assimilation in the pixels and other spaces.

The first clue in the definition of a mathematical space for image is the isolation of discontinuities, which is a pre-processing stage for images. Titaud et al. (2010) defines the space associated with discontinuities in the image as the "space of structures." Discontinuities are well characterized in spectral spaces using familiar tools such as the Fourier, wavelet or curvelet transformations. An example of curvelets is shown in Fig. 3. Another candidate in this category is the levelset method. The assimilation of images then requires two additional operators: the image-to-structure operator and the model-to-structure operator. The first operator converts the images from their original space given by the array of pixels to the space of structures, and the second

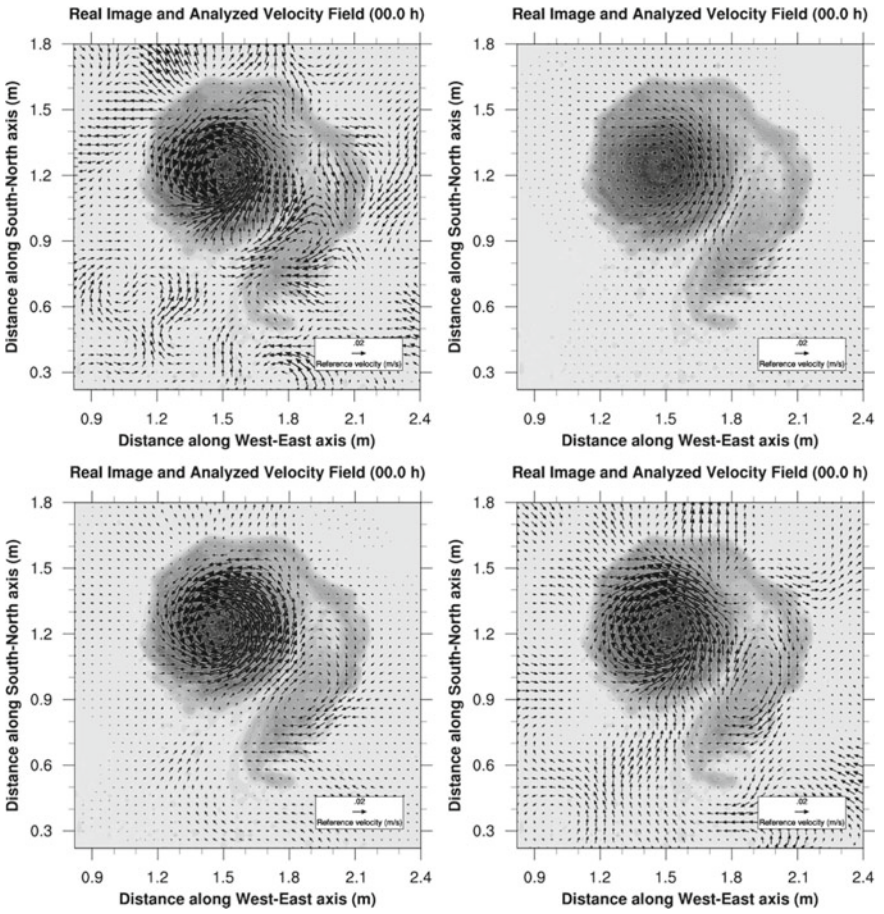

Fig. 2 Analyzed initial velocity field computed by direct image sequence assimilation with different image observation operators: Identity operator (top left); curvelet decomposition and hard thresholding (top right); curvelet decomposition and scale by scale thresholding (bottom left); curvelet decomposition and hard thresholding zeroing coarsest scale (bottom right)

operator converts the model solution to the space of structures. These two operators enable the computation of the image innovations, i.e. the discrepancy between the observed and modeled images, to be minimized in the cost function.

On the model side, the literature identifies three methods to define the model-to-structure operator: advection of passive tracer, advection of structures and Lyapunov exponents. The method of advection of passive tracer extends the model state to include a passive tracer that is advected by the model velocity and its concentration defines the model counterpart of the image. The image-to-structures operator is then used as the observation operator for image observation. The method of advection of

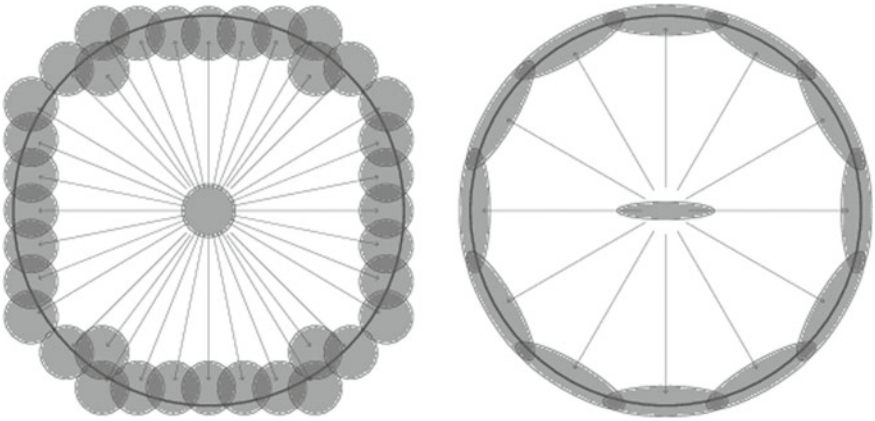

Fig. 3 Schematic view of a single scale approximation of a circle with multiscale decomposition wavelet (left) and curvelet (right)

structures extends the model state with structures of interest and advects those structures with the model velocity. The Lyapunov exponents method defines Lagrangian coherent structures (LCS, Haller, 2015) as the structures in the model and compare them to the structures in the images. In the first two methods, the advection defines the image model and the velocity field provides a coupling between the ocean model and the image model. Advection is not only the coupling mechanism between image propagation and ocean model; it is the dominant dynamical driver (Ren et al. 2011) on the short time scales between consecutive images in a sequence of high temporal frequency. In general, the image model is assumed to be two-dimensional because images are assumed to be of the surface of the ocean.

The assimilation process requires the observation operator and its transpose, so it is important to limit the degree of complexity and nonlinearity in the observation operator as much as possible. Lyapunov exponents and Level sets transformations are complex and nonlinear, and as such present a challenge for the transposition. On the other hand, wavelets and curvelets define a linear transformation and are discussed below. For more information on the Lyapunov exponents in image assimilation, see (Titaud et al. 2011; Le Dimet et al. 2015). Figures 4 and 5 show the potential of the backward Finite-Time Lyapunov Exponents (BFTLE) and the backward Finite-Time Lyapunov Vector (BFTLV) fields (Haller 2001) in extracting structures that are comparable to those present in images.

3.2.2 Multiscale Analysis of Images: Curvelets

Recent years have seen a rapid development of new tools for harmonic analysis. For geophysical flows, there are coherent structures evolving in an incoherent random

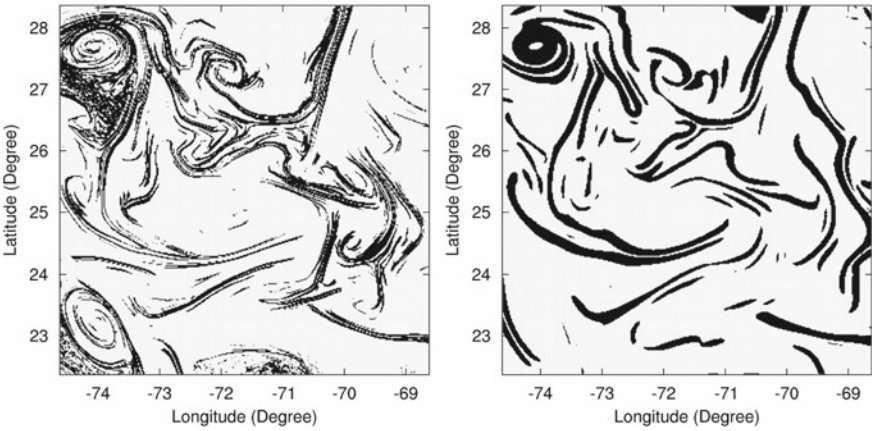

Fig. 4 Backward FTLE (day^{-1} (left) and corresponding Backward FTLV orientations (angular degree) (right) computed from the surface velocity of a simulation of the North Atlantic Ocean

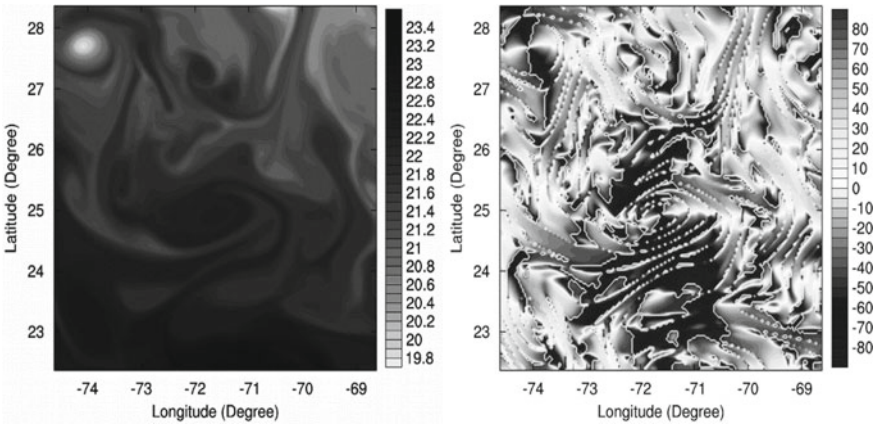

Fig. 5 Sea Surface Temperature field (left) and the corresponding orientations (angular degree) of the gradients (right)

background. If the flow is considered as an ensemble of structures, then the geometrical representation of flow structures might seem to be restricted to a well-defined set of curves along the singularities in the data. The first step in using images as observations in data assimilation is to separate the resolved structures, which are large, coherent and energetic, from the unresolved ones, which are supposed to be small, incoherent and bearing little energy. One of the first studies in this sense (Farge 1992) shows that the coherent flow component is highly concentrated in wavelet space. Wavelet analysis is a particular space-scale representation of signals which in the last few years has found a wide range of applications in physics, signal processing and applied mathematics. The literature is rich regarding wavelets (Mallat

1989; Coifman 1990; Cohen 1992) for example. A major inconvenience of wavelets is that they tend to ignore the geometric properties of the structure and do not account for the regularity of edges. This issue is addressed by the curvelet transform. The curvelet transform is a multiscale directional transform that allows an almost optimal nonadaptive sparse representation of objects with edges (Candès and Donoho 2004, 2005a, b; Candès et al. 2006). In R^2, the curvelet transform allows an optimal representation of structures with C^2-singularities. As curvelets are anisotropic, e.g. Fig. 3, they have a high directional sensitivity and are very efficient in representing vortex edges.

A function $f \in L\left(R^2\right)$ is expressed in terms of curvelets as follows:

$$f = \sum_{j,k,l} \langle f, \psi_{j,l,k} \rangle \psi_{l,j,k} \tag{1}$$

where $\psi_{l,j,k}$ is the curvelet function at scale j, orientation l and spatial position k ($k = (k_1, k_2)$). The orientation parameter is the one that makes the major difference with the wavelet transform. The set of curvelet functions $\psi_{l,j,k}$ does not form an orthonormal basis as it is the case for some families of wavelets. However, the curvelet transform satisfies the Parseval relation so that the L_2-norm of the function f is given by:

$$\|f\|^2 = \sum_{j,k,l} c_{l,j,k} \tag{2}$$

where $c_{l,j,k} = \langle f, \psi_{j,l,k} \rangle$ are the curvelet coefficients.

Figure 2 shows an illustrative comparison of the approximation of a circle by wavelets and by curvelets. The curvelets provide a better approximation of this perfectly anisotropic object. The convergence of curvelets is also better: the best m-term approximation f_m of a function f_m has the representation error

$$\|f - f_m\| \approx m^{-1}$$

for wavelets and

$$\|f - f_m\| \approx C m^{-1} (\ln m)^3$$

for curvelets.

Another interesting property of curvelets in the framework of variational data assimilation is that the adjoint of the curvelet transform is the inverse of the curvelet transform. Therefore, to represent an image, we will consider the truncation of its curvelet development.

3.2.3 Level-Set

The use of the level-set theory has also been proposed for assimilating the information contained in images (Li et al. 2017). Two-dimensional shapes of features (eddies, oil slick) on the ocean surface can be represented by a subdomain Ω whose boundary is defined by the zero level-set of the mapping $\phi : R^2 \rightarrow R$

$$\forall x \in R^2, \begin{cases} \phi(x) < 0, \ x \in \Omega \\ \phi(x) = 0, \ x \in \partial\Omega \\ \phi(x) > 0, \ x \notin \bar{\Omega} \end{cases} \tag{3}$$

with the inclusion of time, the function $\phi(t, x)$ defines the evolution of the shape as advected by a velocity field $v(t, x)$, following the advection–diffusion equation.

$$\frac{\partial \phi}{\partial t} + v.\nabla\phi - \kappa\Delta\phi = 0. \tag{4}$$

The evolution of the subdomain Ω is thus equivalent to the evolution of a concentration given an advecting velocity field. The initial condition for this equation can either be a control variable, of which the first guess is obtained by extracting the shapes in the first image in a sequence, given a threshold of what can be seen in the images. The shape extraction is an "image-to-shape" process that also serves as the "shape" observation operator, i.e. what is now assimilated is the shape or set of shapes extracted from the image. The same process is applied to the evolution of the concentration that simulates what is seen in the images, providing the model-to-shape process. In this case, the discrepancy between the observed and modeled images is expressed as the discrepancy between the shapes extracted from the observed images and those from the evolved concentration. Note that the shape extraction in a nonlinear and non-differentiable process. Some modifications of the process are thus necessary so that it can be linearized and transposed as required by the formulation of the variational assimilation technique.

4 The Cost Function

Once the "image model" or "shape model" has been added to the dynamics, and the image space and the image observation operator (i.e. the relationship between the observed and modeled images) defined, a new cost function can be defined as the extension of the original cost function (for assimilating regular ocean observations) to include the discrepancies between the observed and modeled images. Images should only be assimilated in the context of the extended cost function as it is the only means of constraining the corrections from image observations to the regular ocean observations and dynamics of the ocean model.

The minimization of this cost function can be carried out with the existing algorithms for the assimilation of regular ocean observations. We note here that the use of sequential methods or filters such as the three-dimensional variational (3DVAR) or the ensemble Kalman filter (EnKF) should be avoided. They assume that observations are sampled at the analysis, thus freezing the time dimension in the observations and their underlying dynamics, which is essentially what the assimilation is seeking to extract from images. Methods such as the four-dimensional variational (4DVAR) or the ensemble Kalman smoother (EnKS) are therefore better suited for image assimilation. In the EnKS, the time-dependent cross-correlation between variables of the ocean and image models is inherent to the ensemble covariance and allows corrections from the images to propagate through the ocean model variables and vice versa. In 4dvar that adjoint of the image model allows the image corrections to flow back to the adjoint of ocean model through the adjoint velocity variables, and the ocean corrections also flow to the image model through the forward coupling by the ocean velocity field.

The literature abounds with the formal derivations and algorithms related to the minimization of the cost function, especially with 4DVAR. Those are not repeated here. For detailed formulation of the 4DVAR algorithm for the minimization of the cost function we refer the readers to an excellent academic resource, Bennett (2002) and references therein. Li et al. (2017) also contains derivations using both the physical and the tracer concentration evolution models.

Assimilation of image sequences with 4DVAR requires the implementation of the image evolution model that is coupled with the ocean circulation model through the velocity field. The lateral evolution of the image is assumed to result from advection by the lateral velocity field. The adjoint of the image evolution model will also need to be developed, and both the forward and adjoint of the image evolution model will be integrated with the existing 4DVAR assimilation system. This will enable the propagation of information from the image evolution to all other model variables through the adjoint of the momentum equation.

The 4DVAR data assimilation system of Ngodock and Carrier (2014) is based on the tangent linear and adjoint models of the Navy coastal ocean model (NCOM) (Martin 2000). As a numerical model, NCOM already has components for the evolution of tracer fields such as temperature and salinity. The inclusion of an additional tracer field for simulating the image evolution is thus straightforward in the model dynamics, as well as in the tangent linear and adjoint models. This is how the capability of the NCOM-4DVAR system can be extended to include a tracer evolution component that will be used assimilate ocean images directly, for the purpose of correcting the ocean circulation. The same extension of a 4DVAR system can be done at any research center to include image assimilation.

The assimilation of image sequences results in particular in the update of the velocity variable at high resolution and large coverage. The resulting velocity field can be validated against independent observations of surface currents, especially in coastal areas where such observations are available from high frequency. And, because velocity is correlated to other model variables through advection, the update of velocity also contributes to the update of other model variables. This results from

an implicit cross-correlation between the image model and the dynamical model variables. This cross-correlation is usually provided by the dynamics of the tangent linear and adjoint models in 4DVAR. An interesting question is whether this dynamical cross-correlation is sufficient to propagate dynamical information from the images to all other model variables, or whether additional constraints or regularization terms are necessary to ensure that the assimilation of images provides dynamically consistent corrections of other model variables besides velocity.

5 Conclusion

This paper discussed the assimilation of images, particularly in the context of 4DVAR. The latter is better suited for image assimilation because it takes into account the model dynamics and the timeliness of observations. Images can be assimilated directly or indirectly. In either case, the dynamical model needs to be extended to include an image evolution component. The study is general enough to be applied to many fields besides oceanography. We emphasized the ocean because it is poorly observed and thus can greatly benefit from the assimilation of images arising from the plethora of earth observing satellites. Image assimilation should be an integral part of the future of operational oceanography because in situ observational campaigns are costly and usually very limited in space and time, but satellites with visible bands are increasing in number, coverage and providing images with very high temporal frequency, especially in regions where in situ instruments cannot be deployed. Although images are treated as two-dimensional for the ocean surface, their assimilation within a three-dimensional ocean model yields a correction to other ocean model variables through the coupling provided by the model velocity field. Implementation of image assimilation can be straightforward for research and operational centers that already have a 4DVAR data assimilation system.

Acknowledgements This work was sponsored in part by the Office of Naval Research Program Element 62435N as part of the "Local Analysis Through Tactical Ensonification" project. This paper is NRL paper contribution number NRL/JA/7320-20-xxxx.

References

Adrian RJ (1991) Particle imaging techniques for experimental fluid mechanics. Ann Rev Fluid Mech 23:261–304

Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press

Candes E, Donoho D (2004) New tight frames of curvelets and optimal representations of objects with piecewise-C2 singularities. Comm Pure Appl Math 57:219–266

Candes E, Donoho D (2005) Continuous curvelet transform I. Resolution of the wavefront set. Appl Comput Harmon Anal 19(3):162–197

Candes E, Donoho D (2005) Continuous curvelet transform II. Discretization and frames. Appl Comput Harmon Anal 19(3):198–222
Candes E, Demanet L, Donoho D (2006) Fast discrete curvelet transforms. Multiscale Model Simul 5(3):861–899
Carrier MJ, Ngodock HE, Muscarella PA, Smith SR (2016) Impact of assimilating surface velocity observations on the model sea surface height using the NCOM-4DVAR. Mon Wea Rev 144(3):1051–1068. https://doi.org/10.1175/MWR-D-14-00285.1
Cohen A (1992) Ondelettes et traitement numérique du signal. MASSON
Coifman RR (1990) Wavelet analysis and signal processing, signal processing, Part I: Signal processing theory. Springer, pp 59–68
Fan S, Oey LY, Hamilton P (2004) Assimilation of drifter and satellite data in a model of the Northeastern Gulf of Mexico. Cont Shelf Res 24(9):1001–1013
Farge M (1992) Wavelet transforms and their applications to turbulence. Annu Rev Fluid Mech 24:395–457
Haller G (2001) Distinguished material surfaces and coherent structures in three-dimensional fluid flows. Physica D 149:248–277
Haller G (2015) Lagrangian coherent structures. Annu Rev Fluid Mech 47(1):137–162
Herlin I, Huot E, Berroir JP, Le Dimet FX, Korotaev G (2006) Estimation of a motion field on satellite images from a simplified ocean circulation model. ICIP 2006:1077–1080
Jolliff JK, Lewis MD, Ladner S, Crout RL (2019) Observing the ocean submesoscale with enhanced-color GOES-ABI visible band data. Sensors 19:3900
Le Dimet FX, Souopgui I, Titaud O, Shutyaev V, Hussaini MY (2015) Toward the assimilation of images. Nonlinear Process Geophys 22(1):15–32. https://doi.org/10.5194/npg-22-15-2015
Li L, Le Dimet FX, Ma J, Vidard A (2017) A level-set based image assimilation method: Potential applications for predicting the movement of oil spills. IEEE Trans Geosci Remote Sens, Inst Electr Electron Eng 55(11):6330–6343. https://doi.org/10.1109/TGRS.2017.2726013
Mallat SG (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans Pattern Anal Mach Intell 11(7):7674–7693
Mariano AJ, Griffa A, Ozgokmen TM, Zambianchi E (2002) Lagrangian analysis and predictability of coastal and ocean dynamics 2000. J Atmos Oceanic Technol 19(7):1114–1126
Martin P (2000) Description of the navy coastal ocean model version 1.0. NRL report NRL/FR/7322—00-9961
Muscarella PA, Carrier MJ, Ngodock HE, Smith SR (2015) Do assimilated drifter velocities improve Lagrangian predictability in an operational ocean model? Mon Wea Rev 143:1822–1832. https://doi.org/10.1175/MWR-D-14-00164.1
Ngodock HE, Carrier MJ (2014) A 4DVAR system for the navy coastal ocean model. Part I: system description and assimilation of synthetic observations in Monterey Bay. Mon Wea Rev 142(6):2085–2107. https://doi.org/10.1175/MWR-D-13-00221.1
Ngodock HE, Muscarella PA, Souopgui I, Smith SR (2015) Assimilation of HF radar observations in the Chesapeake-Delaware Bay region using the Navy coastal ocean model (NCOM) and the four-dimensional variational (4DVAR) method. In Liu Y, Kerkering H, Weisberg RH (eds) Coastal ocean observing systems. Elsevier, pp 373–391. https://doi.org/10.1016/B978-0-12-802022-7.00020-1
Nilsson JAU, Dobricic S, Pinardi N, Poulain PM, Pettenuzzo D (2012) Variational assimilation of Lagrangian trajectories in the mediterranean ocean forecasting system. Ocean Sci 8(2):249–259
Ren L, Speer K, Chassignet EP (2011) The mixed layer salinity budget and sea ice in the Southern Ocean. J Geophys Res 116:C08031. https://doi.org/10.1029/2010JC006634
Souopgui I (2010) Assimilation d'images pour les fluides géophysiques. Ph.D thesis.
Schmetz J, Holmlund K, Hoffman J, Strauss B, Mason B Gaertner V, Koch A and Van De Berg L (1993) Operational cloud-motion winds from meteosat infrared images. J Appl Meteor 32(7):1206–1225https://doi.org/10.1175/1520-0450(1993)032<1206:OCMWFM>2.0.CO;2
Taillandier V, Griffa A, Molcard A (2006) A variational approach for the reconstruction of regional scale Eulerian velocity fields from Lagrangian data. Ocean Model 13(1):1–24

Titaud O, Vidard A, Souopgui I, Le Dimet FX (2010) Assimilation of image sequences in numerical models. Tellus A: Dyn Meteorol Ocean 62(1):30–47. https://doi.org/10.1111/j.1600-0870.2009.00416.x

Titaud O, Brankart JM, Verron J (2011) On the use of finite-time Lyapunov exponents and vectors for direct assimilation of tracer images into ocean models. Tellus A: Dyn Meteorol Ocean 63(5):1038–1051. https://doi.org/10.1111/j.1600-0870.2011.00533.x

Yang H, Arnone R, Jolliff J (2015) Estimating advective near-surface currents from ocean color satellite images. Remote Sens Environ 158:1–14

Sensitivity Analysis in Ocean Acoustic Propagation

Hans E. Ngodock, Matthew J. Carrier, Josette Fabre, Robert Zingarelli, Scott Smith, and Innocent Souopgui

Abstract The sensitivity of acoustic pressure to sound speed is investigated through the application of adjoint-based sensitivity analysis using an acoustic propagation model. The sensitivity analysis is extended to temperature and salinity, by deriving the adjoint of the sound polynomial function of temperature and salinity. Numerical experiments using a range dependent model are carried out in a deep and complex environment at the frequency of 300 Hz. It is shown that through the adjoint sensitivity analysis one can infer reasonable variations of sound speed, and thus temperature and salinity. Successful extension of the sensitivity of acoustic pressure to temperature and salinity implies that acoustic pressure observations in a given range-depth plane can be assimilated into an ocean model using the acoustic propagation model as the observation operator.

1 Introduction

The relationship between sound speed and ocean temperature variations has been exploited over the years through acoustic tomography. Underwater acoustic propagation depends nonlinearly on sound speed, which in turn is a nonlinear function of the ocean environment variables, namely temperature and salinity (T and S). Ocean acoustic propagation is modeled in various ways, including using parabolic equations that are generally solved in terms of acoustic pressure. This study investigates the sensitivity of acoustic pressure to temperature and salinity, i.e. how changes in

H. E. Ngodock (✉) · M. J. Carrier · S. Smith
The Naval Research Laboratory Code 7320, 1009, Stennis Space Center, Balch Blvd, MS 39529, USA
e-mail: Hans.Ngodock@nrlssc.navy.mil

J. Fabre · R. Zingarelli
The Naval Research Laboratory Code 7180, 1005, Stennis Space Center, Balch Blvd, MS 39529, USA

I. Souopgui
Department of Marine Sciences, The University of Southern Mississippi, Stennis Space Center, Balch Blvd, MS 102039529, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_16

the latter effect the former, or equivalently how changes in the former can be linearly related to changes in the latter.

Sensitivity analysis can be carried out in two ways: (1) the direct sensitivity analysis method, which analyzes perturbations to the solutions of the acoustic model resulting from perturbations of the temperature and salinity, and (2) the indirect sensitivity method based on the adjoint of the acoustic model. Direct sensitivity analysis is straightforward, but becomes tedious and burdensome when the dimension of the fields or parameters to perturb are large, because a large number of simulations has to be carried out, Lermusiaux et al. (2010). In contrast, adjoint-based sensitivity analysis (Cacuci 1981) only requires a single solution of the adjoint model, when it is available, driven by the derivative of the response function with respect to the prognostic state variables of the modeled phenomenon, Hall (1986), Hall and Cacuci (1983), Hall et al. (1982).

Adjoint modeling in underwater acoustics is mainly used for geoacoustic inversion, Hursky et al. (2004), Meyer and Hermand (2005), Applications of adjoint modeling for sensitivity analysis include the works of Skarsoulis and Cornuelle (2004) who used the adjoint method to compute sensitivity of the travel times in ocean acoustic tomography, who used the adjoint model to compute the derivatives of a waveguide field with respect to several parameters including the sound speed, density and frequency. In this study the adjoint sensitivity analysis is extended from the sensitivity to sound speed back to the sensitivity to temperature and salinity, using both the adjoint of the parabolic equation and the adjoint of the function that relates temperature and salinity to sound speed. For that extension to occur, the groundwork of computing the sensitivity of acoustic pressure to sound speed must first be laid, because the extension is straightforward through the chain rule. Thus, there is a greater emphasis on the derivation and computation of the sensitivity of acoustic pressure to sound speed.

An adjoint model for the range dependent acoustic model (RAM) was developed for the assimilation of acoustic pressure observations, Ngodock et al. (2017). For the sensitivity analysis in the present study, a tangent linear and adjoint of the polynomial function that relate temperature and salinity to sound speed (Chen and Millero 1977) were derived analytically (see appendix) and implemented numerically.

2 The Model

We consider the range-dependent model (RAM) of Collins et al. (1996) which is derived from the reduced wave equation in cylindrical coordinates with a harmonic point source, removing the factor $r^{-1/2}$ from the complex pressure p to handle cylindrical spreading, and assuming azimuthal symmetry to obtain (with a complex wave number to include attenuation)

$$\frac{\partial^2 p}{\partial r^2} + \rho \frac{\partial}{\partial z}\left(\frac{1}{\rho}\frac{\partial p}{\partial z}\right) + k^2 p = 0, \tag{1}$$

where $k = (1 + i\eta\beta)\frac{\omega}{c(r,z)}$ is the wave number, ω is the angular frequency, $c(r, z)$ is the speed of sound in range and depth, β is the attenuation coefficient and $\eta = \left(40\pi \log_{10} e\right)^{-1}$. Factoring the operator in (1) yields

$$\left(\frac{\partial}{\partial r} + ik_0(I + X)^{1/2}\right)\left(\frac{\partial}{\partial r} - ik_0(I + X)^{1/2}\right)p = 0, \tag{2}$$

with

$$X = k_0^{-2}\left(\rho\frac{\partial}{\partial z}\frac{1}{\rho}\frac{\partial}{\partial z} + \left(k^2 - k_0^2\right)I\right), \tag{3}$$

where $k_0 = \omega/c_0$, and c_0 is a representative phase speed. Assuming that outgoing energy dominates backscattered energy, (2) reduces to the outgoing wave equation

$$\frac{\partial p}{\partial r} = ik_0(I + X)^{1/2}p \tag{4}$$

The formal solution of (4) is

$$p(r + \Delta r, z) = \exp(ik_0\Delta r(I + X)^{1/2})p(r, z) \tag{5}$$

where Δr is the range step. By applying an n term rational function to approximate the exponential we have the Padé approximation

$$p(r + \Delta r, z) = \exp(ik_0\Delta r)\prod_{i=1}^{n}\left(\frac{I + \alpha_{j,n}X}{I + \beta_{j,n}X}\right)p(r, z) \tag{6}$$

where I is the identity operator, $\alpha_{j,n}$ and $\beta_{j,n}$ are pre-computed coefficients of the split-step Padé approximation for solving the original wave equation implicitly by separation of variables. The product form in Eq. (6) can also be approximated, without loss of accuracy, by the summation form

$$p(r + \Delta r, z) = \exp(ik_0\Delta r)\left(I + \sum_{j=1}^{n}\frac{\gamma_{j,n}X}{I + \beta_{j,n}X}\right)p(r, z) \tag{7}$$

as shown by Collins et al. (1996).

3 Sensitivity Analysis

For the sake of convenience the model (4) is written is the form

$$\frac{\partial p}{\partial r} = F(X)p \tag{8}$$

The form of the operator F is obvious from (4), and the dependence of (8) on temperature and salinity comes through the sound speed c via the differential operator X in (3). Small perturbations t' and s' of T/S lead to perturbations c' of c, that in turn yield perturbations p' of acoustic pressure governed the first-order Taylor's expansion of (8)

$$\frac{\partial p'}{\partial r} = F(X)p' + \left[\frac{\partial F}{\partial X}\frac{\partial X}{\partial c}c'\right]p \tag{9}$$

where c' relates to t' and s' according to (A6) below. Note that in the appendix the sound speed is denoted by U instead of c. For a given function $G(p)$ of p, the sensitivity G with respect to c deals with relating changes in G to changes in c, and because c is a function of T/S, changes in G can be related to changes in T/S by virtue of the chain rule. In sensitivity analysis $G(p)$ is commonly referred to as the response function. If G is a smooth function of p for which a derivative may readily computed, then $G'(p)p'$ is the change in G resulting from p', a change in p. We may write

$$G'(p)p' = \left(G'(p), p'\right)_p = \left\langle \nabla G, c' \right\rangle_c \tag{10}$$

where the subscripted parentheses and angled brackets represent suitable inner products in the spaces of acoustic pressure and sound speed respectively. Although G' may be computed quite easily, this is not the case if one attempts to express ∇G, since G is not an explicit function of c. In order to exhibit the linear dependence of (10) with respect to c', we introduce a convenient variable λ_p, with which we make the inner product with (9):

$$\begin{aligned}\left(\frac{\partial p'}{\partial r} - F(X)p' - \left[\frac{\partial F}{\partial X}\frac{\partial X}{\partial c}c'\right]p, \lambda_p\right)_p &= \left(-\frac{\partial \lambda_p}{\partial r} - [F(X)]^T\lambda_p, p'\right)_p \\ -\left\langle\left[\frac{\partial F}{\partial X}\frac{\partial X}{\partial c}\right]^T \lambda_p p, c'\right\rangle_c &\end{aligned} \tag{11}$$

It can be shown that if λ_p is the solution of

$$-\frac{\partial \lambda_p}{\partial r} - [F(X)]^T\lambda_p = G'(p) \tag{12}$$

then

$$\nabla G = \left[\frac{\partial F}{\partial X}\frac{\partial X}{\partial c}\right]^T \lambda_p p \tag{13}$$

The computation of ∇G in (13), based on a single solution of (12) (which is called the adjoint model), provides the linear relationship between changes in sound speed (c') and changes in G. In the particular case where G is the identity, i.e. $G(p) = p$, (13) shows how and where acoustic pressure is effected by changes in sound speed, in the depth-range domain. And, considering the dependence of sound speed on T and S, one can compute the sensitivity of acoustic pressure to T/S using the chain rule as described in the appendix. In the following numerical examples, the sensitivity of acoustic pressure to both sound speed (and T and S) is computed using the adjoint model of the range dependent PE model (RAM). Note that the derivation and development of the adjoint model of RAM are described in Ngodock et al. (2017).

4 Numerical Experiments

Numerical experiments are carried out for one simulated radial of 40 km range using a frequency of 300 Hz. The geographic coordinates (longitude and latitude) of the radial are (134.98 E, 36.28 N) and (134.66 E, 36.02 N) for the source and receiver respectively. The sensitivity described results from the integration of the adjoint model, defined as the algebraic transpose of the tangent linear model. The latter requires a model state or solution (also referred to as a background) around which the linearization is performed. The sound speed needed for the background solution is computed using temperature and salinity taken from a NCOM solution with a horizontal resolution of 3 km, and the background solution is shown in Fig. 1, for all

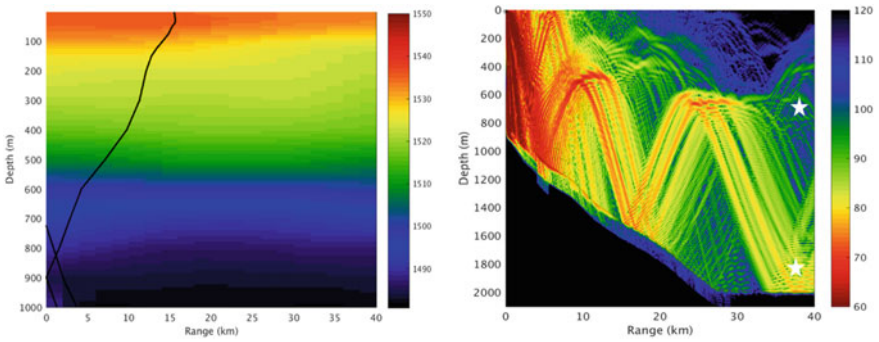

Fig. 1 Sound speed profiles along the radial (left, the black line is a sample profile), and transmission losses (in dB, right) of the acoustic pressure solution around which the model is linearized and transposed for the computation of the sensitivity

three radials, as transmission loss.

The propagation of acoustic pressure along this radial is shown in Fig. 1 in terms of transmission loss, for a source located 10 m below the surface. The sound speed profiles along the radial are also shown in Fig. 1. This case has a weak range dependent duct, the duct is about 30 m deep at the source and weakens to ~10 between 20 and 25 km, there is a strong below layer gradient near the source, that directs the energy towards the bottom, thus trapping little energy in the duct. As the duct weakens with range, the duct and gradients below the duct are strong enough to keep the bottom-bounce energy from returning to the surface. The RAM uses estimates of sound speed, attenuation and density versus depth in the sediment (an elastic bottom) and the attenuation is carried forward in complex wavenumber terms. The sediments for this downslope environment are approximately 100 m thick and slightly lossy, resulting in significant bottom bounce. The strong duct or secondary sound channel (starting around 10 km) prevents the bottom bounce energy from returning to the surface. Thus, a significant amount of acoustic energy is preserved out in range and at depth.

By definition, the adjoint model is integrated backward in range, initialized by the derivative of the response criterion (for which the sensitivity is sought) with respect to acoustic pressure at the range-depth location where the response criterion is defined. The acoustic pressure response is selected two locations of the range-depth plane that are far from the source; the locations are shown as the white stars in Fig. 1. The first location is chosen to be at the range of 35 km (range) and 1850 m (depth). As seen in Fig. 1 this region has relatively low transmission loss of 75 dB. Numerical results from the adjoint sensitivity of acoustic pressure with respect to sound speed, temperature and salinity are shown in Fig. 2. Note that the sensitivity to sound speed is recorded at the same range steps where sound speed is provided to the acoustic model, i.e. every 2 km, resulting in a significantly low spatial resolution in Fig. 2 compared to Fig. 1. It can be seen that the acoustic pressure sensitivity to sound speed is confined along the path of propagation of acoustic energy, and extends from the response region to as far as 7 km back in the range-depth domain. What is meant by path of propagation of acoustic energy is the paths within the waveguide with significant acoustic energy, which is not to be confused with "ray path". The sensitivity is mainly negative, suggesting that an increase in sound speed (along this propagation path) would result in a decrease in acoustic pressure, or, a decrease in sound speed would result in an increase in acoustic pressure. Also, the magnitude of the sensitivity, e.g. 10^{-2} $(\mathrm{m\,s^{-1}})^{-1}$, indicates that a change of 1 $\mathrm{m\,s^{-1}}$ in sound speed would cause a change of 10^{-2} in acoustic pressure.

In order to assess how reasonable these sensitivity estimates are we consider a point in the range-depth domain at 10 km and 500 m. At this location the sensitivity of acoustic pressure to sound speed is approximately –0.05, i.e. $\frac{\partial p}{\partial c} = -0.05$; likewise, the sensitivity of acoustic pressure to temperature is approximately –0.2, i.e. $\frac{\partial p}{\partial T} = -0.2$. By simple application of the chain rule, we get $\frac{\partial c}{\partial T} = \frac{\partial p}{\partial T}\left(\frac{\partial p}{\partial c}\right)^{-1} = 4ms^{-1}K^{-1}$. Thus, a change of 1 K in temperature

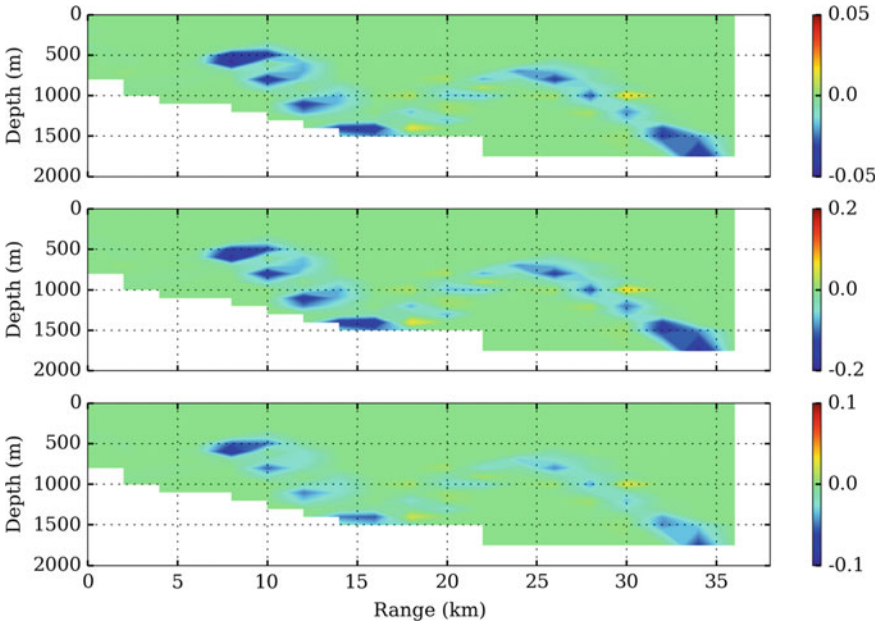

Fig. 2 Sensitivity of sound speed (top), temperature (middle) and salinity (bottom) to acoustic pressure at 35 km range and 1850 m depth

yields a change of approximately 4 m s^{-1} in sound speed, a reasonable estimate that can also be computed directly from the Chen and Millero (1977) formula, see also https://dosits.org/people-and-sound/research-ocean-physics/how-is-sound-used-to-measure-temperature-in-the-ocean/. Therefore, the adjoint of both the acoustic model and the sound speed formula provide an accurate estimate of the variations of acoustic pressure with respect to sound speed, temperature and salinity.

The spatial patterns of the sensitivity of acoustic pressure with respect to temperature and salinity are nearly identical to those of the sensitivity to sound speed. This is because the sound speed is a local function of temperature and salinity, and thus the sensitivity to the former is computed from the sensitivity to the latter. It can also be seen in Fig. 2 that acoustic pressure is more sensitive to temperature than to salinity, resulting from the gradient of sound speed with respect to temperature being greater than the same gradient with respect to salinity

A second sensitivity experiment is carried out along the same radial, with the acoustic pressure response region being at the range of 36 km and depth of 700 m. The transmission loss in this second region is about 100 dB, compared to 60 dB in the first region, and the acoustic energy follows a different path of propagation from the source to this region. Similar to the previous case, the sensitivity of acoustic

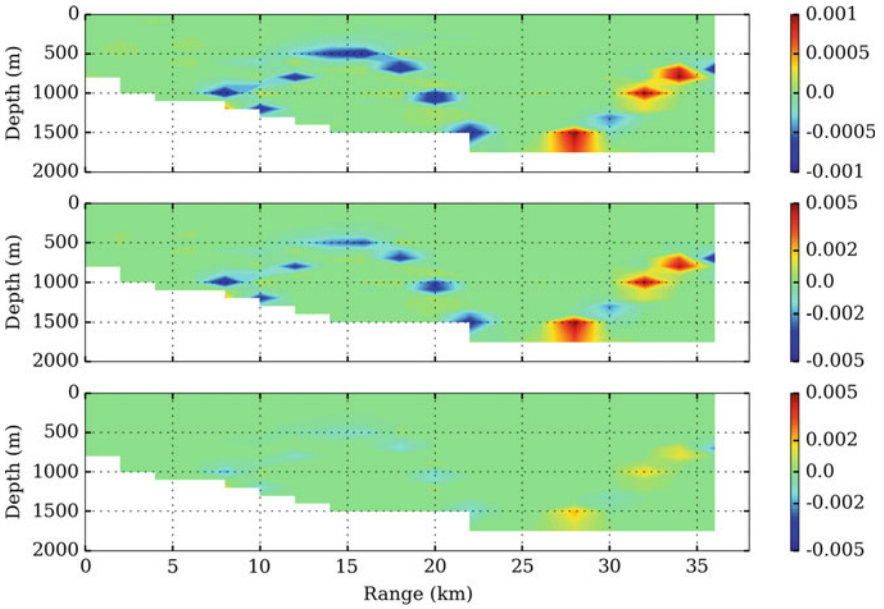

Fig. 3 Same as Fig. 2, except for the response region selected at the range of 36 km and depth of 700 m

pressure with respect to sound speed is confined along the path of propagation of acoustic energy from the source to the response region. Compared to the first case, the sensitivity in this second case is weaker (lower magnitude), a direct consequence of a weaker acoustic energy signal reaching the response region. Here also the sensitivity to temperature and salinity have the same patterns as the sensitivity to sound speed, although the sensitivity to salinity appears to be much weaker than the sensitivity to temperature, compared to what was seen in the Fig. 2 and 3.

5 Discussion and Summary

The derivation of the sensitivity of acoustic pressure to temperature and salinity through adjoint modeling enables the propagation of information from observations of the former back to the latter. Thus, in a coupled acoustics-ocean variational data assimilation system one can infer corrections to the temperature and salinity given observations of acoustic pressure. The corrections to temperature and salinity can be made throughout the depth-range domain, and not only at the locations where profiles of acoustic pressure are observed. This may be particularly useful in situations where observations of temperature and salinity are not readily available.

Simultaneous correction of the acoustic pressure and ocean environmental parameters is also possible when an ensemble of solutions of a coupled ocean-acoustic model is used in the assimilation, e.g. Lermusiaux et al. (2010), because such an ensemble contains the cross-covariance between the ocean and acoustic variables.

This study deals with theoretical derivations and numerical implementation of adjoint sensitivity analysis of acoustic pressure with respect to sound speed. By also deriving the adjoint of the equation of sound speed, the sensitivity analysis is extended to temperature and salinity. It was shown that the sensitivity is usually confined to the path of propagation of acoustic energy, and that acoustic pressure had a higher sensitivity to temperature than salinity. Also, higher sensitivity was detected in the response region that had a stronger acoustic energy signal (lower transmission loss) than in the response region with weaker acoustic energy signal (higher transmission loss). It was shown that the sensitivity computed by the adjoint model yielded accurate estimates of the variations of acoustic pressure with respect to sound speed, temperature and salinity. This study provides the ability to infer corrections to temperature and salinity in a coupled ocean-acoustic variational data assimilation system, given observations of acoustic pressure. This also implies that acoustic pressure observations can be assimilated directly into an ocean model using the acoustic model as the acoustic observation operator.

Acknowledgements This work was sponsored by the Office of Naval Research Program Element 0601153N as part of the "ADARDA" project. This paper is NRL paper contribution number NRL/BC/7320-20-5022.

Appendix: Equation of Sound Speed with Its Tangent Linear and Adjoint

The equation for the speed of sound in seawater in m s^{-1}, given by Chen and Millero (Chen and Millero 1977) is:

$$U(s,t,P) = C_w(t,P) + A(t,P)s + B(t,P)s^{\frac{3}{4}} + D(t,P)s^2 \tag{A1}$$

where s is the salinity in PSS-78, t the temperaure in °C and P the water column pressure in decibars, not to be confused with the acoustic pressure p used in the text above. A, B, C and D are temperature- and pressure-dependent parameters. The term C_w is defined as:

$$\begin{aligned} C_w(t,P) = {} & C_{00} + C_{01}t + C_{02}t^2 + C_{03}t^3 + C_{04}t^4 + C_{05}t^5 \\ & + \left(C_{10} + C_{11}t + C_{12}t^2 + C_{13}t^3 + C_{14}t^4\right)P \\ & + \left(C_{20} + C_{21}t + C_{22}t^2 + C_{23}t^3 + C_{24}t^4\right)P^2 \\ & + \left(C_{30} + C_{31}t + C_{32}t^2\right)P^3 \end{aligned} \tag{A2}$$

The term A is defined as:

$$\begin{aligned} A(t,P) = {} & A_{00} + A_{01}t + A_{02}t^2 + A_{03}t^3 + A_{04}t^4 \\ & + \left(A_{10} + A_{11}t + A_{12}t^2 + A_{13}t^3 + A_{14}t^4\right)P \\ & + \left(A_{20} + A_{21}t + A_{22}t^2 + A_{23}t^3\right)P^2 \\ & + \left(A_{30} + A_{31}t + A_{32}t^2\right)P^3 \end{aligned} \tag{A3}$$

The term B is defined as:

$$B(t,P) = B_{00} + B_{01}t + (B_{10} + B_{11}t)P \tag{A4}$$

The term D is defined as:

$$D(t,P) = D_{00} + D_{10}P \tag{A5}$$

Linearization

Note that in the derivations that follow we have neglected the variations of the water column pressure (P) with temperature and salinity. According the first order Taylor's approximation, the equations (A1)–(A5) above can be linearized as follows, with the prime symbol appended to the linearized variables:

$$\begin{aligned} & U'(s,t,P,s',t') = C_w^{'}(t,P,t') + A'(t,P,t')s + A(t,P)s' \\ & + B'(t,P,t')s^{\frac{3}{4}} + B(t,P)s^{-\frac{1}{4}}s' + 2D(t,P)ss' \end{aligned} \tag{A6}$$

$$\begin{aligned} C_w^{'}(t,P,t') = {} & \left[\left(C_{01} + 2C_{02}t + 3C_{03}t^2 + 4C_{04}t^3 + 5C_{05}t^4\right)\right. \\ & + \left(C_{11} + 2C_{12}t + 3C_{13}t^2 + 4C_{14}t^3\right)P \\ & + \left(C_{21} + 2C_{22}t + 3C_{23}t^2 + 4C_{24}t^3\right)P^2 \\ & \left. + (C_{31} + 2C_{32}t)P^3\right]t' \end{aligned} \tag{A7}$$

$$\begin{aligned} A'(t,P,t') = {} & \left[\left(A_{01} + 2A_{02}t + 3A_{03}t^2 + 4A_{04}t^3\right)\right. \\ & + \left(A_{11} + 2A_{12}t + 3A_{13}t^2 + 4A_{14}t^3\right)P \\ & + \left(A_{21} + 2A_{22}t + 3A_{23}t^2\right)P^2 \\ & \left. + \ (A_{31} + 2A_{32}t)P^3\right]t' \end{aligned} \tag{A8}$$

$$B'(t,P,t') = (B_{01} + B_{11}P)t' \tag{A9}$$

The Adjoint

In the following equation the * symbol is appended to the adjoint variables. Given the adjoint of sound speed as resulting from the adjoint of the acoustic propagation model, the adjoint variables associated to both temperature and salinity are obtained from transposing the equations (A6)–(A9) according the L2 inner product

$$\begin{aligned} s* &= \left[A(t,P) + B(t,P)s^{-\frac{1}{4}} + 2D(t,P)s\right]U* \\ B* &= s^{\frac{3}{4}}U* \\ A* &= sU* \\ C_w^* &= U* \end{aligned} \tag{A10}$$

$$\begin{aligned} t* = t* + \big[&\left(C_{01} + 2C_{02}t + 3C_{03}t^2 + 4C_{04}t^3 + 5C_{05}t^4\right) \\ &+ \left(C_{11} + 2C_{12}t + 3C_{13}t^2 + 4C_{14}t^3\right)P \\ &+ \left(C_{21} + 2C_{22}t + 3C_{23}t^2 + 4C_{24}t^3\right)P^2 \\ &+ (C_{31} + 2C_{32}t)P^3\big]C_w^*(t,P,t') \end{aligned} \tag{A11}$$

$$\begin{aligned} t* = t* + \big[&\left(A_{01} + 2A_{02}t + 3A_{03}t^2 + 4A_{04}t^3\right) \\ &+ \left(A_{11} + 2A_{12}t + 3A_{13}t^2 + 4A_{14}t^3\right)P \\ &+ \left(A_{21} + 2A_{22}t + 3A_{23}t^2\right)P^2 \\ &+ \ (A_{31} + 2A_{32}t)P^3\big]A*(t,P,t') \end{aligned} \tag{A12}$$

$$t* = t* + (B_{01} + B_{11}P)B*(t,P,t') \tag{A13}$$

The coefficients for the above terms are given in Table 1 below.

Table 1 Coefficients of the polynomials (A1)—(A5)

C	A	B	D
$C_{00} = +1402.388$	$A_{00} = +1.389$	$B_{00} = -1.922\text{E-}02$	$D_{00} = +1.727\text{E-}03$
$C_{01} = +5.03711$	$A_{01} = -1.262\text{E-}02$	$B_{01} = -4.42\text{E-}05$	
$C_{02} = -5.80852\text{E-}02$	$A_{02} = +7.164\text{E-}05$		
$C_{03} = +3.3420\text{E-}04$	$A_{03} = +2006\text{E-}06$		
$C_{04} = -1.47800\text{E-}06$	$A_{04} = -3.21\text{E-}08$		
$C_{05} = +3.1464\text{E-}09$			
$C_{10} = +0.153563$	$A_{10} = +9.4742\text{E-}05$	$B_{10} = +7.3637\text{E-}05$	$D_{10} = -7.9836\text{E-}06$
$C_{11} = +6.8982\text{E-}04$	$A_{11} = -1.2580\text{E-}05$	$B_{11} = +1.7945\text{E-}07$	
$C_{12} = -8.1788\text{E-}06$	$A_{12} = -6.4885\text{E-}08$		
$C_{13} = +1.3621\text{E-}07$	$A_{13} = +1.0507\text{E-}08$		
$C_{14} = -6.1185\text{E-}10$	$A_{14} = -2.0122\text{E-}10$		

(continued)

Table 1 (continued)

C	A	B	D
$C_{20} = +3.1260\text{E-}05$	$A_{20} = -3.9064\text{E-}07$		
$C_{21} = -1.7107\text{E-}06$	$A_{21} = +9.1041\text{E-}09$		
$C_{22} = +2.5974\text{E-}08$	$A_{22} = -1.6002\text{E-}10$		
$C_{23} = -2.5335\text{E-}10$	$A_{23} = +7.988\text{E-}12$		
$C_{24} = +1.0405\text{E-}12$			
$C_{30} = -9.7729\text{E-}09$	$A_{30} = +1.100\text{E-}10$		
$C_{31} = +3.8504\text{E-}10$	$A_{31} = +6.649\text{E-}12$		
$C_{32} = -2.3643\text{E-}12$	$A_{32} = -3.389\text{E-}13$		

References

Cacuci DG (1981) Sensitivity theory for nonlinear systems, part 1: nonlinear functional analysis approach. J Math Phys 22(12):2794–2802

Chen CT, Millero FJ (1977) Speed of sound in seawater at high pressures. J Acoust Soc Am 62(5):1129–1135

UNESCO (1983) Algorithms for computation of fundamental properties of seawater. UNESCO Tech Pap Mar Sci 44:1–55

Collins MD, Cederberg RJ, King DB, Chin-Bing SA (1996) Comparison of algorithms for solving parabolic wave equations. J Acoust Soc Am 100(1):178–182

Hall MCG (1986) Application of adjoint sensitivity theory to an atmospheric general circulation model. J Atmos Sci 43:2644–2651

Hall MCG, Cacuci DG (1983) Physical interpretation of the adjoint functions for sensitivity analysis of atmospheric models. J Atmos Sci 40:2537–2546

Hall MCG, Cacuci DG, Schlesinger ME (1982) Sensitivity analysis of a radiative-convective model by the adjoint method. J Atmos Sci 39:2038–2050

Hursky P, Porter MB, Cornuelle BD, Hodgkiss WS, Kuperman WA (2004) Adjoint modeling for acoustic inversion. J Acoust Soc Am 115(2):607–619. https://doi.org/10.1121/1.1636760115,607-619

Lermusiaux PJF, Xu J, Chen C-F, Jan S, Chiu LY, Yiing-Jang Yang Y-J (2010) Coupled ocean-acoustic prediction of transmission loss in a continental Shelfbreak region: predictive skill, uncertainty quantification, and dynamical Sensitivities. IEEE J Oceanic Eng 35(4):895–916. https://doi.org/10.1109/JOE.2010.2068611

Meyer M, Hermand JP (2005) Optimal nonlocal boundary control of the wide-angle parabolic equation for inversion of a waveguide acoustic field. J Acoust Soc Am 117(5):2937–2948. https://doi.org/10.1121/1.1880872

Ngodock H, Carrier M, Fabre J, Zingarelli R, Souopgui I (2017) A variational data assimilation system for the range dependent acoustic model using the representer method: Theoretical derivations. J Acoust Soc Am 142(1):186–194. https://doi.org/10.1121.1.4989541

Skarsoulis EK, Cornuelle BD (2004) Travel-time sensitivity kernels in ocean acoustic tomography. J Acoust Soc Am 116(1):227–238

Difficulty with Sea Surface Height Assimilation When Relying on an Unrepresentative Climatology

John J. Osborne, Matthew J. Carrier, and Hans E. Ngodock

Abstract The United States Navy has two approaches for assimilating sea surface height anomaly (SSHA) data, both relying on climatology. One approach is indirect, with the construction of synthetic temperature (T) and salinity (S) profiles based on observationally-derived climatological covariances between SSHA, T, and S. The other approach is direct via a four-dimensional variational system, but it relies on a mean SSH (here, one constrained by observational climatology) to enable comparisons between observed SSHA and model SSH. Because the approaches rely on observational climatology, they can fail when data are outside that climatology. Such a case is reviewed here. A recent field experiment (Borrione et al. 2017) collected glider T/S profiles along altimeter tracks in the Ligurian Sea (northwest Mediterranean Sea). While SSHA data are similar to observational climatology, T/S data are warmer and saltier. In this study, SSHA and T/S data are independently assimilated in separate experiments. It is found that each experiment fits its assimilated data as expected, but the experiments fail to fit the withheld/unassimilated data. Assimilation mechanisms are found to work as designed. Impacts of climatology on results versus withheld data are discussed.

1 Introduction

Sea surface height anomaly (SSHA) observations are valuable data for ocean forecasting, due to global coverage, all-weather sensing capability, and the ability to infer subsurface structure (e.g., Carnes et al. 1990; Le Traon et al. 2017). SSHA is the only data type that can constrain mesoscale circulation (Oke et al. 2009). It informs about subsurface temperature, salinity, and velocity through a variety of dynamical, statistical, and/or climatological methods (e.g., Le Traon et al. 2017). Because of SSHA's significance, new methods for assimilating SSHA that reduce analysis and forecast errors in SSHA and/or other variables are of interest.

J. J. Osborne (✉) · M. J. Carrier · H. E. Ngodock
Oceanography Division, U.S. Naval Research Laboratory, Stennis Space Center, MS, USA
e-mail: john.osborne@nrlssc.navy.mil

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_17

A recently introduced method for assimilating SSHA is the "direct" method of Ngodock et al. (2016). It is implemented for the Navy Coastal Ocean Model (NCOM; Barron et al. 2006) using four-dimensional variational assimilation (NCOM-4DVAR; Ngodock and Carrier 2014a, b). In the direct method, a mean SSH is added to SSHA to make it comparable to model SSH. Several studies (Carrier et al. 2016a, b; Smith et al. 2017; Ngodock et al. 2017) have used this method and produced low analysis and forecast errors. All of the studies contained strong, non-linear currents (Gulf Loop Current, Kuroshio, and Agulhas), so the method has been tested in dynamically complex regions. However, due to lack of independent *in situ* observations co-located with SSHA, none of the studies compared model temperature and salinity with observed temperature and salinity at the time and location of SSHA observations. Carrier et al. (2016a) used an observation system simulation experiment. It assimilated SSHA and compared experiment temperature and salinity to a nature run. Results showed low forecast errors in temperature and salinity.) Such data are now available in the Mediterranean and model-data comparisons are the focus of the work presented here.

In May and June 2016, the Long-Term Glider Mission for Environmental Characterization field experiment (LOGMEC16, Borrione et al. 2017) was conducted in the Ligurian Sea (northwestern Mediterranean Sea; Fig. 1). LOGMEC16 observations include gliders piloted within a few kilometers of satellite altimeter tracks, typically within five days of an altimeter pass (i.e., roughly co-located in time and space). This kind of data is suitable for validating the subsurface results of direct SSH assimilation. However, the observed temperature and salinity show phenomena that create a challenge for validating direct SSH assimilation.

SSHA data during LOGMEC16 are near or below the 2008–2018 monthly observational climatological means for the study area (Fig. 3; see Sect. 3 for more details on all observations). From a steric-climatological perspective, this indicates salinity near or above climatological mean and water temperatures near or below climatological mean. The glider data do show salinity near or above climatological mean (Fig. 4: raw data, Fig. 5: averages), in agreement with the steric-climatological perspective. However, glider temperatures are also near to above climatological mean (Figs. 4 and 5, raw and average, respectively), disagreeing with the steric-climatological interpretation. This presents a problem, because the NCOM-4DVAR system with direct SSH assimilation is dependent on observational climatology in two ways.

First, direct assimilation of SSH uses a mean SSH, and the mean SSH is taken from a model constrained by observational climatology (see Sect. 2 for more details). In this study, the mean SSH is taken from a multi-year experiment of the U.S. Navy's Global Ocean Forecast System, version 3.0 (GOFS; Metzger et al. 2008, 2010). This approach was also used in earlier studies using the direct SSH method (Carrier et al. 2016a, b; Smith et al. 2017; Ngodock et al. 2017). GOFS 3.0 is dependent on climatology because GOFS assimilates SSHA using the Modular Ocean Data Assimilation System (MODAS; Fox et al. 2002), which depends on observational climatology of temperature, salinity, and SSH. More specifically, MODAS transforms SSHA into profiles of temperature and salinity by assuming SSHA is due to steric variations about observational averages of temperature and salinity. The

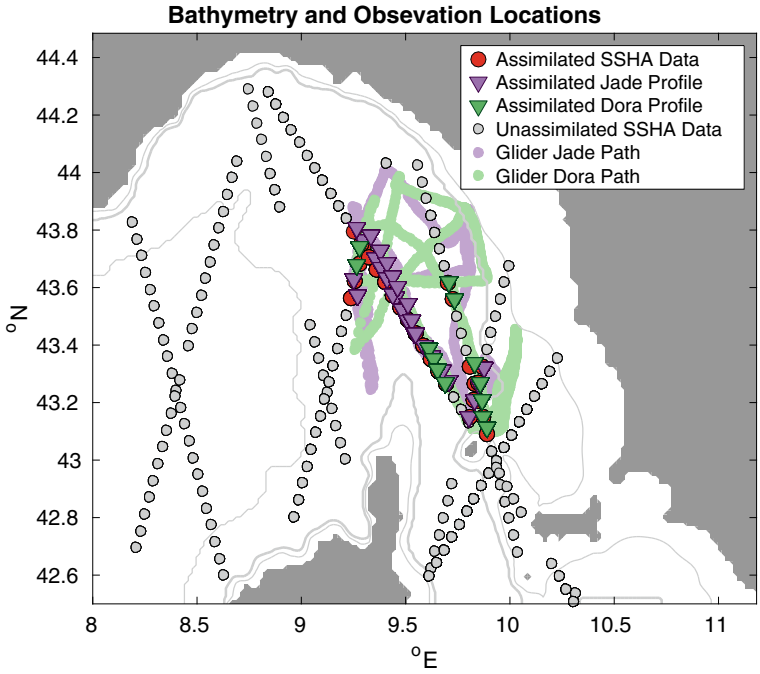

Fig. 1 Model domain with locations of SSHA data (circles; red: assimilated, gray: unassimilated) and glider profile locations (triangles: assimilated, thick lines: unassimilated) during the study period. Assimilated data is mostly in the region between 9°E and 10°E and 43°N and 44°N. Gray lines mark the 100 m, 200 m (thick) and 2000 m isobaths

temperature and salinity averages are from the Global Digital Environmental Model, version 4.0 (GDEM, Carnes et al. 2010. "Model" is a bit of a misnomer as the values in GDEM are derived from observations and observations alone; there is no "model" in the sense of discretized dynamical equations.) Because GOFS assimilates MODAS profiles, and MODAS profiles are variations about climatology, then GOFS' temperature and salinity are variations about climatology, and GOFS' SSH responds accordingly. The second dependency of NCOM-4DVAR on climatology can also be traced back to GOFS: GOFS is the source of the initial conditions and boundary conditions.

At the outset of this effort, the intent was to confirm that the direct SSH method produces an analysis whose subsurface temperature and salinity agree with observations. That is not presented here. Instead, this work examines the ability of a method dependent on climatology (the direct SSH method, dependent on the mean SSH and initial/boundary conditions from a climotologically-constrained GOFS model) to produce uncommon, far-from-climatology, in situ conditions. Unsurprisingly, there is limited ability to do this. Subsequently, the purpose of this work is to highlight the consequences of using an unrepresentative climatology in a method that depends on climatology.

This work is laid out as follows: Model and data assimilation configuration are in Sect. 2. The observations are described in Sect. 3. Experiments and results are presented in Sect. 4. Impacts of climatology are discussed in Sect. 5. Findings are summarized in Sect. 6.

2 Numerical Model: NCOM-4DVAR

2.1 Forward Model

The forward ocean model used here is the Navy Coastal Ocean Model (NCOM; Barron et al. 2006). NCOM is operationally used by the U.S. Navy for regional domains, with resolution from 3.5 km to 500 m (Rowley and Mask 2014). It is a baroclinic, Boussinesq, free surface, hydrostatic, primitive equation model. Turbulence closure is achieved with the Mellor-Yamada 2.5 scheme (Mellor and Yamada 1982). Horizontal diffusion is handled via the Smagorinsky scheme (Smagorinsky 1963). Time stepping is done with a leapfrog scheme and an Asselin filter (Asselin 1972) to prevent time splitting. The time step used here is 120 s.

The domain covers the Ligurian Sea, 8°E to 11.2°E by 42.5°N to 44.5°N (Fig. 1; 258 km by 222 km). Model resolution is 1.5 km in the horizontal (172 by 148 points). NCOM uses a hybrid σ-z coordinate system in the vertical. This domain has 25 terrain-following σ-layers above 126 m depth and 19 fixed z-levels below 126 m. Bathymetry is sourced from the U.S. Naval Research Laboratory's Digital Bathymetry Data Base, an on-going project with 2 min resolution (NRL DBDB2). Model results are output hourly.

Based on the availability of SSHA and glider data, the study period is 8 May through 10 June 2016 (additional details in Sect. 3). The model is initialized on 1 April 2016 and spun up through 7 May 2016. Initial and boundary conditions are sourced from GOFS Version 3.1 (experiment 92.7, 1/12° horizontal resolution; Metzger et al. 2017). GOFS uses HYCOM as its ocean model (Bleck 2002). The nesting ratio is greater than 3:1, and some error is to be expected in initial conditions and boundary conditions, particularly in small-scale features near the boundary. GOFS assimilates SSHA, sea surface temperature, and temperature and salinity profiles via the Navy Coupled Ocean Data Assimilation system (NCODA; Cummings 2005; Cummings and Smedstad 2014). GOFS uses NCODA with a 3DVAR technique (Daley and Baker 2001) adapted for the ocean. Boundary conditions are applied every 3 h and treated with the Flather (1976) and Chapman (1985) conditions. Tides in the Ligurian Sea are generally weak (e.g., Alberola et al. 1995; Tsimplis et al. 1995) and no tidal forcing is applied. Forcing from the Arno River is applied (Barron and Smedstad 2002). Atmospheric forcing is sourced from the Navy Global Environmental Model (NAVGEM; Hogan et al. 2014) with 0.5° resolution and applied every 3 h. Atmospheric forcing includes diurnal variability.

2.2 Data Assimilation Configuration

Data is assimilated with the NCOM-4DVAR system (Ngodock and Carrier 2014a, b), which is a weak constraint, indirect representer 4DVAR method (Bennett 1992, 2002; Chua and Bennett 2001). The system is cycled every 24 h at 0000 UTC with a 24 h forecast length. A 48 hr assimilation window (i.e., the period for the adjoint and tangent linear model to act) is used. The data window is 120 h (all prior to the analysis time). Data in the data window but prior to the assimilation window (i.e., data at 49 to 120 h before the analysis time) are shifted to the beginning of the assimilation window. Combined observation and representation error is set to 0.05 m, 0.224 °C and 0.1 PSU for SSH, temperature, and salinity respectively. For temperature and salinity, these are mean values, with larger errors near the surface and smaller errors at depth (Ngodock and Carrier 2014a). Using these errors, SSH innovations are projected in the vertical via the adjoint model. Horizontal correlation length scales are a function of the first mode Rossby radius of deformation and average 16 km in this region. During the study period, to better isolate the impact of the data of interest, only the data of interest are assimilated.

As discussed above, the direct SSH method requires a mean SSH to compare SSHA observations with model SSH. The selected mean SSH must

1. Have the same vertical datum as the study area model (lest they be dynamically unbalanced at the boundary, which then propagates into the interior) and
2. Accurately represent the response of the sea surface to circulation (including temperature, salinity, and velocity), averaged over a long period.

Regarding the second aspect, mean SSH is like mean dynamic topography. They should have similar gradients, representing the major circulation features of the ocean. They are not identical, however. The mean dynamic topography is the level of the mean sea surface (as measured by altimeters) above or below the earth's geoid (e.g., Le Traon 2011) while the mean SSH is in reference to a vertical datum (e.g., mean sea level, mean lower low water). Additionally, a mean SSH derived from a model will be limited by factors including model resolution (and parameterized processes at subgridscale resolution) and forcing (e.g. tides, surface forcing from the atmosphere, and for regional models, boundary conditions).

Here, the mean SSH used is from GOFS experiment 90.9, run from January 2011 through August 2013 (https://www.hycom.org/data/glba0pt08/expt-90pt9). The first condition on mean SSH is satisfied by construction, as both GOFS and NCOM models are built with the same bathymetry data set (DBDB2) with the same vertical datum (mean sea level). The second condition is also satisfied. GOFS results have low error against profile observations (Metzger et al. 2010, 2017). For GOFS experiment 90.9, the period is relatively short (32 months) and thus possibly unrepresentative of a long-term mean. Recently, a 22-year GOFS 3.1 re-analysis experiment became available (experiment 53.X, Helber et al. 2013, https://www.hycom.org/dataserver/gofs-3pt1/reanalysis). The mean SSH from experiment 90.9 and from experiment 53.X are shown for the study area in Figs. 2a and b. Also shown is the difference

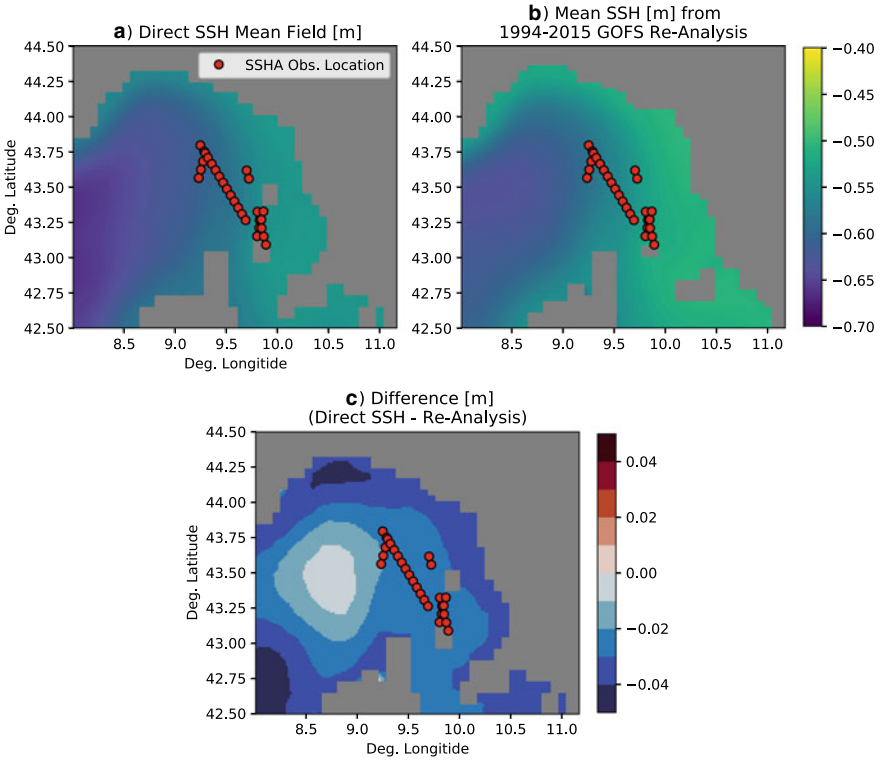

Fig. 2 **a** Mean SSH used here for the SSH Experiment, **b** Mean SSH from a 1994–2015 GOFS re-analysis experiment, and **c** the difference. Red markers indicate the position of SSHA observations

between them (Fig. 2c). The 90.9 mean is broadly similar to the 53.X mean. In particular, at the SSHA observations (red markers), where the choice of mean SSH matters, the 90.9 mean is about 2–3 cm lower than the 53.X mean. This difference is less than the combined observation and representation error for SSH (0.05 m), and thus relatively small. As will be discussed, the difference between the 90.9 mean and the 53.X mean does not qualitatively change results.

3 Observations: SSHA and Glider Temperature and Salinity

During the study period, the assimilated observations are SSHA data and in situ temperature and salinity data from underwater gliders. The gliders were mostly guided along satellite ground tracks (Fig. 1). Details on these observations are now given, including the construction of a one-to-one pairing of SSHA data and glider profiles. Observations are compared against observational climatology. It will be

shown that SSHA is near average, but observed temperature and salinity are generally greater (i.e., hotter and saltier) than observational climatology.

3.1 SSHA Observations

SSHA data are from Jason-2 (OSTM/Jason-2 Products Handbook) and SARAL/AltiKa (SARAL/AltiKa Products Handbook; Verron et al. 2015). The data are further processed by the ALtimetry Processing System (ALPS; Jacobs et al. 2002). Both altimeters are on their historic/repeat ground tracks, not moving to geodetic orbits until July 2016 for SARAL/AltiKa and October 2016 for Jason-2. SARAL/AltiKa data are available near the gliders on 16 May, 21 May, and 1 June 2016. Jason-2 data are available near the gliders on 8 May, 28 May, and 7 June 2016. While Jason-2 did pass over the Ligurian Sea on 18 May 2016, no data are available in the central part of the domain, near the gliders. Locations of SSHA data are shown in Fig. 1, with red circles for the location of assimilated data and gray circles for unassimilated data. Timing of altimeter passes is in Table 1, along with the number of assimilated observations. For assimilation purposes, the time of observations is rounded to the nearest hour.

To understand how SSHA compares against recent observational climatology, the distribution of SSHA values assimilated in this study and the distribution of SSHA from 2008 to 2018 are compared. ALPS data are available from 2008, following the last update for mean range/sea level for ALPS data. Making this comparison shows if the SSHA data are below, near, or above average. Assuming a steric relationship between SSHA, temperature, and salinity, knowledge of SSHA also indicates if temperature and salinity are below, near, or above average.

Distributions of study-period SSHA and climatology are shown by month in Fig. 3. Seasonal variability is evident. Climatological SSHA is lowest in March and highest in October. This is similar to results from studies of the entire Mediterranean (e.g., Fukumori et al. 2007; Landerer and Volkov 2013). The mean of the assimilated May 2016 observations (red) is similar to the average over May 2008–2018 (black). The mean of the assimilated June 2016 observations is lower than the 2008–2018 June

Table 1 Timing of satellite altimeter passes during the study period. Number of assimilated observations varies between assimilation-forecast cycles due glider movement (Sect. 3.3)

SSH observation date and time (UTC)	Satellite	Number of assimilated observations
8 May 2016 10:46	Jason 2	8–12
16 May 2016 17:58	SARAL/AltiKa	4
21 May 2016 04:50	SARAL/AltiKa	5–7
28 May 2016 06:43	Jason 2	2–9
1 June 2016 17:55	SARAL/AltiKa	3–4
7 June 2016 04:41	Jason 2	4

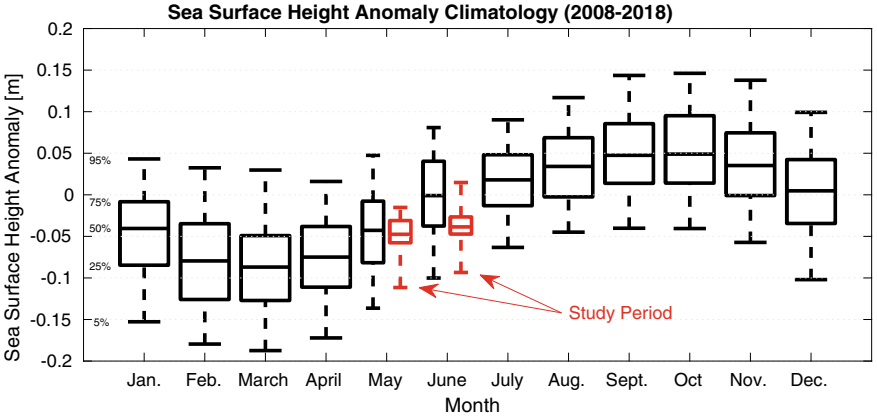

Fig. 3 Box-and-whisker plot showing the monthly climatology (2008–2018) of SSHA [m] in the study area (Fig. 2). Boxes show the range between the 25th and 75th percentile (with the 50th percentile marked inside each box) and the whiskers extend to the 5th and 95th percentiles. For May and June, the red box-and-whiskers show the range of assimilated values during the study period

climatology by about 0.05 m. Assuming the steric relationship, these SSHA suggest the following about sub-surface temperature and salinity near the SSHA data:

(i) In May, temperature and salinity will generally be near average, if perhaps slightly colder and saltier than climatology
(ii) In June, temperature and salinity will generally be colder and saltier than climatology.

These hypotheses are now checked against the glider data.

3.2 Glider Data

LOGMEC16 (Borrione et al. 2017) deployed "Slocum" gliders (e.g., Davis et al. 2003; Rudnick et al. 2004) in the Ligurian Sea from 3 May through 27 June 2016. Gliders were rated for 900 m and collected temperature and salinity data. Two gliders were deployed from 3 May through 11 June 2016 ("Dora" and "Jade") and piloted along satellite altimeter tracks (Fig. 1, green and purple markers, respectively). In Fig. 1, the location of assimilated glider profiles is shown with bold green and purple triangles and unassimilated profile locations in muted green/purple. Glider Jade was lost around 11 June, hence limiting the study period. For assimilation purposes, glider data are sorted into profiles, each profile with fixed longitude, latitude, and time. Observation times are rounded to the nearest hour, as with SSHA.

Glider-observed salinity and potential temperature are shown in Fig. 4 as functions of depth and time. Figure 5 shows mean profiles ± 1 standard deviation of

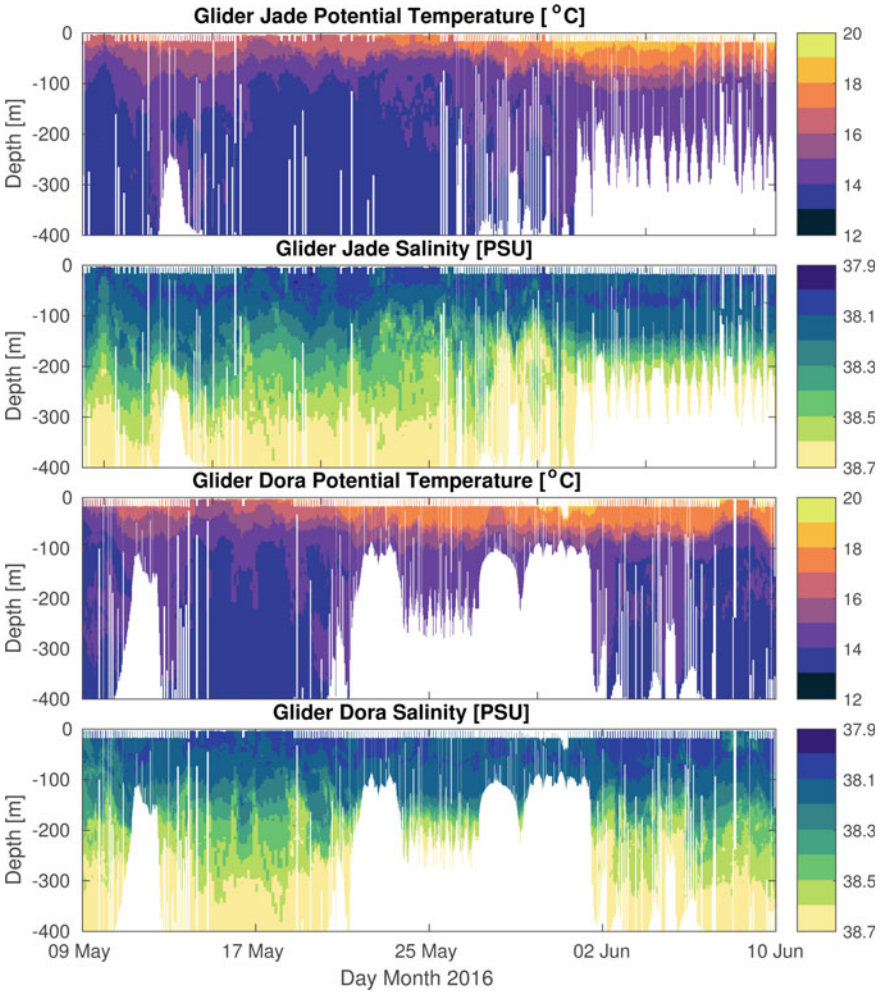

Fig. 4 Potential temperature (°C) and salinity (PSU) data from gliders Jade (top two panels) and Dora (bottom two panels). These plots included unassimilated data, which are the majority of the data. Below 400 m, there is less horizontal variability in glider observations than at shallower depths (Fig. 5) and observations are not shown

LOGMEC16 temperature and salinity (red lines and light red field) as well as GDEM4 climatological mean profiles ±1 standard deviation (black lines and gray field). GDEM4 is constructed from the U.S. Navy's Master Oceanographic Observation Data Set (Bauer 1982; Jugan and Beresford 1991), the World Ocean Database 2005 (Boyer et al. 2006), and ARGO profiles (as of 11 October 2007; https://www.usgodae.org/argo/argo.html). The three data sets contain thousands of profiles in the study area dating back to 1925. To construct the GDEM4 climatological means and standard deviations, GDEM4 is sampled at each LOGMEC16 profile location,

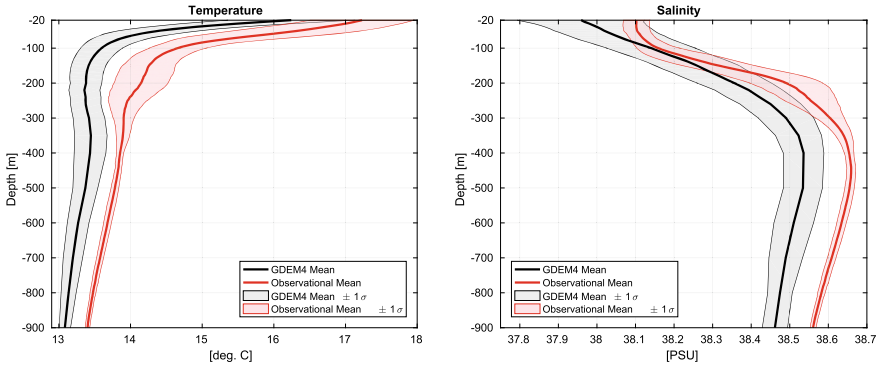

Fig. 5 Panel A: Comparison of GDEM4 climatological potential temperatures (°C) at LOGMEC16 glider locations (black line showing the mean and gray showing ±1 standard deviation) and LOGMEC16 glider temperature observations (red line showing the mean and pink showing ±1 standard deviation). Panel B: The same as A, but for salinity (PSU)

retrieving profiles of both mean and standard deviation. The profiles are then averaged. Mean standard deviation is computed as the mean of the standard deviations (not the standard deviation of the means, which have little spatial variability in this area).

The observed mean temperature is warmer than GDEM4 climatology at all depths, up to ~1 °C warmer. In terms of GDEM4 temperature standard deviation, observed mean temperature is warmer than climatology by ~2 standard deviations throughout the water column. The temperature data do not support the steric inference about temperature from SSHA, instead, the observations show below-average SSHA generally coincident with above-average temperatures.

Observed salinities are somewhat similar to GDEM4 climatology above 200 m (though observations have a sharper vertical gradient between 100 and 200 m depth). Below 300 m, observed salinities are greater than GDEM4 climatology by up to ~0.1 PSU, or up to twice the GDEM salinity standard deviation (depending on depth). The salinity data do support the steric inference about salinity from SSHA, that below-average SSHA indicates above-average salinity.

3.3 Co-location of Assimilated Data

For assimilation experiments, data sets with comparable spatio-temporal coverage are used. Doing so allows differences in results to be attributed to the type of data rather than the coverage (or some combination thereof). To create the data sets, SSHA observations and glider profiles are paired. Data pairs are within 5 km and 120 h of each other (i.e., in Fig. 1, the red circles and green and purple triangles). No SSHA observation is matched with multiple glider profiles or vice versa; there

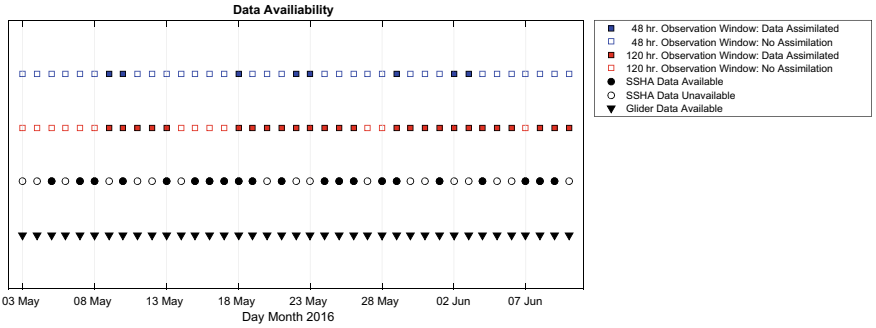

Fig. 6 Time lines indicating which days have glider data (triangles, bottom row), SSHA data (circles, second from bottom), and which days do/do not assimilate data (top two rows). Red squares show which days assimilate data based on a 5 day observation window and maximum 5 km distance between glider profiles and altimetry observations (filled: data, empty: no data). Blue squares are similar, using a 2 day observation window. Due to the greater number of days with data, the 5 day observation window is used

is a one-to-one matching. The inertial period at these latitudes is about 17.5 h and mesoscale phenomena are not expected to significantly change during the 120 h. The 5 km proximity requirement ensures that a glider profile and an SSHA observation are well within a first-mode baroclinic Rossby radius of each other (about 10 km (Escudier et al. 2013)), and thus should be representative of similar phenomena.

Based on the availability of paired data, the study period is limited to 9 May through 10 June 2016 (33 days). Figure 6 graphically represents this. Triangles (bottom row) indicate days when gliders made observations and circles (second from bottom) indicate days with SSHA observations. The filled and empty red squares (second from top) indicate which days do and do not assimilate data, respectively. On days with data, there are between 2 and 12 SSHA observations/glider profiles. Using a shorter data window greatly reduces the number of days with collocated observations. A 48 h data window would have just eight days with observations between 9 May and 10 June (blue squares, filled and unfilled, top row). Conducting assimilation experiments with so few data would not be valuable.

For some forecast cycles, the altimeter track is long enough that it takes a few days for a glider to travel the track. This changes which observations are assimilated from cycle to cycle. Figure 7 shows an example of this variability during 9–13 May. During the rest of the study period, there are several other instances of this kind of selection in assimilated data.

4 Experiment Setup and Results

In this section, results from three experiments are presented in order to understand the impact of using an unrepresentative climatology.

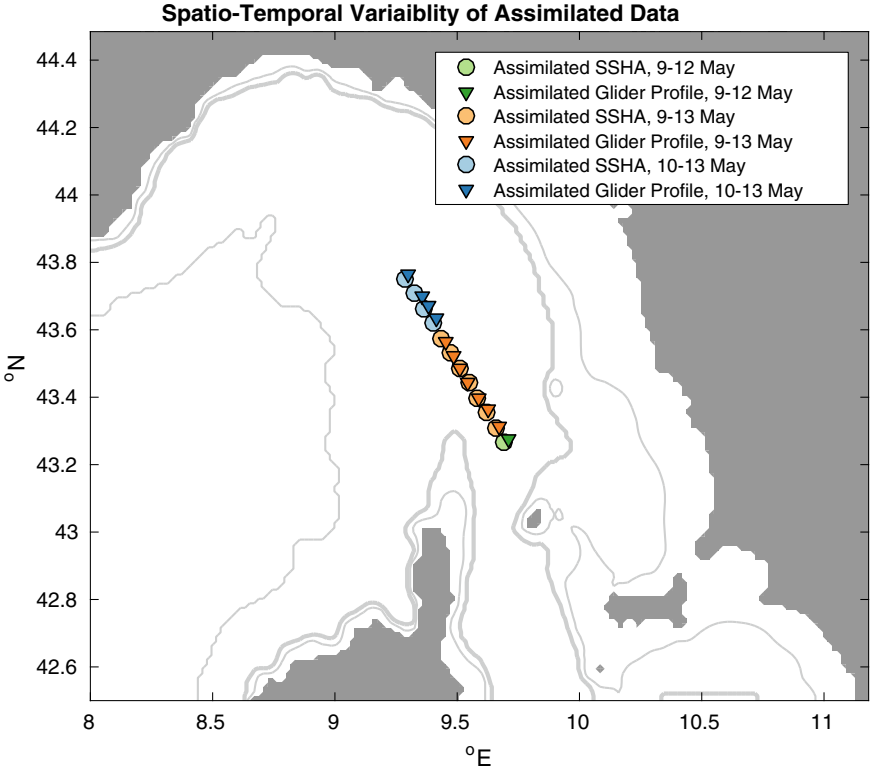

Fig. 7 An example of how assimilated data varies between forecast cycles. Jason-2 passes over the Ligurian Sea on 8 May 2016. Observation locations are marked with circles; only locations near glider data are shown. Similarly, only the location of assimilated glider profiles (triangles) near SSHA observations are shown. Glider Jade is in the vicinity of the pass on 8–10 May and moves from south to north. Subsequently, different data are assimilated on different days. The two green markers show the position of assimilated SSHA (circle) and assimilated glider (triangle) observations assimilated 9–12 May. Orange markers show observation locations of assimilated data for 9–13 May and blue markers show 10–13 May. The thick gray line marks the 200 m isobath. The thin gray lines mark the 50 m, 100 m, and 2000 m isobaths

- Experiment "SSH" assimilates SSHA data at the co-located positions and withholds glider data. This will show that, when using a method dependent on an observational climatology that is unrepresentative of current conditions, there is limited ability to produce an analysis that matches withheld data.
- Experiment "Glider" assimilates glider data at the co-located positions and withholds SSHA data. This tests if assimilation of temperature and salinity can produce model SSH that matches recovered SSH (SSHA observations plus mean SSH). This is an opportunistic experiment, with climatology's impact on the SSH Experiment being the focus.
- A "Free Run" experiment assimilates nothing and serves as a baseline.

All three experiments use the same initial and boundary conditions.

Experiments are compared against data in several ways. First, as a basic diagnostic, experiments are compared against assimilated data, to confirm that their analysis states fit assimilated data (which they do). Second, experiments are compared against withheld data. Accurately modeling unobserved or indirectly observed quantities is a goal of the ocean forecasting process. In this work, the experiments do not fit withheld data. Additional investigations are made to understand why withheld data are not fit and how climatology plays a role.

4.1 Fit to Assimilated Data

Checking if data assimilation experiments fit assimilated data is a primary, basic diagnostic. If the analysis does not fit assimilated data, this indicates issues with the assimilation system. To quantify how experiments' analysis states fit assimilated data, the "JFit" metric is used (e.g., Carrier et al. 2016a; Ngodock et al. 2016):

$$\text{JFit} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - H_i X|}{\varepsilon_i}$$

where N is the number of observations, y_i denotes the ith observation value, H the observation operator, X is the analysis state, and ε is the observation error. JFit is computed for each analysis cycle (daily). Values near 1 are expected, because the assimilation is designed to fit assimilated observations within one standard deviation of observation error.

Figure 8a shows results for the SSH Experiment (green line) and Glider Experiment (orange line) using the experiments' analysis states and assimilated data. The SSH Experiment generally fits recovered SSH with JFit between 0.5 and 1. For the Glider Experiment, JFit with respect to assimilated temperature and salinity varies over time. During the middle of the study period, JFit is near or below 1, indicating an accurate fit to the assimilated data. During the beginning and end of the experiment, the Glider Experiment has JFit between 1.5 and 2. These values are relatively large. In comparison, in Ngodock et al. (2017), JFit for analysis temperature was typically between 1 and 1.2. However, those experiments assimilated much more temperature data, including SST and all available in situ profile data (e.g., ARGO floats). Here, the small amount of data may be limiting results.

4.2 Fit to Unassimilated Data

To quantify how experiments fit unassimilated data, the mean absolute error metric is used (MAE; Willmott and Matsuura 2005; Willmott et al. 2017),

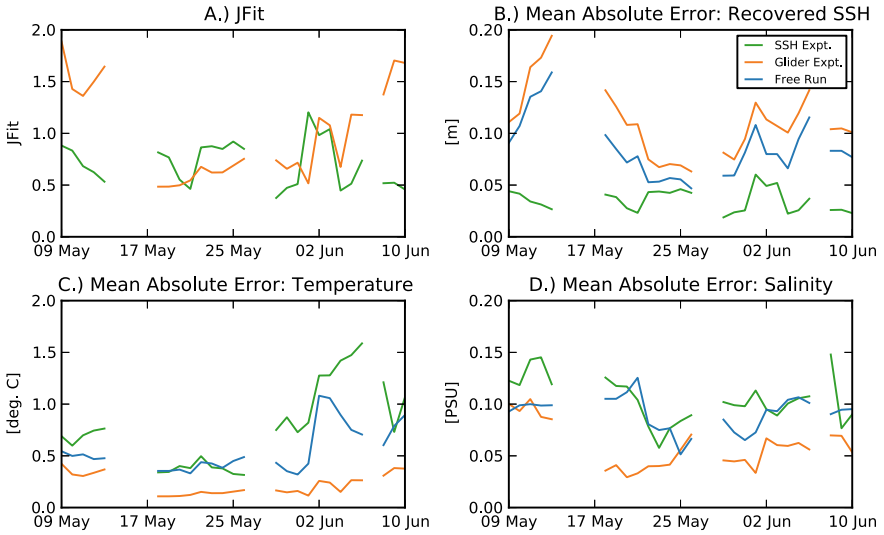

Fig. 8 Panel A (top left): JFit for SSH Experiment with respect to recovered SSH (green) and JFit for Glider Experiment with respect to temperature (orange). Panel B (top right): Mean absolute error with respect to recovered SSH for the SSH Experiment (green), Glider Experiment (orange), and Free Run (blue). Panel C (bottom left): Mean absolute error in assimilated temperature. Panel D (bottom right): Mean absolute error in assimilated salinity

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - H_i X|$$

with variables defined as above. (MAE is JFit without error normalization. Errors are not normalized because data has not been assimilated and no assumptions have been made about the errors.) Here, the purpose is to see if assimilation reduces errors relative to the Free Run, which is expected for multi-variate systems like 4DVAR. Results are shown for recovered SSH, temperature, and salinity (Figs. 8b–d, respectively). As in Fig. 8a, the green line represents the SSH Experiment and the orange line represents the Glider Experiment. The blue line represents the Free Run. Table 2 summarizes the results. For a best-case/lowest-error reference, the experiment that assimilates a given data type is also shown.

For each of the three data types, the experiment assimilating that data type has lowest average MAE during the study period (e.g., for temperature data, the Glider Experiment does best, Fig. 8c and Table 2). The second-lowest average MAE is achieved by the Free Run, regardless of data type. Finally, the experiment with highest average MAE is the experiment assimilating other data (e.g, against temperature, the SSH Experiment does worst). This shows that, for this study, assimilation of one data type does not improve errors against the withheld data type. This is unexpected, given that NCOM-4DVAR has been shown to fit withheld data (Ngodock and Carrier 2014b).

Table 2 Ranking of experiments by average MAE for each of the assimilated data types. Rank of 1 indicates lowest error; rank of 3 indicates highest error. For all three data types, the experiment assimilating that data type ranks best, the Free Run is second, and the experiment assimilating other data ranks worst

Ranking of experiments by average MAE for different data types			
Rank	Recovered SSH	Temperature (T)	Salinity (S)
1	SSH Expt., 0.04 m *Assimilates recovered SSH*	Glider Expt., 0.22 K *Assimilates T/S*	Glider Expt., 0.06 PSU *Assimilates T/S*
2	Free Run, 0.09 m *No assimilation*	Free Run, 0.55 K *No assimilation*	Free Run, 0.09 PSU *No assimilation*
3	Glider Expt., 0.11 m *Assimilates T/S, but not recovered SSH*	SSH Expt., 0.77 K *Assimilates recovered SSH, but not T/S*	SSH Expt., 0.11 PSU *Assimilates recovered SSH, but not T/S*

In two of the MAE data sets, there is somewhat similar variability between the Free Run and the experiment not assimilating the given variable. For recovered SSH MAE (Fig. 8b), the Free Run and Glider Experiment vary similarly (correlation coefficient of $r^2 = 0.94$). For temperature MAE (Fig. 8c), the Free Run and SSH Experiment correlation coefficient is $r^2 = 0.57$. The similarities in these time series may be due to errors in initial conditions, boundary conditions, atmospheric forcing, common across the experiments. However, the correlation coefficient is $r^2 = 0.20$ for the Free Run's and SSH Experiment's salinity MAE (Fig. 8d). This does not support the common error hypothesis. On the other hand, the observation error is 0.1 PSU. The Free Run's MAE is generally below that and the SSH Experiment's study-period averaged MAE is 0.11 PSU, so it is relatively low. Errors for an assimilative experiment are expected to be random, so it not surprising that the correlation coefficient is low. In other words, while the direct assimilation of SSHA data is generally unable to reduce salinity errors relative to the Free Run, the errors are relatively small to begin with.

Some value of ocean data assimilation is in the estimation of unobserved phenomena. This is expected for a multivariate system. Fitting the assimilated data with low errors is expected and achieved here. Fitting unassimilated data no better than the Free Run (true in all three comparisons here; Figs. 8b–d and Table 2) is worth investigating, given prior results (Carrier et al. 2016a, b; Smith et al. 2017; Ngodock et al. 2017). Results here must be considered in the context that the glider observations show water warmer than climatology (all observed depths) and saltier than climatology (300–900 m). Investigation continues by examining the exact nature of the misfits, the innovations, and the analysis increments.

4.3 Direct Comparison of Experiment Results Against Glider Jade Temperature and Salinity

To understand the SSH Experiment's misfits to glider data, it is compared against glider Jade data (results against glider Dora data are similar and not shown). This will show where and when errors are being produced. For additional context, the Free Run and Glider Experiments are also compared to observations. Comparisons are made by (1) interpolating experiment analysis state to the glider profile locations/times and (2) subtracting the observed temperature and salinity. Results are shown in Fig. 9.

For temperature errors, the Free Run and SSH Experiment have similar patterns (Figs. 9a, c), with the SSH Experiment having larger error magnitudes. The similar error patterns suggest a common cause, perhaps errors or bias in initial and/or boundary conditions. The two experiments are generally colder than observed throughout the water column (but not always, e.g., 17–22 May). In particular, the SSH Experiment is up to 4.0 °C colder near 50 m depth between 2 and 10 June. The Glider Experiment is usually colder than observed below 300 m (Fig. 9e), like the Free Run and SSH Experiment, but not to the same magnitude nor as consistently in time. Above 300 m, the Glider Experiment can be warmer or colder than observed. This contrasts with the Free Run and SSH Experiment, which are almost entirely colder. That all experiments, particularly the non-assimilative Free Run, are colder than observed below 300 m suggests that the initial and boundary conditions, identical across experiments, are colder than reality. Consistent with Fig. 8c, assimilation of glider temperature data generally reduces temperature errors while direct assimilation of SSHA generally increases such errors.

With respect to observed salinity, all experiments are broadly similar in error patterns but with varying magnitudes (Figs. 9b, d, f): Above 200 m depth, all are generally saltier than observed. Below 200 m depth, all are generally fresher than observed. The Glider Experiment generally has the smallest error magnitudes (and is assimilating some glider Jade data), the SSH Experiment generally has the largest error magnitudes, and the Free Run is between the two. This is consistent with results shown in Fig. 8d. Like with temperature errors, the similar error patterns between the experiments suggests also suggests errors in the initial and boundary conditions.

Collectively, these results show that the SSH Experiment tends to amplify errors in subsurface temperature and salinity. The Glider Experiment reduces such errors, which is to be expected since it is assimilates subsurface temperature and salinity.

4.4 Comparison Against Recovered SSH

Figure 9 shows that the SSH Experiment is producing temperatures that are colder than observed throughout the water column and salinities that are generally saltier than observed above 200 m and fresher than observed below 300 m. Since the temperature errors are relatively large, the innovations and increments that produce the

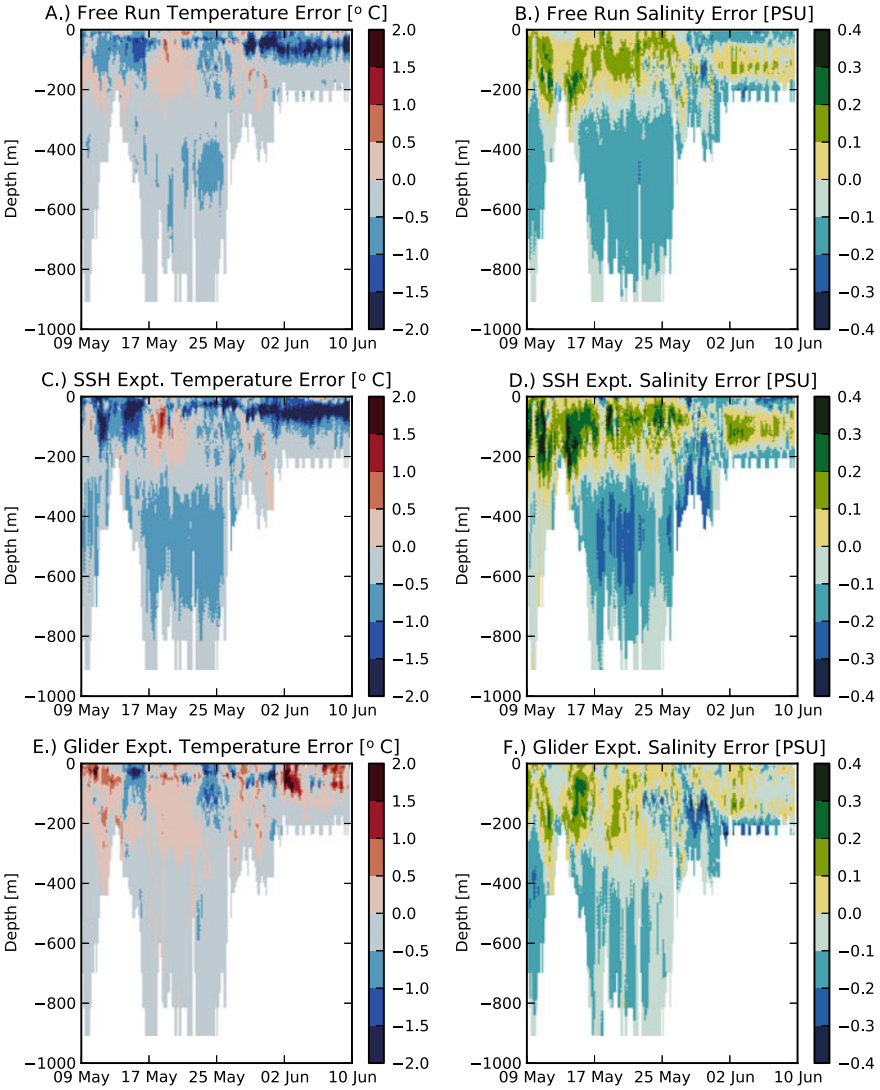

Fig. 9 Errors in experiments with respect to glider Jade observations (computed as observations less experiment); top: Free Run, middle: SSH Experiment, bottom: Glider Experiment; left: temperature, with red indicating experiment hotter than observed; right: salinity, with green indicating experiment saltier than observed)

temperature errors are now investigated. SSH innovations are considered from two perspectives: First, by examining the SSH in both the background and analysis states of a forecast cycle (Fig. 10) and, second, by direct inspection of the SSH Experiment's innovations (Fig. 11).

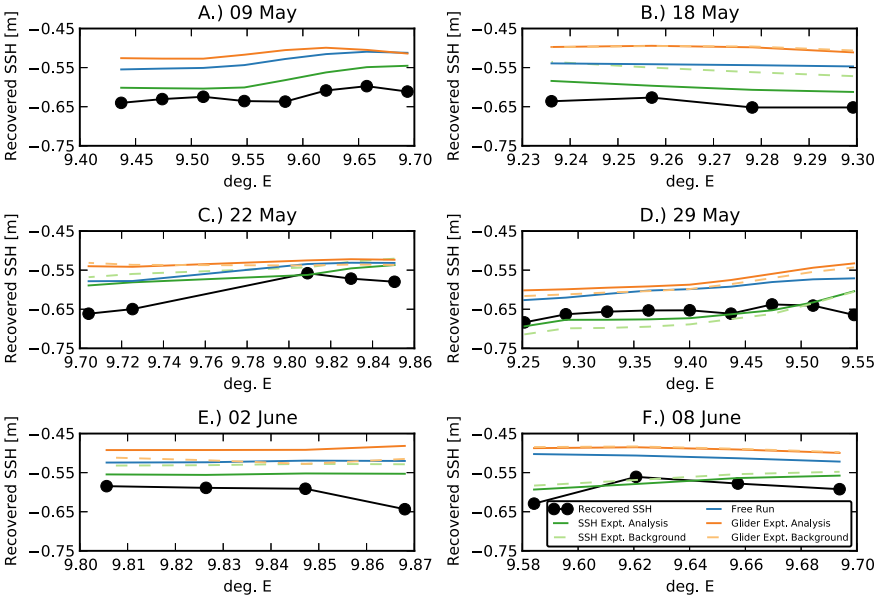

Fig. 10 For selected forecast cycles, recovered SSH (black line, with circles marking observation longitudes and values) and experiment SSH. The Free Run is shown in blue, the SSH Experiment is shown in green (solid: analysis, dashed: background), and the Glider Experiment is shown in orange (solid: analysis, dashed: background). Note that for 9 May (Fig. 10a), the Free Run is exactly the background for the assimilation experiments. Also note that for several cycles, the Glider Experiment background SSH is very similar to analysis SSH and may be difficult to visually distinguish

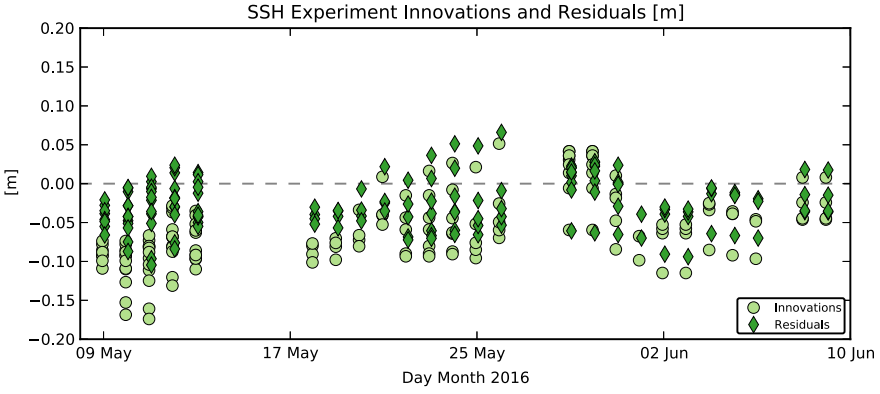

Fig. 11 SSH Experiment innovations (light green circles; calculated as recovered SSH less experiment background) and residuals (dark green diamonds; calculated as recovered SSH less experiment analysis), plotted for every forecast cycle. For visual clarity, for each forecast cycle, innovations are staggered slightly to the left and residuals slightly to the right. Negative values indicate that the experiment SSH is higher than the recovered SSH

Modeled SSH and recovered SSH are shown in Fig. 10, taken as along-track data (recovered SSH shown as black dots). For brevity, results are shown for the first time a particular track is assimilated (refer to Table 1 for timing details). For the assimilation experiments, the background state and analysis state are shown to in order to reveal the impact of assimilation on SSH (dashed line: background, solid line: analysis; colors as before). Note that for 9 May (Fig. 10a), the Free Run is the background state for the assimilation experiments and the dashed lines are not shown. Figure 11 focuses solely on the SSH Experiment and shows both innovations (light green circles) and residuals (dark green diamonds) for every forecast cycle (for visual clarity, diamonds and circles for each forecast cycle are horizontally offset).

For the SSH Experiment, Fig. 10 generally shows that its background SSH is persistently greater than recovered SSH (dashed light green line vs. solid black line) and that the analysis SSH (solid green line) is closer to the observations than the background, as expected. This is also seen in Fig. 11—the residuals are closer to zero than the innovations. For the Free Run, its SSH is always greater than recovered SSH (Fig. 10), again evidence of potential error in the initial and/or boundary conditions. For the Glider Experiment, both its analysis and background SSH are greater than observed.

To see the impact of SSH innovations on the SSH Experiment's temperature field, cross sections of temperature increments along altimeter tracks are shown in Figs. 12a–f (track locations shown in Fig. 12g). Increments are shown at the time observations are assimilated (corresponding with Fig. 10). For the 9 May analysis, all the SSH innovations are negative and the resulting along-track temperature increments are mostly negative (cf. Fig. 11 vs. Fig. 12a), following the covariance between SSH and temperature provided by the 4DVAR's linearized dynamics in the adjoint and tangent linear models. This is seen again in subsequent analysis cycles. For the 29 May analysis, the innovations are relatively small and have mixed sign (Fig. 11) and the resulting temperature increments are also relatively small and have mixed sign (Fig. 12d, observations from 28 May). For the 2 June forecast cycle, the observations on 1 June produce mostly negative innovations (Figs. 10e and 11) and there is cooling near the observations (Fig. 12f).

4.5 Comparison Against MODAS Profiles

An additional check of the direct SSH method is now made by analyzing the SSHA data with MODAS (Fox et al. 2002) which produces synthetic temperature and salinity profiles. This check is done because MODAS is in operational use by the U.S. Navy and has been tested with NCOM-4DVAR (Smith et al. 2017). MODAS is found to be capable of producing low errors in temperature and salinity (Metzger et al. 2010). Thus, this check provides additional context for the direct SSH method. MODAS profiles are computed from SSHA data used in the SSH Experiment and SST from the SSH Experiment. Figure 13a shows mean MODAS temperature profiles ± 1 standard deviation (purple). Also shown are the SSH Experiment's analysis state

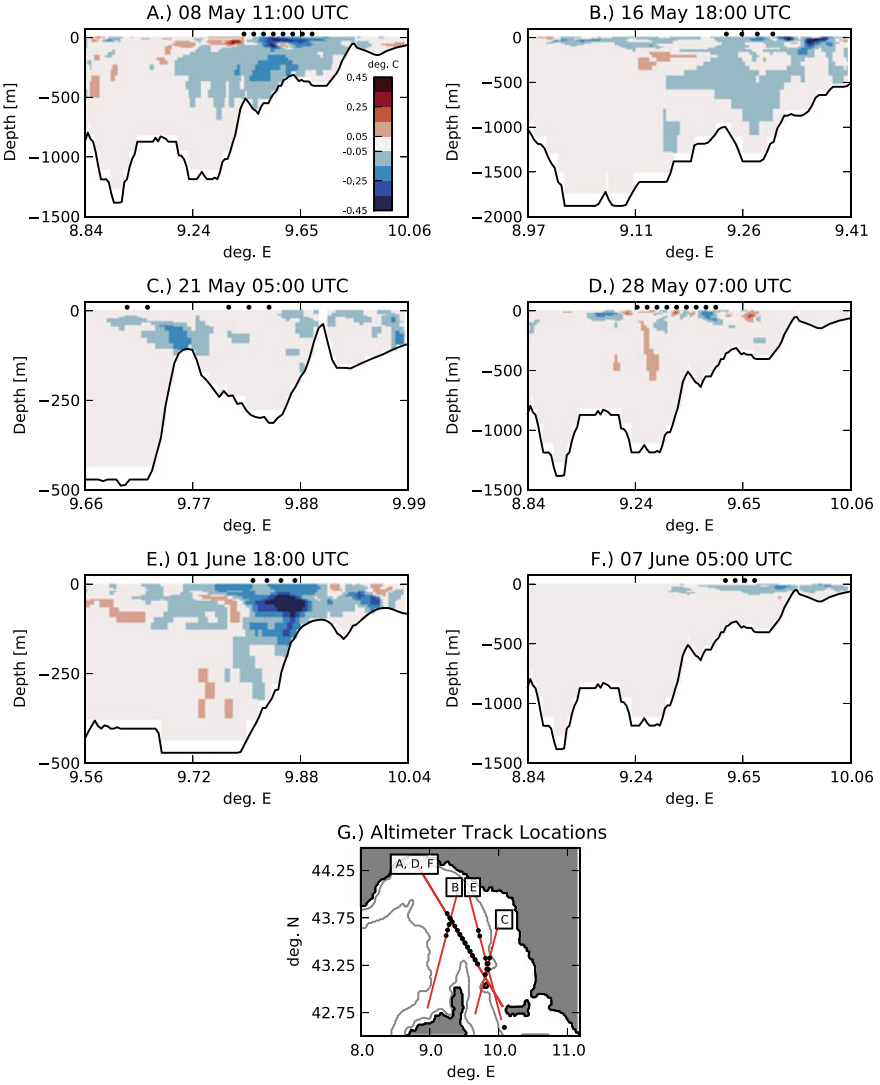

Fig. 12 Panels A–F: Cross sections of temperature increments for the SSH Experiment along altimeter passes at observation times. Blue indicates negative temperature increments and red indicates positive temperature increments. Black circles near 0 m depth mark observation location. Cross sections are viewed from the south, with the west on the left and east on the right. Note that the depth and length of each cross section varies. Panel G: Location of the altimeter track/cross sections (red lines) and SSHA observations (black circles). Cross sections are labeled with the corresponding panel label. Gray lines mark the 200 and 2000 m isobaths

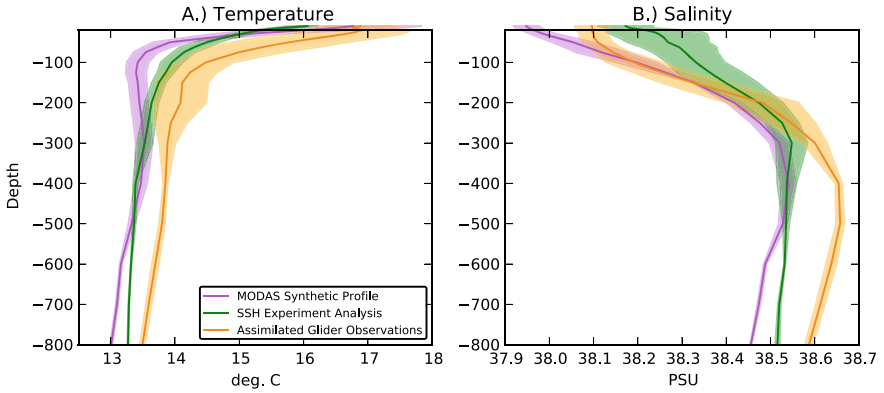

Fig. 13 Panel A: Similar to Fig. 5, a comparison of mean temperature profiles (line) ±1 standard deviation (shaded field) between MODAS synthetic profiles (purple), the SSH Experiment analysis state (green), and assimilated glider observations (orange; after vertically interpolating to the model levels). Panel B: As in Panel A, but for salinity

mean temperature profile ±1 standard deviation (green) sampled at SSHA observation locations, and assimilated glider profiles mean temperature profile ±1 standard deviation (orange). Figure 13b is similar, but for salinity. At most depths, the mean MODAS profiles are colder and fresher than the SSH Experiment analysis state. This comparison is significant in two ways. First, while the direct SSH method is erroneous with respect to the glider data, the MODAS synthetic profiles are even more erroneous (except for salinity above 200 m) and assimilation of them would produce a worse result than the SSH Experiment. Second, while the direct SSH method offers some improvement in temperature and salinity errors, both methods result in an ocean that is like climatology (Fig. 5, black lines): colder than observed (all depths) and fresher than observed (below 300 m). As MODAS depends on the GDEM4 climatology, it is unsurprising that its results are also colder and fresher than the assimilated glider data. Given that MODAS is capable of producing low temperature and salinity errors, the MODAS errors show that using a unrepresentative climatology will produce an erroneous analysis.

The preceding investigation shows how the direct assimilation of SSHA is working in the SSH Experiment: negative SSH innovations (Fig. 11) result in (generally) negative temperature increments (Fig. 12). In the resulting analysis, SSH better matches the observations than the background SSH (Figs. 10 and 11). However, the analysis state is generally colder than the glider observations (Figs. 8c and 9c) and generally increases salinity errors over the Free Run (Figs. 8d and 9b, d). For the Glider Experiment, assimilation reduces errors in temperature and salinity but does not produce an SSH that matches recovered SSH.

5 Discussion

Observational climatology may be affecting results in two ways, in influencing initial and boundary conditions and in influencing the mean SSH. Additionally, climatology is unable to capture shifts in dynamics, temporary or permanent, which may be significant for the mean SSH.

With initial and boundary condition errors, there does appear to be a cold, fresh bias, as seen in the Free Run's temperature and salinity errors versus glider data (Figs. 9a, b). The conditions come from a climatologically-constrained global model (GOFS 3.1). Using different initial and boundary conditions could provide different results, but that would not directly test the impact of climatology on the parent model—to test the impact of climatology, the parent model would need to be run with a different climatology, which is not practical. The mean SSH may be erroneous, possibly because it is from a relative brief experiment (32 months). However, it is similar to a 22-year reanalysis product (Fig. 2). If the mean SSH were changed to the re-analysis product, SSH innovations would still be negative, particularly in the first forecast cycles. They would still produce negative temperature increments, which would cool the model state and increase errors over the Free Run.

The mean SSH is derived from the same climatologically constrained global model. The global model is assimilating MODAS profiles in every cycle, and MODAS profiles are steric variations about climatological temperature and salinity in response to an input SSHA. This forces the global model's SSH about climatological mean SSH, as derived from the climatological temperature and salinity. Thus, the mean SSH is reflective of the climatology and influences every SSHA observation assimilated by the SSH Experiment.

The mean SSH may be problematic in an additional climatologically linked way. The SSH in the Mediterranean Sea varies across a range of time scales, both long-term multi-year (Calafat et al. 2012), and brief few-month (Landerer and Volkov 2013) periods. These SSH variations can be up to 10 cm apiece or combined up to 20 cm. This exceeds the difference between Free Run SSH and recovered SSH (Fig. 10) and could account for errors in the mean SSH. While a climatology can capture the combined range of these modes, it cannot recognize if conditions are in some kind of low SSH mode or high SSH mode. Therefore, what may appear to be a below-average SSHA observation can in fact be an above-average SSHA observation, if the Mediterranean is in a low SSH mode. It is unclear what, if any, SSH variations were present during the LOGMEC16 study period or in the GOFS experiments supplying the mean SSH and the initial/boundary conditions. Thus, a bias may be present between the mean SSH used here and with the "true" mean SSH during the study period. As the Mediterranean Sea exhibits SSH variability, SSHA observations may need to be considered in the context of low SSH and high SSH modes, not just a long-term mean. This could be tested by adjusting the mean SSH by the difference in SSHA observations and climatology over the entire Mediterranean, as averaged over some time interval. Such testing is beyond the scope of this study.

6 Summary

SSHA data have been assimilated using a direct SSH 4DVAR method in the Ligurian Sea. The SSHA data are co-located with temperature and salinity profiles from gliders. The assimilation method uses inputs (a mean SSH, initial and boundary conditions) taken from a climatologically constrained global model. While SSHA data are near climatological mean, observed temperature and salinity are not, being both warmer and saltier than observed.

When assimilating only SSHA data, the analysis fits the assimilated data within 1 standard deviation of observation error, which is expected. However, comparison against withheld subsurface temperature, co-located with the SSHA data, shows relatively large errors. The SSH Experiment is generally colder than observed. In fact, the SSH Experiment is colder than a non-assimilative Free Run experiment. Comparing the SSH Experiment's salinity against co-located subsurface salinity observations shows that the SSH Experiment is generally saltier than observed above 200 m depth and fresher than observed below. However, the salinity errors are of similar magnitude to the Free Run.

The process by which the SSH Experiment adjusts temperature has been investigated and found to operate in the appropriate manner. SSH innovations are generally negative throughout the study period, and this results in temperature innovations that are generally negative.

A second approach for assimilating SSHA data, MODAS, was also investigated. MODAS transforms SSHA into synthetic profiles of temperature and salinity by assuming SSHA is due to steric variations in temperature and salinity about climatology. Here, MODAS profiles are colder than direct SSH method profiles, and hence are even more erroneous relative to the observations. Given MODAS' general ability to produce accurate temperature and salinity, this shows that using an unrepresentative climatology produces an erroneous analysis.

An additional experiment has been conducted using only temperature and salinity data co-located with SSHA data. It fit assimilated temperature and salinity between 1 and 2 standard deviations of observation error. However, it did not fit withheld SSHA data. In fact, its SSH was more erroneous than the Free Run relative to observations.

The influence of climatology on the SSH Experiment results is identified in several ways. The global model that provides initial/boundary conditions and the mean SSH is climatologically constrained due its use of synthetic temperature and salinity profiles. This results in sea surface height, subsurface temperature, and subsurface salinity varying about climatology. Additionally, sea surface height in the Mediterranean Sea can vary on both decadal and monthly time scales (accounting for seasonal variability), at up 10 cm each, resulting in periods of SSH higher or lower than mean SSH. The combined effect (20 cm) is greater than the difference in SSH between the Free Run and the SSH Experiment and could account for the errors in the mean SSH.

References

Alberola C, Rousseau S, Millot C, Astraldi M, Font J, Garcialafuente J, Gasparini G-P, Send U, Vangriesheim A (1995) Tidal currents in the western Mediterranean sea. Oceanol Acta 18(2):273–284

Asselin R (1972) Frequency filter for time integrations. Mon Weather Rev 100(6):487–490

Barron CN, Smedstad LF (2002) Global river inflow with Navy Coastal Ocean Model. In: OCEANS'02 MTS/IEEE, Biloxi, MI, USA, vol 3, pp 1472–1479. https://doi.org/10.1109/OCEANS.2002.1191855

Barron CN, Birol Kara A, Martin PJ, Rhodes RC, Smedstad L (2006) Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). Ocean Modell 11:347–375. https://doi.org/10.1016/j.ocemod.2005.01.004

Bauer R (1982) Functional description: master oceanographic data set (MOODS), Compass Systems Inc.

Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press

Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press

Bleck R (2002) An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. Ocean Model 4(1):55–88. https://doi.org/10.1016/S1463-5003(01)00012-9

Borrione I, Oddo P, Russo A, Coelho E (2017) Interpreting operational altimetry signals in near-coastal areas using underwater autonomous vehicles and remotely sensed ocean colour data. Abstract EGU2017-19400, presented at the EGU General Assembly 2017, Vienna, Austria, 23–28 April 2017

Boyer TP, Antonov JI, Garcia HE, Johnson DR, Locarnini RA, Mishonov AV, Pitcher MT, Baranova OK, Smolyar IV (2006) World ocean database 2005. In: Levitus S (ed) NOAA Atlas NESDIS 60, U.S. Government Printing Office, Washington, D.C., 190 pp., DVDs

Calafat FM, Chambers DP, Tsimplis MN (2012) Mechanisms of decadal sea level variability in the eastern North Atlantic and the Mediterranean Sea. J Geophys Res 117:C09022. https://doi.org/10.1029/2012JC008285

Carnes MR, Mitchell JL, deWitt PW (1990) Synthetic temperature profiles derived from Geosat altimetry: comparison with air-dropped expendable bathythermograph profiles. J Geophys Res 95(C10):17979–17992. https://doi.org/10.1029/JC095iC10p17979

Carnes M, Helber RW, Barron CN, Dastugue JM (2010) Validation test report for GDEM4. NRL Report NRL/MR/7330—10-9271. https://apps.dtic.mil/dtic/tr/fulltext/u2/a530343.pdf. Accessed 7 December 2018

Carrier MJ, Ngodock HE, Muscarella P, Smith S (2016a) Impact of assimilating surface velocity observations on the model sea surface height using the NCOM-4DVAR. Mon Weather Rev 144:1051–1068. https://doi.org/10.1175/MWR-D-14-00285.1

Carrier MJ, Ngodock HE, Smith SR, Souopgui I, Bartels B (2016b) Examining the potential impact of SWOT observations in an ocean analysis-forecasting system. Mon Weather Rev 144. https://doi.org/10.1175/MWR-D-15-0361.1

Chapman DC (1985) Numerical treatment of cross-shelf open boundaries in a barotropic coastal ocean model. J Phys Oceanogr 15(8):1060–1075

Chua BS, Bennett AF (2001) An inverse ocean modeling system. Ocean Model 3(3–4):137–165

Cummings JA (2005) Operational multivariate ocean data assimilation. Quart J Roy Meteor Soc 131:3583–3604. https://doi.org/10.1256/qj.05.105

Cummings JA, Smedstad OM (2014) Ocean data impacts in global HYCOM. J Atmos Oceanic Tech 31(8):1771–1791

Daley R, Barker E (2001) NAVDAS: Formulation and diagnostics. Mon Weather Rev 129(4):869–883

Davis RE, Eriksen CE, Jones CP (2003) Autonomous Buoyancy-driven underwater gliders. In: Griffiths G (ed) Technology and applications of autonomous underwater vehicles, Chap. 3. Taylor and Francis, pp 37–58

Escudier R, Bouffard J, Pascual A, Poulain P-M, Pujol M-I (2013) Improvement of coastal and mesoscale observation from space: application to the northwestern Mediterranean sea. Geophys Res Lett 40:2148–2153. https://doi.org/10.1002/grl.50324

Flather RA (1976) A tidal model of the north-west European continental shelf. Mem Soc R Sci Liege 10:141–164

Fox DN, Teague WJ, Barron CN, Carnes MR, Lee CM (2002) The modular ocean data assimilation system (MODAS). J Atmos Oceanic Technol 19:240–252. https://doi.org/10.1175/1520-0426(2002)019,0240:TMODAS.2.0.CO;2

Fukumori I, Menemenlis D, Lee T (2007) A near-uniform basin-wide sea level fluctuation of the Mediterranean Sea. J Phys Oceanogr 37(2):338–358. https://doi.org/10.1175/JPO3016.1

Helber RW, Townsend TL, Barron CN, Dastugue JM, Carnes MR (2013) Validation test report for the improved synthetic ocean profile (ISOP) system, Part I: Synthetic profile methods and algorithm. NRL Memo. Report, NRL/MR/7320—13-9364

Hogan TF, Liu M, Ridout JA, Peng MS, Whitcomb TR, Ruston BC, Reynolds CA, Eckermann SD, Moskaitis JR, Baker NL, McCormack JP, Viner KC, McLay JG, Flatau MK, Xu L, Chen C, Chang SW (2014) The navy global environmental model. Oceanography 27(3):116–125

Jacobs GA, Barron CN, Fox DN, Whitmer KR, Klingenberger S, May D, Blaha JP (2002) Operational altimeter sea level products. Oceanography 15:13–21. https://doi.org/10.5670/oceanog.2002.32

Jugan MJ, Beresford H (1991) Editing approach for the Navy's master oceanographic observation data set. In: Published in proceedings of MTS'91, An Ocean Cooperative: Industry, Government, and Academia, vol II

Landerer FW, Volkov DL (2013) The anatomy of recent large sea level fluctuations in the Mediterranean sea. Geophys Res Lett 40:553–557. https://doi.org/10.1002/grl.50140

Le Traon PY (2011) Satellites and operational oceanography. In: Schiller A, Brassington GB (eds) Operational oceanography for the 21st century, 2011, Chap. 2, pp 29–54. https://doi.org/10.1007/978-94-007-0332-2_2.

Le Traon PY, Dibarboure G, Jacobs G, Martin M, Rémy E, Schiller A (2017) Use of satellite altimetry for operational oceanography. In: Satellite altimetry over oceans and land surfaces. CRC Press, pp 581–608

Mellor GL, Yamada T (1982) Development of a turbulence closure model for geophysical fluid problems. Rev Geophys 20(4):851–875. https://doi.org/10.1029/RG020i004p00851

Metzger EJ, Smedstad OM, Thoppil PG, Hurlburt HE, Wallcraft AJ, Franklin DS, Shriver JF, Smedstad LF (2008) Validation test report for the global ocean prediction system V3.0 – 1/12° HYCOM/NCODA: Phase I. NRL Memorandum Report NRL/MR/7320—08-9148, 85 pp. https://apps.dtic.mil/dtic/tr/fulltext/u2/a490855.pdf. Accessed 7 December 2018

Metzger EJ, Smedstad OM, Thoppil PG, Hurlburt HE, Franklin DS, Peggion G, Shriver JF, Wallcraft AJ (2010) Validation test report for the global ocean forecast system V3.0–1/12° HYCOM/NCODA: Phase II. NRL Memorandum Report NRL/MR/7320—10-9236, 76 pp. https://apps.dtic.mil/dtic/tr/fulltext/u2/a518693.pdf. Accessed 7 December 2018

Metzger EJ, Helber RW, Hogan PJ, Posey PG, Thoppil PG, Townsend TL, Wallcraft AJ, Smedstad OM, Franklin DS, Zumudio-Lopez L, Phelps MW (2017) Global ocean forecast system 3.1 validation testing. NRL Report NRL/MR/7320—17-9722. https://apps.dtic.mil/dtic/tr/fulltext/u2/1034517.pdf. Accessed 7 December 2018

Ngodock H, Carrier M (2014a) A 4DVAR system for the navy coastal ocean model. Part I: System description and assimilation of synthetic observations in Monterey Bay. Mon Weather Rev 142(6):2085–2107

Ngodock H, Carrier M (2014b) A 4DVAR system for the navy coastal ocean model. Part II: Strong and weak constraint assimilation experiments with real observations in Monterey Bay. Mon Weather Rev 142(6):2108–2117

Ngodock H, Carrier M, Souopgui I, Smith S, Martin P, Muscarella P, Jacobs G (2016) On the direct assimilation of along-track sea-surface height observations into a free-surface ocean model

using a weak constraints four-dimensional variational (4D-Var) method. Quart J Roy Meteor Soc 142:1160–1170. https://doi.org/10.1002/qj.2721

Ngodock H, Carrier M, Smith S, Souopgui I (2017) Weak and strong constraints variational data assimilation with the NCOM-4DVAR in the Agulhas region using the representer method. Mon Weather Rev 145:1755–1764. https://doi.org/10.1175/MWR-D-16-0264.1

NRL DBDB2—Global 2-Minute topography. http://www7320.nrlssc.navy.mil/DBDB2_WWW/. Accessed on 15 December 2018

Oke PR, Balmaseda MA, Benkiran M, Cummings JA, Dombrowsky E, Fujii Y, Guinehut S, Larnicol G, Le Traon P-Y, Martin MJ (2009) Observing system evaluations using GODAE systems. Oceanography 22(3):144–153. https://doi.org/10.5670/oceanog.2009.72

OSTM/Jason-2 Products Handbook, CNES: SALP-MU-M-OP-15815-CN, EUMETSAT: EUM/OPS-JAS/MAN/08/0041, JPL: OSTM-29-1237 NOAA/NESDIS: Polar Series/OSTM J400. Issue 1 rev 11, 13 Jan 2017. https://www.aviso.altimetry.fr/fileadmin/documents/data/tools/hdbk_j2.pdf

Rowley C, Mask A (2014) Regional and coastal prediction with the relocatable ocean nowcast/forecast system. Oceanography 27(3):44–55. https://doi.org/10.5670/oceanog.2014.67

Rudnick DL, Davis RE, Eriksen CE, Fratantoni DM, Perry MJ (2004) Underwater gliders for ocean research. Mar Technol Soc J 38(1):48–59

SARAL/AltiKa Products Handbook. CNES: SALP-MU-M-OP-15984-CN. Issue 2 rev 5m 1 July 2016. https://www.aviso.altimetry.fr/fileadmin/documents/data/tools/SARAL_Altika_products_handbook_01.pdf

Smagorinsky J (1963) General circulation experiments with the primitive equations: I. The basic experiment. Mon Weather Rev 91(3):99–164

Smith S, Ngodock H, Carrier M, Shriver J, Muscarella P, Suoupgui I (2017) Validation and operational implementation of the navy coastal ocean model four dimensional variational data assimilation system (NCOM 4DVAR) in the Okinawa trough. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, vol III. Springer International Publishing

Tsimplis MN, Proctor R, Flather RA (1995) A two-dimensional tidal model for the Mediterranean Sea. J Geophys Res 100(C8):16223–16239. https://doi.org/10.1029/95JC01671

Verron J, Sengenes P, Lambin J, Noubel J, Steunou N, Guillot A, Picot N, Coutin-Faye S, Sharma R, Gairola RM, Raghava Murthy DVA, Richman JG, Griffin D, Pascual A, Rémy F, Gupta PK (2015) The SARAL/AltiKa altimetry satellite mission. Mar Geodesy 38(sup1):2–21. https://doi.org/10.1080/01490419.2014.1000471

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82. https://doi.org/10.3354/cr030079

Willmott CJ, Robeson SM, Matsuura K (2017) Climate and other models may be more accurate than reported, Eos, 98. https://doi.org/10.1029/2017EO074939. Published on 21 June 2017. Accessed 30 Jan 2019

Theoretical and Practical Aspects of Strongly Coupled Aerosol-Atmosphere Data Assimilation

Ting-Chi Wu, Milija Zupanski, Anton Kliewer, Lewis Grasso, and Leah D. Grant

Abstract Coupled data assimilation is one of the most active research areas in recent years because of its potential for improving the prediction of coupled modeling systems. Among various coupling options, strongly coupled data assimilation is the most efficient option for processing the information from observations. At the same time, coupled aerosol-atmosphere modeling is steadily gaining more interest due to its relevance to air quality, aviation, solar energy, and climate. It is well known that aerosols play an important role in Earth's radiation balance. Aerosol-atmosphere interaction is clearly multi-scale, from large-scale stratospheric impact to small-scale aerosol-cloud interaction. Such complex prediction system requires advanced data assimilation methodology that can deal with multi-scale interactions and observation information flow. In this chapter we address theoretical and practical aspects of strongly coupled data assimilation in application to aerosol-atmosphere coupling. We describe major aspects of developing strongly coupled data assimilation and related challenges. We also show results from a case study using a recently developed regional aerosol-atmosphere coupled data assimilation system. Finally, a general discussion on the future needs of strongly coupled data assimilation is provided.

T.-C. Wu (✉) · M. Zupanski · A. Kliewer · L. Grasso
Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO 80523, USA
e-mail: ting-chi.wu@colostate.edu

A. Kliewer
NOAA Earth System Research Laboratory, Global System Laboratory, Boulder, CO, USA

L. D. Grant
Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_18

1 Introduction

The primary role of coupling of prediction systems is to allow more realistic interactions between previously independent components, and therefore have a more accurate representation of relevant dynamical and physical processes. Since data assimilation has been an integral part of numerical weather prediction (NWP), there is a need for developing data assimilation for coupled prediction systems, often referred to as *coupled data assimilation.* A commonly used classification of coupled data assimilation includes *weakly* and *strongly* coupled data assimilation (Penny et al. 2017; Zupanski 2017). In a *weakly* coupled system, each component (e.g., atmosphere, chemistry, aerosol) has its own independent data assimilation system and analysis. In a *strongly* coupled system, all coupled system components are included in a holistic data assimilation system that can simultaneously assimilate observations from all components.

1.1 Background on Coupled Data Assimilation System

In this chapter we are primarily interested in describing strongly coupled data assimilation in aerosol-atmosphere coupled prediction system. Commonly used coupled aerosol-atmosphere prediction systems include the Goddard Earth Observing System Version 5 (GEOS-5), Navy Global Environmental Model/Navy Aerosol Analysis and Prediction System (NAVGEM/NAAPS), European Centre for Medium-Range Weather Forecasts/Copernicus Atmosphere Monitoring Service (ECMWF/CAMS), National Oceanic and Atmospheric Administration (NOAA) Global Forecast System (GFS), the Weather Research and Forecasting-Chemistry (WRF-Chem), and the Regional Atmospheric *Modeling* System (*RAMS*) *Model* (Molod et al. 2012; Hogan et al. 2014; Morcrette et al. 2009; Putman and Lin 2007; Chen et al. 2013; Grell et al. 2005; Fast et al. 2006; Saleeby and van den Heever 2013). Although some aspects presented here may be of general importance for data assimilation, they are mainly relevant to commonly used variational, ensemble, and hybrid variational-ensemble data assimilation systems (Parrish and Derber 1992; Rabier et al. 1999; Houtekamer and Mitchell 2001; Whitaker and Hamill 2002; Kleist and Ide 2015). In those systems, the background (or sometimes referred to as forecast or prior) error covariance is a key element of successful data assimilation analysis (e.g., Lorenc 1986; Kalnay 2003), which directly implies that coupled background error covariance plays a fundamental role in coupled data assimilation. Further, cross-covariance between components in a coupled system has the same relevance as the cross-covariance between variables in a standalone system. For example, it is well known that there exists a physical relationship between atmospheric temperature and wind. Data assimilation that includes such correlations (or cross-covariance) between wind and temperature in its background error covariance will produce more accurate analysis than a standalone data assimilation for wind and for temperature. Similarly, if correlations between an aerosol

and an atmospheric variable exist, a data assimilation that includes such correlations in the coupled background error covariance will produce more accurate analysis.

Another benefit of strongly coupled data assimilation is that it provides a mechanism for transferring observation information between coupled components. This may be especially relevant for coupled aerosol-atmosphere system. Given that there are generally fewer aerosol observations than atmospheric observations, assimilation of atmospheric observations can potentially improve aerosol initial conditions. Atmospheric observations can also be beneficial for improving the vertical distribution of the aerosol initial conditions, even when aerosol observations are assimilated. The most widely available aerosol observations are in terms of Aerosol Optical Depth (AOD), which are a vertically integrated quantity and therefore do not produce a vertical distribution of aerosol. In that situation, using observed atmospheric profiles can provide additional information about vertical distribution of aerosol through strongly coupled data assimilation.

1.2 Theoretical Description of Coupled Data Assimilation System

In order to illustrate the impact of coupled data assimilation, we consider a two-variable, one-point, aerosol-atmosphere coupled system. As shown in Zupanski (2017), when atmospheric component is observed under such system, the Kalman filter analysis equation can be written as follows:

$$x^a_{atm} = x^b_{atm} + \frac{\varepsilon^2}{1+\varepsilon^2}\left[y_{atm} - x^b_{atm}\right] \tag{1}$$

$$x^a_{aero} = x^b_{aero} + \rho\left(\frac{\sigma_{aero}}{\sigma_{atm}}\right)\frac{\varepsilon^2}{1+\varepsilon^2}\left[y_{atm} - x^b_{atm}\right] \tag{2}$$

$$\varepsilon = \frac{\sigma_{atm}}{r_{atm}} \tag{3}$$

In the above equations, subscripts *atm* and *aero* refer to atmospheric and aerosol components, respectively, superscripts a and b denote analysis and background, respectively, x is state, y is observation, σ and r denote background and observation errors, respectively, and ρ is the correlation between atmospheric and aerosol variables. Equation (1) is a standalone analysis for the atmospheric component, which means that when only atmospheric variables are observed, the coupled atmospheric analysis is identical to the standalone atmospheric analysis. Equation (2) represents the aerosol analysis, which critically depends on the correlation between atmospheric and aerosol variables (Eq. 3). When the correlations between atmospheric and aerosol variables are non-existent or negligible, aerosol analysis is the same as the guess,

meaning no change from the assimilation. However, when the correlations exist the aerosol analysis can be updated from assimilating atmospheric observations.

The above discussion illustrates the main motivation for using the formalism of strongly coupled data assimilation instead of weakly coupled data assimilation: strongly coupled data assimilation is more general as it includes weakly coupled assimilation as an option. When correlations between variables are naturally negligible, a strongly coupled system will still correctly produce the analyses approximately equal to standalone analyses. When correlations are relevant, the strongly coupled system will update all variables, effectively increasing the utility of observations. The implied assumption for achieving the desired impact of strongly coupled data assimilation is that the estimated cross-correlations are reliable.

One critical issue in strongly coupled data assimilation is related to spatial and temporal scales of coupled processes. Although further understanding of the impact of having different spatial and temporal scales between a coupled system on the estimate of the background error covariance is necessary, it is likely that in an idealized data assimilation scenario where error covariances are exact and full-rank, all correlations (temporal, spatial, cross-variable, and cross-components) will be accurately accounted for. This is because in that situation the covariance would accurately represent the interactions between uncertainties of coupled components, and therefore implicitly address the scale differences. In practical applications, however, the coupled error covariance may not be able to account for different scales of coupled components (e.g., aerosol and atmosphere) with sufficient accuracy, in particular the temporal scales. While there is no commonly accepted solution to this problem, a possible strategy in such situations could be to modify existing background error covariance to reflect the different temporal scales between coupled components. For example, one could enforce covariance localization in time using pre-defined characteristic correlation scales or one could also use a covariance averaged over several previous data assimilation cycles. That said, the aerosol and atmosphere time scales may not be as different as the time scale differences between other coupled system such as the land surface and the atmosphere. As such, accounting for temporal correlations may not be a concern in an aerosol-atmosphere coupled system. In any case, incorporating different time scales in coupled error covariance is an important next step in making strongly coupled aerosol-atmosphere data assimilation more reliable and effective.

When using variational data assimilation, in which error covariance is approximated by a mathematical function, satisfactory modeling the correlations between coupled components may be difficult to achieve (Ménard et al. 2019). However, aerosol and chemistry data assimilation with four-dimensional variational (4D-Var) methods may offer new possibilities. For example, Hakami et al. (2005) found that adjoint inverse modeling in 4D-Var helps in constraining various inputs for chemical transport models, while Sandu et al. (2005) concluded that 4D-Var is a feasible approach for carbon-cycle aerosol assimilation. As a smoother, 4D-Var has the advantage of automatically accounting for time correlations during the data assimilation process. On the contrary, time correlations have to be fully imposed in sequential data assimilation, i.e. filters. With that, 4D-Var can be an advantageous option

for coupled aerosol-atmosphere data assimilation since the interaction between the different time scales of aerosol and atmosphere will be more realistic in 4D-Var compared to the interactions in filters. This certainly opens additional avenues for strongly coupled aerosol-atmosphere data assimilation research directed towards using smoothers instead of filters.

When using ensemble data assimilation, however, all correlations come naturally from ensemble forecasting. A potential difficulty may be that small ensemble size does not produce reliable estimates of correlations, which then requires additional attention. Considering the above possibilities, it seems that using a strongly coupled data assimilation formalism has more advantages than disadvantages. Most importantly, strongly coupled formalism potentially allows a more efficient use of observations, eventually leading to an improved analysis and prediction.

1.3 Single Observation Experiment

One of the main advantages of using an ensemble data assimilation algorithm is the flow-dependent background error covariance. Created by ensemble of model forecasts it is time-dependent and includes complex correlations between variables. For aerosol data assimilation the correlations between atmospheric and aerosol variables have the most significance. In principle, the correlations allow observations of one component to impact the analysis of another component. This also helps in the areas where AOD observations may have insufficient coverage, by indirectly providing additional information through cross-correlation. Atmospheric observations also provide additional information about the three-dimensional structure of aerosol, through the flow-dependent correlations.

To illustrate this impact, we conduct two single observation experiments using a regional coupled chemistry-aerosol-atmosphere WRF-Chem model, with the Goddard Chemistry Aerosol Radiation and Transport (GOCART) aerosol module. The data assimilation interval is 6 h, and model grid spacing is 9 km with a total of 50 vertical layers.

In the first experiment, we assimilate a single east–west wind component (u wind) observation at 25°N, 53°E and near the model surface. In Fig. 1, we show the impact of such assimilation on the DUST_3 (2.4 μm) variable from the GOCART aerosol module. Note that in a less advanced, uncoupled data assimilation system, the impact of assimilating a single wind observation on dust variable would be equal to zero. In Fig. 1a, one can see negative increments of dust, in both horizontal and vertical directions, suggesting that increasing westerly wind in that area will produce a decrease of dust concentration. In Fig. 1b, one can also notice that the impact of wind observation on dust is limited in the vertical direction and is generally confined to lower levels where the observation was located.

In the second experiment, we assimilate a single DUST_3 observation in the same place located at 25°N, 35°E. The impact of assimilating such observation to near-surface wind is shown in Fig. 2. One can notice a dominant negative response, which

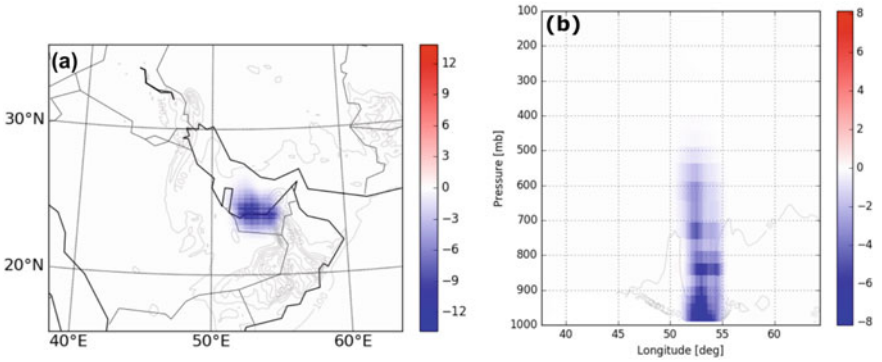

Fig. 1 Analysis increments (i.e., analysis minus background) of DUST_3 (ug $kg^{-1}{}_{dry\ air}$) in response to a single east–west wind observation (u component wind), valid at 00 UTC on August 4, 2016: **a** horizontal distribution at surface and **b** vertical cross section along 25°N

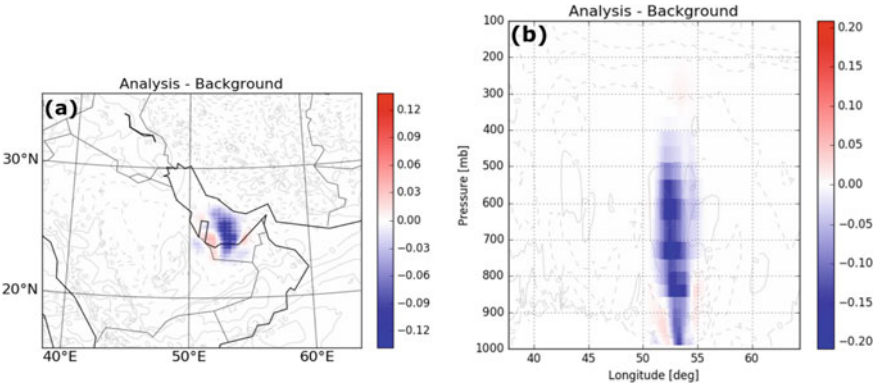

Fig. 2 Similar to Fig. 1, except for analysis increments of u component wind (m s^{-1}) in response to a single DUST_3 observation, valid at 00 UTC on August 4, 2016: **a** horizontal distribution near surface and **b** vertical cross section along 25°N

is consistent with the findings in Fig. 1. The analysis response of dust is also limited in both vertical and horizontal directions, as anticipated due to the use of covariance localization.

The rest of the chapter is organized as follows. We begin by describing the current status of aerosol-atmosphere coupled data assimilation in Sect. 2, followed by aerosol observations and observation operator in Sect. 3. Challenges of strongly coupled data assimilation are discussed in Sect. 4, with numerical experiments of a case study and results presented in Sect. 5. Summary and future directions are given in Sect. 6.

2 Current Status on Aerosol-Atmosphere Coupled Data Assimilation

Before an overview on the current status of aerosol-atmosphere coupled data assimilation is given, a brief discussion on a prerequisite topic regarding online versus offline approaches for weather and aerosol *forecasting* is provided herein. An offline approach involves an aerosol model run that is driven by meteorological fields produced by an atmospheric model run (e.g., Sekiyama et al. 2010; Rubin et al. 2017). As a result, interactions between the atmospheric and the aerosol processes are restricted to one-way. That is, the meteorological fields from an atmospheric model are used to initialize the aerosol model, but the outcome from the aerosol model is not fed back to the atmospheric model. On the other hand, an online (sometimes also referred to as inline) approach involves an integrated model run of both atmospheric and aerosol components (e.g., Liu et al. 2011; Lee et al. 2017), in which a two-way interaction of atmospheric and aerosol components is allowed. As indicated in Grell and Baklanov (2011), major advantages of using an online approach as opposed to an offline approach include a more realistic presentation of the atmosphere, a more numerically consistent treatment of both components, and improved forecast via improved assimilation. Nevertheless, the reduced computational cost and more flexibility in ensemble forecasting makes the offline approach still rather appealing, especially for regulatory agencies.

As mentioned in the introduction, there exist two general approaches for aerosol-atmosphere coupling from the *data assimilation* perspective. As discussed earlier, the two approaches are (i) a weakly coupled data assimilation and (ii) a strongly coupled data assimilation. A weakly coupled data assimilation system performs data assimilation of each component independently, although the updated analysis of both meteorological and aerosol fields can be used to initialize a coupled aerosol-atmosphere forecast. Since the individual component is treated separately, there does not exist cross-component elements in the background error covariance matrix, which is essential for the data assimilation update. In contrast, a strongly coupled data assimilation system performs data assimilation and forecast of both aerosol and atmospheric components simultaneously, treating the coupled system as a single integrated system. As such, there exist cross-component elements in the background error covariance matrix, which allows observational information from one component to potentially influence the other component within a coupled data assimilation update. Based on the varying degrees of data assimilation update, weakly (strongly) coupled data assimilation can be further classified into quasi weakly (strongly) and weakly (strongly). Interested readers are redirected to Penny et al. (2017) and Penny and Hamill (2017) for more details.

2.1 Operational Centers and Research Community

With the increased computational power, many NWP centers have reconsidered the *online* approach over the more common *offline* approach for weather and aerosol *forecasting*. For example, the ECMWF Integrated Forecast System (IFS) (Morcrette et al. 2008), the Japan Meteorological Agency (JMA) Model of Aerosol Species in the Global Atmosphere (MASINGAR) (Tanaka and Chiba 2005), and the UK Met Office (UKMO) Unified Model (UM) (Collins et al. 2011). Nevertheless, several NWP centers favor the offline approach and that include the US Navy Fleet Numerical Meteorology and Oceanography Center (FNMOC) Naval Research Laboratory (NRL) NAAPS (Lynch et al. 2016) and the Météo-France Modèle de Chimie Atmospherique à Grande Echelle (MOCAGE) (Guth et al. 2016). A summary of the current status of global NWP efforts on aerosol forecasting is provided by Xian et al. (2019). Among these efforts, the ECMWF IFS system is considered a strongly coupled aerosol-atmosphere data assimilation system because a single data assimilation algorithm is employed to update both aerosol and atmospheric states (Benedetti et al. 2009). Although the JMA MASINGAR is an inline forecast model of aerosol that is coupled to an atmospheric model, data assimilation of aerosol into MASINGAR is performed separately from the atmospheric data assimilation (Yumimoto et al. 2018).

In addition to operational efforts, numerous research efforts have addressed the assimilation of aerosol and/or chemistry data into research forecast models for the improvement of weather and air quality simulations (Collins et al. 2001; Weaver et al. 2007; Wang and Niu 2013; Zhang et al. 2014; Lee et al. 2017; Eltahan and Alahmadi 2019). Among them, U.S. National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) provides global reanalysis dataset of both atmospheric and aerosol fields using their GEOS-5 (Randles et al. 2017). Unlike GEOS-5, the WRF-Chem (Grell et al. 2005), which is developed and maintained by the National Center for Atmospheric Research (NCAR), is a widely used research model for regional aerosol, air quality, and atmospheric studies. Similar to WRF-Chem, the RAMS model is also a research model developed for studying regional aerosol-atmosphere interactions.

2.2 Global Versus Regional Applications

Unlike global applications, specifying realistic lateral boundary conditions is critical to regional simulations and data assimilation, in general (Chikhar and Gauthier 2017). A study by Tang et al. (2009) examined the impact of specifying lateral boundary conditions from six different sources on the simulation of tropospheric ozone over the continental U.S., which include a fixed ozone profile, three time-varying ozone profiles derived from global models, and two time-varying ozone profiles derived from soundings. Their results suggest that specifying lateral boundary conditions

with those derived from global models improves the simulation most significantly; however, they found that uncertainties associated with the global models can also translate to the corresponding regional simulations. In addition, Chikhar and Gauthier (2017) pointed out that biases can emerge from the differences in spatial resolution as well as physical parameterizations used between the regional model and the global model, which provides lateral boundary conditions for the regional simulations. Such an issue can be reduced by using a unified system where a regional model and its global version are used together to provide lateral boundary conditions.

3 Aerosol Observation and Forward Operator

3.1 Retrievals Versus Direct Measurements

For analyses and therefore model forecasts to benefit from coupled aerosol-atmosphere assimilation, aerosol observations must be available similarly to atmospheric variables. These observations generally fall into two categories: direct assimilation of aerosol-affected satellite radiances or the assimilation of retrieved aerosol products. Both approaches carry distinct strengths and weaknesses. For example, direct assimilation would necessitate complex radiative transfer code which would lead to costly computational time. On the other hand, retrieved observations inherently make assumptions related to the physical characteristics of aerosols. These include species type, shape, size (bulk or binned categorization), and refractive indices. With that, retrieved products must then be matched to a particular model. Even with these challenges assimilation of retrieved aerosol products is the current operational approach as it affords the availability of quality observations with estimates of uncertainty. The following subsections briefly describe currently available aerosol observations.

3.1.1 Aerosol Optical Depth

An example of a retrieved aerosol product is the aforementioned AOD. As the name suggests, AOD is a quantity that measures the loss of light due to scattering and absorption through a vertical column. This quantity depends on the type and physical characteristics of the aerosols that are present. Ground-, airborne-, and spaceborne-based AOD observations have been used in a variety of data assimilation systems (variational, ensemble, hybrid) at National Center for Environmental Prediction (NCEP), ECMWF and NRL. Liu et al. (2011) showed that 3D-Var assimilation of AOD from the Moderate Resolution Imaging Spectroradiometer (MODIS) improved both aerosol analyses and aerosol forecasting. Further, Benedetti et al. (2019) utilized 4D-Var to assimilate MODIS AOD observations and demonstrated improvement in dust analyses and forecasts for up to 48 h in East Asia. Examples

of ensemble-based assimilation of aerosols can also be seen in Pagowski and Grell (2012), Rubin et al. (2016), and Schwartz et al. (2014). Hybrid data assimilation has also been shown to be effective in aerosol analyses and forecasts (Schwartz et al. 2014; Choi et al. 2020).

3.1.2 Satellite Radiances Affected by Aerosols

Visible, ultraviolet (UV) and near-infrared wavelengths could very well be the future of aerosol assimilation. This has been shown possible (Weaver et al. 2007) but several challenges have prevented this from becoming operationally viable. These include the speed and complexity of the available radiative transfer codes, complexity of the model, and how polarization would be addressed. A benefit of direct assimilation would be the ability to assimilate from different satellite instruments. Currently attempts are underway at ECMWF to assimilate two aerosol visible radiances from MODIS and have been shown to be effective in representing plumes in the 4D-Var analyses comparable to the available observations. While direct assimilation of aerosol-affected satellite radiances has been shown viable, future research is still required for this to become operationally feasible.

3.1.3 LIDAR

Light Detection and Ranging (LIDAR) instruments use a pulsed laser to generate three-dimensional observational imagery of the Earth's atmosphere and surface characteristics. This is done by observing the backscatter from molecules and particles. LIDAR instruments can retrieve profiles describing the composition of atmosphere in regard to water content and aerosols and also determine wind fields. One such example is the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite. CALIPSO utilizes LIDAR along with infrared and visible imagers to observe clouds and aerosols and is part of the "A-Train" satellite constellation. The vertical profiles retrieved by CALIPSO have been able to provide highly-accurate cloud heights and high thin cirrus clouds which have been difficult to observe previously.

Given the utility and quality of LIDAR observations, there are other space-borne instruments that are either in the pre-launch design phase or that have recently been launched and are now used operationally. The EarthCARE satellite is part of the European Space Agency's Earth Explorer Programme and scheduled to launch in 2022. EarthCARE will carry LIDAR, radar, radiometers, and imagers with the goal of producing high-resolution horizontal and vertical profiles of aerosols, liquid water, cloud distribution, and atmospheric radiative heating and cooling. These new datasets of highly-variable parameters are expected to improve forecasting and climate modelling.

Aeolus, another space-borne satellite, was launched in 2018 and has been used operationally at ECMWF since January 2020. Aeolus employs a LIDAR instrument

capable of observing the Doppler shift of atmospheric molecules and particles to retrieve highly precise wind profiles. While the wind profiles are currently used to improve numerical weather prediction, Aeolus also has the ability to retrieve aerosol optical properties such as extinction and optical depth. The value and use of these aerosol profiles has yet to be fully explored.

3.1.4 AERONET

The AERONET (AErosol RObotic NETwork) program is a network of ground-based sun photometers capable of measuring atmospheric aerosol properties. By measuring sun and sky radiances at a fixed number of wavelengths in the visible and near-infrared spectrum, precipitable water and aerosol properties such as AOD, single scattering albedo, aerosol scattering phase function, and aerosol volume size distribution can be retrieved. This global network has grown to over 600 sites as of 2018. AERONET thus provides a vast database of ground-truth calibration data for current and future satellite instruments which is a crucial component of utilizing new observations in data assimilation to improve numerical weather prediction.

3.2 AOD Observation Operator

To assimilate AOD observations a data assimilation system must include a forward operator (also known as observation operator) that computes a model-equivalent value of AOD. This operator will be unique to the numerical prediction model as it depends on the represented aerosol species. Each aerosol species has specific physical properties including effective radius and wavelength-dependent indices of refraction. These characteristics must be known to calculate the mass extinction coefficient via Mie theory (Bohren and Huffman 1983). To account for hygroscopic growth -Köhler theory, Petters and Kreidenweis (2007) grows each particle to equilibrium per the ambient relative humidity. Since the Mie calculations can be expensive, look-up-tables can be created offline for quick reference of a species' humidity-dependent mass extinction coefficient (Eq. 4). This technique has been applied in this study (see Sect. 5.3). Total-column AOD is then computed by summing over all species and model levels following Liu et al. (2011) and Pagowski et al. (2014). The calculation of AOD at a given wavelength λ (nm) is expressed as

$$AOD(\lambda) = const \cdot \sum_{i=1}^{N_{aero}} \sum_{k=1}^{k_{top}} E_{ext}(\lambda, n_{ri}, r_{effi}) \cdot c_{ik} \cdot \frac{\Delta p_k}{g} \tag{4}$$

where *AOD(λ)* represents the spectrally dependent AOD operator (unit less), i is the index for aerosol species, N_{aero} is the total number of aerosol categories that contribute to the AOD calculation, k is the index for model vertical levels, and k_{top} is

the model top level. E_{ext} is the spectrally dependent mass extinction coefficient (m^2 g^{-1}), which is a function of the index of refraction n_r and effective radius r_{eff} (nm) of a given aerosol species, c_i, in the form of mass mixing ratio (g of aerosol/kg of dry air). Δp_k is the pressure difference (mb) between two vertical levels k and $k + 1$, and g is the acceleration due to gravity (m s^{-2}). *const* is a constant of 10^5, as a result of unit conversion (Eq. 4).

3.3 AOD Error and Bias Estimation

AOD observations include both a quality flag and a definition of an observational error which depends on the retrieval algorithm, e.g. MODIS Dark Target (Levy et al. 2013) versus Deep Blue (Hsu et al. 2006). Ideally these definitions would extend to error covariances which would describe correlations between different aerosol products in both space and time. Moreover, to improve assimilation of these observations, an estimation of bias and the ability to correct for it, is also desired. These bias correction procedures can be generally categorized as either static (offline) or variational. The static bias correction scheme (Eyre 1992) considers differences in the observations and the model state over a period of time and defines bias predictors using satellite scan angle along with several atmospheric variables (e.g. skin temperature, total column water, etc.). This is carried out offline for each satellite sensor and band and is frequently updated. The bias correction is then applied to the observations in the data assimilation system. Variational bias correction methods include bias coefficients within the state vector of the minimized cost function. Therefore, these coefficients are continuously updated, along with the state vector itself, during each data assimilation cycle. The bias is defined as a linear combination of predictors, similar to the static scheme, using scan angle along with atmospheric variables. More details can be found in (Derber et al. 1991; Parrish and Derber 1992; Derber and Wu 1998; Dee 2005; Auligné et al. 2007).

4 Challenges

4.1 Choice of Control Variables

The choice of control variables is directly related to the background error covariance, which plays a fundamental role in data assimilation. Control variables can be defined as a subset of variables of an NWP system that can potentially impact its prediction. In general, control variables include not only the initial conditions of prognostic variables of an NWP system, but also non-prognostic variables, empirical model parameters, and model error bias. The particular choice of control variables depends on both the feature of interest (e.g., tropical cyclones, thunderstorms, and blowing

dust) and the type of observed data to be assimilated (e.g., satellite radiances, radar data, and satellite retrieved quantities). Over the previous decades, research efforts focused on improving the forecast of severe thunderstorms, as a result, doppler radar data was assimilated; therefore, a choice of control variables would be the horizontal component of the prognostic wind (e.g., Sun 2006; Hu and Xue 2007). Another example of a feature of interest is blowing dust. Progress has also been made in the assimilation of airborne dust using aerosol-atmosphere coupling.

An important aspect of a strongly coupled aerosol-atmosphere data assimilation system is to have a set of control variables that cover the state of both components (Pagowski et al. 2014). Control variables associated with the aerosol component include the initial conditions of dust, sea salt, carbon particles from agricultural or wildfire burning, and sulphate from agricultural and industrial sources, which are some of the typical aerosol species. In addition, control variables associated with the atmospheric component typically include the initial conditions of temperature, pressure, all components of the three-dimensional wind vectors, and water vapor mixing ratio. Having a set of control variables, which covers both components, allows the information of assimilated observations to be spread into relevant variables via the background error covariance matrix, which includes cross-component correlations (to be discussed in Sect. 4.2).

Additional information is now given to a specific scenario: airborne dust, which results from high winds over semi-arid surfaces. Similar to cloud microphysical schemes, aerosol solvers/models predict moments of aerosol species. In particular, a single-moment scheme predicts only the mass mixing ratio (ug of dust per kg of air) of a given aerosol species (e.g., WRF-Chem; Grell et al. 2005); whereas a double-moment scheme predicts both mass mixing ratio and number concentration (number of dust particle per kg of air) of a given aerosol species (e.g., RAMS aerosol module; Saleeby and van den Heever 2013). Consequently, a double-moment scheme allows three-dimensional variability in particle size, because particle size is a function of both mass and number concentration. Efforts to advance the field of dust assimilation have focused on the first moment, the mass field, as a first step. After the assimilation of dust with the first moment becomes better understood, the next step is to include the second moment, number concentration, in the set of control variables.

There are important challenges when including only the first moment as a control variable. As a result of altering only the first moment, during the data assimilation process, mass may appear in a region devoid of number concentration. As previously stated, particle size depends on both moments. Consequently, if there is a region in a numerical domain with non-zero mass and non-existence number concentration, then calculation of particle size becomes problematic. An additional challenge is that a forecast from an analysis, which contains an inconsistency between the two moments, will cause numerical errors. Although focus was placed on the first moment (mass) and the second moment (number concentration), the above discussion applies equally well to any double-moment prognostic variables like the first and second moments of cloud microphysics (Cotton et al. 2003; Saleeby and Cotton 2004).

In preparation for a discussion of background error covariance (Sect. 4.2), additional care should be exercised in choosing control variables. Because the background

error covariance matrix is computed from the set of control variables, choice of the set of control variables has fundamental impact on the efficiency and success of assimilation (Xie and MacDonald 2012; Sun et al. 2015).

4.2 Background Error Covariance

As mentioned in the previous section, background error covariance matrix provides a mechanism for spreading the information from assimilated observations to control variables represented by grid points (both horizontally and vertically) of all coupled components (Fisher 2003). In addition, background error covariance not only allows observation of different types to act in synergy, but also helps maintain the analysis state closer to balance (Bannister 2008a). Having chosen a set of control variables does not naturally guarantee a corresponding background error covariance matrix that can accurately represent the associated actual error. Careful tuning and possibly modeling of background error covariance is required for any effective data assimilation schemes that include variational, ensemble, as well as hybrid methods.

Due to its prohibitive size (NWP system has a large dimensional state space ~ 10^8), the use of the explicit form of background error covariance matrix is impossible. Instead, several techniques have been developed to measure characteristics of background error statistics for modeling and specifying realistic background error covariance matrix. A review of measuring and modeling background error covariance in the context of atmospheric data assimilation systems was provided in Fisher (2003) and Bannister (2008a, 2008b). Methods to measure the background error statistics include the following: analysis of innovations, differences between forecasts of different lengths that verify at the same time (i.e., the National Meteorological Center (NMC) method; Parrish and Derber 1992), the lagged NMC methods, and the ensemble-based Monte Carlo method. In particular, the NMC method is widely used by several NWP centers due to its advantage of low computational cost. However, the NMC method was often found to overestimate covariances due to the use of longer forecast lengths, e.g., 24 h and 48 h, to estimate errors of the background, which is usually a 6 h forecast. Following the measurement of background error statistics, the modeling of background error covariance can be achieved via spectral/wavelet methods (Fisher 2006) and control variable transform (Bannister 2008b), both of which seek to simplify the representation of the background error covariance matrix and were developed for variational-based schemes.

Benedetti and Fisher (2007) and Kahnert (2008) were the first to apply the NMC method to estimating background error statistics of aerosols. With that, a satisfactory background error covariance matrix was constructed with the use of a wavelet modeling approach without the need to prescribe the vertical and horizontal correlation (Benedetti et al. 2009). In addition, a generalized background error covariance matrix model was developed by Descombes et al. (2015) as a community tool to be used beyond atmospheric applications (e.g. geophysical, chemistry, etc.).

In the context of ensemble data assimilation, background error covariance can be created with the use of ensemble of model forecasts. As such, the ensemble background error covariance matrix is time-dependent and includes embedded correlations between control variables from the model. Nevertheless, additional care is still required to fine tune the ensemble background error covariance to avoid filter divergence as well as spurious correlation due to the use of a much smaller ensemble (i.e., reduced rank). In general, a good practice to visualize the structure functions of the background error covariance (Thépaut et al. 1996) can be achieved via examining analysis increments of control variables resulted from assimilating a single observation of the kind of a control variable in a pre-specified grid point (i.e., single observation experiment; see Sect. 1.3).

4.3 Non-Gaussianity and Non-Linearity

Many variational and ensemble-based data assimilation and retrieval systems assume that the observational and model errors come from a Gaussian distribution. Previous research has indicated this is not necessarily true for variables that are not from a Gaussian distribution, e.g. variables that are positive definite such as humidity or total precipitable water. Recent research has sought to address this limitation by introducing a cost function based on a mixed Gaussian-lognormal distribution (Fletcher and Jones 2014). Here the incremental 3D and 4DVAR formulations of the mixed distribution cost function is derived and improved performance is shown with experiments based on the Lorenz 1963 toy model. This formulation has also been shown to improve 1DVAR water vapor mixing ratio retrievals (Kliewer et al. 2016) as this variable is certainly positive-definite. Another recent approach that avoids any assumption of probability distribution is with the application of particle filters (Van Leeuwen 2010) however these methods have not been found to be operationally viable as of yet due to their computational cost.

The non-Gaussian nature of AOD can certainly have an impact on the quality of the coupled data assimilation. As previously described the forward operator for AOD observations is certainly non-linear since it incorporates hygroscopic growth as a function of relative humidity. Preliminary experiments have confirmed this by noting that the distribution of innovations during assimilation is often positively or negatively skewed. While this issue can have impact on the data assimilation analyses and the subsequent NWP forecasts, this is out of the scope for what is presented here and is not addressed within these experiments.

4.4 Insufficient Data for Independent Verification

A standard way of measuring the success of data assimilation performance is to compare its analysis and background in observation space against independent observations, i.e., observations not assimilated. Benedetti et al. (2018) describe several observation types that can be used for verifying chemistry and aerosol data assimilation. However, there are situations in which the number of observations available for assimilation is limited and/or their representativeness is inadequate (a few pointwise observations to validate global—over all points—data assimilation). This is especially relevant to aerosol data assimilation, and in particular to regional aerosol data assimilation. Commonly used verification data include AERONET and CALIPSO. Although proven useful, there are some concerns when using them for verification of aerosol data assimilation, and in particular regional aerosol data assimilation, related to their limited spatial and temporal coverage. Given that typical Gaussian data assimilation involves some kind of optimization over all grid points and observations, having a few pointwise observations such as AERONET is not sufficient for verifying data assimilation. Similar is true for CALIPSO, which produces a high-resolution but narrow-swath vertical cross-section of aerosol.

One can also think of additional issues that may become important. For example, a new satellite sensitive to a particular aerosol variable that is rarely observed is launched with a goal of demonstrating the usefulness of new observation type in data assimilation. Under these assumptions there are likely no other, independent observations similar to the new satellite and therefore direct verification is not possible. Another example may be the limitation introduced by choosing the area of interest that is sporadically observed, such as polar regions, oceans, and deserts. Without sufficient statistically independent observations such studies may never be properly verified. Although a particular research may be of great scientific interest, not having independent observations to verify data assimilation performance could preclude efforts to assimilate these observations.

Described scenarios may be more common in regional data assimilation applications, but they could happen in global applications as well. This is because observation operators that transform control variables to observed variables often only have local impact, especially in the horizontal directions. Imagine a case when a special type of observation is available and assimilated only over a small area of a global domain while the verifying independent observations are not available, it will not be possible to reliably assess the impact of assimilated data.

All of the above suggests that there is a need to address alternative verifications for data assimilation in general, and particularly for aerosol, without using independent observations. The main underlying premise of such an approach is that a data assimilation algorithm contains additional information that is overlooked and consequently not used for its verification.

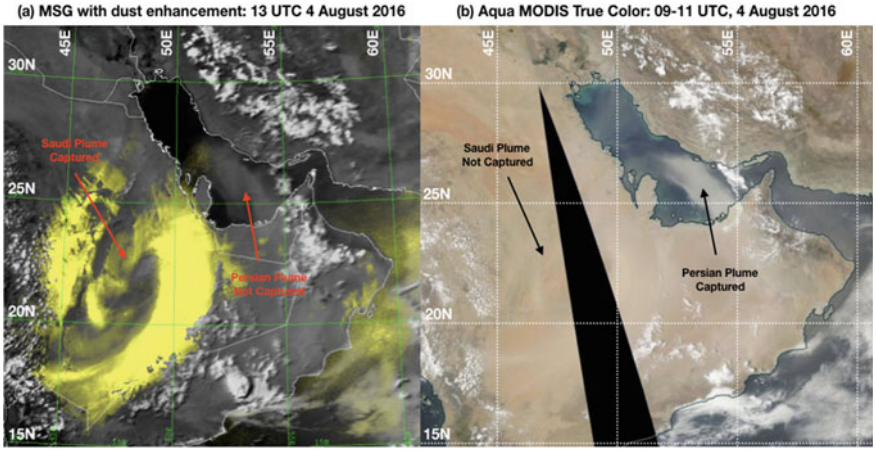

Fig. 3 Satellite imagery of the two dust plumes over the Arabian Peninsula on 4 August 2016: **a** Meteosat Second Generation (MSG) imagery with dust enhancement applied (showing dust in yellow) and **b** Aqua MODIS true color imagery

5 Experiments and Results

5.1 Case Study

A dust storm case over the Arabian Peninsula, one of the major dust sources of the world and the so-called dust belt (Jish Prakash et al. 2015), occurred on 4 August 2016 (Miller et al. 2019; Saleeby et al. 2019) and was chosen to illustrate the utility of a strongly coupled aerosol-atmosphere data assimilation system. On 4 August 2016, two distinct dust plumes occurred (Fig. 3), in which one plume advected offshore of the United Arab Emirates (UAE) to the central portion of the Persian Gulf (referred to as the Persian Plume; Fig. 3a), which was detected by the Spinning Enhanced Visible and Infrared Imager (SEVIRI) onboard Meteosat Second Generation (MSG) - 8 imagery with dust enhancement algorithm applied, and the other plume was located in interior regions of Saudi Arabia (referred to as the Saudi Plume; Fig. 3b), which was detected by Aqua MODIS true color imagery. As discussed in Miller et al. (2019), the environment of the Saudi Plume was characterized by values of total precipitable water (TPW) less than approximately 25 mm whereas the Persian Plume was in an environment characterized by values of TPW in excess of 45 mm.

5.2 Overview of the RAMS-MLEF System

In order to demonstrate the utility of a strongly coupled aerosol-atmosphere data assimilation system, an NWP model was interfaced to a data assimilation system.

That is, RAMS (Cotton et al. 2003) was interfaced with the Maximum Likelihood Ensemble Filter (MLEF; Zupanski 2005; Zupanski et al. 2008), hereafter referred to as the RAMS-MLEF system, to conduct experiments for the 4 August 2016 case. Before the experimental setup is described, a brief introduction to RAMS, MLEF, and the RAMS-MLEF system is provided.

RAMS is a multi-purpose mesoscale numerical prediction model that was developed at CSU. Throughout the years, RAMS has undergone multiple upgrades that include improvements to its microphysics via the implementation of a bimodal and double-moment cloud water scheme (Saleeby and Cotton 2004), an improved capability to assimilate lightning data (Federico et al. 2017), and the development of an interactive aerosol module (Saleeby and van den Heever 2013). Of these recent upgrades, the development of a RAMS aerosol module is directly related to the study herein. There are a total of nine aerosol categories represented by the aerosol module in RAMS: (i) submicrometer sulphate, (ii) supermicrometer sulphate, (iii) submicrometer mineral dust, (iv) supermicrometer mineral dust, (v) film-mode sea salt, (vi) jet drop-mode sea salt, (vii) spume-mode sea salt, (viii) submicrometer regenerated aerosols, and (ix) supermicrometer regenerated aerosols. For each aerosol category, the size is represented by a lognormal distribution given by

$$n(r) = \frac{N}{r\sqrt{2\pi}\ ln(\sigma_g)} exp\left[-\frac{\left(ln\frac{r}{r_g}\right)^2}{2ln^2\sigma_g} \right] \tag{5}$$

where $n(r)$ is number concentration of aerosols of dry radius r, N is total number concentration of aerosols, r_g is lognormal distribution geometric median radius, and σ_g is lognormal distribution geometric standard deviation. Although the shape of the size distribution as described in Eq. (5) is fixed during a simulation, the distribution is allowed to translate in the direction of r. That is, as a result of sources and sinks of aerosol mass during a simulation, the size distribution given in Eq. (5) is allowed to shift toward larger or smaller values of r. In addition, the width of the size distribution is determined by σ_g, which behaves like a dispersion parameter in a Gamma size distribution used in microphysical development.

MLEF is a hybrid data assimilation algorithm with both variational and ensemble features. Similar to other data assimilation methods (e.g., Evensen 1994; Houtekamer and Mitchell 2001; Anderson 2001; Bishop et al. 2001; Whitaker and Hamill 2002), a generalized flow chart of MLEF also consists of a forecast step and an analysis step. During the forecast step, MLEF generates an ensemble of forecasts to estimate the flow-dependent background/forecast error covariance. After completion of the forecast step, minimization of a prescribed cost function occurs during the analysis step, see Fig. 4, where $\mathbf{x}$ and $\mathbf{y}$ represent the state vector and the observation vector, respectively; subscript $\mathbf{f}$ denotes the forecast (or background) and subscript $\mathbf{a}$ denotes the analysis; $\mathbf{P_f}$ is the flow-dependent background/forecast error covariance matrix and $\mathbf{P_a}$ is the analysis error covariance matrix; superscript $\mathbf{t}$ denotes time; $\boldsymbol{h}$ denotes a collection of observation operators; $\boldsymbol{m}$ represents a forecast model. Unlike pure

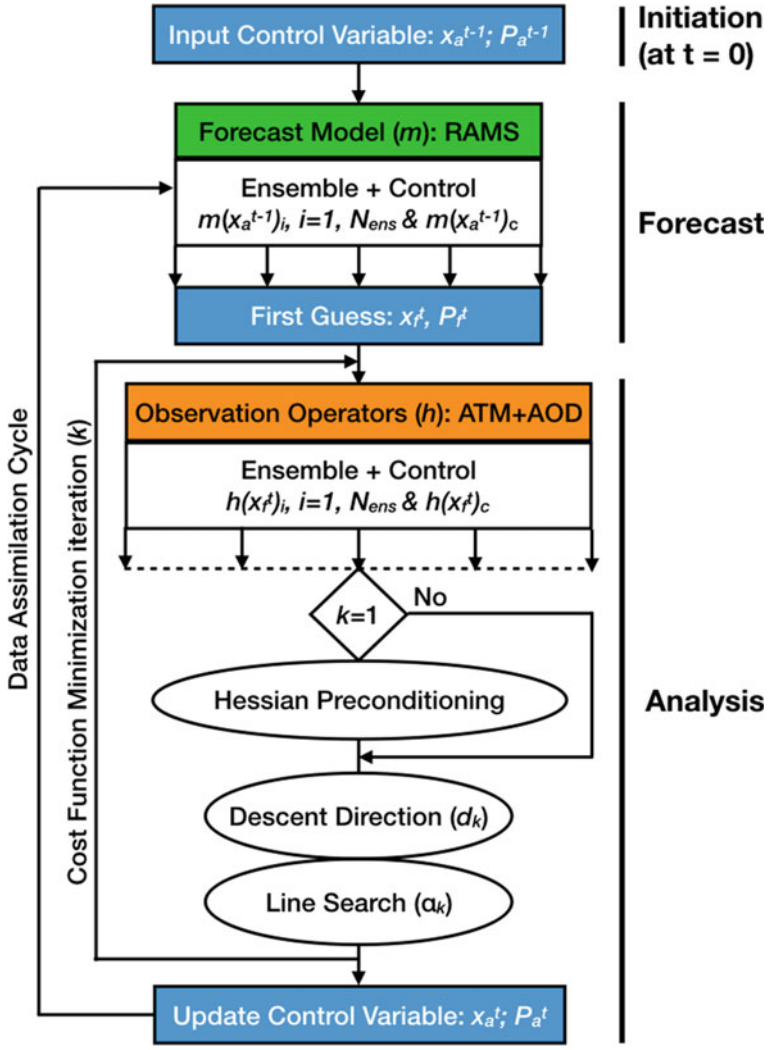

Fig. 4 A flow chart of the RAMS-MLEF system. Interfaces between MLEF and RAMS are highlighted in the following colors: blue boxes represent interfaces for Input/Output (I/O) between MLEF and RAMS, the green box represents the interface as a driver to call and run RAMS, and the orange box represents the interface for observation operators, which require input from RAMS

variational methods (e.g., Parrish and Derber 1992; Zupanski 1993; Rabier et al. 1999), MLEF, a hybrid system, solves the prescribed cost function, Eq. (6), with Hessian preconditioning in the ensemble space,

$$J(x) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_b})^T P_b^{-1}(\boldsymbol{x} - \boldsymbol{x_b}) + \frac{1}{2}[y - h(x)]^T R^{-1}[y - h(x)] \qquad (6)$$

where **R** is the observation error covariance, which is often a diagonal matrix following the assumption that observations are not spatially correlated. Although any forecast model, as indicated by ***m*** in Fig. 4, can be interfaced with MLEF, this study utilizes RAMS.

A schematic diagram shown in Fig. 4 outlines the components of the RAMS-MLEF system. Specifically, three interfaces are implemented in MLEF and they are (1) Input/Output (I/O) interfaces between MLEF and RAMS, (2) an interface that acts as a driver to call and run RAMS, and (3) an interface for observation operators that utilize RAMS output to compute the first guess of assimilated quantities as part of the innovation of data assimilation. In MLEF, observation operators for atmospheric observations are adapted from the forward component of the Gridpoint Statistical Interpolation (GSI; Wu et al. 2002; Kleist et al. 2009) through a module as illustrated by ATM in the orange box of Fig. 4. With that, atmospheric observations that are provided by NCEP, such as the conventional observations (e.g., radiosonde, surface station, buoy, etc) within the NCEP Prepared Binary Universal Form for the Representation of meteorological data (PrepBUFR) dataset and non-conventional atmospheric observations provided by satellite radiances data from various platforms, can be assimilated by MLEF, which is consistent with operations at NCEP. However, the AOD observation operator that is embedded in the Community Radiative Transfer Model (CRTM; Han et al. 2006), which is one of the observation operators within GSI, was specifically designed for the GOCART (Chin et al. 2000) aerosol species. Similarly, an AOD observation operator was developed specifically for the RAMS aerosol module within the RAMS-MLEF system.

In the RAMS-MLEF system, an observation operator for AOD specific for the RAMS aerosol module was developed in accordance with Eq. (4). Out of the nine aerosol categories, eight of them are used, i.e., $N_{aero} = 8$, to calculate AOD for this study. Supermicrometer sulphate is not used due to its little contribution to the total AOD. The optical properties of the eight aerosol categories at 0.55 μm under dry conditions are provided in Table 1. The mass extinction coefficient is computed using Mie theory, in which the spherical assumption of aerosol particles is required. For each of the aerosol categories, particles are first grown hygroscopically to equilibrium with ambient relative humidity using κ-Köhler theory (Petters and Kreidenweis 2007) and the refractive index is adjusted based on volume mixing with water. To reduce computational expense, a lookup table of the mass extinction coefficient as a function of ambient relative humidity (RH, %) for each of the eight aerosol categories at 0.55 μm is prepared. A 1% interval of RH is used in the lookup table, which is plotted in Fig. 5. For a simulated RH with a value that falls between two integer numbers (e.g., 85.6%), the integer value that is closer to the simulated value will be used (e.g., 86%).

Configuration of the RAMS-MLEF used for this study is now described. A time-lagged methodology (Suzuki and Zupanski 2018) is used to generate an initial set of N ensemble RAMS forecasts, which are valid at a prescribed initial time (0000 UTC 03 August 2016 is used for this study). As mentioned in Suzuki and Zupanski (2018), the so-called time-lagged methodology involves running a single deterministic or control forecast centered at the initial time ($t = 0$) of data assimilation, i.e., from $t =$

Table 1 Optical properties for the RAMS aerosol categories under dry conditions and their hygroscopicity parameters

Aerosol categories	Geometric median diameter (μm)	Effective radius (μm)	Geometric standard eeviation (μm)	Density (g cm^{-3})	Index of refraction at 0.55 μm (real)	Index of refraction at 0.55 μm (imaginary)	Mass extinction coefficient (m^2 g^{-1}) at 0.55 μm	Hygroscopicity parameter (kappa)
sub-micrometer sulfate	0.08	0.095	1.8	1.76	1.524	1.00E–07	2.44	0.55
sub-micrometer mineral dust	1.398	1.658	1.8	2.5	1.5	0.0005	0.411	0.04
super-micrometer mineral dust	5.9	6.997	1.8	2.65	1.5	0.0005	0.0854	0.05
film-mode seasalt	0.2	0.237	1.8	2.2	1.5	1.00E–08	3.04	0.8
jet drop-mode sea salt	2.0	2.372	1.8	2.2	1.5	1.00E–08	0.312	0.8
spume-mode sea salt	12.0	14.232	1.8	2.2	1.5	1.00E–08	0.0514	0.8
sub-micrometer regenerated aerosol	0.02	0.024	1.8	2.4	1.524	6.00E–03	0.166	0.17
super-micrometer regenerated aerosol	2.0	2.372	1.8	2.4	1.524	6.00E–03	0.298	0.17

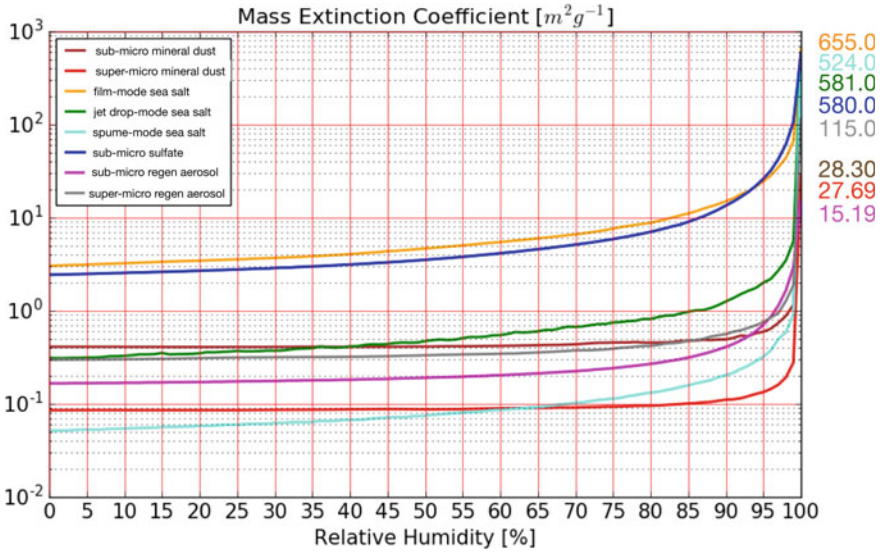

Fig. 5 Mass extinction coefficient (m^2 g^{-1}) as a function of relative humidity (RH ; %) at 0.55 μm for the eight RAMS aerosol categories listed in Table 1. Colored numbers on the right-hand side of the figure indicate values of mass extinction coefficient at RH = 100%

$-T$ to $t = +T$, where T is a specified assimilation window ($T = 6$ h is used in this study). During this deterministic forecast, RAMS is configured to generate output at every *2 T/N* step and thus creating $N + 1$ output, where N denotes the size of the ensemble ($N = 32$ for the August 2016 study). Out of the total of $N + 1$ output, the output that is valid at $t = 0$ is denoted by an Mx1 column matrix $\boldsymbol{x_c}$, where M is the total number of control variables times grid points of a RAMS domain and c indicates the control member. The other N outputs are used to define ensemble perturbations ($\boldsymbol{p_i}$, $i = 1, N$) at $t = 0$ by calculating $\boldsymbol{p_i} = \frac{1}{\sqrt{N}}(\boldsymbol{x_i}\text{-}\boldsymbol{x_c})$, where $\boldsymbol{x_i}$ is the state from an ensemble member and $\boldsymbol{p_i}$ is one column of a matrix whose square is $\mathbf{P_f}$.

Each assimilation cycle of the RAMS-MLEF system begins with a 6 h ensemble and control forecasts and ends with a control analysis along with the associated analysis error covariance, $\mathbf{P_a}$. At the end of the ensemble and control forecasts of any cycle, $\mathbf{P_f}$, which contains the cross-component ingredients for strongly coupled data assimilation, is re-computed and used as part of the cost function for the assimilation of observational data. Results at the end of a cycle include an updated $\boldsymbol{x_c}$, i.e., the analysis field, and the associated analysis error covariance, which is used to characterize the uncertainty of the analysis field.

Covariance inflation is used to increase the ensemble spread during each assimilation cycle. Due to the use of identical lateral boundary conditions, ensemble members may collapse. One way to avoid ensemble members from collapsing is to use the covariance inflation methodologies described in Zhang et al. (2004) and Whitaker

and Hamill (2012), which act to increase the ensemble spread in order to account for unrepresented error of sources. In the RAMS-MLEF system, a linear combination of these two methods are used, where 50% of weight is given to the method described in Whitaker and Hamill (2012) and 50% of weight is given to the method described by Zhang et al. (2004).

As mentioned earlier, success of a coupled data assimilation system is highly dependent on the choice of control variables. A set of control variables used in the RAMS-MLEF system includes the following: the three-dimensional wind components (*u*, v, and *w*), perturbation Exner function (*pi*), ice-liquid water potential temperature (θ_{il}), water vapor mixing ratio (*rv*), and the mass mixing ratio of the sub- and super-micrometer mineral dust (*md1mp* and *md2mp*). Because RAMS uses a leapfrog time stepping scheme, two temporal solutions, t_1 and t_2, exist only for the *u*, *v*, *w*, and *pi* prognostic variables, where an Asselin filter (Cotton et al. 2003) is used to prevent the two temporal solutions from diverging via damping the computational mode. In order to preserve the difference of the two temporal solutions for *u*, *v*, *w*, and *pi*, the RAMS-MLEF system stores the differences before assimilation occurs, and then only alters the t_1 solution of *u*, *v*, *w*, and *pi* during the assimilation. After assimilation, the t_2 solution will be updated through the use of the stored differences. As a consequence, the differences between t_1 and t_2 stay the same before and after data assimilation even through both time solutions are changed. Note that the RAMS aerosol module (Saleeby and van den Heever 2013) uses a double-moment scheme, which predicts both mass mixing ratio and number concentration for all 9 aerosol categories.

As stated in Sect. 4.1, prediction of both mass (first moment) and number concentration (second moment) of dust may be included into a data assimilation study. Mass and number concentration for both the sub- (*md1mp and md1np)* and super-micrometer mineral dust (*md2mp and md2np*) are predicted by the RAMS aerosol module. Dust mass and numbers are predicted for two different particle sizes; one for the sub-micrometer (~ 0.41 μm radius) mineral dust, second for the super-micrometer (~ 1.74 μm radius) mineral dust. In other words, mass and number for sub-micrometer (super-micrometer) mineral dust is referred to as dust bin 1 (dust bin (2)). As stated above, only mass in each dust bin is updated during assimilation of observed quantities of dust, which results in an inconsistency between dust mass and numbers for each dust bin of an analysis. One method to rectify the inconsistency between mass and number in an analysis is to assume an average dust particle size for each dust bin and recompute the number concentration of each dust bin from the updated mass field and assumed particle size. Consequently, both mass and numbers in each dust bin within an analysis become consistent with one another. Since *u*, *v*, *w*, *pi*, θ_{il}, *rv*, and both moments of each dust bin have been updated, the next assimilation cycle begins with the forecast initialized from the analysis.

5.3 *Application of the RAMS-MLEF System*

One RAMS-MLEF experiment named ATMAOD is carried out from 0000 UTC 03 August to 0600 UTC 04 August 2016 with a 6-hourly data assimilation cycle (total of 6 cycles). In this ATMAOD experiment, both the conventional atmospheric observations from NCEP PrepBUFR dataset and the 0.55 μm MODIS AOD retrievals are assimilated. There is only one domain used and the domain is composed of 400 east–west, 225 north–south, and 50 vertical grid points. In Fig. 6, the NCEP PrepBUFR dataset used in the ATMAOD experiment is displayed. Note that the majority of the dataset is only available at the surface and is indicated by green, blue, and orange symbols. Red symbols indicate the location of rawinsondes, which are the only source of conventional data that provide information from the surface to approximately the lower stratosphere of the atmosphere.

Due to the availability of MODIS data that is used to produce AOD retrievals, AOD retrievals are only assimilated at the cycle 2 (0600 UTC 03 August), cycle 3 (1200 UTC 03 August) and cycle 6 (0600 UTC 04 August) of the ATMAOD experiment (Fig. 7a). For the study herein, an observation error value for the AOD retrievals, a unitless quantity, is 0.1. Similar to Remer et al. (2005) and Liu et al. (2011), AOD observation error (*Err*) is increased by 5% (15%) for ocean (land) scenes (see Eq. 7).

$$\mathrm{Err}_{\mathrm{ocean}} = 0.1 + 0.05 * \mathrm{AOD}$$

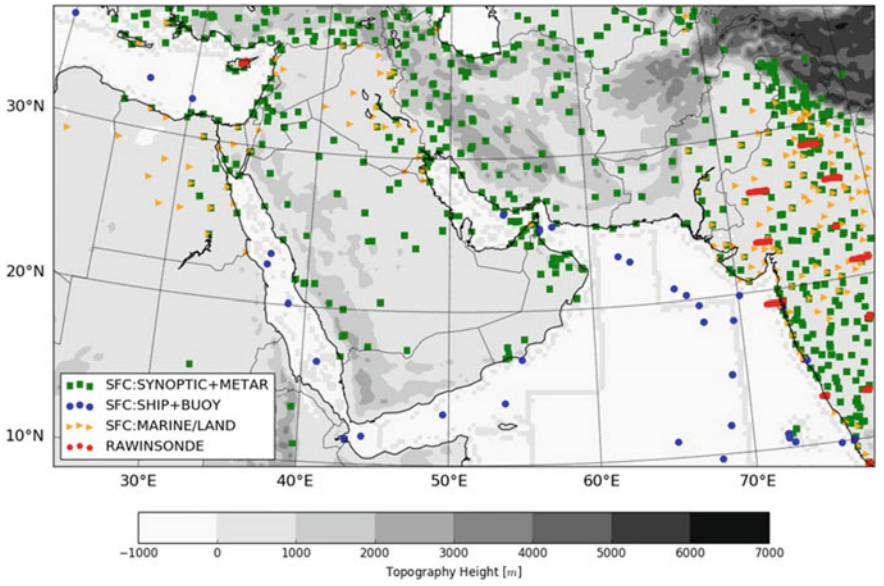

Fig. 6 NCEP PrepBUFR dataset that was assimilated into the ATMAOD experiment over the RAMS domain that covers the Arabian Peninsula. Topographic height (m) is plotted in gray scale

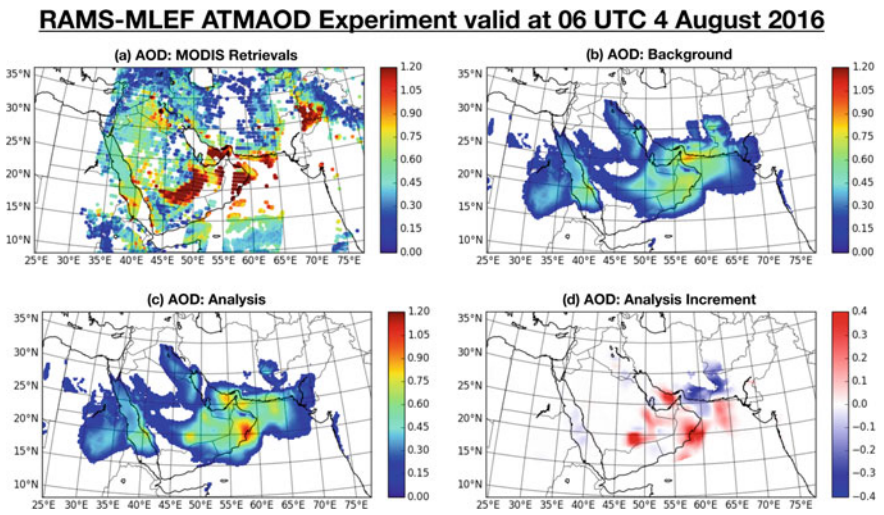

Fig. 7 Horizontal distribution of AOD: **a** retrievals from MODIS and **b**–**d** RAMS simulated AOD field computed from cycle 6 of the RAMS-MLEF ATMAOD experiment: **b** background, **c** analysis, and **d** analysis increment, i.e., analysis minus background: **c**–**b**. Note that AOD is a unitless quantity. Valid time is 0600 UTC 4 August 2016

$$\mathrm{Err}_{\mathrm{land}} = 0.1 + 0.15 * \mathrm{AOD} \tag{7}$$

In order to reduce the effects of spatially correlated observation error, data thinning is applied to the AOD retrievals prior to the actual assimilation. For a given cycle, AOD retrievals are first thinned such that every fifth pixel of a given retrieval image is excluded from assimilation and used for verification. Once spatial thinning is completed, the next step is quality control. During the quality control procedure, the so-called gross check is applied to remove large differences (usually three times the prescribed observation error, where observation error is one standard deviation for the assumed Gaussian distribution) between the AOD retrievals and the first guess.

In Fig. 7, assimilated MODIS AOD retrievals (thinned and passed quality control) are presented along with simulated AOD computed from the background and analysis field of cycle 06 of the ATMAOD experiment along with the difference between the analysis and background AOD (i.e. analysis increment). Both the background and analysis appears to have captured the general distribution of AOD (Fig. 7b, c), however, with slightly smaller magnitude compared to the retrievals (Fig. 7a). Nevertheless, after assimilating the AOD retrievals, the representation of the Persian plume (around 55°E and 26°N) and the Saudi Plume (from 45°E and 18°N to 52°E and 23°N) (see Fig. 3) is improved from the background in the analysis of ATMAOD experiment. The analysis increment of AOD further confirms that by assimilating MODIS AOD retrievals, the magnitude of AOD of both plumes are increased from background to analysis to reflect the assimilation.

In addition to the ATMAOD experiment, an AODONLY experiment, in which only AOD retrievals from MODIS were assimilated, was utilized to examine the role of assimilating atmospheric observations in the RAMS-MLEF system. The AODONLY experiment was performed by running a 6 h forecast from the analysis of cycle 5 of the ATMAOD experiment, and then assimilating AOD retrievals into the 6 h forecast valid at 0600 UTC 04 August 2016 for a resulting AODONLY analysis valid at the same time. Another 6 h forecast was run from the AODONLY analysis and was valid at 1200 UTC 04 August 2016. Differences between the two experiments were examined in order to understand impact of assimilating atmospheric observations on simulated dust; that is, variables from the AODONLY experiment were subtracted from the same variables from the ATMAOD experiment. Specifically, total dust ($md1mp + md2mp$) difference at the lowest model level between the two experiments is shown in Fig. 8. Since there were few atmospheric observations over the region of interest (e.g., the Saudi and the Persian plumes), their impact is limited to Persian Gulf coastal areas. In Fig. 8a, where total dust difference at the cycle 06 analysis is shown, one can notice a positive difference in the southeast part of the Persian Gulf, i.e. an increase of total dust due to assimilated atmospheric observations, and

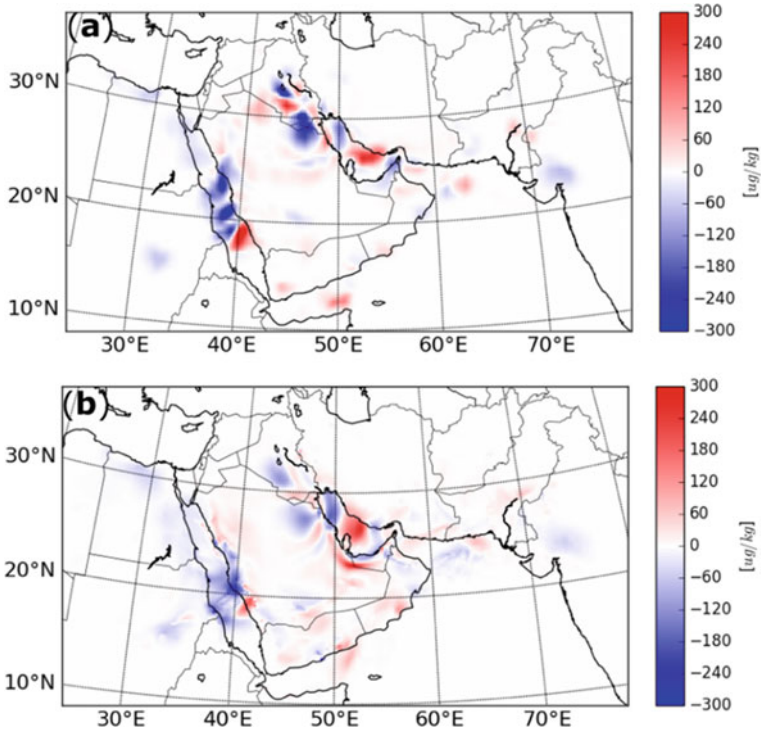

Fig. 8 Total dust (ug kg^{-1}) difference, ATMAOD experiment minus AODONLY experiment, at the lowest model level for **a** the analysis at cycle 6, valid 0600 UTC 4 August 2016 and **b** the 6-h forecast initialized from the analysis valid at 1200 UTC 4 August 2016

a negative difference in the northwest part of the Persian Gulf, indicating a decrease of total dust due to atmospheric observations. A 6 h forecast difference valid at 1200 UTC 04 August 2016 (Fig. 8b) also shows that the analysis differences are generally retained in the forecast. There is subtle change in the magnitude and the pattern of the total dust difference, but it is possible to identify and follow the movement of these changes over the 6 h time period. Such a result indicates that data assimilation was able to transform the information from atmospheric observations to dust initial conditions in such a way that it is supported by coupled model dynamics. More importantly, this result suggests that ensemble cross-covariance in strongly coupled data assimilation can have a satisfactory structure, which is encouraging for future applications.

5.4 Synthetic Geostationary Satellite Imagery.

Since dust is included in the RAMS-MLEF system, a new way to visualize output is needed. In Sect. 5.2, reference was made to the CRTM, which is part of GSI. Brightness temperatures (Tbs) of NWP data, void of dust, are computed by the CRTM, which are used by GSI in an assimilation process. However, since the RAMS-MLEF system contains dust, an AOD observation operator, distinct from the CRTM, was developed for the RAMS-MLEF system, which is dependent on solar reflection at 0.55 μm (see Sects. 3.2 and 5.2). A method is sought to visualize increments, which are independent of the AOD observation operator within the RAMS-MLEF system. To this end, Tbs for the SEVIRI instrument onboard MSG-08 (see Sect. 5.1) were computed, from output of the RAMS-MLEF system, with a radiative transfer model (RTM; Grasso et al. 2008), which was designed to include both moments of each of the two dust bins in the RAMS-MLEF system. Computed satellite imagery hereafter is referred to as synthetic imagery.

Several variables are needed in order to compute synthetic SEVIRI imagery. For this study, synthetic imagery was computed at both 10.80 μm and 12.00 μm, since values of Tb(10.80 μm)–Tb(12.00 μm) are useful to examine increments of simulated dust. Thus, the following two-dimensional variables were required: Latitude, longitude, and surface temperatures of both land and water bodies. Both latitude and longitude were used to compute the spectrally dependent two- dimensional surface emissivity from a monthly global dataset (Seemann et al. 2008) for the two wavelengths 10.80 and 12.00 μm. Furthermore, the following three-dimensional variables were also required: Pressure, temperature, water vapor mixing ratio along with the mass and number concentration of each dust bin. Although cloud condensate is present in RAMS-MLEF, synthetic imagery will focus exclusively on dust to avoid instances of modeled cloud layers covering and/or mixing with dust. Additional information is also needed to compute synthetic SEVIRI imagery.

In addition to modeled variables, spectrally and size dependent optical properties of dust were also required. Specifically, values of the complex index of refraction of dust, at 10.80 μm and 12.00 μm, were acquired from the Aerosol Refractive

Index Archive (ARIA; http://eodg.atm.ox.ac.uk/ARIA/, last access: 25 August 2020). Values of the complex index of refraction were used by Mie theory (Bohren and Huffman 1983) to compute the following optical properties for both wavelengths and each dust bin: Mass extinction, single-scattering albedo, and an asymmetry factor. That is, two sets of optical properties were computed; one set for *md1mp* and a second set for *md2mp*. In order for the RTM to generate synthetic MSG-08 SEVIRI imagery, the two sets of optical properties must be combined into one set, which will be referred to as the bulk set of optical properties.

Use was made of the second moment of each dust bin in order to compute the bulk set of optical properties. For example, the bulk single-scattering albedo, B_{ssa}, was computed by adding the product of the number concentration of bin 1, *md1np*, and single-scattering albedo of bin 1, ssa1, to the product of the number concentration of bin 2, *md2np*, and single-scattering albedo of bin 2, ssa2; the result was divided the sum of $md1np + md2np$, see Eq. (8).

$$B_{ssa} = \frac{md1np \cdot ssa1 + md2np \cdot ssa2}{md1np + md2np} \tag{8}$$

A similar number concentration weighted mean of the asymmetry factor resulted in the bulk asymmetry factor. Computation of the bulk mass extinction was slightly more involved. Values of the mass extinction coefficient for bin 1 and bin 2, from Mie, were multiplied by the mass of dust in bin 1 and bin 2, respectively to yield mass extinction. Bulk values of the mass extinction were then computed from a number concentration weighting mean of the mass extinction of each dust bin. All values of the bulk optical properties along with two- and three-dimensional variables from RAMS-MLEF were used by the RTM to generate MSG-08 SEVIRI synthetic imagery for each wavelength. Synthetic MSG-08 SEVIRI imagery at 10.80 and 12.00 μm was computed by the RTM for both the background and analysis fields. One advantage of synthetic imagery is the ability to visualize increments, which is a difference between background and analysis fields, with and without simulated dust; something that is impossible to achieve with observed imagery.

In order to evaluate model output, a comparison of simulated RAMS output with observations is necessary. Data from CALIOP (Winker et al. 2009), onboard CALIPSO, was used to produce a Vertical Feature Mask (VFM), which displays different scattering objects in the atmosphere of the Earth. For the August 2016 case herein, a descending CALIPSO ground track, white contour oriented north-northeast to south-southwest with arrows, valid about 2225 UTC 03 August 2016, is superimposed on true-color imagery from MODIS, valid near 2220 UTC 03 August 2016 (Fig. 9a). Corresponding to the CALIPSO ground track is the VFM, within which different atmospheric constituents are identified, from CALIOP (Fig. 9b). As indicated by the VFM, observed dust extended from the surface to a height of approximately 6.0 km; which is indicated by a horizontal dashed red contour. Total simulated dust mass, $md1mp + md2mp$, within a vertical cross section from RAMS, valid 0600 UTC 04 August 2016, green line in Fig. 9a, exhibited dust from the surface

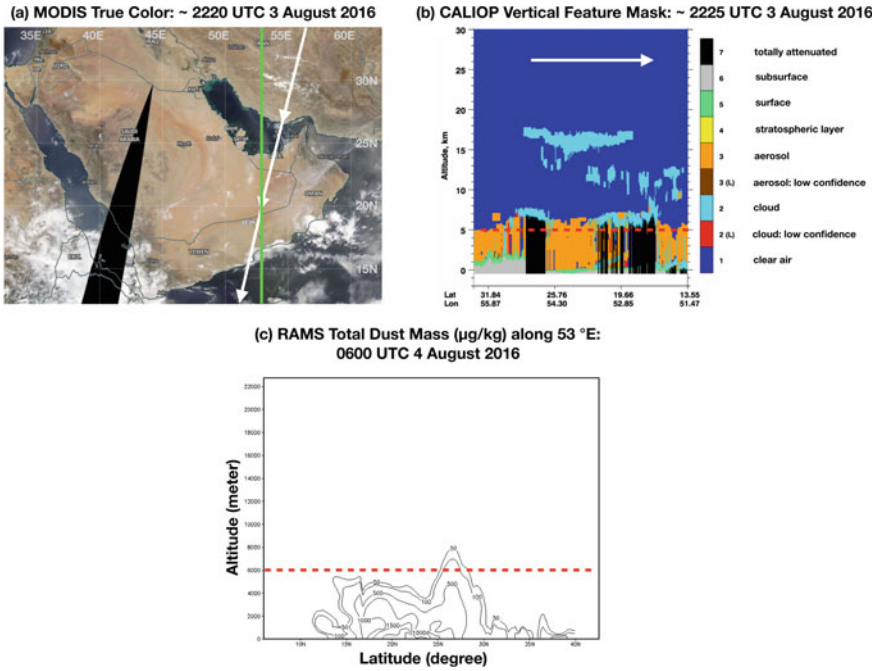

Fig. 9 **a** Composite true-color imagery from MODIS; the portion of the composite east of 45 E is valid at approximately 2220 UTC 3 August 2016. A white line segment with arrows denotes the ground track and motion for CALIPSO at approximately 2225 UTC 3 August 2016. A green line segment is used to denote the location of a vertical cross section from RAMS. **b** VFM from CALIOP along the CALIPSO ground track in (**a**); observed dust extended from the surface to about 6.0 km; a broken red line segment denotes a constant height of 6.0 km. **c** vertical cross section, along the green line in **a**, of the total simulated dust mass, *md1mp* + *md2mp,* which extended from the surface to about 6.0 km, valid at 0600 UTC 4 August 2016

to approximately 6.0 km (Fig. 9c). That is, the depth of observed dust supported the depth of dust simulated by RAMS.

Physical interpretation of increments of synthetic imagery is aided by increments of the total simulated dust mass. Dust mass of *md1mp* and *md2mp* of the background were added and then summed in the vertical throughout the depth of the simulated domain; a similar procedure was applied to the total dust mass of the analysis. Subtraction of the background dust field from the analysis dust field formed the dust increment shown in Fig. 10a. Positive (negative) regions in Fig. 10a indicated regions where dust mass was increased (decreased) as a result of the assimilation of observed AOD from MODIS. In addition to changes in *md1mp* and *md2mp* in the RAMS-MLEF assimilation system, the following three thermodynamic variables were also changed as a result of assimilation of observed AOD: Pressure, temperature, and water vapor mixing ratio (see Sect. 5.3). In order to examine the impact of increments of the three thermodynamic variables, RTM imagery at 10.80 μm (Fig. 10b) and 12.00 μm (Fig. 10c) were first produced with dust absent. Although the patterns

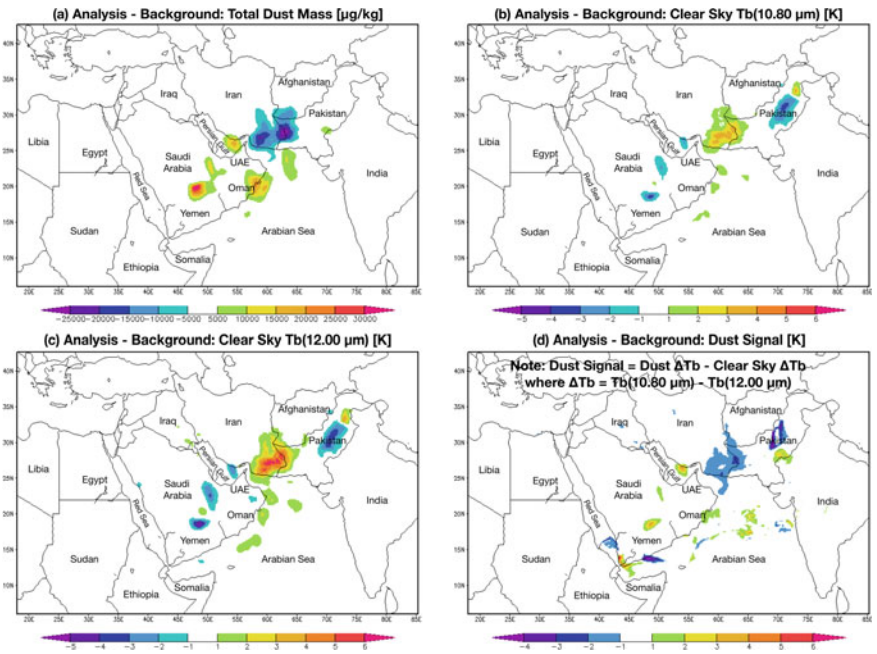

Fig. 10 Increments from RAMS-MLEF output valid at 0600 UTC 4 August 2016. Total simulated dust mass increment is shown in **a**; synthetic MGS-08 SEVIRI increments at 10.80 and 12.00 μm are displayed in **b** and **c**, respectively; increments in the dust signal are shown in **d**

evident in both Fig. 10b, c are similar, the amplitude of values were larger in synthetic imagery at 12.00 μm. Note also the opposite behavior between patterns in the total dust mass increment (Fig. 10a) and patterns in the increments of synthetic imagery at both 10.80 μm and 12.00 μm. In particular, a decrease (increase) of total dust mass resulted in an increase (decrease) of values of Tbs in imagery at both wavelengths. There were, however, regions in the synthetic increments that exhibited a lack of any relation to the dust increments; for example, central Pakistan. One possible reason for non-zero increments in synthetic imagery, that is independent of increments in dust, is a consequence of the background error covariance matrix. The background error covariance matrix spreads assimilated observations across variables and model grid points. Subsequently, increments in imagery can result as a consequence of a change in of one or more non-dust variables.

Unlike the opposite behavior between increments in total dust and increments in synthetic imagery, a similar behavior was evident between total dust increments and increments in the dust signal (Fig. 10d). In order to understand the physical interpretation of the dust signal in Fig. 10d, an explanation of how the dust signal was computed is warranted. Values of the channel difference, $\Delta Tb = Tb(10.80\,\mu m) - Tb(12.00\,\mu m)$, may be used to detect dust; however, if the clear-sky surface is desert, then dust detection with the channel difference may be a challenge since a dust signal

may blend in with the clear-sky desert surface. One strategy, proposed herein, to isolate the dust signal is to subtract the clear-sky channel difference from the dust channel difference; that is, the dust signal is equal to $\Delta Tb_{dust} - \Delta Tb_{clear-sky}$. As a consequence, the increment in the dust signal is the dust signal of the background subtracted from the dust signal of the analysis (Fig. 10d). Regions where the dust increment in Fig. 10a increased (decreased) corresponded in an increase (decrease) in the increment of the dust signal in Fig. 10d. In particular, when the assimilation of observed AOD increased dust mass, there was a corresponding increase in the dust signal; for example, along the northern coast of the Persian Gulf, interior Sadia Arabia, border of Pakistan and India, and along the coast of Oman. In response to a reduction of total dust mass over the border of Iran and Pakistan, values of the dust signal decreased in the same region. There were also regions of values of the increment of the dust signal that showed little relationship to increments in the total dust mass. For example, there was a negative increment of the dust signal over central Pakistan, which may be a result of the background error covariance matrix. As a way to link this section with Sect. 5.3, patterns of increments of the dust signal (Fig. 10b) were similar to patterns of increments in AOD (Fig. 7d in Sect. 5.3).

5.5 Model Response to Adjustments from Data Assimilation.

As discussed in Sect. 5.3, the ATMAOD experiment assimilates both atmospheric and aerosol observations and updates a list of control variables as part of the analysis step of each six-hourly assimilation cycle. Other RAMS prognostic variables will respond to the changes in the control variables throughout the forecast step of the next data assimilation cycle by the model dynamical core and physical parameterizations (e.g. microphysical scheme, radiation scheme, etc.). With that, this section focuses on shedding light on the following question: What is the difference between a short-term forecast from a background initial state; that is, prior to AOD assimilation, and a forecast from an analysis initial state; that is, after AOD assimilation? In particular, this section discusses the influence of the modified total dust mass (*md1mp* + *md2mp*), which resulted from AOD assimilation, on the hydrometeor condensate field and shortwave outgoing energy. To this end, a few definitions are in order: Two simulations were conducted: (1) a simulation initialized from an analysis, which resulted from the assimilation of AOD, and is referred to as the Assimilation Forecast (AF), and (2) a simulation initialized from a background, from which an analysis is derived, and is referred to as the Background Forecast (BF). Both the AF and BF began at 0600 UTC 04 August 2016. Focus will be given to values of the Vertically Integrated Total Dust Mass (VITDM) of the BF subtracted from values of the VITDM of the AF (shaded in Fig. 11). Thus, positive values of the VITDM in Fig. 11 indicated that the assimilation of AOD increased the total dust mass in the AF compared to the BF. There are five regions in Fig. 11, within which the influence of assimilation of AOD on total condensate is discussed presently.

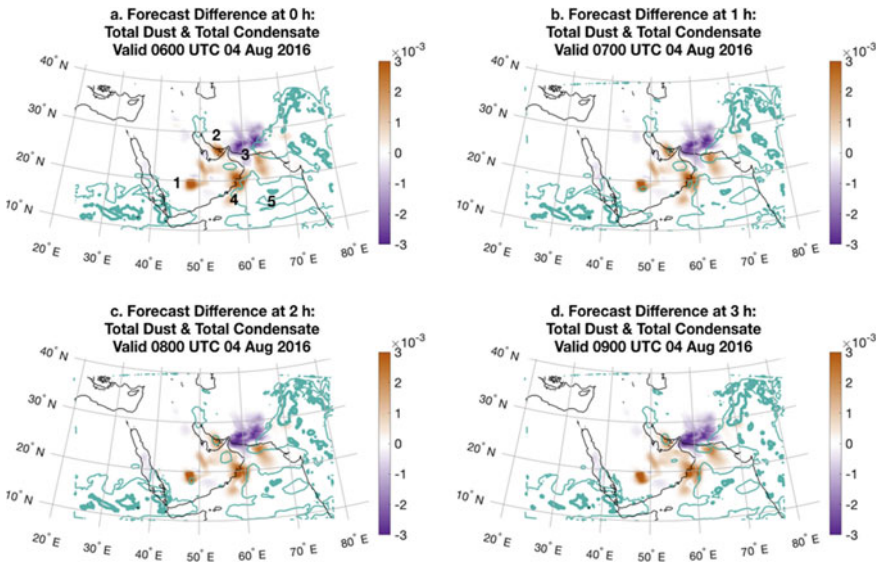

Fig. 11 Vertically integrated total dust mass (*md1mp* + *md2mp*; kg m^{-2}) difference between the BF and AF simulations (shaded; AF minus BF) at **a** 0 h forecast, **b** 1 h forecast, **c** 2 h forecast, and **d** 3 h forecast initialized from 0600 UTC 04 August 2016. The two green contours are used to indicate values of vertically integrated total condensate mass (mm) of a simulation initialized with an analysis field: thin for 0.1 mm; thick for 1.0 mm

Two responses of the assimilation of AOD on simulated total condensate are identified: direct and indirect. Focus will be given to regions 1, 2 (the Saudi plume and the Persian plume, respectively (see Sects. 5.1, 5.3, 5.4, and 5.5) in Fig. 11a. A plausible direct response occurred in regions 1, 2, 3, and 4 while a plausible indirect response occurred in region 5. A direct response occurred from the following: Assimilation of AOD resulted in an increase of values of the *md1mp* + *md2mp* in the AF, which subsequently leads to a modification of the total number concentration, since only dust mass is a control variable (see Sect. 5.3), which then resulted in an increase of the population of Cloud Condensation Nuclei (CCN). That is, given a fixed dust particle size, an increase in dust mass, due to assimilation of AOD, will cause an increase in the dust number concentration. Development of simulated condensate occurs in RAMS when supersaturation increases above a critical value. Supersaturation is a function of upward vertical motion; therefore, when upward vertical motion occurs, supersaturation may increase above a critical value. Once supersaturations increase above a critical value, a certain percentage of the CCN population is activated to become cloud droplets, which begins a complex interaction of simulated microphysical habit types. One simulated hour after the AF simulation began, 0.1 mm of vertically integrated total condensate developed in regions 1, 2, and 4 at 0700 UTC 04 August 2016 (Fig. 11b). A progression occurred in region 3 where the 0.1 mm contour moved westward, bounding a local maximum of total

dust mass, by 0800 UTC (Fig. 11c), followed by a closed contour of 0.1 mm of vertically integrated total condensate at 0900 UTC (Fig. 11d). Notice in Fig. 11, region 5 was characterized by small changes in values of the VITDM. In response to complex changes of control variables, in region 5, through the horizontal spread of information from the flow-dependent background error covariance matrix during the assimilation of AOD at 0600 UTC 04 August 2016, temporal changes in simulated total condensate (Figs. 11 a–d), in region 5, occurred as an example of an indirect response to the assimilation of AOD. In the interest of brevity, a plausible explanation of both the direct and indirect response to the assimilation of AOD on total condensate was provided above.

Direct and indirect responses of the assimilation of AOD on simulated solar reflection are also identified. An additional consequence of increased VITDM (Fig. 11 shaded) in the AF compared to the BF was an increase in the outgoing shortwave energy (Fig. 12). That is, direct and indirect responses of the assimilation of AOD on the simulated energy budget are presently discussed. Although values of the control variable θ_{il} are prognostic, values of surface potential temperature are diagnostic. Consequently, a forecast must begin in order for the surface potential temperature to be diagnosed; thus, the time of 0610 UTC in Figs. 12 a and b. This discussion will

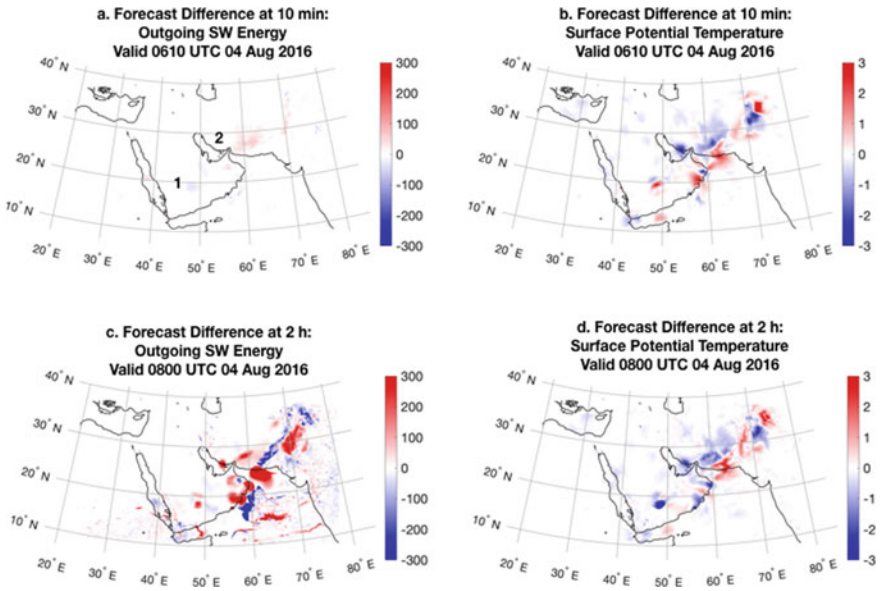

Fig. 12 **a** Values of the difference of outgoing simulated shortwave energy (W m^{-2}) computed from the BF subtracted from the AF at 10 min forecast initialized from 0600 UTC 4 August 2016. Positive values (red) indicate more outgoing shortwave from the AF compared to the BF simulations. **b** Values of the difference of simulated surface potential temperature (K) from the BF subtracted from the AF also at 10 min forecast initialized from 0600 UTC 4 August 2016. Negative values (blue) indicate cooler surface potential temperature from the AF compared to the BF simulations. **c–d** same as **a–b**, except for 2 h forecast

focus primarily on the Saudi plume and Persian plume, regions 1 and 2, respectively (Fig. 12a). At 0610 UTC slight variations of outgoing shortwave energy resulted from the assimilation of AOD (Fig. 12a). However, rather significant changes of surface potential temperature were already evident ten minutes into the AF simulation throughout the domain in regions away from 1 and 2 (Fig. 12b). As seen in Fig. 12b, the pattern of changes in surface potential temperature exhibited little resemblance to the pattern seen in Fig. 12a. A lack of similarity in patterns between Figs. 12a, b suggests that the influence of the flow-dependent background error covariance matrix may have been responsible for the patterns in surface potential temperature differences between the AF and BF simulations; that is, an indirect response of the assimilation of AOD on surface potential temperatures. In time, the enhanced reflection of shortwave energy from the dust mass in regions 1 and 2, evident in Fig. 12c, caused a reduction, or cooling, of the surface potential temperature at 1800 UTC. In other words, the loss of solar energy from the AF simulation, compared to the BF simulation, resulted in surface cooling below the enhanced VITDM for both regions 1 and 2 (Fig. 12 d); that is, a direct response of the assimilation of AOD on surface potential temperatures.

Although the above explanations are speculative, a more detailed analysis is, unfortunately, beyond the scope of this chapter. That is, demonstrating a link between cross-component control variables would require a thorough analysis on the role of the flow-dependent background error covariance matrix, which is responsible for updating values of control variables. That said, efforts in this section focused on providing plausible explanations for direct and indirect responses of the condensate and shortwave radiation fields to changes in total dust mass (and number concentration diagnosed afterwards) due to the assimilation of AOD.

6 Summary and Future Directions

As pointed out in Carrassi et al. (2018), coupled data assimilation is one of the major areas of active research in the field of geosciences and is expected to be advanced quickly in the coming future. In this chapter, theoretical and practical aspects of strongly coupled data assimilation with a focus on the aerosol and atmosphere coupling are discussed. We began this chapter by providing an overview and description of coupled data assimilation followed by an example from a single observation experiment of an aerosol-atmosphere coupled data assimilation using WRF-Chem. In Sect. 2, the current status of aerosol-atmosphere coupled data assimilation in both operational and research communities are reviewed in detail. Next, a description of available observational data of aerosols from various measurements such as AOD, satellite radiances, LIDAR backscattering, etc., along with a discussion of observational errors is given in Sect. 3. In Sect. 4, we present several major challenges associated with coupled data assimilation with a focus on aerosol applications. For example, the choice of control variable and the associated background error covariance is essential for the result of a successful coupled data assimilation. We

further provide a brief discussion on extending coupled data assimilation to include non-Gaussian and/or non-linear features as aerosols and their associated errors are known to behave as such. In addition, unlike meteorological observations, aerosols are under sampled. The lack of independent observations that can be used to verify the result from assimilating available aerosol observations is an issue that remains to be addressed by an improved observation network. Finally, we introduced the newly developed RAMS-MLEF, a strongly coupled aerosol-atmosphere data assimilation system, for the first time to study the impact of assimilating AOD under a strongly coupled system. A well-explored dust storm event over the Arabian Peninsula that occurred on 3–4 August 2016 was used as a case study to demonstrate the utility of the RAMS-MLEF system. In addition to examining analysis increments, which is a common practice in data assimilation, we use synthetic satellite imagery to further highlight the impact of aerosols from the viewpoint of satellite. Since short-term forecast is part of a typical data assimilation cycle, we also look into the response of aerosols adjustment from data assimilation during the short-term forecast. To end this chapter, a few future directions for research are provided.

Overall, more detailed assessments on the value of strongly coupled aerosol-atmosphere data assimilation is required. In particular, it is important for such assessments to be conducted under operational settings in order to examine more case studies with realistic configuration. In doing so, there is an urgent need to further address possibilities to improve the estimation of coupled background error covariance. While estimating coupled background error covariance under ensemble based framework may be straightforward, more work is required in order to accurately represent cross-component and cross-variable correlations for the variational aspect of hybrid based data assimilation methods (Ménard et al. 2019). In addition, using information theory to diagnose the degrees of coupling strength between any pairs of selected model variables within a coupled system can help choose control variables that are more relevant to the coupled system. Knowing the degrees of coupling strength can also benefit the efficiency of coupled data assimilation via simplifying portions of the background error covariance matrix due to low coupling strength and thus reducing computational cost. Provided that the background error covariance dictates the analysis increments, understanding the characteristics of the spatial and temporal scales of the physical processes within a coupled system is critical for assigning proper localization lengths between cross-component and cross-variable terms in the background error covariance matrix. As data assimilation methodologies advance, observations of aerosols and their corresponding observation operators also require more further development. For example, a recent study by Zhang et al. (2019) explored the use of artificial light sources for aiding AOD retrievals over nighttime. In the meantime, increasing temporal observation frequency as well as deploying instruments that allow observations of fine vertical distribution of aerosols are of critical values for improving our understanding of the spatiotemporal distribution of aerosol. There also exists a need to investigate the pros and cons of assimilation of satellite radiances sensitive to aerosols versus assimilation of retrieved quantities. Given that Artificial Intelligence (AI) techniques have shown promising results on emulating the atmosphere with sufficient training and data, there is potential to use

AI to facilitate and speed up the performance of aerosol assimilation via improved observation operators. Last but not least, verification of aerosol analysis and forecast using independent observations will benefit most from the availability of new types of observations and dense observational networks of aerosols.

References

Anderson JL (2001) An Ensemble Adjustment Kalman Filter for Data Assimilation. Mon Weather Rev 129:2884–2903

Auligné T, McNally AP, Dee DP (2007) Adaptive bias correction for satellite data in a numerical weather prediction system. Q J R Meteorol Soc 133:631–642. https://doi.org/10.1002/qj.56

Bannister RN (2008a) A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. Q J R Meteorol Soc 134:1951–1970.https://doi.org/10.1002/qj.339

Bannister A (2008b) A review of forecast error covariance statistics in atmospheric variational data assimilation. II: modelling the forecast error covariance statistics. Q J R Meteorol Soc 134:1971–1996. https://doi.org/10.1002/qj.340

Benedetti A, Coauthors (2009) Aerosol analysis and forecast in the European centre for medium-range weather forecasts integrated forecast system: 2. data assimilation. J Geophys Res Atmos 114:1–18.https://doi.org/10.1029/2008JD011115

Benedetti A, Coauthors (2018) Status and future of numerical atmospheric aerosol prediction with a focus on data requirements. Atmos Chem Phys 18:10615–10643.https://doi.org/10.5194/acp-18-10615-2018

Benedetti A, Fisher M (2007) Background error statistics for aerosols. Q J R Meteorol Soc 133:391–405. https://doi.org/10.1002/qj.37

Benedetti A, Di Giuseppe F, Jones L, Peuch VH, Rémy S, Zhang X (2019) The value of satellite observations in the analysis and short-range prediction of Asian dust. Atmos Chem Phys 19:987–998.https://doi.org/10.5194/acp-19-987-2019

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter Part I: theoretical aspects. Mon Weather Rev 129:420–436. https://doi.org/10.1175/1520-0493(2001)129%3c0420:ASWTET%3e2.0.CO;2

Bohren CF, Huffman DR (1983) Absorption and Scattering of Light by Small Particles. Wiley, Hoboken, p 544

Carrassi A, Bocquet M, Bertino L, Evensen G (2018) Data assimilation in the geosciences: an overview of methods, issues, and perspectives. Wiley Interdiscip Rev Clim Chang 9.https://doi.org/10.1002/wcc.535

Chen X, Andronova N, Van Leer B, Penner JE, Boyd JP, Jablonowski C, Lin SJ (2013) A control-volume model of the compressible Euler equations with a vertical lagrangian coordinate. Mon Weather Rev 141:2526–2544.https://doi.org/10.1175/MWR-D-12-00129.1

Chikhar K, Gauthier P (2017) Impact of lateral boundary conditions on regional analyses. Mon Weather Rev 145:1361–1379. https://doi.org/10.1175/MWR-D-16-0245.1

Chin M, Rood RB, Lin SJ, Müller JF, Thompson AM (2000) Atmospheric sulfur cycle simulated in the global model GOCART: model description and global properties. J Geophys Res Atmos 105:24671–24687. https://doi.org/10.1029/2000JD900384

Choi Y, Chen SH, Huang CC, Earl K, Chen CY, Schwartz CS, Matsui T (2020) Evaluating the impact of assimilating aerosol optical depth observations on dust forecasts over North Africa and the East Atlantic using different data assimilation methods. J Adv Model Earth Syst 12:1–30. https://doi.org/10.1029/2019MS001890

Collins WJ, Coauthors (2011) Development and evaluation of an earth-system model—HadGEM2. Geosci Model Dev 4:1051–1075.https://doi.org/10.5194/gmd-4-1051-2011

Collins WD, Rasch PJ, Eaton BE, Khattatov BV, Lamarque JF, Zender CS (2001) Simulating aerosols using a chemical transport model with assimilation of satellite aerosol retrievals: methodology for INDOEX. J Geophys Res Atmos 106:7313–7336. https://doi.org/10.1029/2000JD900507

Cotton WR, Coauthors (2003) RAMS 2001: current status and future directions. Meteorol Atmos Phys 82:5–29.https://doi.org/10.1007/s00703-001-0584-9

Dee DP (2005) Bias and data assimilation. Q J R Meteorol Soc 131:3323–3343. https://doi.org/10.1256/qj.05.137

Derber JC, Wu W-S (1998) The Use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. Mon Weather Rev 126:2287–2299. https://doi.org/10.1175/1520-0493(1998)126%3c2287:TUOTCC%3e2.0.CO;2

Derber JC, Wu W-S, Parrish DF, Lord SJ (1991) The new global operational analysis system at the national meteorological center. Wea Forecast 6:538–547

Descombes G, Auligné T, Vandenberghe F, Barker DM, Barré J (2015) Generalized background error covariance matrix model (GEN-BE v2.0). Geosci Model Dev 8:669–696. https://doi.org/10.5194/gmd-8-669-2015

Eltahan M, Alahmadi S (2019) Numerical dust storm simulation using modified geographical domain and data Assimilation: 3DVAR and 4DVAR (WRF-Chem/WRFDA). IEEE Access 7:128980–128989. https://doi.org/10.1109/ACCESS.2019.2930812

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. J Geophys Res 99:10143–10162

Eyre JR (1992) A bias correction scheme for simulated TOVS brightness temperatures

Fast JD, Gustafson WI, Easter RC, Zaveri RA, Barnard JC, Chapman EG, Grell GA, Peckham SE (2006) Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of Houston using a fully coupled meteorology-chemistry-aerosol model. J Geophys Res Atmos 111:1–29. https://doi.org/10.1029/2005JD006721

Federico S, Petracca M, Panegrossi G, Dietrich S (2017) Improvement of RAMS precipitation forecast at the short-range through lightning data assimilation. Nat Hazards Earth Syst Sci 17:61–76. https://doi.org/10.5194/nhess-17-61-2017

Fisher M (2003) Background error covariance modelling. Semin Recent Dev Data Assim 45–63

Fisher M (2006) Wavelet Jb—a new way to model the statistics of background errors. ECMWF Newsl Winter 200:23–28. https://doi.org/10.21957/ef5601ak43

Fletcher SJ, Jones AS (2014) Multiplicative and additive incremental variational data assimilation for mixed lognormal-Gaussian errors. Mon Weather Rev 142:2521–2544. https://doi.org/10.1175/MWR-D-13-00136.1

Grasso LD, Sengupta M, Dostalek JF, Brummer R, Demaria M (2008) Synthetic satellite imagery for current and future environmental satellites. Int J Remote Sens 29:4373–4384. https://doi.org/10.1080/01431160801891820

Grell G, Baklanov A (2011) Integrated modeling for forecasting weather and air quality: a call for fully coupled approaches. Atmos Environ 45:6845–6851. https://doi.org/10.1016/j.atmosenv.2011.01.017

Grell GA, Peckham SE, Schmitz R, McKeen SA, Frost G, Skamarock WC, Eder B (2005) Fully coupled online chemistry within the WRF model. Atmos Environ 39:6957–6975. https://doi.org/10.1016/j.atmosenv.2005.04.027

Guth J, Josse B, Marécal V, Joly M, Hamer P (2016) First implementation of secondary inorganic aerosols in the MOCAGE version R2.15.0 chemistry transport model. Geosci Model Dev 9:137–160. https://doi.org/10.5194/gmd-9-137-2016

Hakami A, Henze DK, Seinfeld JH, Chai T, Tang Y, Carmichael GR, Sandu A (2005) Adjoint inverse modeling of black carbon during the Asian Pacific Regional Aerosol Characterization Experiment. J Geophys Res D Atmos 110:1–17. https://doi.org/10.1029/2004JD005671

Han Y, van Delst P, Liu Q, Weng F, Yan B, Treadon R, Derber J (2006) NOAA technical report NESDIS 122 JCSDA Community Radiative Transfer Model (CRTM)—Version 1. NOAA Tech Rep

Hogan TF, Coauthors (2014) The navy global environmental model. Ocenography 27:116–125.https://doi.org/10.1017/CBO9781107415324.004
Houtekamer PL, Mitchell HL (2001) A sequential ensemble Kalman filter for atmospheric data assimilation. Mon Weather Rev 129:123–137. https://doi.org/10.1175/1520-0493(2001)129%3c0123:ASEKFF%3e2.0.CO;2
Hsu NC, Tsay SC, King MD, Herman JR (2006) Deep Blue retrievals of Asian aerosol properties during ACE-Asia. IEEE Trans Geosci Remote Sens 44:3180–3195. https://doi.org/10.1109/TGRS.2006.879540
Hu M, Xue M (2007) Impact of configurations of rapid intermittent assimilation of WSR-88D radar data for the 8 May 2003 Oklahoma City tornadic thunderstorm case. Mon Weather Rev 135:507–525. https://doi.org/10.1175/MWR3313.1
Jish Prakash P, Stenchikov G, Kalenderski S, Osipov S, Bangalath H (2015) The impact of dust storms on the Arabian Peninsula and the Red Sea. Atmos Chem Phys 15:199–222. https://doi.org/10.5194/acp-15-199-2015
Kahnert M (2008) Variational data analysis of aerosol species in a regional CTM: background error covariance constraint and aerosol optical observation operators. Tellus B 60:753–770
Kalnay E (2003) Atmospheric modeling data assimilation and predictability. Cambridge University Press, Cambridge, p 341
Kleist DT, Ide K (2015) An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. Mon Weather Rev 143:452–470. https://doi.org/10.1175/MWR-D-13-00350.1
Kleist DT, Ide K, Parrish DF, Derber JC, Treadon R, Wu W-S, Lord S (2009) Introduction of the GSI into the NCEP Global Data Assimilation System. Weather Forecast 24:1691–1705.https://doi.org/10.1175/2009WAF2222201.1
Kliewer AJ, Fletcher SJ, Jones AS, Forsythe JM (2016) Comparison of gaussian, logarithmic transform and mixed Gaussian-log-normal distribution based 1DVAR microwave temperature-water-vapour mixing ratio retrievals. Q J R Meteorol Soc 142:274–286. https://doi.org/10.1002/qj.2651
Lee E, Županski M, Županski D, Park SK (2017) Impact of the OMI aerosol optical depth on analysis increments through coupled meteorology–aerosol data assimilation for an Asian dust storm. Remote Sens Environ 193:38–53. https://doi.org/10.1016/j.rse.2017.02.013
Levy RC, Mattoo S, Munchak LA, Remer LA, Sayer AM, Patadia F, Hsu NC (2013) The collection 6 MODIS aerosol products over land and ocean. Atmos Meas Tech 6:2989–3034. https://doi.org/10.5194/amt-6-2989-2013
Liu Z, Liu Q, Lin HC, Schwartz CS, Lee YH, Wang T (2011) Three-dimensional variational assimilation of MODIS aerosol optical depth: Implementation and application to a dust storm over East Asia. J Geophys Res Atmos 116:1–19. https://doi.org/10.1029/2011JD016159
Lorenc AC (1986) Analysis methods for numerical weather prediction. Q J R Meteorol Soc 112:1177–1194. https://doi.org/10.1002/qj.49711247414
Lynch P, Coauthors (2016) An 11-year global gridded aerosol optical thickness reanalysis (v1.0) for atmospheric and climate sciences. Geosci Model Dev 9:1489–1522.https://doi.org/10.5194/gmd-9-1489-2016
Ménard R, Gauthier P, Rochon Y, Robichaud A, de Grandpré J, Yang Y, Charrette C, Chabrillat S (2019) Coupled stratospheric chemistry-meteorology data assimilation. Part II: weak and strong coupling. Atmosphere (Basel) 10. https://doi.org/10.3390/ATMOS10120798
Miller SD, Coauthors 2019 A tale of two dust storms: analysis of a complex dust event in the Middle East. Atmos Meas Tech 12:5101–5118.https://doi.org/10.5194/amt-12-5101-2019
Molod A, Takacs L, Suarez M, Bacmeister J, Song I-S, Eichmann A (2012) The GEOS-5 atmospheric general circulation model: mean climate and development from MERRA to fortuna
Morcrette JJ, Beljaars A, Benedetti A, Jones L, Boucher O (2008) Sea-salt and dust aerosols in the ECMWF IFS model. Geophys Res Lett 35:1–5. https://doi.org/10.1029/2008GL036041

Morcrette JJ, Beljaars A, Benedetti A, Jones L, Boucher O, Coauthors (2009) Aerosol analysis and forecast in the european centre for medium-range weather forecasts integrated forecast system: Forward modeling. J Geophys Res Atmos 114:1–17. https://doi.org/10.1029/2008JD011235

Pagowski M, Grell GA (2012) Experiments with the assimilation of fine aerosols using an ensemble Kalman filter. J Geophys Res Atmos 117:1–15. https://doi.org/10.1029/2012JD018333

Pagowski M, Liu Z, Grell GA, Hu M, Lin HC, Schwartz CS (2014) Implementation of aerosol assimilation in gridpoint statistical interpolation (v. 3.2) and WRF-Chem (v. 3.4.1). Geosci Model Dev 7:1621–1627. https://doi.org/10.5194/gmd-7-1621-2014

Parrish DF, Derber JC (1992) The national meteorological center's spectral statistical-interpolation analysis system. Mon Weather Rev 120:1747–1763. https://doi.org/10.1175/1520-0493(1992)120%3c1747:TNMCSS%3e2.0.CO;2

Penny SG, Hamill TM (2017) Coupled data assimilation for integrated earth system analysis and prediction. Bull Am Meteorol Soc 98:169–172. https://doi.org/10.1175/BAMS-D-17-0036.1

Penny SG, Hamill TM, Coauthors (2017) Coupled data assimilation for integrated earth system analysis and prediction: goals, challenges and recommendations, p 50. www.wmo.int/pages/prog/arep/wwrp/new/documents/Final_WWRP_2017_3_27_July.pdf

Petters MD, Kreidenweis SM (2007) A single parameter representation of hygroscopic growth and cloud condensation nucleus activity-Part 3: including surfactant partitioning. Atmos Chem Phys 7:1961–1971. https://doi.org/10.5194/acp-13-1081-2013

Putman WM, Lin SJ (2007) Finite-volume transport on various cubed-sphere grids. J Comput Phys 227:55–78. https://doi.org/10.1016/j.jcp.2007.07.022

Rabier F, Jarvinen H, Klinker E, Mahfouf J-F, Simmons A (1999) The ECMWF operational implementationof four dimensional variational assimilation. Part I: experimental results with simplified physics

Randles CA, Coauthors (2017) The MERRA-2 aerosol reanalysis, 1980 onward. Part I: description and data assimilation evaluation. J Clim 30: 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1

Remer LA, Coauthors (2005) The MODIS aerosol algorithm, products, and validation. J Atmos Sci 62:947–973.https://doi.org/10.1175/JAS3385.1

Rubin JI, Coauthors (2016) Development of the Ensemble Navy Aerosol Analysis Prediction System (ENAAPS) and its application of the Data Assimilation Research Testbed (DART) in support of aerosol forecasting. Atmos Chem Phys 16:3927–3951.https://doi.org/10.5194/acp-16-3927-2016

Rubin JI, Reid JS, Hansen JA, Anderson JL, Holben BN, Xian P, Westphal DL, Zhang J (2017) Assimilation of AERONET and MODIS AOT observations using variational and ensemble data assimilation methods and its impact on aerosol forecasting skill. J Geophys Res 122: 4967–4992.https://doi.org/10.1002/2016JD026067

Saleeby SM, Cotton WR (2004) A large-droplet mode and prognostic number concentration of cloud droplets in the colorado state university regional atmospheric modeling system (RAMS). Part I: module descriptions and supercell test simulations. J Appl Meteorol 43:182–195. https://doi.org/10.1175/1520-0450(2004)043%3c0182:ALMAPN%3e2.0.CO;2

Saleeby SM, van den Heever SC (2013) Developments in the CSU-RAMS aerosol model: emissions, nucleation, regeneration, deposition, and radiation. J Appl Meteorol Climatol 52:2601–2622.https://doi.org/10.1175/JAMC-D-12-0312.1

Saleeby SM, Cotton WR, van den Heever SC, Coauthors (2019) The influence of simulated surface dust lofting erodible fraction on radiative forcing. Atmos Chem Phys 19:10279–10301.https://doi.org/10.5194/acp-19-10279-2019

Sandu A, Liao W, Carmichael GR, Henze DK, Seinfeld JH (2005) Inverse modeling of aerosol dynamics using adjoints: theoretical and numerical considerations. Aerosol Sci Technol 39:677–694. https://doi.org/10.1080/02786820500182289

Schwartz CS, Liu Z, Lin H-C, Cetola JD (2014) Assimilating aerosol observations with a "hybrid" variational-ensemble data assimilation system. J Geophys Res Atmos 119:4043–4069.https://doi.org/10.1002/2013JD020937.Received

Seemann SW, Borbas EE, Knuteson RO, Stephenson GR, Huang HL (2008) Development of a global infrared land surface emissivity database for application to clear sky sounding retrievals from multispectral satellite radiance measurements. J Appl Meteorol Climatol 47:108–123. https://doi.org/10.1175/2007JAMC1590.1

Sekiyama TT, Tanaka TY, Shimizu A, Miyoshi T (2010) Data assimilation of CALIPSO aerosol observations. Atmos Chem Phys 10:39–49. https://doi.org/10.5194/acp-10-39-2010

Sun J (2006) Convective-scale assimilation of radar data: Progress and challenges. Q J R Meteorol Soc 131:3439–3463. https://doi.org/10.1256/qj.05.149

Sun J, Wang H, Tong W, Zhang Y, Lin C-Y, Xu D (2015) Comparison of the Impacts of momentum control variables on high-resolution variational data assimilation and precipitation forecasting. Mon Weather Rev 144:149–169.https://doi.org/10.1175/MWR-D-14-00205.1

Suzuki K, Zupanski M (2018) Uncertainty in solid precipitation and snow depth prediction for Siberia using the Noah and Noah-MP land surface models. Front Earth Sci 12:672–682

Tanaka TY, Chiba M (2005) Global simulation of dust aerosol with a chemical transport model MASINGAR. J Meteorol Soc Japan 83:255–278. https://doi.org/10.2151/jmsj.83a.255

Tang Y, Coauthors (2009) The impact of chemical lateral boundary conditions on CMAQ predictions of tropospheric ozone over the continental United States Environ Fluid Mech 9:43–58.https://doi.org/10.1007/s10652-008-9092-5

Thépaut J-N, Courtier P, Belaud G, Lemaître G (1996) Dynamical structure functions in a four-dimensional variational assimilation: a case study. Q J R Meteorol Soc 122:535–561. https://doi.org/10.1002/qj.49712253012

Leeuwen Van PJ (2010) Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Q J R Meteorol Soc 136:1991–1999. https://doi.org/10.1002/qj.699

Wang H, Niu T (2013) Sensitivity studies of aerosol data assimilation and direct radiative feedbacks in modeling dust aerosols. Atmos Environ 64:208–218. https://doi.org/10.1016/j.atmosenv.2012.09.066

Weaver C, Coauthors (2007) Direct insertion of MODIS radiances in a global aerosol transport model. J Atmos Sci 64 808–827.https://doi.org/10.1175/JAS3838.1

Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 1913–1924

Whitaker JS, Hamill TM (2012) Evaluating methods to account for system errors in ensemble data assimilation. Mon Weather Rev 140:3078–3089.https://doi.org/10.1175/MWR-D-11-00276.1

Winker DM, Vaughan MA, Omar A, Hu Y, Powell KA, Liu Z, Hunt WH, Young SA (2009) Overview of the CALIPSO mission and CALIOP data processing algorithms. J Atmos Ocean Technol 26:2310–2323. https://doi.org/10.1175/2009JTECHA1281.1

Wu W-S, Purser RJ, Parrish DF (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. Mon Weather Rev 130:2905–2916. https://doi.org/10.1175/1520-0493(2002)130%3c2905:TDVAWS%3e2.0.CO;2

Xian P, Coauthors (2019) Current State of the global operational aerosol multi-model ensemble: an update from the International Cooperative for Aerosol Prediction (ICAP) Q J R Meteorol Soc 176–209https://doi.org/10.1002/qj.3497

Xie Y, MacDonald AE (2012) Selection of momentum variables for a three-dimensional variational analysis. Pure Appl Geophys 169:335–351. https://doi.org/10.1007/s00024-011-0374-3

Yumimoto K, Tanaka TY, Yoshida M, Kikuchi M, Nagao TM, Murakami H, Maki T (2018) Assimilation and forecasting experiment for heavy Siberian wildfire smoke in May 2016 with himawari-8 aerosol optical thickness. J Meteorol. Soc Japan 96B:133. https://doi.org/10.2151/jmsj.2018-035

Zhang F, Snyder C Sun J (2004) Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble kalman filter. Mon Weather Rev 1238–1253

Zhang J, Campbell JR, Hyer EJ, Reid JS, Westphal DL, Johnson RS (2014) Evaluating the impact of multisensor data assimilation on a global aerosol particle transport model. J Geophys Res Atmos 119:4674–4689. https://doi.org/10.1002/2013JD020975.Received

Zhang J, Campbell JR, Hyer EJ, Reid JS, Westphal DL, Johnson RS, Jaker SL, Reid JS, Miller SD, Solbrig J, Toth TD (2019) Characterization and application of artificial light sources for nighttime

aerosol optical depth retrievals using the visible infrared imager radiometer suite day/night band. Atmos Mea 12:3209–3222

Zupanski D (1993) The effects of discontinuities in the Betts–Miller cumulus convection scheme on four-dimensional variational data assimilation. Tellus A. https://doi.org/10.3402/tellusa.v45i5.15053

Zupanski M (2005) Maximum likelihood ensemble filter: theoretical aspects. Mon Weather Rev 133:1710–1726. https://doi.org/10.1175/MWR2946.1

Zupanski M (2017) Data assimilation for coupled modeling systems. In: Park S-K, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, vol 3. Springer, Berlin, Heidelberg, pp 1–553

Zupanski M, Fletcher SJ, Navon IM, Uzunoglu B, Heikes RP, Randall DA, Ringler TD, Daescu D (2006) Initiation of ensemble data assimilation. Tellus Ser A Dyn Meteorol Oceanogr 58:159–170. https://doi.org/10.1007/s00024-011-0374-3

Zupanski M, Navon IM, Zupanski D (2008) The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. Q J R Meteorol Soc 134:1039–1050

Improving Near-Surface Weather Forecasts with Strongly Coupled Land–Atmosphere Data Assimilation

Zhaoxia Pu

Abstract Near-surface weather forecasts are critical for protecting life and human activities. However, they remain a challenging problem in modern numerical weather prediction (NWP) due to difficulties in surface data assimilation and uncertainties in representing complicated land–atmosphere interactions in numerical models. This chapter summarizes recent developments from the author's research team to understand and develop effective data assimilation methods that enhance near-surface weather forecasts. Results from several recent journal publications are summarized and presented to introduce strongly coupled land–atmosphere data assimilation in the context of land–atmosphere interaction. The first part of the mentioned work evaluated the association between near-surface variables and soil moisture with observations, coupled land–atmosphere model, and data assimilation systems. Results indicated a strong coupling between soil moisture and the low-level atmosphere, especially the atmospheric boundary layer. Then, the weakly and strongly coupled land–atmosphere data assimilation methods were compared regarding their influence on the prediction of near-surface atmospheric conditions. Results showed that strongly coupled land–atmosphere data assimilation, with simultaneous corrections to the land and atmospheric conditions, outperformed weakly coupled data assimilation. Finally, strongly coupled land–atmosphere data assimilation in an ensemble Kalman filter data assimilation system was implemented with an NWP model. Its positive impacts on predicting both atmosphere and land states were demonstrated. The potential of strongly coupled land–atmosphere data assimilation for future developments and applications is discussed in the concluding remarks.

Z. Pu (✉)
Department of Atmospheric Sciences, University of Utah, 135 S 1460 E, Rm. 819, Salt Lake City, UT 84112, USA
e-mail: Zhaoxia.Pu@utah.edu

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_19

1 Introduction

Near-surface weather forecasts are critical for protecting life and property, economic and operational activities, and routine day-to-day planning. Aviation, military, wind energy, and energy consumption operations rely on accurate near-surface forecasts, and even small forecast errors can have major consequences. Temperature, wind, and precipitation are some of the more important variables to forecast, but visibility-reducing phenomena, such as dust, fog, and smog, also need to be accurately forecast. Therefore, improving forecasts for any of these near-surface variables has far-reaching significance.

However, previous studies of numerical models have demonstrated the unavoidable errors of near-surface atmospheric forecasts (e.g., Liu et al. 2008a, b; Mass et al. 2002; Zhang et al. 2013; Pu 2017). It has also been found that forecast errors in near-surface atmospheric variables (such as 2-m temperature and 10-m winds) are quite large, even in many cases when forecasts in the middle and upper levels of the atmosphere are reasonable. The problem is more serious over complex terrain (Zhang et al. 2013; Pu 2017). These uncertainties in near-surface atmospheric conditions can contribute to inaccurate near-surface weather forecasts (e.g., fog, inversion, etc.) and mesoscale/synoptic-scale weather forecasts in general. More importantly, errors in near-surface atmospheric conditions also present a forecasting challenge at operational NWP centers with both mesoscale and global models. Specifically, near-surface temperature errors in NWP models have been observed in many different modeling systems throughout the world (Werth and Garrett 2011; Viterbo and Betts 1999; García-Díez et al. 2013).

Moreover, near-surface forecast errors also make it difficult to represent land–atmosphere interaction in numerical models, since the near-surface atmosphere is a transition area between the land and the atmosphere. These errors in near-surface atmospheric conditions interact with and influence both soil states and atmospheric boundary layer conditions through numerical model integration processes and contribute to the detriment of short- and medium-range weather forecasting, as well as prediction at sub-seasonal to seasonal and climate scales for climate models. Consequently, this prevents the use of numerical simulations to study the processes, especially the atmospheric boundary layer processes, related to severe weather systems. Meanwhile, it has been found that inaccurate forecasts of near-surface variables are associated with uncertainties in soil state, such as soil moisture. Commonly, uncertainties in representing land use, soil moisture, and terrain conditions on the underlying surface, which affect the land-atmosphere interaction directly, are identified as the major sources of error in near-surface weather forecasting (Massey et al. 2014; Zhang et al. 2013; Ren et al. 2018). Hence, the impacts of underlying surface characteristics and bias corrections on NWP have been investigated in recent years (e.g., Fan and van den Dool 2011; Massey et al. 2016; Chen et al. 2017; Lin et al. 2017). Results indicated that the bias correction of soil moisture could help near-surface temperature prediction in those case studies. Moreover, notable progress has been made recently in studying land–atmosphere interactions at

regional and local scales as well as short-range weather forecasting (e.g., Santanello et al. 2005, 2007, 2016, 2018). All these previous works motivate us to seek a way to improve near-surface weather forecasting through improved representation of soil moisture and land–atmosphere interactions in land models.

In current research and operational practices, remotely sensed soil moisture data are typically incorporated into advanced NWP models under a framework of weakly coupled data assimilation, with a land surface analysis scheme independent of the atmospheric analysis component; thus, the land and atmospheric analyses are performed separately (Kumar et al. 2015; Santanello et al. 2018; Xia et al. 2019). So far, there has been limited progress in NWP models with coupled land–atmosphere data assimilation (e.g., de Rosnay et al. 2014).

In order to improve near surface weather forecasting, the author and her research team have made significant progress with observations, numerical simulations, and data assimilation in understanding the correlations between soil moisture and near-surface atmospheric variables as well as the characteristics of their error covariances in coupled data assimilation. This chapter summarizes outcomes from a series of those studies, mostly results from four published journal papers (Lin and Pu 2018, 2019, 2020; Liu and Pu 2019), to introduce the concept of strongly coupled land–atmosphere data assimilation and demonstrate its promise in improving near-surface weather forecasting. Challenges and future developments are also discussed.

2 The Relationship Between Soil Moisture and Near-Surface Atmospheric Conditions

Although earlier studies in the community have demonstrated that soil moisture has an influence on near-surface temperature, no study has yet made it clear whether soil moisture and near-surface temperature are correlated or to what degree they are associated in short-range weather prediction. In Liu and Pu (2019), the relationship between soil moisture and temperature at 2-m height (2-m temperature) was first examined with long-term meteorological and soil observations during 2008–2016 from 16 stations over the United States in four different land cover types, including Shrub and Grassland, Grassland, Shrubland, and Forest. Meteorological observations included surface Mesonet data, and soundings were obtained from the MesoWest Network (http://mesowest.utah.edu) and University of Wyoming Network (http://weather.uwyo.edu/upperair/sounding.html), respectively. Soil moisture data included five layers (5, 10, 20, 50, and 100 cm) from in situ observations from the Climate Reference Network and Soil Climate Analysis Network (https://www.drought.gov/drought/soil-moisture-map).

With the correlation statistics and an information flow analysis method (Liang 2014, 2015; also see details in Liu and Pu 2019), we found that soil moisture at all levels and the near-surface atmospheric temperature had weak to moderate causality with seasonal variability. The distribution of soil moisture depended on land use and

land cover, and the dependence decreased with soil depth. Although the correlations between soil moisture and near-surface temperature was moderate, with a correlation coefficient of less than 0.6, there was strong interaction between the top soil layer and the atmosphere, implying that the impact of soil moisture on near-surface temperature was significant.

Two meteorological sounding stations collocated with soil moisture measurements were also used to investigate the relationship between atmospheric profiles and near-surface temperature (i.e., 2-m temperature). It was found that the causality between wind profile and near-surface temperature was retained in most weather conditions. Correlations between near-surface temperature and boundary layer temperature profiles were quite strong, especially during the warm season. Meanwhile, the correlations decreased with height through the atmosphere. Furthermore, correlations between near-surface temperature and upper atmospheric conditions had seasonal variability and also varied with land use and land cover.

The findings from long-term observations were further proved by a series of sensitivity experiments in Liu and Pu (2019) with a single column model (SCM, Hacker et al. 2007) based on the mesoscale community Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) coupled with the Noah land surface model (Chen and Dudhia 2001; Ek et al. 2003). The impact of changes in soil moisture on short-range forecasts (up to 48 h) of near-surface temperature and atmospheric profiles was examined.

A control experiment was conducted with the average soil state, and two other sensitivity experiments were performed with an increase or decrease in soil moisture of 25% (e.g., within the seasonal variation range of soil moisture), respectively. Results showed that the impact of soil moisture on temperature was often focused on the lower levels of the atmospheric boundary layer. An increase (decrease) in soil moisture resulted in cooler (warmer) near-surface 2-m temperature through the redistribution of surface heat flux. Meanwhile, there was seasonal variation, since changes in temperature with soil moisture fluctuations were more obvious during summer and autumn. In general, an increase in soil moisture caused a temperature inversion to appear earlier and disappear later, resulting in longer inversion duration. A decrease in soil moisture had the opposite effect. Moreover, changes in near-surface temperature caused by soil moisture in all seasons were mainly from near-surface (top) soil levels. The evolution of soil thermodynamic characteristics associated with changes in soil moisture could affect surface energy distribution and influence near-surface temperature directly (see details in Liu and Pu 2019).

3 Strongly Coupled Versus Weakly Coupled Land–Atmosphere Data Assimilation

The observational analysis and single column model study mentioned above indicated that soil moisture and near-surface atmospheric conditions were strongly coupled

and influenced each other. The results from Liu and Pu (2019) implied that realistic soil moisture states in land surface models could benefit the accurate prediction of near-surface and atmospheric boundary layer conditions. The findings from this study encouraged us to explore using coupled land–atmosphere data assimilation to improve numerical weather prediction.

Coupled data assimilation can be done in two different ways (Lin and Pu 2019): weakly or strongly coupled. With weakly coupled data assimilation, assimilating observations into a model does not affect the control states of the other coupled model(s) during the analysis. Land data assimilation (e.g., Kumar et al. 2014, 2015) and atmospheric data assimilation are done separately, and the analysis results are then input into coupled land–atmosphere data and interact during model integration. Therefore, the impact of weakly coupled data assimilation on the entire domain is seen only via model integration. In contrast, strongly coupled data assimilation (Penny and Hamill 2017; Penny et al. 2017; Lin and Pu 2019, 2020) requires the estimation of error covariance of the control states in all the coupled models and the simultaneous computation of the analysis across the entire domain. So far, most coupled land–atmosphere data assimilation has been done with weak coupling (Mahfouf 2010; Mahfouf and Bliznak 2011; Schneider et al. 2014; Duerinckx et al. 2017; Santanello et al. 2016; Seto et al. 2016; Lin et al. 2017). Almost none of these studies addressed the land–atmosphere data assimilation problems with strongly coupled data assimilation before Lin and Pu (2018, 2019, 2020). However, results in Liu and Pu (2019) indicated a strong response of atmospheric conditions to changes in soil moisture, suggesting that strongly coupled data assimilation is necessary for land–atmosphere data assimilation.

3.1 Characteristics of Background Error Covariance of Soil Moisture and Atmospheric States in Strongly Coupled Land–Atmosphere Data Assimilation

To explore the methodology of strongly coupled data assimilation, a deep understanding of the error covariance between soil moisture and atmospheric states within a strongly coupled land–atmosphere model is the first step. An early study by Zupanski (2017) has formulated that two-component coupled system data assimilation could be implemented through the coupled forecast error covariances cross the variables in different coupling components (e.g., land–atmosphere or aerosol–atmosphere). He conducted a single observation experiment to understand and illustrate the structure of forecast error covariance in both coupled land–atmosphere and atmosphere-chemistry models. Results indicated that the cross-component correlations have a potential to increase the utility of observations in data assimilation by spreading the information throughout the components. Following Zupanski (2017), Suzuki et al. (2017) investigated forecast error covariance and correlation structures between land and atmospheric variables by applying the Maximum Likelihood Ensemble Filter

(MLEF) data assimilation method with a coupled atmosphere–land surface model through a series of single observation experiments. They demonstrated that coupled error covariance methods improve the efficiency of information transfer between the atmosphere and the land surface by allowing the well-observed atmosphere to influence land surface variables.

Different from these previous studies, in our study, instead of using single observation experiments, we used a completed variational framework as an example to examine the error covariance between soil moisture and atmospheric states within a strongly coupled land–atmosphere model (see details in Lin and Pu 2018). A classic one- and three-dimensional variational data assimilation (1D- and 3D-Var) system computes optimal states by minimizing the following cost function (J) in an incremental form (Ide et al. 1997; Courtier et al. 1998):

$$J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}(\mathbf{H}\delta\mathbf{x} - \mathbf{d})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \tag{1}$$

where $\delta\mathbf{x}$ is a vector of the analysis increment, with $\delta\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{a}} - \mathbf{x}^{\mathrm{b}}$ at the minimum of the cost function, in which $\mathbf{x}^{\mathrm{b}}$ and $\mathbf{x}^{\mathrm{a}}$ denote the vectors of the background and analysis, respectively; $\mathbf{H}$ denotes the linear form of an operator that projects the analysis variables onto the observation space; $\mathbf{d}$ is the innovation vector,$\mathbf{d} = \mathbf{y}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}$, in which $\mathbf{y}^{\mathrm{o}}$ is a vector of observations; $\mathbf{B}$ represents the background error covariance matrix; and $\mathbf{R}$ is the observation error covariance matrix. For implementing a variational method in NWP, the estimation of $\mathbf{B}$ is necessary and important. The $\mathbf{B}$-matrix contains information about the weights of the control states and multivariate error correlation, which allows the balanced spread of the information from the observations to the control states.

In weakly coupled data assimilation, the $\mathbf{B}$-matrix contains only the background error covariance information for either soil states or atmospheric variables because separate data assimilation procedures are used for the land and atmosphere. However, in strongly coupled data assimilation, the $\mathbf{B}$-matrix contains error covariance information for both soil states and atmospheric variables. Let us first use top-layer soil moisture (SM_1) and bottom-layer atmospheric states (T_1, Q_1, U_1, and V_1) as an example. With these five variables, the symmetric and positive definite $\mathbf{B}$-matrix of a given pixel can be described as follows:

$$\boldsymbol{B} = \begin{bmatrix} \sigma^2_{\eta_{SM_1}} & - & - & - & - \\ cov(\eta_{T_1}, \eta_{SM_1}) & \sigma^2_{\eta_{T_1}} & - & - & - \\ cov(\eta_{Q_1}, \eta_{SM_1}) & cov(\eta_{Q_1}, \eta_{T_1}) & \sigma^2_{\eta_{Q_1}} & - & - \\ cov(\eta_{U_1}, \eta_{SM_1}) & cov(\eta_{U_1}, \eta_{T_1}) & cov(\eta_{U_1}, \eta_{Q_1}) & \sigma^2_{\eta_{U_1}} & - \\ cov(\eta_{V_1}, \eta_{SM_1}) & cov(\eta_{V_1}, \eta_{T_1}) & cov(\eta_{V_1}, \eta_{Q_1}) & cov(\eta_{V_1}, \eta_{V_1}) & \sigma^2_{\eta_{V_1}} \end{bmatrix}, \tag{2}$$

where the diagonal elements are the auto-covariance of the forecast error of the explained variables and the off-diagonal elements are the covariance.

Using a variational data assimilation framework and the mesoscale community WRF model (Skamarock et al. 2008), Lin and Pu (2018) estimated the WRF-Noah (i.e., the WRF model coupled with the Noah land surface model) background error covariance between the surface soil moisture and atmospheric states. WRF version 3.9.1 (Skamarock et al. 2008; Powers et al. 2017), with the Advanced Research version of the WRF (ARW) solver, was used with WRF's CONUS physics suite. It included the new Thompson microphysics scheme, the Rapid Radiative Transfer Model (RRTM) longwave and shortwave schemes, the Monin–Obukhov-based Eta similarity surface-layer scheme, the Noah land surface model, the Mellor–Yamada–Janji´ć planetary boundary layer scheme, and the Tiedtke cumulus parameterization scheme (see details in Skamarock et al. 2008). A single domain of the Lambert conformal projection was configured with grid spacing of 9 km and 602 × 392 grids horizontally. The Noah land surface model had four soil layers as the default, with thicknesses of 10, 30, 60, and 100 cm from top to bottom. Lookup tables were used for the prescribed parameters of land use (vegetation) and soil types. The study domain covered the entire contiguous United States. The NMC method (Parrish and Derber 1992) was used to compute the **B**-matrix:

$$\mathbf{B} = \overline{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}, \tag{3}$$

where $\boldsymbol{\eta}$ is the difference in paired forecasts that have different initialization times but are valid at the same time, and the overbar denotes an average of forecast error samples. In a regional application (e.g., WRFDA), $\boldsymbol{\eta}$ is often obtained from paired 12 and 24 h forecasts, as follows:

$$\boldsymbol{\eta} = \mathbf{x}^{\mathrm{f}}_{\mathrm{t+24|t}} - \mathbf{x}^{\mathrm{f}}_{\mathrm{t+24|t+12}} \tag{4}$$

where each of the components on the right-hand side denotes the samples of 24 and 12 h forecasts with bias adjustment with respect to each control state.

To compute the **B**-matrix, we initialized WRF-Noah simulations at 0000 and 1200 UTC nearly every day from 2015 to 2017 to obtain 12 and 24 h forecasts. Every month, we computed the **B**-matrix by using 54 pairs of 12 and 24 h forecasts to show the "all-time" results. For the daytime (nighttime) results, we obtained 27 pairs from forecasts valid at 00 UTC (12 UTC). The 00 UTC corresponds to 6 pm Central Standard Time locally over the United States, and we considered that the forecasts valid at 00 UTC would contain the model errors during the daytime from 6 am to 6 pm local time.

Detailed results are documented in Lin and Pu (2018). Notably, these results indicated that the forecast errors in top-10 cm soil moisture and near-surface air potential temperature and specific humidity were correlated and relatively large during the daytime in the summer. The magnitude of the error correlation between surface soil moisture, temperature, and humidity was comparable, which suggests that (1) part of the error in surface soil moisture comes from atmospheric forcing, and (2) atmospheric initial conditions could potentially be corrected via soil moisture data

assimilation. Specifically, the results showed a negative error correlation between soil moisture and potential temperature but a positive correlation between soil moisture and air humidity. In general, the correlation was seen nearly everywhere over the study domain, and the daytime correlation was larger than the nighttime correlation. These results not only suggested strong coupling between soil moisture and the atmosphere, but also identified the correlation structures between soil moisture and atmospheric variables, notably in the near-surface and boundary layer atmosphere (see Figs. 1 and 2, also Lin and Pu 2018).

3.2 Soil Moisture Data Assimilation: Weakly Versus Strongly Coupled Data Assimilation

In subsequent studies, Lin and Pu (2019) implemented the strongly coupled land–atmosphere data assimilation in Lin and Pu (2018) to study the relative effect of assimilating soil moisture data on weather forecasts under a framework of weakly and strongly coupled land–atmosphere data assimilation. Specifically, experiments aimed to quantify the additional impact on lower-troposphere atmospheric forecasts via direct analysis (i.e., a strongly coupled case) relative to the impact on forecasts via the dynamics of land–atmosphere interactions (i.e., a weakly coupled case) when soil moisture data were assimilated. The study used the Noah land surface model coupled with the WRF model and conducted experiments in the summer over the continental United States. The NASA Soil Moisture Active Passive (SMAP) satellite-derived soil moisture data products, SMAP 9 km level-2 enhanced soil moisture retrievals (O'Neill et al. 2016), were assimilated.

In the variational data assimilation framework, strongly coupled data assimilation adopted the background error covariance estimated from Lin and Pu (2018). The results of the numerical experiments during July 2016 showed that strongly coupled data assimilation could provide additional benefits to forecasts of air temperature and humidity compared to weakly coupled data assimilation. Over the U.S. Great Plains, on average, assimilation of SMAP data under weakly coupled data assimilation reduced a warm bias in temperature and a dry bias in humidity by 7.3% and 19.3%, respectively, while strongly coupled data assimilation contributed an *additional* bias reduction of 2.2% (temperature) and 3.3% (humidity). More importantly, improvements in precipitation forecasts and near-surface atmospheric conditions were also found with strongly coupled data assimilation compared with weakly coupled data assimilation (see details in Lin and Pu 2019).

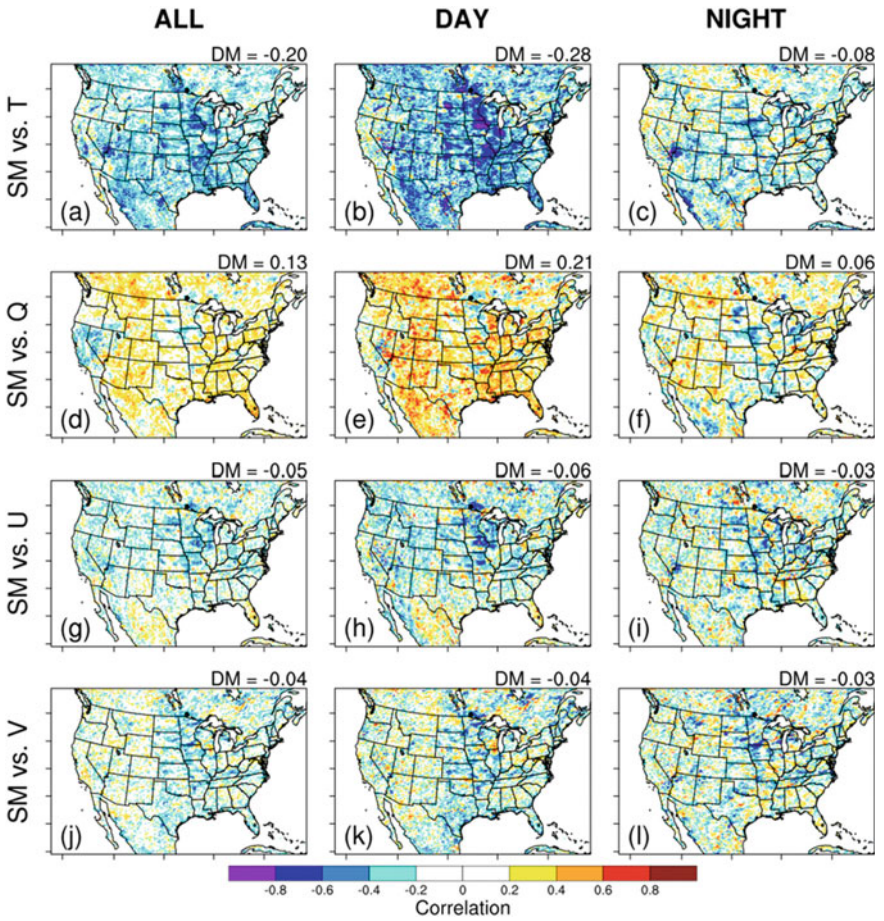

Fig. 1 Error correlations between top-layer soil moisture (SM) and bottom-layer **a–c** potential temperature T, **d–f** specific humidity Q, **g–i** zonal wind U, and **j–l** meridional wind V during July 2016. The forecast samples valid at 0000 and 1200 UTC are categorized as "DAY" and "NIGHT" results, and the "ALL" results are computed based on all the samples. Domain-mean (DM) values are computed based on the results of warm land pixels without considering the 10-grid-wide boundary. The 95% confidence intervals of the DM values vary from DM ± 0.001 to DM ± 0.0015 (From Lin and Pu 2018)

4 Enhanced Near-Surface Weather Forecasts Using Strongly Coupled Land–Atmosphere Data Assimilation

Following the outcomes from Lin and Pu (2018) and Lin and Pu (2019), a strongly coupled land–atmosphere data assimilation system was implemented by Lin and Pu (2020) using the U.S. National Centers for Environmental Prediction (NCEP)'s Gridpoint Statistical Interpolation (GSI)-based ensemble Kalman filter (EnKF) data

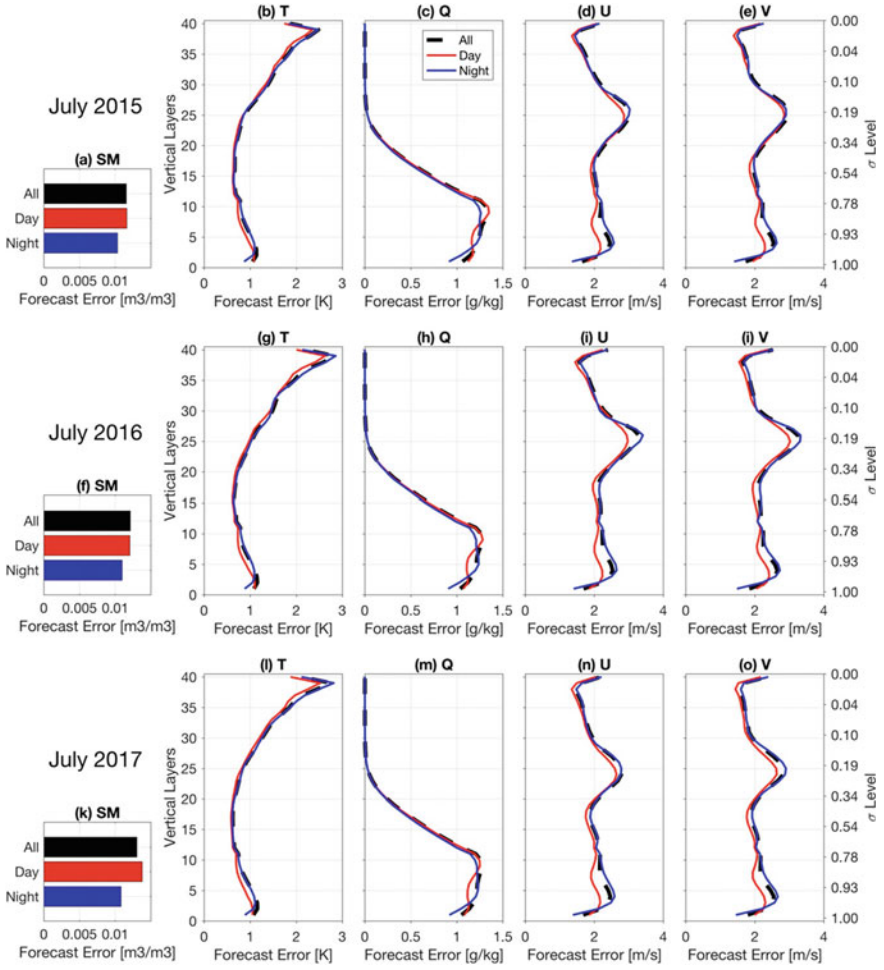

Fig. 2 Horizontal domain mean (DM) values for the error correlation between the top-10 cm soil moisture and atmospheric states including T, Q, U, and V during July from 2015 to 2017 (From Lin and Pu 2018)

assimilation system (GSI-EnKF, a community data assimilation system maintained by the NCAR Developmental Testbed Center). The model used was the WRF model coupled with the Noah land surface model. Two unique implementations enabled the incorporation of soil moisture observations via strongly coupled land–atmosphere data assimilation in our system. The first step was to include soil moisture as a control state, along with the common control analysis states in GSI-EnKF, including potential temperature, specific humidity, zonal and meridional winds, and surface dry air pressure. To enable soil moisture as a control state that was compatible with GSI-EnKF localization, we set all four layers of soil moisture inside the Noah

land surface model as a control state in the GSI analysis. In the second step, we added soil moisture as a new type of conventional observation. Overall, this strongly coupled land–atmosphere data assimilation could assimilate both soil moisture and atmospheric observations simultaneously, with consistent cross-model error covariance from ensemble forecasts of land and atmospheric components of the numerical weather prediction model (e.g., WRF).

For the experiment, an ensemble size of 40 was used, as a sample size of around 40 is quite common in regional ensemble-based studies (e.g., Pu et al. 2013; Zhang and Pu 2014; Schwartz et al. 2015; McNicholas and Mass 2018). To keep a reasonable ensemble spread and avoid filter divergence, a tunable inflation coefficient can be set to adjust the posterior ensemble spread to match the prior ensemble spread (relaxation-to-prior spread; Whitaker and Hamill 2012). The inflation coefficient ranges from 0 (no inflation) to 1 (i.e., both prior and posterior ensemble spread are of the same magnitude). Tests with the assimilation of in-situ soil moisture data and all other conventional atmospheric observations indicated that this GSI-EnKF based strongly coupled data assimilation system could simultaneously adjust atmospheric and soil moisture states through assimilating atmospheric observations and soil moisture data.

Key findings included the following: (1) including soil moisture as a control variable in GSI-EnKF resulted in significant reduction of analysis errors in near-surface atmospheric variables, such as temperature and humidity; (2) with the strongly coupled system, soil moisture analysis errors were reduced significantly when soil moisture data were assimilated with all other available atmospheric observations; (3) combined assimilation of soil moisture and atmospheric observations in a strongly coupled data assimilation system resulted in improved analysis and forecasts in an NWP framework. Specifically, strongly coupled land–atmosphere data assimilation led to improved near-surface weather forecasting (See details in Lin and Pu 2020).

5 Discussion and Concluding Remarks

5.1 Summary and Discussion

Near-surface weather forecasts present a challenging problem in modern NWP. A series of studies from the author's research team, as summarized above, led to a significant understanding of the problem and made it clear that coupled land–atmosphere data assimilation, especially strongly coupled land–atmosphere data assimilation, is a promising way to improve near-surface weather forecasting.

In summary, the observational analyses showed significant correlations between soil moisture in the top soil layer and surface 2-m temperature. Sensitivity experiments with a single column model indicated that near-surface weather conditions

responded to soil moisture changes, suggesting a strong coupling between soil moisture and the near-surface atmosphere. Results encouraged us to assimilate soil moisture data into a land model for improved near-surface weather forecasting. Strongly coupled land–atmosphere data assimilation was then evaluated in a variational data assimilation framework. It was found that the error covariances between soil moisture and near-surface temperature and humidity were significant during the daytime and warm season in the boundary layer. Based on the error covariance structures and correlations, we can expect that soil moisture changes could cause adjustments in near-surface and atmospheric boundary layer conditions. The increments in atmospheric conditions would lead to changes in soil moisture. The subsequent data assimilation with the WRF-Noah model indicated that strongly coupled land–atmosphere data assimilation in this variational framework successfully assimilated SMAP soil moisture data. More importantly, with simultaneous corrections in both atmosphere and land variables, this strongly coupled data assimilation method outperformed weakly coupled data assimilation. Finally, strongly coupled data assimilation was implemented in an ensemble Kalman filter data assimilation system. Results proved that this strongly coupled data assimilation could indeed improve prediction of both soil and atmospheric states.

Although variational data assimilation is different from the ensemble Kalman filter, both methods have proven the success of strongly coupled data assimilation. In reality, according to Lin and Pu (2020), the structure of ensemble spreads from the strongly coupled system in the GSI-EnKF system (Fig. 3) was very similar to the structure in Lin and Pu (2018) (e.g., Fig. 2), implying that strongly coupled land–atmosphere data assimilation is capable of representing the strong coupling between soil moisture and the near-surface and boundary layer atmospheric states in the EnKF data assimilation system; thus it has great potential to be implemented into NWP models for many forecast applications.

5.2 Concluding Remarks

Land–atmosphere interaction is an essential process in weather and climate systems. Coupled land–atmosphere models and land surface parameterizations are necessary to represent land–atmosphere interactions in weather and climate models. Due to lack of observations, our limited understanding of and capability to accurately represent land–atmosphere interaction in coupled models or parameterizations, and the errors in initial and boundary conditions, uncertainties in weather and climate prediction present significant challenges in weather forecasting and climate prediction. Notably, the near-surface atmosphere and atmospheric boundary layer interact with the land surface directly. Because of the complexity of the water and energy budget in the interface of the land and atmosphere, uncertainties in numerical model parameterizations and initial conditions are significant. As a consequence, near-surface weather forecasting remains a significant challenge in numerical weather prediction (Pu 2017). In light of the strong interaction between the near-surface atmosphere

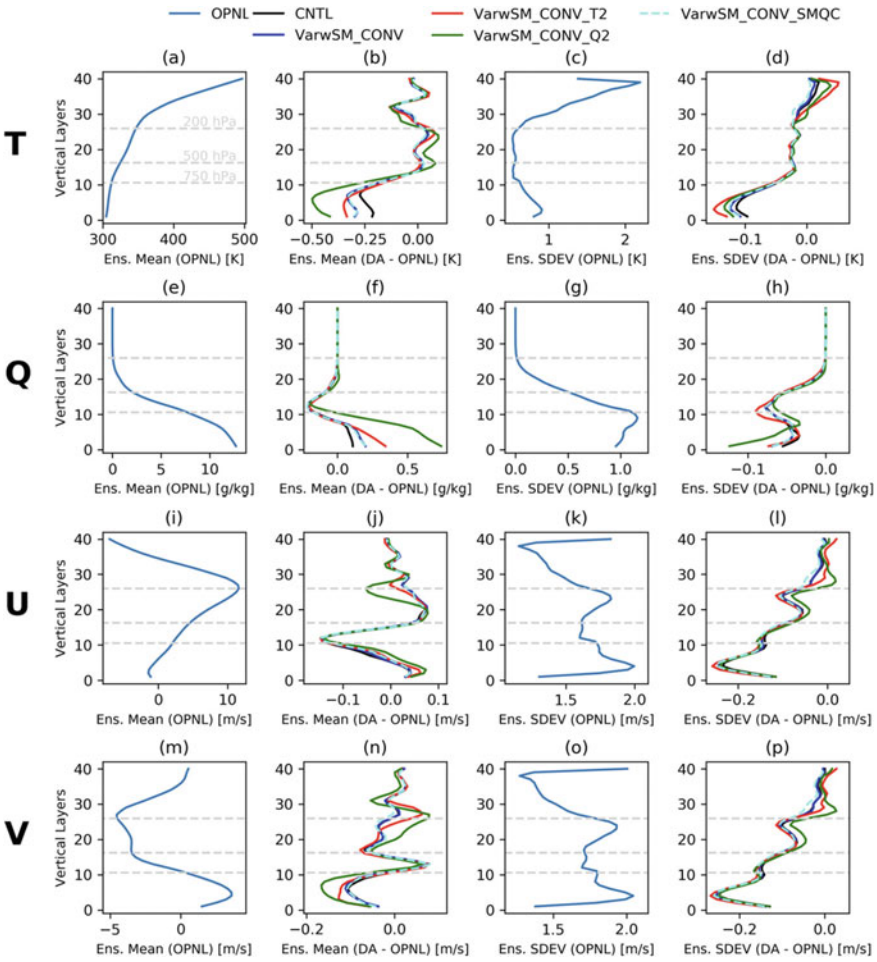

Fig. 3 Ensemble mean and spread (SDEV or standard deviation) of atmospheric first guesses for potential temperature (T), specific humidity (Q), zonal wind (U), and meridional wind (V) during the study period (July 2018) (From Lin and Pu 2020)

and land surface, four research articles (Liu and Pu 2019; Lin and Pu 2018, 2019, 2020) are summarized in this chapter dedicated to understanding the fundamentals of improving near-surface weather forecasting using coupled data assimilation.

The observational analyses showed significant correlations between soil moisture in the top soil layer and surface 2-m temperature. However, the correlation coefficient between soil moisture and 2-m temperature was less than 0.6, implying that soil moisture is not the sole factor that influences near-surface weather conditions. Given the heterogeneous nature of land use and land cover as well as soil types, there are many other factors that could influence land–atmosphere interactions that need to be studied in future work to examine their influence on near-surface weather

prediction. Nevertheless, the notable correlation between soil moisture and near-surface atmospheric conditions provides a direct way to implement coupled land–atmosphere data assimilation. More complicated data assimilation systems are still needed in order to fully resolve or mitigate the uncertainties associated with land–atmosphere interactions in weather and climate models.

In addition, near-surface temperature and soil moisture could also influence the atmospheric boundary layer, even upper atmospheric conditions. Most of this influence is in the local atmospheric boundary layer in short-range weather prediction. However, through the integration of atmospheric models with time, the soil moisture influence could propagate to the entire atmospheric column and over a large region. Therefore, the influence of soil moisture in medium-range weather forecasting and sub-seasonal to seasonal climate prediction should be expected. From these perspectives, strongly coupled land–atmosphere data assimilation should be an active research area, not only for weather forecasting but also for climate prediction. Nevertheless, since the temporal scales between land and atmosphere variabilities are not the same, strategies to adjust the temporal scales of land and atmospheric variables during the coupled data assimilation could be another important problem to explore in future studies.

Furthermore, many severe weather and climate events are associated with land–atmosphere interactions, such as hurricane evolution after landfall, floods, droughts, etc. Considering the need to improve forecasts and public warnings for these high-impact weather and climate events, we can foresee the utility of strongly coupled data assimilation in many research applications.

Acknowledgements This study is supported by NOAA Award NA16NWS4680015 and NASA CYGNSS science team award through University of Michigan.

References

Chen F, Dudhia J (2001) Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon Weather Rev 129(4):569–585

Chen L, Ma Z, Mahmood R, Zhao T, Li Z, Li Y (2017) Recent land cover changes and sensitivity of the model simulations to various land cover datasets for China. Meteorol Atmos Phys 129(4):395–408

Courtier P, Andersson E, McNally A, Rabier F et al (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: formulation. Q J R Meteorol Soc 124:1783–1807

de Rosnay P, Balsamo G, Albergel C, Muñoz-Sabater J, Isaksen L (2014) Initialisation of land surface variables for numerical weather prediction. Surv Geophys 35(3):607–621

Duerinckx A, Hamdi R, Deckmyn A, Djebbar A, Mahfouf J-F, Termonia P (2017) Combining an EKF soil analysis with a 3D-Var upper-air assimilation in a limited-area NWP model. Q J R Meteorol Soc 143:2999–3013

Ek MB, Mitchell KE, Lin Y, Rogers E, Grummann P, Koren V, Gayno G, Tarpley JD (2003) Implementation of Noah land surface model advances in the national centers for environmental prediction operational mesoscale Eta model. J Geophys Res 108:8851

Fan Y, van den Dool H (2011) Bias correction and forecast skill of NCEP GFS ensemble week-1 and week-2 precipitation, 2-m surface air temperature, and soil moisture forecasts. Weather Forecast 26(3):355–370

García-Díez M, Fernández J, Fita L, Yagüe C (2013) Seasonal dependence of WRF model biases and sensitivity to PBL schemes over Europe. Q J R Meteorol Soc 139:501–514

Hacker JP, Anderson JL, Pagowski M (2007) Improved vertical covariance estimates for ensemble-filter assimilation of near surface observations. Mon Weather Rev 135(3):1021–1036. https://doi.org/10.1175/MWR3333.1

Ide K, Courtier P, Ghil M, Lorenc AC (1997) Unified notation for data assimilation: operational, sequential, and variational. J Meteorol Soc Jpn 75:181–189

Kumar P, Kishtawal CM, Pal PK (2014) Impact of satellite rainfall assimilation on weather research and forecasting model predictions over the Indian region. J Geophys Res 119:2017–2031

Kumar P, Peters-Lidard CD, Santanello JA, Reichle RH, Draper CS, Koster RD, Nearing G, Jasinski MF (2015) Evaluating the utility of satellite soil moisture retrievals over irrigated areas and the ability of land data assimilation methods to correct for unmodeled processes. Hydrol Earth Syst Sci 19:4463–4478

Liang X (2014) Unraveling the cause-effect relation between time series. Phys Rev E 90(5). https://doi.org/10.1103/PhysRevE.90.052150

Liang X (2015) Normalizing the causality between time series. Phys Rev E 92(2). https://doi.org/10.1103/PhysRevE.92.022126

Lin L-F, Ebtehaj AM, Wang J, Bras RL (2017) Soil moisture background error covariance and data assimilation in a coupled land atmosphere model. Water Resour Res 53:1309–1335. https://doi.org/10.1002/2015WR017548

Lin L-F, Pu Z (2018) Characteristics of background error covariance of soil moisture and atmospheric states in strongly coupled land atmosphere data assimilation. J Appl Meteorol Climatol 57(11):2507–2529

Lin L-F, Pu Z (2019) Examining the impact of SMAP soil moisture retrievals on short-range weather prediction under weakly- and strongly-coupled data assimilation with WRF-Noah. Mon Weather Rev 147:4345–4366

Lin L-F, Pu Z (2020) Improving near-surface short-range weather forecasts using strongly coupled land-atmosphere data assimilation with GSI-EnKF. Mon Weather Rev 148:2863–2888

Liu J, Pu Z (2019) Does soil moisture have an influence on near-surface temperature? J Geophys Res Atmos 124:6444–6466

Liu Y et al (2008) The operational meso gamma-scale analysis and forecast system of the U.S. Army test and evaluation command. Part I: overview of the modeling system, the forecast products, and how the products are used. J Appl Meteorol Climatol 47:1077–1092

Liu Y et al (2008) The operational meso gamma-scale analysis and forecast system of the U.S. Army test and evaluation command. Part II: interrange comparison of the accuracy of model analyses and forecasts. J Appl Meteor Climatol 47:1093–1104

Mahfouf J-F (2010) Assimilation of satellite-derived soil moisture from ASCAT in a limited-area NWP model. Q J R Meteorol Soc 136:784–798. https://doi.org/10.1002/qj.602

Mahfouf J-F, Bliznak V (2011) Combined assimilation of screen-level observations and radar-derived precipitation for soil moisture analysis. Q J R Meteorol Soc 137:709–722. https://doi.org/10.1002/qj.791

Mass CF, Ovens D, Westrick K, Colle BA (2002) Does increasing horizontal resolution produce more skillful forecasts? Bull Am Meteorol Soc 83:407–430

Massey JD, Steenburgh WJ, Hoch SW, Knievel JC (2014) Sensitivity of near-surface temperature forecasts to soil properties over a sparsely vegetated dryland region. J Appl Meteorol Climatol 53:1976–1995

Massey JD, Steenburgh WJ, Knievel JC, Cheng WYY (2016) Regional soil moisture biases and their influence on WRF model temperature forecasts over the Intermountain West. Weather Forecast 31(1):197–216

McNicholas C, Mass CF (2018) Impacts of assimilating smartphone pressure observations on forecast skill during two case studies in the Pacific Northwest. Weather Forecast 33:1375–1396. https://doi.org/10.1175/WAF-D-18-0085.1

O'Neill PE, Chan S, Njoku EG, Jackson T, Bindlish R (2016) SMAP enhanced L2 radiometer half-orbit 9 km EASEGrid soil moisture, version 1. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/CE0K6JS5WQMM. Accessed 1 Dec 2020

Parrish DF, Derber JC (1992) The National Meteorological Center's spectral statistical-interpolation analysis system. Mon Weather Rev 120:1747–1763

Penny SG, Hamill TM (2017) Coupled data assimilation for integrated earth system analysis and prediction. Bull Am Meteorol Soc 98:ES169–172. https://doi.org/10.1175/BAMS-D-17-0036.1

Penny SG, Akella S, Alves O, Bishop C, Buehner M, Chevallier M, Counillon F, Draper C, Frolov S, Fujii Y, Karspeck A, Kumar A, Laloyaux P, Mahfouf J-F, Martin M, Pena M, de Rosnay P, Subramanian A, Tardif R, Wang Y, Wu X (2017) Coupled data assimilation for integrated earth system analysis and prediction: goals, challenges and recommendations. WWRP 2017-3, 50 pp

Powers JG et al (2017) The weather research and forecasting model. Bull Am Meteorol Soc 98:1717–1737

Pu Z (2017) Surface data assimilation and near-surface weather prediction over complex terrain. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, vol III. Springer, Berlin, pp 219–240. https://doi.org/10.1007/978-3-319-43415-5_10

Pu Z, Zhang H, Anderson JA (2013) Ensemble Kalman filter assimilation of near-surface observations over complex terrain: comparison with 3DVAR for short-range forecasts. Tellus 65A:19620

Ren Y, Li Y, Pu Z, Zhang T, Duan H, Wang W (2018) Effect of updated RegCM4 land use data on near-surface temperature simulation in China. J Meteorol Res 32:758–767

Santanello JA Jr, Friedl M, Kustas W (2005) Empirical investigation of convective planetary boundary layer evolution and its relationship with the land surface. J Appl Meteorol 44:917–932

Santanello JA Jr, Friedl M, Ek M (2007) Convective planetary boundary layer interactions with the land surface at diurnal time scales: diagnostics and feedbacks. J Hydrometeorol 8:1082–1097

Santanello JA Jr, Kumar SV, Peters-Lidard CD, Lawston PM (2016) Impact of soil moisture assimilation on land surface model spinup and coupled land-atmosphere prediction. J Hydrometeorol 17:517–540

Santanello JA Jr et al (2018) Land–atmosphere interactions: the LoCo perspective. Bull Am Meteorol Soc 99:1253–1272

Schneider S, Wang Y, Wanger W, Mahfouf J-F (2014) Impact of ASCAT soil moisture assimilation on regional precipitation forecasts: a case study for Austria. Mon Weather Rev 142:1525–1541. https://doi.org/10.1175/MWR-D-12-00311.1

Schwartz CS, Romine GS, Sobash RA, Fossell KR, Weisman ML (2015) NCAR's experimental real-time convection-allowing ensemble prediction system. Weather Forecast 30:1645–1654. https://doi.org/10.1175/WAF-D-15-0103.1

Seto R, Koike T, Rasmy M (2016) Heavy rainfall prediction applying satellite-based cloud data assimilation over land. J Geophys Res Atmos 121:9737–9755. https://doi.org/10.1002/2016JD025291

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang XY, Wang W, Powers JG (2008) A description of the advanced research WRF version 2. NCAR technical note, NCAR/TN-468+STR (available at NCAR, Boulder, CO 80300)

Suzuki K, Zupanski M, Zupanski D (2017) A case study involving single observation experiments performed over snowy Siberia using a coupled atmosphere-land modeling system. Atmos Sci Lett 18:106–111. https://doi.org/10.1002/asl.730

Viterbo P, Betts AK (1999) The impact of the ECMWF reanalysis soil water on forecasts of the July 1993 Mississippi Flood. J Geophys Res 104:19361–19366

Werth D, Garrett A (2011) Patterns of land surface errors and biases in the Global Forecast System. Mon Weather Rev 139:1569–1582

Whitaker JS, Hamill TM (2012) Evaluating methods to account for system errors in ensemble data assimilation. Mon Weather Rev 140:3078–3089. https://doi.org/10.1175/MWR-D-11-00276.1

Xia YL, Hao ZC, Shi CX et al (2019) Regional and global land data assimilation systems: innovations, challenges, and prospects. J Meteorol Res 33:159–189. https://doi.org/10.1007/s13351-019-8172-4

Zhang H, Pu Z (2014) Influence of assimilating surface observations on numerical prediction of landfalls of Hurricane Katrina (2005) with an ensemble Kalman filter. Mon Weather Rev 142:2915–2934

Zhang H, Pu Z, Zhang X (2013) Examination of errors in near-surface temperature and wind from WRF numerical simulations in regions of complex terrain. Weather Forecast 28:893–914

Zupanski M (2017) Data assimilation for coupled modeling systems. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic, and hydrologic applications, vol III. Springer, Berlin, pp 55–70, 553 pp. https://doi.org/10.1007/978-3-319-43415-5

Ensemble Kalman Filter Experiments at 112-km and 28-km Resolution for the Record-Breaking Rainfall Event in Japan in July 2018

Koji Terasaki and Takemasa Miyoshi

Abstract A record-breaking heavy rainfall event occurred in western Japan in July 2018, associated with an intensified Baiu front between the Okhotsk high and the Pacific high. A cold air inflow from the Okhotsk high was an important factor to intensify the heavy rainfall. The Okhotsk high moved southward after Typhoon Prapiroon passed through the north of western Japan. The purpose of this study is to investigate the resolution dependence of the predictability of this rainfall event focusing on the Okhotsk high. We performed data assimilation and forecast experiments using a global numerical weather prediction system consisting of the Nonhydrostatic ICosahedral Atmospheric Model (NICAM) and the Local Ensemble Transform Kalman Filter (LETKF) at 112-km and 28-km horizontal resolution. The results showed that both experiments successfully reproduced the heavy rain, but the 28-km forecast outperformed the 112-km forecast for both location and intensity of the heavy rainfall. The 112-km forecast showed the location erroneously shifted northward because it failed to predict the retention of the Okhotsk high over Hokkaido. The results suggest that the precise prediction of the Okhotsk high be essential to predict the location of the associated Baiu front. The data assimilation cycles at 28-km resolution requires about 64 times more computations than that at 112-km resolution. Therefore, downscaled 28-km forecasts initialized by the 112-km LETKF analyses were also performed. The results showed that data assimilation at 28-km resolution was important in this case.

K. Terasaki (✉) · T. Miyoshi
RIKEN Center for Computational Science, Kobe, Japan
e-mail: koji.terasaki@riken.jp

T. Miyoshi
RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program, Kobe, Japan

Prediction Science Laboratory, RIKEN Cluster for Pioneering Research, Kobe, Japan
e-mail: takemasa.miyoshi@riken.jp

University of Maryland, College Park, Maryland, USA

Japan Agency for Marine–Earth Science and Technology, Yokohama, Japan

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_20

1 Introduction

A record-breaking heavy rainfall event occurred in western Japan in July 2018, with devastating 237 fatalities as of 9 January 2019. Figure 1 shows 72-h accumulated rainfall obtained by the Japan Meteorological Agency (JMA)'s radar analysis from

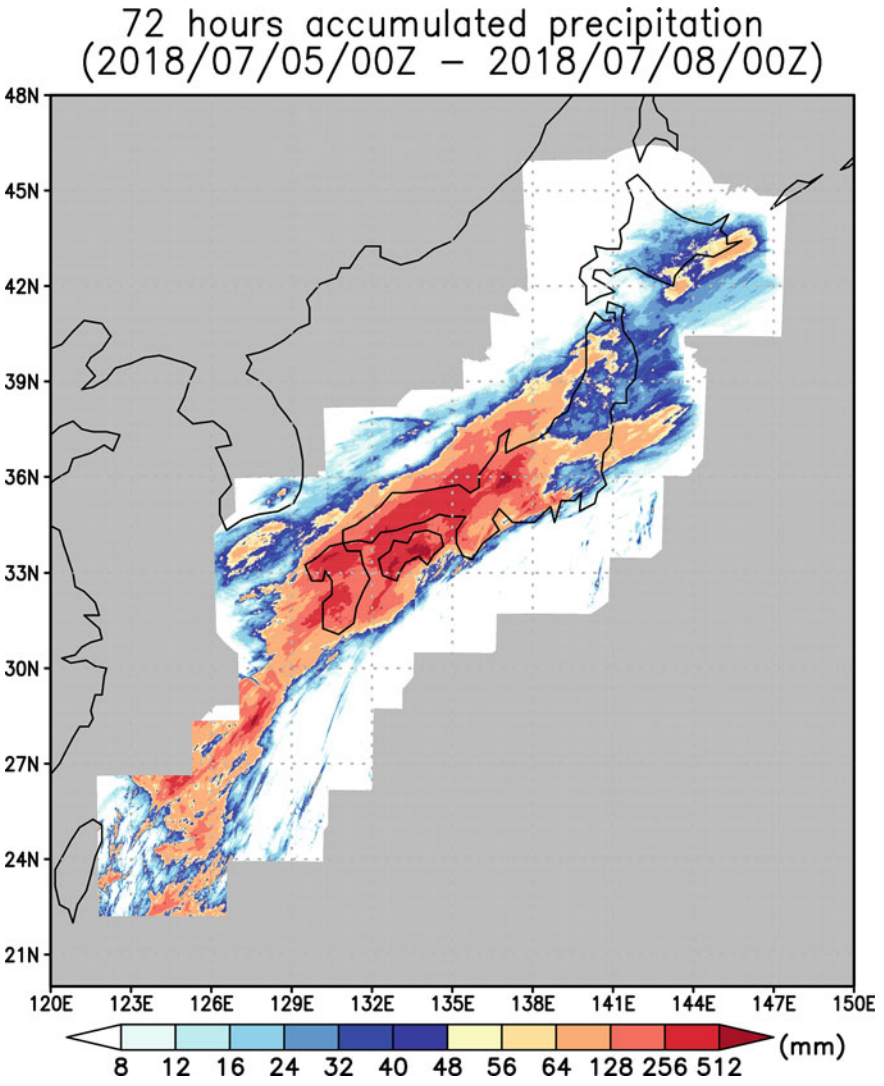

Fig. 1 72-hour accumulated rainfall by JMA's radar analysis from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018

0000 UTC 5 July 2018 to 0000 UTC 8 July 2018. More detailed and accurate numerical weather prediction is essential for better evacuation behaviors. JMA reported that this extreme heavy rainfall event was caused by the following reasons:

- extremely high moisture convergence from the Pacific high and the East China Sea,
- cold air inflow in the lower troposphere from the Okhotsk high
- enhanced convection due to the upper-level trough.

Sekizawa et al. (2019) and Shimpo et al. (2019) reported that the extremely large moisture transport caused the heavy rainfall. Matsunobu and Matsueda (2019) pointed out that an accurate prediction of the Pacific high was necessary for an accurate prediction of the heavy rainfall. Moteki (2019) focused on the role of Typhoon Prapiroon on the process of the Baiu front formation using the Japanese 55-year Reanalysis (JRA-55). A Baiu front was formed between the Okhotsk high and the Pacific high, and a cold air inflow from the Okhotsk high was an important factor to intensify the Baiu front activities. Moteki (2019) revealed that the strong cold air inflow from the Okhotsk high made the location of the Baiu front stationary at approximately 35°N.

In this study, we used a global data assimilation system which comprises the nonhydrostatic icosahedral atmospheric model (NICAM; Satoh et al. 2008, 2014) and the local ensemble transform Kalman filter (Hunt et al. 2007; Miyoshi and Yamane 2007). The NICAM-LETKF system has been developed to assimilate various types of observations including conventional observations, Global Satellite Mapping of Precipitation (GSMaP; Kubota et al. 2007) data and satellite radiances (Terasaki et al. 2015; Terasaki and Miyoshi 2017; Kotsuki et al. 2017a; Yashiro et al. 2016). Based on these achievements, an operational system called NEXRA, standing for the NICAM-LETKF JAXA (Japan Aerospace Exploration Agency) Research Analysis (Kotsuki et al. 2019), was developed to run the NICAM-LETKF continuously in near real time at 112-km resolution with 100 ensemble members. Using the NEXRA system, Kotsuki et al. (2019) investigated the predictability of the heavy rainfall event in western Japan in July 2018 and revealed that a generation of low-pressure system in the middle of the Baiu front contributed to increasing the rainfall in western Japan based on the ensemble correlation analysis. However, the horizontal resolution of 112 km was relatively low to represent the heavy rainfall and Baiu front.

The purpose of this study is to investigate how higher resolution at 28 km helps improve the predictability of this heavy rainfall event, focusing on the cold air inflow from the Okhotsk high. We performed data assimilation and forecast experiments with the NICAM-LETKF system at 112-km and 28-km resolution and compared the analyses and forecasts.

2 NICAM-LETKF System and Experimental Settings

NICAM is a non-hydrostatic global atmospheric model with an icosahedral grid, and the horizontal resolution is defined by the number of times the original icosahedron is divided. In NICAM, this number is called the grid division level or simply "glevel." The horizontal resolution is increased by dividing one triangle into four triangles. Table 1 shows the horizontal resolution corresponding each glevel. Glevel-0 corresponds to the original icosahedron.

The LETKF was first developed by Hunt et al. (2007). The LETKF updates the background state at each model grid point independently, so that it is suitable for parallel computation. The LETKF is an advanced data assimilation method that includes the flow-dependent error structure and generates ensemble perturbations representing the analysis errors. The global numerical weather prediction system NICAM-LETKF has been developed by combining the NICAM and LETKF (Terasaki et al. 2015, 2019; Terasaki and Miyoshi 2017).

This study performs data assimilation and forecast experiments using the NICAM-LETKF system at glevel-6 (112-km resolution) and glevel-8 (28-km resolution) with 38 vertical levels. The model top is set to 40 km in both experiments. The 112-km experiment applies the prognostic Arakawa-Schubert scheme for cumulus parameterization (Arakawa and Schubert 1974) and the large-scale condensation scheme for microphysics processes. The 28-km experiment explicitly computes cloud microphysics processes with a single moment bulk scheme (Tomita 2008; Roh and Satoh 2014) without any cumulus parameterization scheme.

The assimilated observations are conventional observations known as PREPBUFR from the National Centers for Environmental Prediction (NCEP) and satellite-borne Advanced Microwave Sounding Unit–A (AMSU-A) radiances. The channels 6, 7, and 8 of AMSU-A radiances are assimilated, which are sensitive from the middle troposphere to the lower stratosphere. Spatial thinning is essential when the observation errors are correlated in space. In this experiment, spatial thinning is applied to AMSU-A radiances to achieve a spatial density of 250 km.

Table 1 Horizontal resolution for each grid division level (glevel) of NICAM

Glevel	Horizontal resolution (km)
5	224
6	112
7	56
8	28
9	14
10	7
11	3.5
12	1.7
13	0.87

A relaxation to prior spread (RTPS; Whitaker and Hamill 2012) with a fixed relaxation parameter of 0.95 which was manually tuned by Kotsuki et al. (2017b) is applied. Covariance localization is an essential technique to avoid sampling noise due to a limited ensemble size in an ensemble Kalman filter. In this study, covariance localization is achieved by inflating the observation error for distant observations from an analysis grid, the method known as observation localization (Hunt et al. 2007; Greybush et al. 2011). A Gaussian function is used as the localization function, but it is replaced by zero beyond $2\sqrt{10/3}$ times the standard deviation to zero-out the distant noise. This also helps reduce the memory usage and computational cost in the LETKF analysis computation. The optimal localization scale depends on experimental settings such as the ensemble size, observation density, and model resolution. The horizontal localization scale defined by the standard deviation of a Gaussian function are chosen to be 400 km for the 112-km experiment following Terasaki and Miyoshi (2017). As for the 28-km experiment, narrower localization of 250 km is chosen after manual tuning. The vertical localization scale is set to 0.4 natural logarithms of pressure in both experiments.

The ensemble size is fixed at 32. The 32 initial conditions were chosen from the NCEP FNL (Final) operational global analysis data. Here, the analyses at every other day from 0000 UTC 1 May 2008 to 29 May 2008, from 0000 UTC 1 June 2008 to 29 June 2008, and 0000 UTC 1 and 3 May 2009 were chosen. Data assimilation cycles are performed every 6 h from 0000 UTC 10 June 2018 to 0000 UTC 9 July 2018 at both 112-km and 28-km resolution. The ensemble forecast experiments were initialized every day at 0000 UTC on 2–5 July 2018 for both experiments to investigate the predictability of this heavy rainfall event during 72 h from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018. Table 2 summarizes the experimental settings.

Table 2 Experimental settings for glevel-6 and glevel-8

Horizontal resolution	Glevel-6 (112 km)	Glevel-8 (28 km)
Vertical resolution	38 layers (model top: 40 km)	
Cumulus parameterization	Prognostic Arakawa-Schubert scheme	Not used
Cloud microphysics scheme	Not used	1-moment bulk scheme
Large scale condensation scheme	Used	Not used
Data assimilation	Local ensemble transform Kalman filter (LETKF)	
Ensemble size	32	
Assimilated observations	Conventional observations and AMSU-A radiances	
Period (DA)	From 0000 UTC 10 June 2018 to 0000 UTC 9 July 2018	
Ensemble forecast	Initial time: 0000 UTC 2 July 2018 (6-day forecast) Initial time: 0000 UTC 3 July 2018 (5-day forecast) Initial time: 0000 UTC 4 July 2018 (4-day forecast) Initial time: 0000 UTC 5 July 2018 (3-day forecast)	

The high-resolution data assimilation experiment at 28-km resolution requires about 64 times more computations than that at 112-km resolution. It would be beneficial if we have a similar forecast skill with a downscaled high-resolution forecast initialized by a lower resolution analysis. Therefore, we perform three kinds of deterministic forecast experiments initialized at 0000 UTC 4 July 2018: (1) 112-km forecast initialized by 112-km analysis, (2) 28-km forecast initialized by 28-km analysis, and (3) 28-km forecast initialized by 112-km analysis.

3 Results

3.1 Analysis

First, we compare the analysis sea level pressure (SLP) in both experiments. Figure 2 shows the analysis ensemble mean of SLP with glevel-8 (a–e) and glevel-6 (f–j). Both experiments reproduce the location of Typhoon Prapiroon well. However, the intensity of Prapiroon is underestimated in the 112-km experiment because it is too coarse to resolve the typhoon structure well. Alternatively, the 28-km experiment could resolve the typhoon structure better than the 112-km experiment and improved its location and intensity. According to the JMA's best track analysis, the minimum SLP of Prapiroon reached to 960 hPa at 1800 UTC 2 July 2018 (obtained from https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/besttrack.html). The minimum SLP of Prapiroon in the NICAM-LETKF analyses are 988.9 hPa in glevel-6 and 962.5 hPa in glevel-8. Prapiroon moves northeastward and transforms to an extratropical cyclone at 0600 UTC 4 July 2018 before passing through the Tsugaru Strait further northeastward. After the extratropical cyclone moves to the east of Hokkaido, the northerly wind from the western side of the cyclone brings strong cold air and moves the Okhotsk high southward. In both experiments, the synoptic fields are generally well reproduced in the analyses.

3.2 Ensemble Forecast

We perform ensemble forecast experiments initialized every day at 0000 UTC on 2–5 July 2018 to investigate the predictability of the heavy rainfall event. The prediction of heavy rainfall has a large uncertainty in its intensity and location. Ensemble forecasts provide information on the predictability of the heavy rainfall. First, we look at the ensemble-mean precipitation and the maximum precipitation in ensemble members.

Figure 3 shows the ensemble-mean forecasts of 72-h accumulated rainfall from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018 with four different initial times. All forecasts generally show wet conditions over a broad area in Japan. In general, forecasts initialized at earlier times show broader rainfall distributions with lower

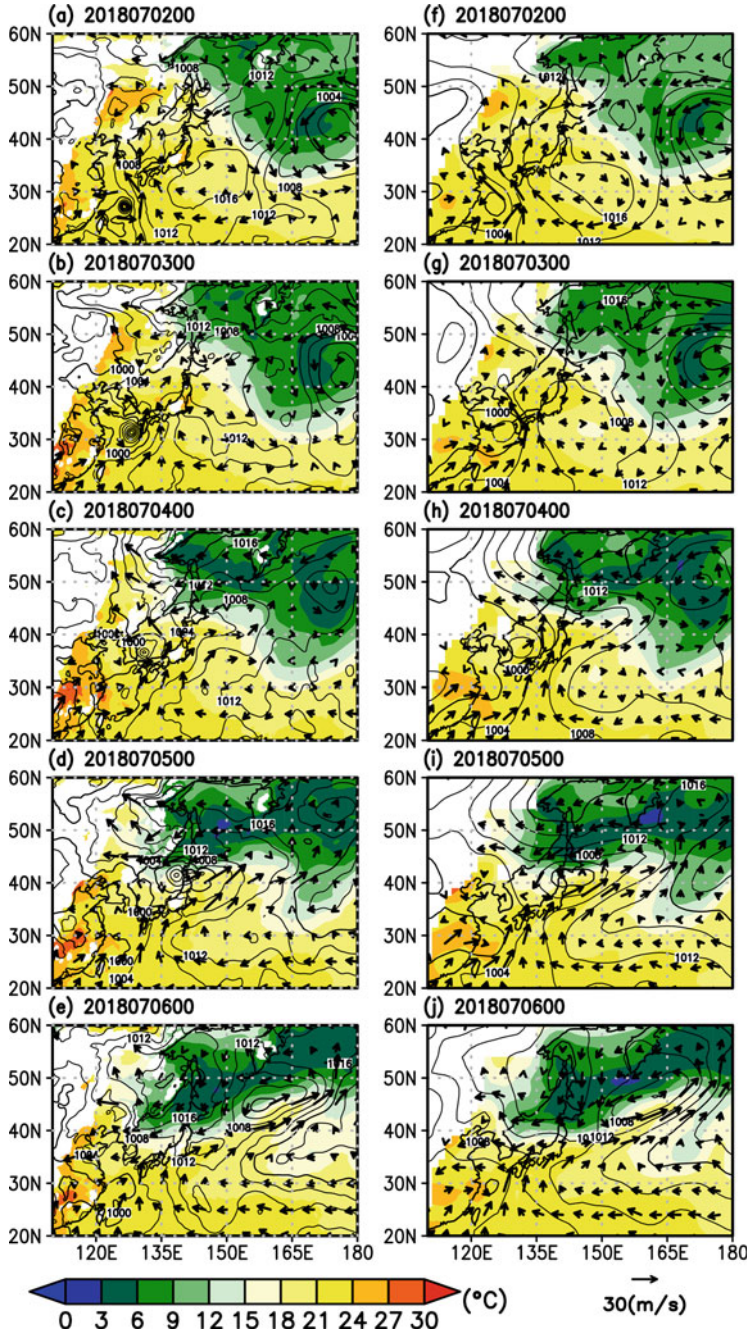

Fig. 2 Ensemble-mean analysis sea level pressure (hPa: contours), temperature (K: shades), and horizontal winds (vectors) at 925 hPa for every day at 0000 UTC from 2 to 6 July 2018 for **a**–**e** 28-km and **f**–**j** 112-km experiments

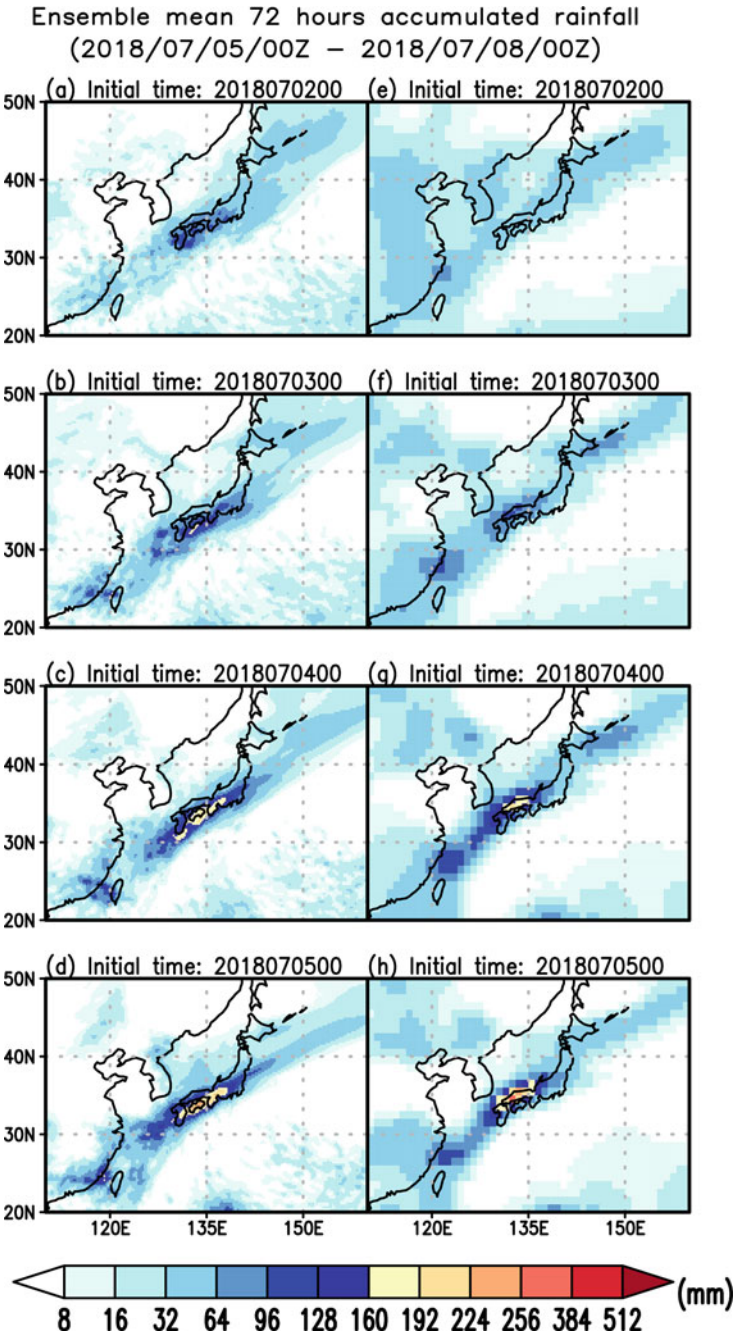

Fig. 3 Ensemble-mean forecast of 72-h accumulated rainfall from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018 for **a**–**d** 28-km and **e**–**h** 112-km experiments, initialized at **a**, **e** 0000 UTC 2 July 2018, **b**, **f** 0000 UTC 3 July 2018, **c**, **g** 0000 UTC 4 July 2018, and **d**, **h** 0000 UTC 5 July 2018

peak values. This is reasonable since the ensemble spread is generally larger in longer forecasts, so that ensemble averaging makes spatial patterns broader with lower peaks. As the initial time becomes later, the rainfall distributions become narrower with higher peak values. The forecasts initialized on 2 and 3 July 2018 at 112-km resolution (Fig. 3e, f) show that the rainfall area spreads over the Korean Peninsula and eastern China. However, the 28-km forecasts (Fig. 3a, b) have less rainfall in these regions, indicating the better forecast skill at 28-km resolution. The forecasts initialized on 4 and 5 July 2018 capture the heavy rainfall in general, but the peak location at 112-km resolution is shifted to the north compared with the JMA's observation and the 28-km experiment.

Using the ensemble forecasts, we investigate the maximum 72-h accumulated rainfall among all ensemble members (Fig. 4). This shows the potential risk and corresponds to potentially the worst-case scenario captured by the ensemble forecasts. The 28-km forecasts (Fig. 4a–d) show better predictions of the heavy rainfall in its location and intensity than the 112-km forecasts (Fig. 4e–h). The forecasts initialized on 2 July 2018 at 112-km resolution (Fig. 4e) fail in predicting the potential of heavy rainfall, but those initialized on and after 3 July 2018 are improved (Fig. 4f–h). The forecasts at 28-km resolution initialized on 2 July 2018 (Fig. 4a) predict the possible occurrence of the heavy rainfall in western Japan. As the initial time becomes later, the heavy rainfall area becomes smaller with higher peak values, indicating more certainty of the precipitation forecast. The results suggest that the ensemble forecasts be useful to capture the potential risk of the heavy rainfall compared with ensemble-mean forecast.

Next, we investigate the role of the typhoon track and the Okhotsk high on the heavy rainfall. Figure 5 shows the JMA's best track and ensemble forecasts of typhoon tracks. The 112-km forecasts initialized on 2 July 2018 predict the track of Prapiroon northwestward compared with the best track and has significant uncertainty (Fig. 5d). The forecasts initialized on and after 3 July 2018 at 112-km resolution improve the typhoon track forecast. The track forecasts at 28-km resolution are more accurate than those at 112-km resolution. Both experiments tend to predict the location of Prapiroon eastward. According to the analysis field, Prapiroon moves to the southern side of eastern Hokkaido around Nemuro Peninsula. Figure 6 shows the forecast SLP fields at 0000 UTC 6 July 2018 with four different initial times on 2–5 July 2018 at 28-km resolution (Fig. 6a–d) and 112-km resolution (Fig. 6e–h). The extratropical cyclone, which is transformed from Prapiroon, moves southeastward for the 28-km experiment as the initial time becomes later, and it causes a southeastward shift of the Okhotsk high. The SLP ensemble spread around the extratropical cycle is large in the forecasts initialized on 2 July 2018 (Fig. 6a), with large uncertainties in the track forecast. However, the ensemble forecasts become more certain with later initial times. As the track forecast becomes better, the location of the Okhotsk high is also predicted well.

The location and intensity of the extratropical cyclone affect the location of the Okhotsk high and may have altered the amount of the cold air inflow toward the Baiu front. The cold air inflow from the Okhotsk high is one of the important factors to cause this heavy rainfall by enhancing the Baiu front activity. Figure 7 shows

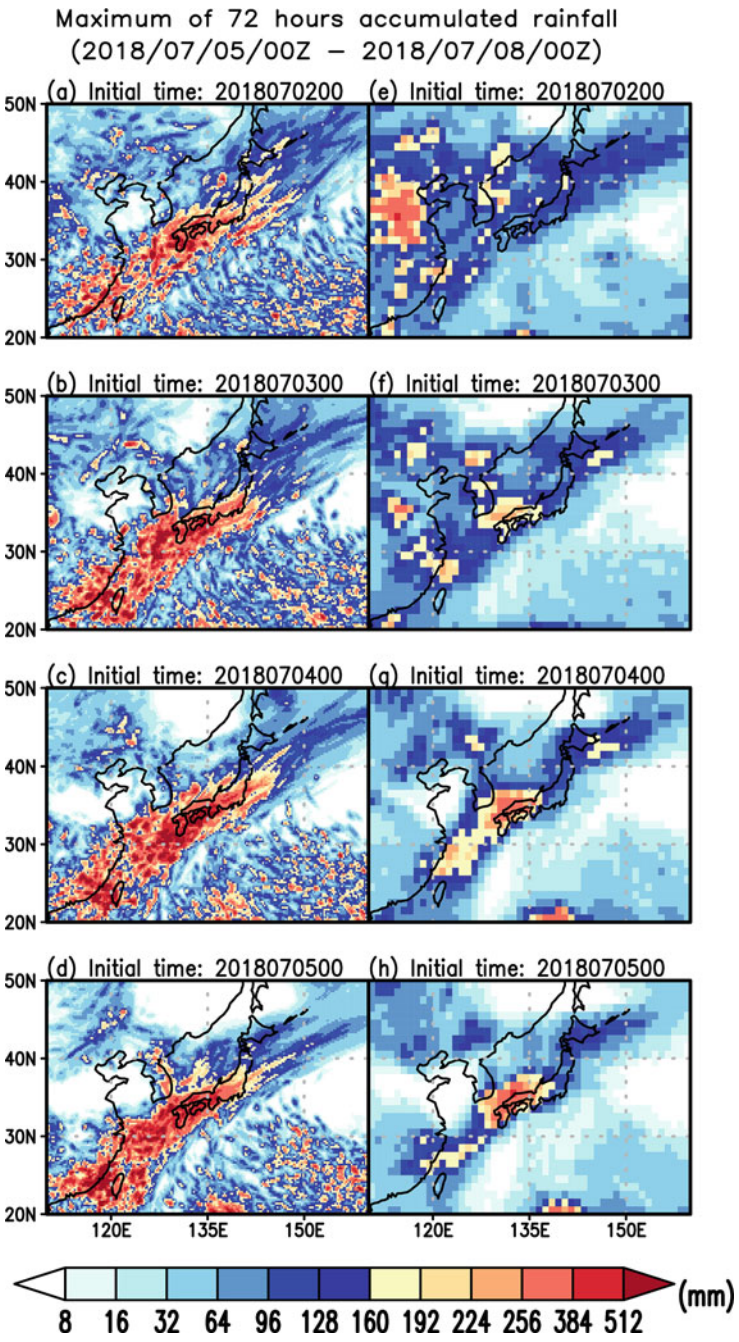

Fig. 4 Maximum 72-h accumulated rainfall from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018 among all ensemble members at **a**–**d** 28-km and **e**–**h** 112-km forecasts, initialized at 0000 UTC on **a**, **e** 2 July 2018, **b**, **f** 3 July 2018, **c**, **g** 4 July 2018, and **d**, **h** 5 July 2018

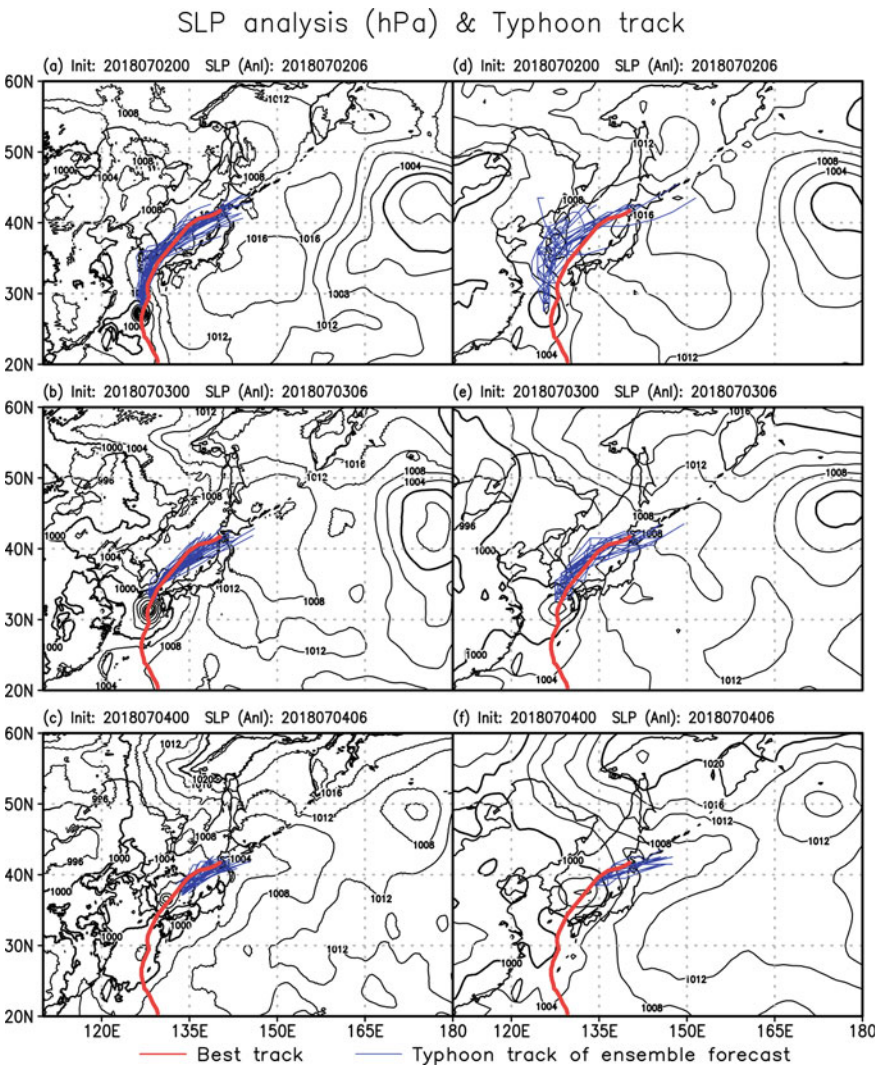

Fig. 5 Track forecast of Prapiroon at **a**–**c** 28-km and **d**–**f** 112-km resolution. Red line indicates the JMA best track from 0000 UTC 30 June 2018 to 0600 UTC 5 July 2018, and blue lines indicate the tracks of ensemble forecasts from 6 h after the initial times to 0600 UTC 5 July 2018. The forecasts are initialized at 0000 UTC on **a**, **d** 2 July 2018, **b**, **e** 3 July 2018, and **c**, **f** 4 July 2018. Contours show the analysis sea level pressure at 6 h later of each initial time

the time-series of the regional averaged temperature at 925 hPa around western Japan (132.5–137.5°E, 37.5–42.5°N). To verify the accuracy of the NICAM-LETKF analysis, we compare the results with the ERA-interim (Dee et al. 2011) and National Centers for Environmental Prediction (NCEP) NCEP FNL analyses. After Prapiroon passes through the Japan islands, the cold air is brought from the Okhotsk high.

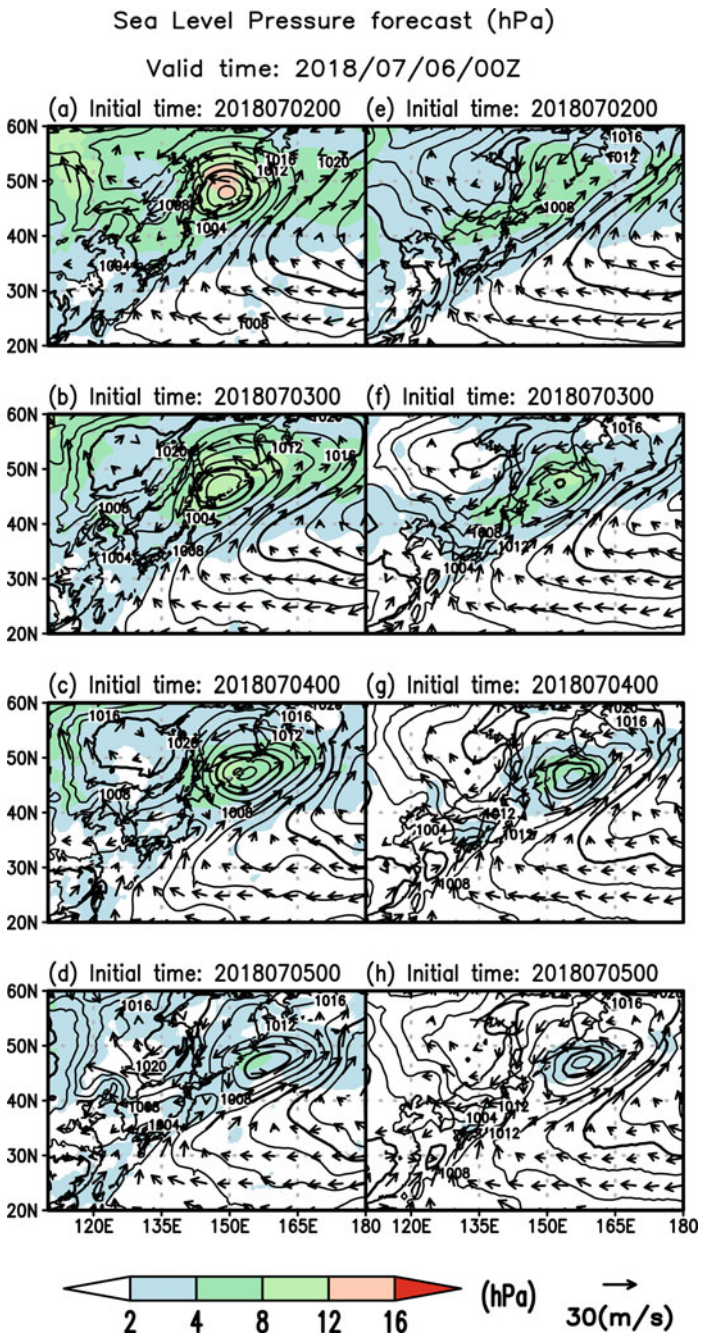

Fig. 6 Forecast ensemble mean (contours) and spread (shades) for SLP (hPa) and forecast ensemble mean for 925-hPa-level horizontal winds (m/s, vectors) valid at 0000 UTC 6 July 2018 for **a**–**d** 28-km and **e**–**h** 112-km experiments. The forecasts are initialized at 0000UTC on **a**, **e** 2 July 2018, **b**, **f** 3 July 2018, **c**, **g** 4 July 2018, and **d**, **h** 5 July 2018

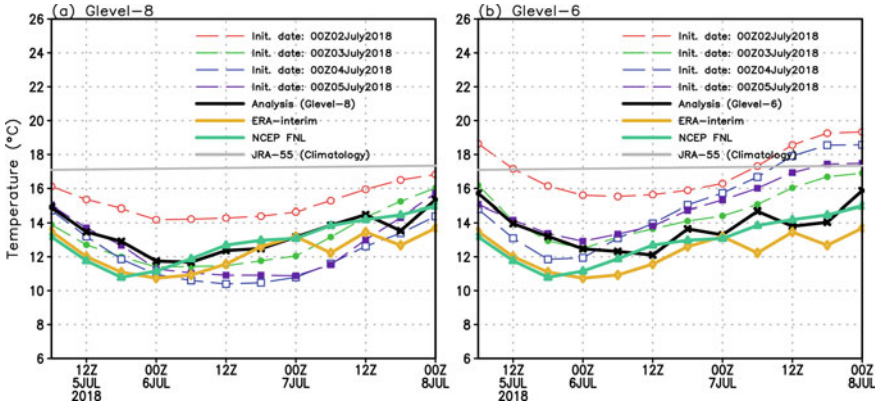

Fig. 7 Time-series of the regional averaged (132.5–137.5°E, 37.5–42.5°N) temperature at 925 hPa for **a** 28-km and **b** 112-km experiments. Dashed lines indicate the forecast experiments with different initial dates. Thick black lines indicate the analyses temperature for **a** 28-km and **b** 112-km experiments. Orange and light green lines show temperature of ERA-interim and NCEP FNL, respectively. The gray line shows the climatology of the JRA-55

The regional averaged temperature drops to around 12 °C, much lower than the climatology by JRA-55 (gray line in Fig. 7). There are slight differences between the NICAM-LETKF analysis and the other reanalysis datasets. Namely, the NICAM-LETKF system reproduces the cold air inflow from the Okhotsk high well. In both experiments, the forecasts initialized on 2 July 2018 predict higher temperature than the analyses. The poor forecasts of the typhoon track initialized on 2 July 2018 lead to the poor reproducibility of the cold air inflow from the Okhotsk high. The weak cold air inflow from the Okhotsk high characterizes a possible reason why the forecasts initialized on 2 July 2018 fail to predict the heavy rainfall. The forecasts initialized on and after 3 July 2018 at 28-km resolution reproduce the cold air inflow from the Okhotsk high well. However, the 112-km forecasts show the temperature increase after 7 July 2018 and could not predict the duration of the cold air inflow.

Figure 8 shows the forecast SLP and temperature at 925 hPa initialized at 0000 UTC 5 July 2018. The 28-km forecasts predict the location and intensity of the Okhotsk high well (Fig. 8a–d). The Okhotsk high persistently locates over the north of Hokkaido, and it leads to the persistent cold air inflow (purple dashed line in Fig. 7a). The 112-km forecasts also successfully predict the Okhotsk high until 1200 UTC 6 July 2018 (Fig. 8e–f), but the Okhotsk high moves eastward from 0000 UTC 7 July 2018 (Fig. 8g–h). As a result, the 112-km forecasts could not predict the retention of the cold air inflow from the Okhotsk high (purple dashed line in Fig. 7b). These make the Baiu front shifted to the north and cause the northward shift of the heavy rainfall in the 112-km forecasts. The retained Okhotsk high is a critical factor to keep the location of the Baiu front, resulting in long-lasting heavy rainfall.

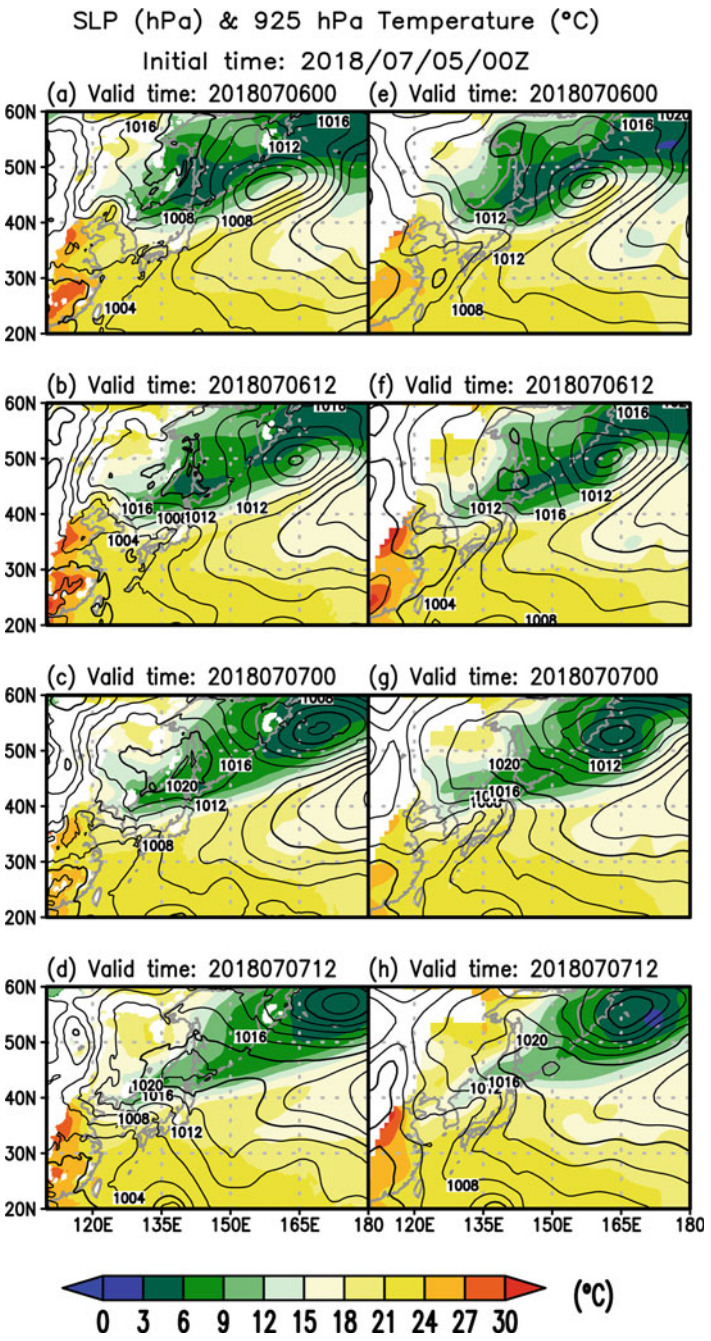

Fig. 8 Forecast ensemble mean SLP (hPa: contours) and temperature at 925 hPa (color shades) at **a**–**d** 28-km and **e**–**h** 112-km experiments initialized at 0000 UTC 5 July 2018 and valid at **a**, **e** 0000 UTC 6 July 2018, **b**, **f** 1200 UTC 6 July 2018, **c**, **g** 0000 UTC 7 July 2018, and **d**, **g** 1200 UTC 7 July 2018

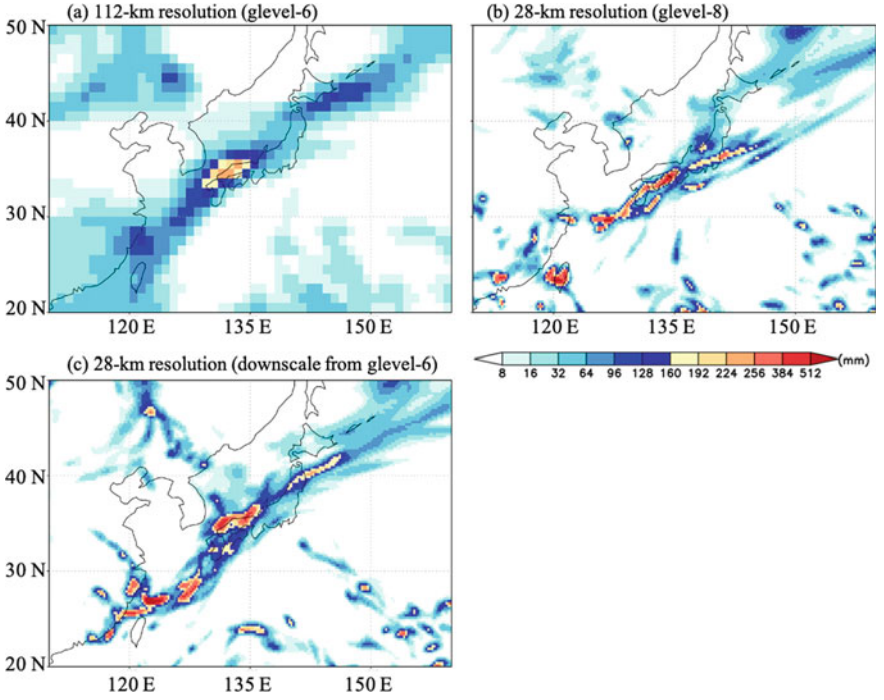

Fig. 9 72-h accumulated rainfall from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018 for the ensemble mean forecasts of **a** 112-km resolution, **b** 28-km resolution, and **c** the experiment downscaled from 112-km resolution to 28-km resolution initiated at 0000 UTC 4 July 2018

3.3 Downscaled Forecast

Finally, we show the results of deterministic forecasts. Figure 9 shows 72-h accumulated rainfall from 0000 UTC 5 July 2018 to 0000 UTC 8 July 2018 initialized at 0000 UTC 4 July 2018. In general, the heavy rainfall was predicted well in western Japan. The 112-km forecast (Fig. 9a) shows much lower peak rainfall than the two 28-km forecasts (Fig. 9b, c). Both Fig. 9a, c use the same initial condition at 112-km resolution, but the downscaled forecast (Fig. 9c) shows the finer structure of precipitation with higher peak values. However, the location of the heavy rainfall is similar. Namely, both Fig. 9a, c shows peak rainfall areas northward compared with the 28-km experiment (Fig. 9b). Therefore, the initial condition at 112-km resolution would not be optimal to predict the location of the heavy rainfall accurately. This is probably because the initial conditions at 112-km resolution make the Okhotsk high and Baiu front shifted northwestward, while the initial conditions 28-km resolution improve these features. In addition, the cloud physics schemes used in 112-km and

28-km resolution are different, and the initial conditions downscaled from 112-km to 28-km resolution would not be optimal for the 28-km resolution. The results suggest that a high-resolution data assimilation be important for accurate prediction of the heavy rainfall in this case.

4 Summary and Concluding Remarks

This study investigated the predictability of the heavy rainfall event occurred in July 2018 in western Japan using the NICAM-LETKF system at two different horizontal resolution of 112 km and 28 km. A rare combination of various atmospheric conditions brought this record-breaking heavy rainfall. The results indicated that the location of the Okhotsk high and the associated cold air inflow were important factors to determine the location and intensity of the Baiu front. The location of the Okhotsk high was affected by the track of the extratropical cycle changed from Typhoon Prapiroon and the northerly winds from the western side of the extratropical cyclone.

According to the results of the ensemble forecast experiments, the track forecast of Prapiroon was important to predict the location of the Okhotsk high. Both experiments well reproduced the typhoon track initialized at and after 0000 UTC 3 July 2018. However, the 28-km forecasts predicted the strength and duration of the cold air inflow from the Okhotsk high better than the 112-km forecasts. This study also performed the downscaled forecast at 28-km resolution initialized by analysis at 112-km resolution. Running a low-resolution data assimilation cycle helped reduce the computational cost. However, the location of the heavy rainfall was not improved because 112-km analyses led to the northwestward shift of the Okhotsk high even though 28-km model was used for prediction. The results suggest that running the data assimilation cycles at high resolution be beneficial.

It is important for disaster prevention and mitigation for heavy rain events to have a longer lead time and more accurate predictions. Besides, we performed forecast experiments with a relatively low resolution of 28 km to resolve the torrential rainfall. More detailed predictions at higher resolution are needed for more accurate prediction and more direct link to evacuation behaviors. Lateral boundary conditions are very important in regional numerical weather prediction. Using the ensemble prediction data obtained from the global ensemble data assimilation system, it becomes possible to perform high-resolution predictions with proper uncertainties of the lateral boundary conditions for regional ensemble data assimilation and forecasts.

References

Arakawa A, Schubert WH (1974) Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. J Atmos Sci 31:674–701

Dee DP et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quart J R Meteorol Soc 137:553–597. https://doi.org/10.1002/qj.828

Greybush SJ, Kalnay E, Miyoshi T, Ide K, Hunt B (2011) Balance and ensemble Kalman filter localization techniques. Mon Wea Rev 139:511–522. https://doi.org/10.1175/2010MWR3328.1

Hunt BR, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126

Kubota T, Shige S, Hashizume H, Aonashi K, Takahashi N, Seto S, Hirose M, Takayabu YN, Nakagawa K, Iwanami K, Ushio T, Kachi M, Okamoto K (2007) Global precipitation map using satellite-borne microwave radiometers by the GSMaP Project: Production and validation. IEEE Trans Geosci Remote Sens 45(7):2259–2275

Kotsuki S, Miyoshi T, Terasaki K, Lien GY, Kalnay E (2017a) Assimilating the global satellite mapping of precipitation data with the nonhydrostatic icosahedral atmospheric model NICAM. J Geophys Res Atmos 122:1–20. https://doi.org/10.1002/2016JD025355

Kotsuki S, Ota Y, Miyoshi T (2017b) Adaptive covariance relaxation methods for ensemble data assimilation: experiments in the real atmosphere. Quart J Roy Meteor Soc 143:2001–2015. https://doi.org/10.1002/qj.3060

Kotsuki S, Terasaki K, Kanemaru K, Satoh M, Kubota T, Miyoshi T (2019) Predictability of record-breaking rainfall in Japan in July 2018: ensemble forecast experiments with the near-real-time global atmospheric data assimilation system NEXRA. SOLA 15A:1–7. https://doi.org/10.2151/sola.15A-001

Matsunobu T, Matsueda M (2019) Assessing the predictability of heavy rainfall events in Japan in early July 2018 on medium-range timescales. SOLA 15A:1–7. https://doi.org/10.2151/sola.15A-004

Miyoshi T, Yamane S (2007) Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution. Mon Wea Rev 135:3841–3861. https://doi.org/10.1175/2007MWR1873.1

Moteki Q (2019) Role of Typhoon Prapiroon (Typhoon No. 7) on the formation process of the Baiu front inducing heavy rain in July 2018 in western Japan. SOLA 15A:37–42. https://doi.org/10.2151/sola.15A-007

Roh W, Satoh M (2014) Evaluation of precipitating hydrometeor parameterizations in a single-moment bulk microphysics scheme for deep convective systems over the tropical open ocean. J Atmos Sci 71:2654–2673. https://doi.org/10.1175/JAS-D-13-0252.1

Satoh M, Matsuno T, Tomita H, Miura H, Nasuno T, Iga S (2008) Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations. J Comput Phys (special Issue on Predicting Weather, Climate and Extreme Events) 227:3486–3514. https://doi.org/10.1016/j.jcp.2007.02.006

Satoh M, Tomita H, Yashiro H, Miura H, Kodama C, Seiki T, Noda AT, Yamada Y, Goto D, Sawada M, Miyoshi T, Niwa Y, Hara M, Ohno T, Iga S, Arakawa T, Inoue T, Kubokawa H (2014) The non-hydrostatic icosahedral atmospheric model: Description and development. Prog Earth Planet Sci 1:18. https://doi.org/10.1186/s40645-014-0018-1

Sekizawa S, Miyasaka T, Nakamura H, Shimpo A, Takemura K, Maeda S (2019) Anomalous moisture transport and oceanic evaporation during a torrential rainfall event over western Japan in early July 2018. SOLA 15A:22–25. https://doi.org/10.2151/sola.15A-005

Shimpo A et al (2019) Primary factors behind the heavy rain event of July 2018 and the subsequent heat wave in Japan. SOLA 15A:13–18. https://doi.org/10.2151/sola.15A-003

Terasaki K, Sawada M, Miyoshi T (2015) Local ensemble transform Kalman filter experiments with the nonhydrostatic icosahedral atmospheric model NICAM. SOLA 11:23–26. https://doi.org/10.2151/sola.2015-006

Terasaki K, Miyoshi T (2017) Assimilating advanced microwave sounding Unit-A satellite radiances with the NICAM-LETKF. J Meteorol Soc J 95:433–446. https://doi.org/10.2151/jmsj.2017-028

Terasaki K, Kotsuki S, Miyoshi T (2019) Multi-year analysis using the NICAM-LETKF data assimilation system. SOLA 15:41–46. https://doi.org/10.2151/sola.2019-009

Tomita H (2008) New microphysical schemes with five and six categories by diagnostic generation of cloud ice. J Meteor Soc Jpn 86A:121–142

Whitaker JS, Hamill TM (2012) Evaluating methods to account for system errors in ensemble data assimilation. Mon Wea Rev 140:3078–3089. https://doi.org/10.1175/MWR-D-11-00276.1

Yashiro H, Terasaki K, Miyoshi T, Tomita H (2016) Performance evaluation of throughput-aware framework for ensemble data assimilation: the case of NICAM-LETKF. Geosci Model Dev 9:2293–2300. https://doi.org/10.5194/gmd-9-2293-2016

Convective-Scale Data Assimilation and Precipitation Prediction with a Local Ensemble Transform Kalman Filter Radar Assimilation System Over Complex Terrain: A Thorough Investigation with the Heavy Rainfall in Taiwan on 16 June 2008

Shu-Chih Yang, Hsiang-Wen Cheng, Pin-Ying Wu, Zih-Mao Huang, and Chih-Chien Tsai

Abstract Convective-scale heavy rainfall prediction in Taiwan is very challenging due to the complexity of the precipitation systems and the topography. The WRF-LETKF Radar Assimilation System (WLRAS) has been an important component of convective-scale data assimilation with the purpose of short-term precipitation prediction. This chapter reviews how assimilating radar radial velocity and reflectivity data impacts the prediction of a heavy rainfall event over the coast of southwestern Taiwan on 16 June 2008 and further affect the prediction of the initialization and development of an afternoon thunderstorm downstream over northern Taiwan. The characteristics and attributions of the sampling error due to insufficient ensemble members are further identified with the same case. Finally, improving the moisture analysis accuracy is relatively restrictive in WLRAS since the moisture variable is not directly observed in radar data. We illustrate how assimilating the ground-based GNSS ZTD data can complement radar data in convective-scale data assimilation and precipitation prediction.

S.-C. Yang (✉)
Department of Atmospheric Sciences, National Central University, Taoyuan, Taiwan
e-mail: shuchih.yang@atm.ncu.edu.tw

H.-W. Cheng
National Space Organization, Hsinchu, Taiwan

P.-Y. Wu
Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan

Z.-M. Huang
Central Weather Bureau, Taipei, Taiwan

C.-C. Tsai
National Science and Technology Center for Disaster Reduction, New Taipei, Taiwan

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_21

1 Introduction

Heavy rainfall has always been a primary concern among high-impact weather events due to the enormous societal and economic impacts. Thus, accurate heavy rainfall prediction is critical for disaster prevention and decision making. Combining high-resolution numerical weather prediction (NWP) models with data assimilation (DA) systems (Yano et al. 2018) is one of the state-of-the-art approaches to improving heavy rainfall prediction. However, heavy rainfall prediction over complex orography is even more challenging due to the complicated dynamical and thermodynamic processes and their interactions with the terrain. This review chapter discusses applications of the radar data assimilation to heavy rainfall prediction in Taiwan, which has complex terrain (Fig. 1a), and the associated challenges and remedies.

DA methods, which estimate the optimal initial conditions for numerical modeling, synergize the information provided by numerical model simulation and observations (Kalnay 2003). For convective-scale NWP models, radar DA based on methodologies of variational analysis or the Ensemble Kalman filter has been widely utilized to construct a realistic convective-scale structure because of the high spatial and temporal resolution of radar data (Sun and Wang 2013; Aksoy et al. 2009; Zhang et al. 2009; Schraff et al. 2016). With the advantage of using flow-dependent background error statistics (Chung et al. 2013; Ménétrier et al. 2014), ensemble-based radar data assimilation (EnRDA) has been extensively applied in operational centers and research works (e.g., Tong and Xue 2005; Gao and Xue 2008; Dowell et al. 2011; Chang et al. 2014, Wheatley et al. 2015). EnRDA has demonstrated important applications in improving precipitation prediction accuracy (Bick et al. 2016; Yokota et al. 2018; Gastaldo et al. 2018) and probabilistic predictions for convective storms (Snook et al. 2015; Yussouf et al. 2016; Putnam et al. 2019). Since the milestone of the first EnRDA work by Snyder and Zhang (2003), the use of assimilation strategies and the assimilation of data from new-generation radar systems have advanced the applications of EnRDA in convective-scale prediction. For example, assimilating non-precipitating information suppresses spurious precipitation (Tong and Xue 2005; Aksoy et al. 2009; Bick et al. 2016). Using rapid update cycling with an interval shorter than one hour can capture the development of convective systems (Bick et al. 2016). The assimilation of polarimetric radar data contributes to improving the analysis accuracy of model variables involving cloud microphysics (Jung et al. 2010; Li et al. 2017; Carlin et al. 2017; Putnam et al. 2019), while the assimilation of the phase-array-radar data at a very rapid interval (less than 1 min) can capture the development of fast-evolving convection (Miyoshi et al. 2016; Huang et al. 2020).

Despite the success of EnRDA in convective-scale precipitation prediction, there are also challenges since it inherits the properties of EnKF and carries the limitations of the radar measurement. A common issue is that using an insufficient ensemble size introduces sampling error, which can contaminate the estimation of the background error covariance, resulting in spurious analysis corrections that degrade the performance of EnRDA. It is essential to apply covariance localization, which tapers the

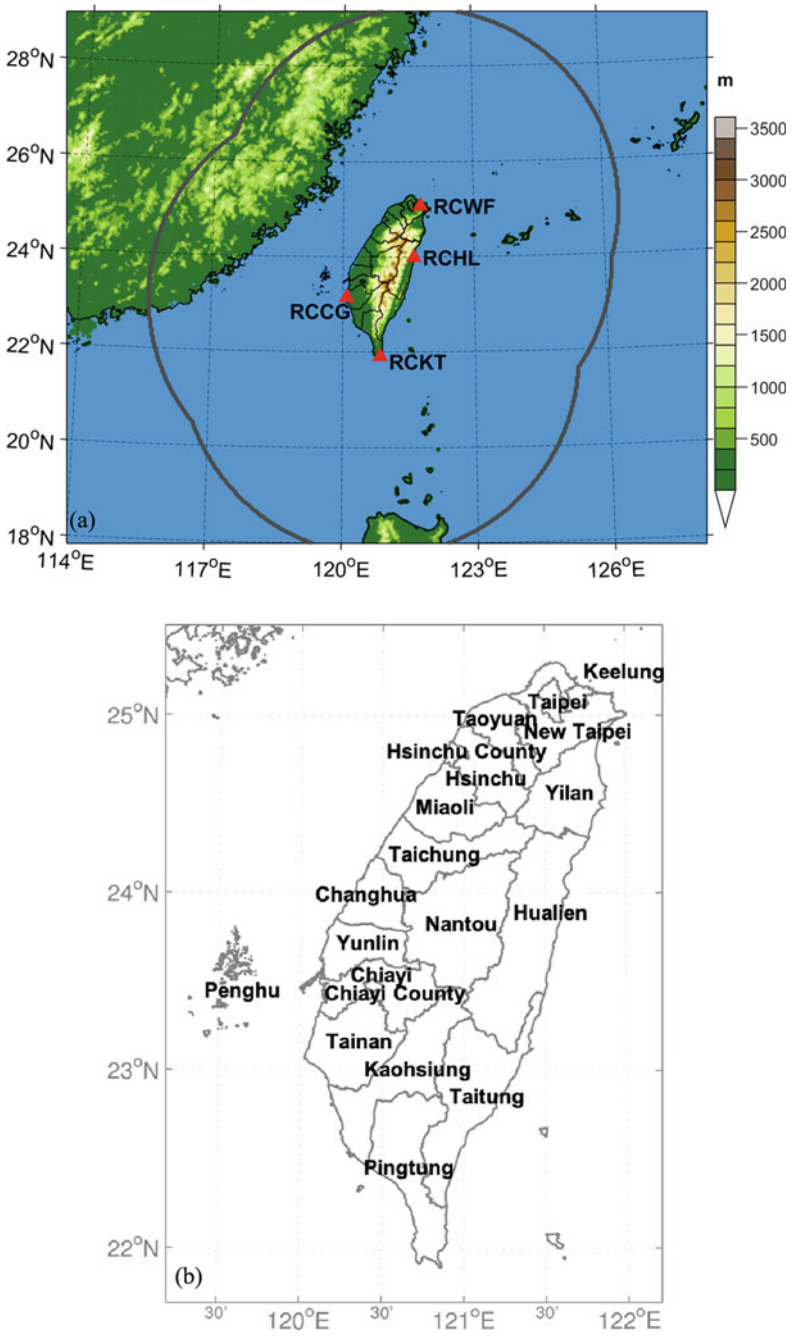

Fig. 1 **a** Locations of the radar sites used in this study. Color shading indicates the topography, and the gray line indicates the observing range of the radar (230 km). **b** Administrative divisions of Taiwan

covariance values and limits the analysis corrections at long distances, to avoid the degradation from sampling error (Hamill et al. 2001; Miyoshi et al. 2014). Optimizing the use of covariance localization is critical to keep the impact of radar data, given the consideration of the spatial characteristic of the prognostic variables (Zhang et al. 2009; Tsai et al. 2014) and model error (Sobash and Stensrud 2013). Another common issue is the dispersiveness of the ensemble spread, whose amplitude is important to manifest the observational impact. The covariance inflation method, such as multiplicative inflation, additive noise, and relaxation methods, is adopted to increase the spread (Zeng et al. 2018; Sobash and Wicker 2015; Maldonado et al. 2020). Under the framework of observation system simulation experiments with EnRDA, Maldonado et al. (2020) report that the impact of horizontal localization on short-range ensemble forecasts was larger compared to inflation. Furthermore, correcting the unobserved variables during the assimilation process, such as the moisture and temperature fields, is achieved through the cross-variable background error covariance. The characteristics of these fields are critical for convection initialization and development. However, the precondition of convection is not observed by conventional S-, C- and X-band radars, and the data are available only after the rain starts. Besides, there is no near-surface information due to the limitation of the scanning angle and ground clusters. Therefore, it is restrictive to improve the initialization of the convections by assimilating radar data alone. At last, EnKF assumes Gaussian probability distributions for background and observation error. However, nonlinearities in the dynamics of convective storms and observation operators can result in non-Gaussian error probability distributions and may limit the effectiveness of data assimilation techniques in convective-scale analysis and prediction (Vukicevic and Posselt 2008; Posselt and Bishop 2012). In particular, the Gaussian framework could be inappropriate for variables such as water vapor and other classes of hydrometeors (Posselt et al. 2014), which are related to the radar observed quantities.

Focusing on the applications of EnRDA, this chapter reviews the performance and challenges of EnRDA on convective-scale analysis and precipitation prediction over complex terrain with the WRF-Radar Assimilation Local Ensemble Kalman Filter System (WRLAS). WLRAS combines the local ensemble transform Kalman filter (LETKF, Hunt et al. 2007) with the Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) and assimilate data from the radar network in Taiwan (Tsai et al. 2014, hereafter TYL14). The distribution of radar systems provides good coverage of Taiwan, and the observing range partially overlaps (Fig. 1a). Although severe weather systems can be well observed when they approach Taiwan with heavy precipitation, the application of WLRAS to precipitation prediction is still very challenging given the complexity of the heavy rainfall and the topography of the Taiwan. This chapter gives a thorough review of the impact of assimilating radar data for predicting the heavy rainfall in Taiwan on 16 June 2008. We discuss the characteristics and attributions of convective-scale sampling errors and a potential strategy to modify the covariance inflation under the influence of prevailing winds. Furthermore, assimilating the ground-based GNSS Zenith Total Delay (ZTD), which represents the moisture information, is proposed as a remedy for the limitation of WLRAS. The impact of assimilating GNSS ZTD in addition to radar data is presented based

on the prediction of two heavy rainfall events with very different synoptic weather conditions.

The chapter is organized as follows. Section 2 briefly introduces WLRAS and the assimilation parameters. Section 3 illustrates the impact of assimilating the radar data on short-term precipitation prediction on 16 June 2008. Section 4 discusses the limitations of WLRAS and the added benefit from assimilating the ground-based GNSS ZTD observations. Finally, a summary and conclusions are given in Sect. 5.

2 WRF-Local Ensemble Transform Kalman Filter Radar Assimilation System (WLRAS)

The WLRAS combines the radar data assimilation and high-resolution numerical weather prediction model with the advantage of using a flow-dependent background error covariance over complex terrain. The analysis ensemble can be used for convective-scale ensemble prediction, providing products such as probability quantitative precipitation forecasts.

WLRAS assimilates the radar radial velocity (V_r) and reflectivity (Z_h) from four radars in Taiwan. In general, four S-band radars at Wufenshen (RCWF), Hualien (RCHL), Chigu (RCCG), and Kent-din (RCKT) covers Taiwan well (Fig. 1a). The horizontal scanning range at the lowest elevation is 230 km for V_r and 460 km for Z_h, and there is a total of 9 scanning elevations. The radar data have been processed with different quality control steps to consider non-meteorological echoes, such as ground/sea clutter, beam blocking, and attenuation (Chang et al. 2009; Zhang et al. 2011). The original resolution of the radar data is 250 m horizontally every 7.5 min. We adopt the superobbing strategy (Lindskog et al. 2004) to thin the data and avoid spatial correlations between observations. For each radar, fan-shaped areas are defined with 2-km and 2-degree intervals in the radial and azimuthal directions, respectively. The center of each fan-shaped area is the location of a superobservation, and the value is obtained by summing the radar data within this area with a distance-based Gaussian weighting factor. The observation errors are 3 m s^{-1} for V_r and 5 dBZ for Z_h (TYL14). Note that the fan-shaped area is adjustable, depending on the analysis grid spacing. For the assimilation intervalΔt_a, the superobservations are processed to be available at every Δt_a using the radar data spanning within a period $\pm\ \Delta t_a/2$.

The observation operator for simulating Z_h is in accordance with the Goddard cumulus ensemble (GCE) microphysics scheme (Tao et al. 2003), in which rainwater, snow, and graupel are the hydrometeor species observable by radar. In comparison with TYL14, which considered only the reflectivity factor from rainwater, the current operator incorporates the contributions of snow and graupel considers cold-rain processes (Dowell et al. 2011) and the ice-phase hydrometeors above the bright band can be better simulated. Thus, it is feasible to assimilate the reflectivity data

above the melting layer (Gao and Stensrud 2012). The total reflectivity (Z_h) is calculated as ten times the logarithm to base 10 of the sum of the reflectivity factors from rainwater (Z_r), snow (Z_s) and graupel (Z_g):

$$Z_h = 10 \log_{10}(Z_r + Z_s + Z_g), \tag{1}$$

where the individual reflectivity factors can be derived under the single-moment assumptions of the GCE microphysics scheme as:

$$Z_r = 3.63 \times 10^9 (\rho_a q_r)^{1.75}, \tag{2}$$

$$Z_g = 1.12 \times 10^9 (\rho_a q_g)^{1.75}, \tag{3}$$

$$Z_s = 2.79 \times 10^8 (\rho_a q_s)^{1.75}, \quad when \ T \leq 0°C. \tag{4a}$$

$$Z_s = 1.21 \times 10^{11} (\rho_a q_s)^{1.75}, \quad when \ T > 0°C. \tag{4b}$$

In Eqs. (2)–(4), ρ_a is the air density (kg m^{-3}), and q_r, q_s, and q_g is the mixing ratios (kg k g^{-1}) of rainwater, graupel, and snow from the model, respectively.

The radial velocity is calculated by summing the projections of the wind and estimated terminal velocity from the model onto the radar beam:

$$V_r = \left[ux + uy + (w - v_t)z\right](x^2 + y^2 + z^2)^{-\frac{1}{2}}, \tag{5}$$

where (u, v, w) are the zonal, meridional and vertical components of the wind, $v_t$is the estimated terminal velocity, and (x, y, z) is the position based on the Cartesian coordinates with the radar site as the origin. V_t(ms^{-1}) is the terminal velocity calculated by $v_t = 5.4(p_0/\bar{p})^{0.4}(\rho_a q_r)^{0.125}$, where p_0 is the surface pressure (Pa), $\bar{p}$ is the base-state pressure (Pa). The observation localization (R-localization, Hunt et al. 2007) is adopted by multiplying the observation error variance with a distance-dependent weighting function w (Eq. 6):

$$w = \exp\left(d_h^2/\sigma_h^2 + d_v^2/\sigma_v^2\right). \tag{6}$$

Here, d and σ are the distance (km) between the observation and the model analysis grid point, and the localization scale (km), respectively. Subscripts h and v denote the horizontal and vertical direction, respectively. Observations located farther than the cut-off scale (triple the value of σ_h or σ_v) from each model grid point will not be used to update the model state.

WLRAS has options to use different covariance inflation strategies, including multiplicative inflation and relaxation to prior spread (RTPS, Whitaker et al. 2012). WLRAS is performed with the rapid-update cycle with an assimilation interval of 15 min with the consideration to capture the error growth of the convective-scale systems. The assimilation interval is chosen to ensure dispersiveness and flow-dependency of the ensemble.

With an ensemble size of 40 members and a typhoon-related heavy rainfall event, TYL14 demonstrated the flow-dependency of the background error covariance in WLRAS in relating to the development of rain band (Fig. 10 in TYL14). WLRAS has been evaluated for rainfall prediction associated with typhoons and Meiyu cases and has demonstrated a useful nowcasting skill (Shao 2015; Tsai et al. 2016; Cheng et al. 2019, 2020; Yang et al. 2020).

3 Impact of Assimilating the Radar Data on Short-Term Precipitation Prediction

a. Direct impact of the radar data assimilation

In this subsection, we demonstrate the impact of assimilating radar data on precipitation prediction over Taiwan and discuss how the assimilation of radial velocity and reflectivity leads to changes in simulating the location and intensity of heavy rainfall. The illustration case is a heavy rainfall event associated with a mesoscale convective system (MCS) that developed during the prefrontal condition on 16 June 2008 (IOP#8 during SoMEX). This case is characterized by its long duration and heavy precipitation near the coast of southwestern Taiwan. This event is a classical illustration of the continuous development of strong convections under interactions between the synoptic flow, MCS, and topography. Observation (Fig. 2a), provided by Wind Synthesis System using Doppler Measurement (WISSDOM, Liou et al. 2012), shows that the southwesterly wind prevails offshore southwestern Taiwan and the wind turns southerly at the slope due to orographic blocking effects. A convergence zone (shading) is established from offshore to the coast. In this area, reflectivity analysis from Quantitative Precipitation Estimation and Segregation Using Multiple Sensor (QPESUMS) from the Central Weather Bureau (CWB) exhibits a west-oriented intense reflectivity band (Fig. 3a). Correspondingly, a near-surface cold pool induced from evaporation cooling is established and further enhanced by the land breeze (Tu et al. 2014). The convergence zone at the coast is sustained by the prevailing southwesterly offshore and outflow of the cold pool at the coast, and is the key for the long duration of heavy rainfall (Xu et al. 2012).

The ensemble for the following experiments was initialized at 1800 UTC 15 June 2008 by downscaling the 36 analysis ensemble members derived from regional ensemble data assimilation (Yang et al. 2014). The ensemble is spun-up for 6 hours to derive the convective-scale perturbations with a triply nesting domain, and then,

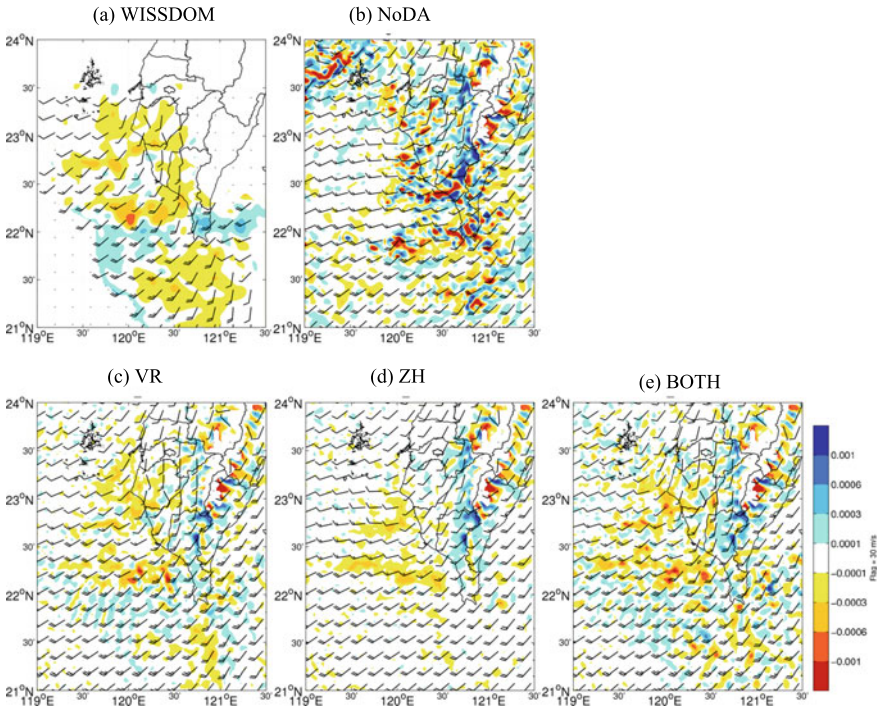

Fig. 2 Wind at the level of 2 km from **a** WISSDOM, **b** the NoDA forecast, the **c** VR, **d** ZH, and **e** BOTH analyses at 0200 UTC (last analysis cycle) 16 June 2008. Color shading indicates the convergence

the data from the RCCG and RCKT radars were assimilated for the inner domain (3 km) every 15 minutes from 0000 UTC to 0200 UTC 16 June 2008. The horizontal and vertical localization scales are 12 and 4 km, respectively, as suggested by TYL14 with 40 members. The VR, ZH, and BOTH experiments assimilate radial velocity, reflectivity, and both data, respectively. Deterministic forecasts initialized from the analysis mean at 0200 UTC are used to compare the impact of radar data assimilation. A baseline taken from the forecast initialized from the ensemble mean at 1800 UTC 15 June 2008, before performing the radar data assimilation, is used to identify the impact of radar data assimilation on the forecast skill. This experiment is referred to as NoDA.

Without assimilating radar data, a westerly dominates offshore of southwestern Taiwan (Fig. 2b). The simulated reflectivity is stronger but mostly appears over the coast due to the strong westerly impinging Taiwan. Assimilating the radial velocity corrects the wind field effectively, including the southwesterly offshore and southerlies at the coast. Thus, the VR and BOTH wind analyses can exhibit the convergence (Fig. 2c, e) shown in the observation. Although positive impacts are also derived from assimilating reflectivity, the convergence is weaker (Fig. 2d). When the reflectivity data is assimilated, ZH and BOTH (Fig. 3d, e) can establish the west-oriented

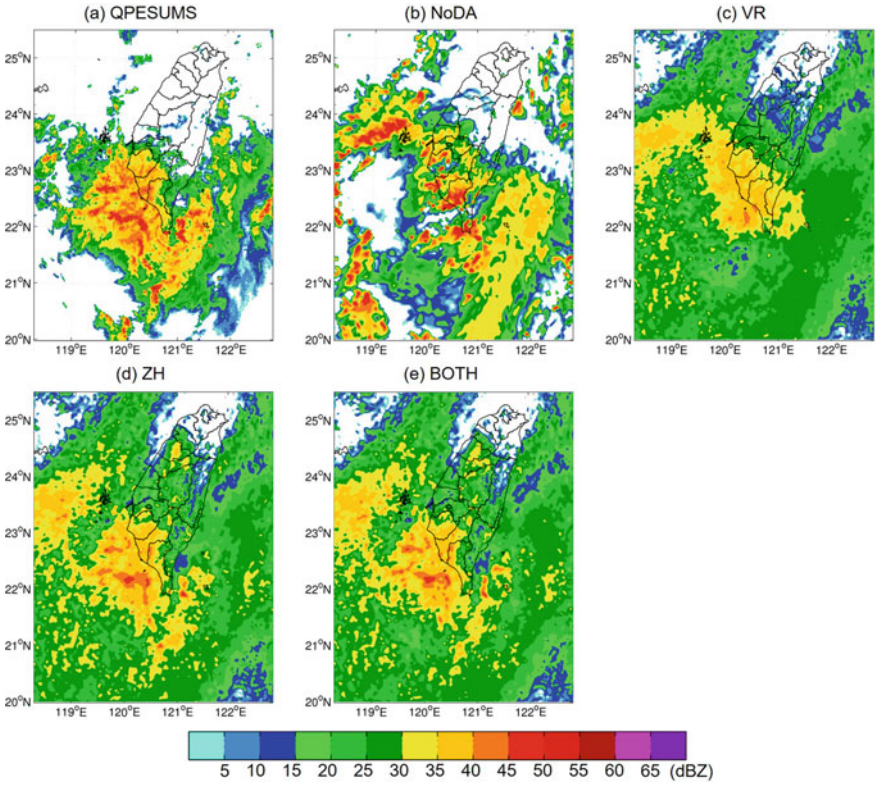

Fig. 3 Reflectivity analysis from **a** QPESUMS, **b** the NoDA forecast, the **b** VR, **c** ZH, and **d** BOTH analyses at 0200 UTC 16 June 2008

intense reflectivity band offshore, while VR does not have such a feature (Fig. 3c). The assimilation of reflectivity provides greater adjustment to the thermodynamic fields, including great moisture enhancement offshore of southwestern Taiwan (Fig. 4c, d vs. Fig. 4a, b), and a stronger and broadly extended cold pool near the surface (Fig. 5c, d vs. Fig. 5a, b), in comparison with NoDA and VR. The BOTH analysis, that assimilates both radial velocity and reflectivity, can well capture the dynamical and thermodynamic conditions associated with the orography effect.

The differences in the initial conditions lead to great differences in precipitation prediction in space and time. The location of the heavy rainfall maximum in NoDA (Fig. 6b) is south of that shown in the observation due to the lack of a southerly component near the coast, and the heavy rainfall over southwestern Taiwan did not last long (Fig. 7a). With more rainwater and moisture than those in VR, both the ZH and BOTH experiments exhibit heavy rainfall over southwestern Taiwan (Fig. 6d, e). Although the ZH experiment exhibits a reasonable precipitation prediction pattern, most of the rain takes place during the first three hours, and the rainfall intensity becomes very weak afterward. Such a result indicates that the rainwater brought

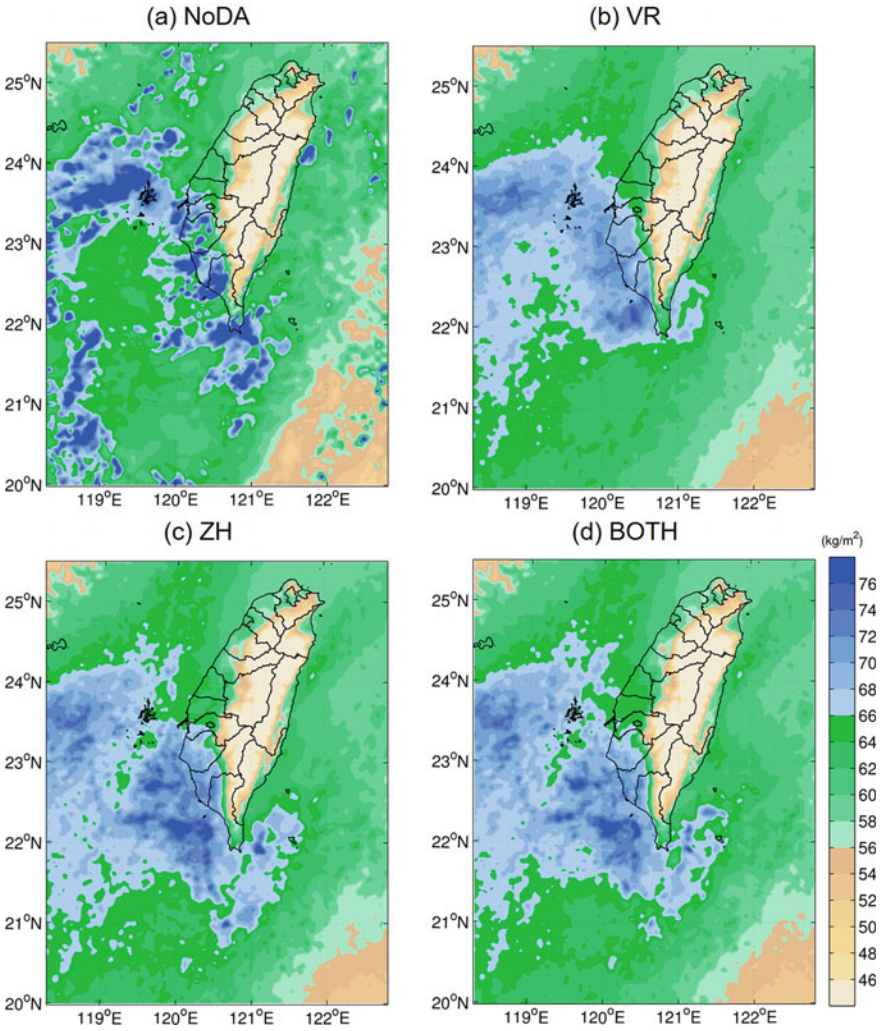

Fig. 4 Total precipitable water from **a** the NoDA forecast, the **b** VR, **c** ZH, and **d** BOTH analyses at 0200 UTC 16 June 2008

in through the assimilation of reflectivity precipitates quickly, since the dynamical field (wind) is not adjusted correspondingly to support the mass field. VR has the lowest skill with the least amount of rainfall over southwestern Taiwan, and the heavy rainfall is located offshore instead of moving onshore (Fig. 6c). The hourly rainfall rate does pick up some signal at 1500 LST (0700 UTC), but the intensity is too weak. The later response with VR has been pointed out in TYL14, which suggests that the response in the very short-term rainfall forecast to the adjustment of winds is slower than the adjustment of hydrometeors. It is evident that BOTH has

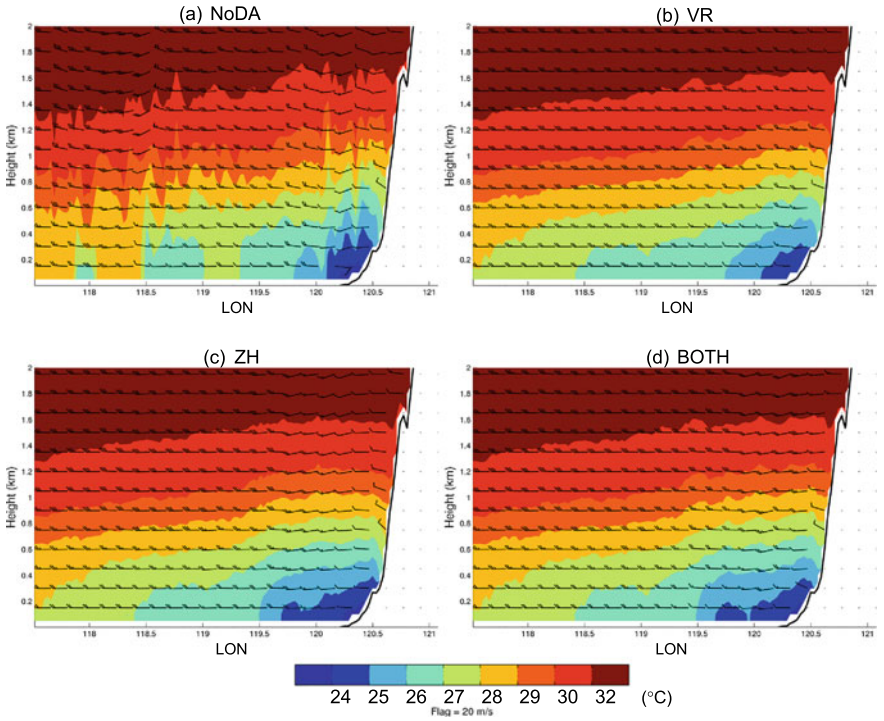

Fig. 5 Vertical cross-section of potential temperature at 0200 UTC 16 June 2008 from **a** the NoDA forecast and the **b** VR, **c** ZH, and **d** BOTH analyses

the best forecast performance in terms of pattern and intensity. It is also the only experiment that can simulate the rainfall at the coast. The convergence zone in the BOTH forecast is the best maintained since the wind offshore of southwestern Taiwan can be corrected by assimilating V_r, and the outflow from the cold pool is enhanced by assimilating Z_h. The improvements in the dynamical and thermodynamic fields are essential to capture the pattern and variation of heavy rainfall over the terrain. Given that BOTH agrees very well with the observation during the first 7 hours, it could provide very useful information for disaster prevention and decision-making regarding the location of heavy rainfall and the corresponding temporal variations.

Figures 6 and 7a demonstrate that the assimilation of radar data can be very beneficial for very short-term rainfall prediction, even only with either radial velocity or reflectivity. However, to capture the realistic behavior, the wind and mass adjustments need to be complementary to each other, providing proper dynamical conditions to support the changes in moisture and hydrometeor fields over complex terrain. For such type of a long-lived precipitation event, the moisture supply plays an important role in the rainfall intensity and duration. If the moisture field is not updated during the analysis step and the change in moisture is merely through model integration,

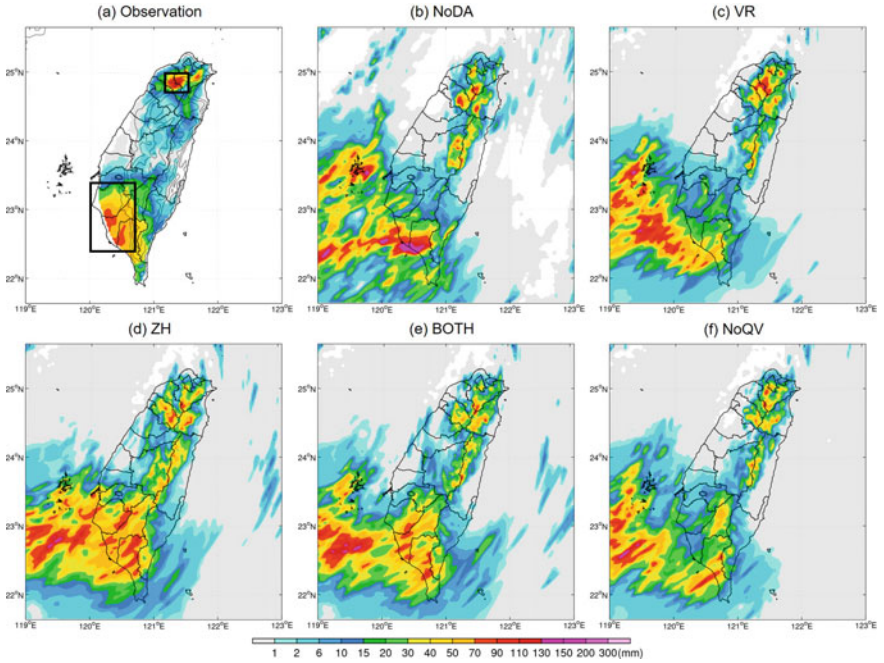

Fig. 6 Six-hour rainfall accumulation from 0200 to 0800 UTC 16 June 2008 from **a** observation (rain gauge) and the **b** NoDA, **c** VR, **d** ZH, **e** BOTH and **f** NoQV forecasts

the precipitation prediction ends up losing the heavy rainfall in southwestern Taiwan (Fig. 6f), even if Z_h is assimilated.

Although the impact of radar data could last for seven hours, forecast errors of all the experiments are eventually dominated by errors in the environment of the convection. All the forecasts have too-strong westerlies over the coast of southwestern Taiwan and excessively dry conditions there at later lead times. Without the required moisture transport, the amount of precipitation after 6 hours reduces rapidly.

b. Impact of radar data assimilation on the onset of the afternoon thunderstorm

In addition to the direct impact on heavy rainfall prediction in Sect. 3 a, assimilating radar data can affect the prediction of the onset of convection downstream the observed area. On the same day, a short-lived afternoon thunderstorm developed under an unstable thermodynamic condition over northern Taiwan, downstream of the prevailing southwesterly (Fig. 6a). This subsection appeared in an earlier form in Cheng et al. (2020), which investigated the sensitivity of predicting this short-lived heavy rainfall event with the assimilation of radar data in the upstream conditions based on the same experiments in Sect. 3 a. Although the area of interest is within the observing range of RCWF, there is no radar data available in this area during the initiation of the afternoon thunderstorm. Nevertheless, assimilating the radar data

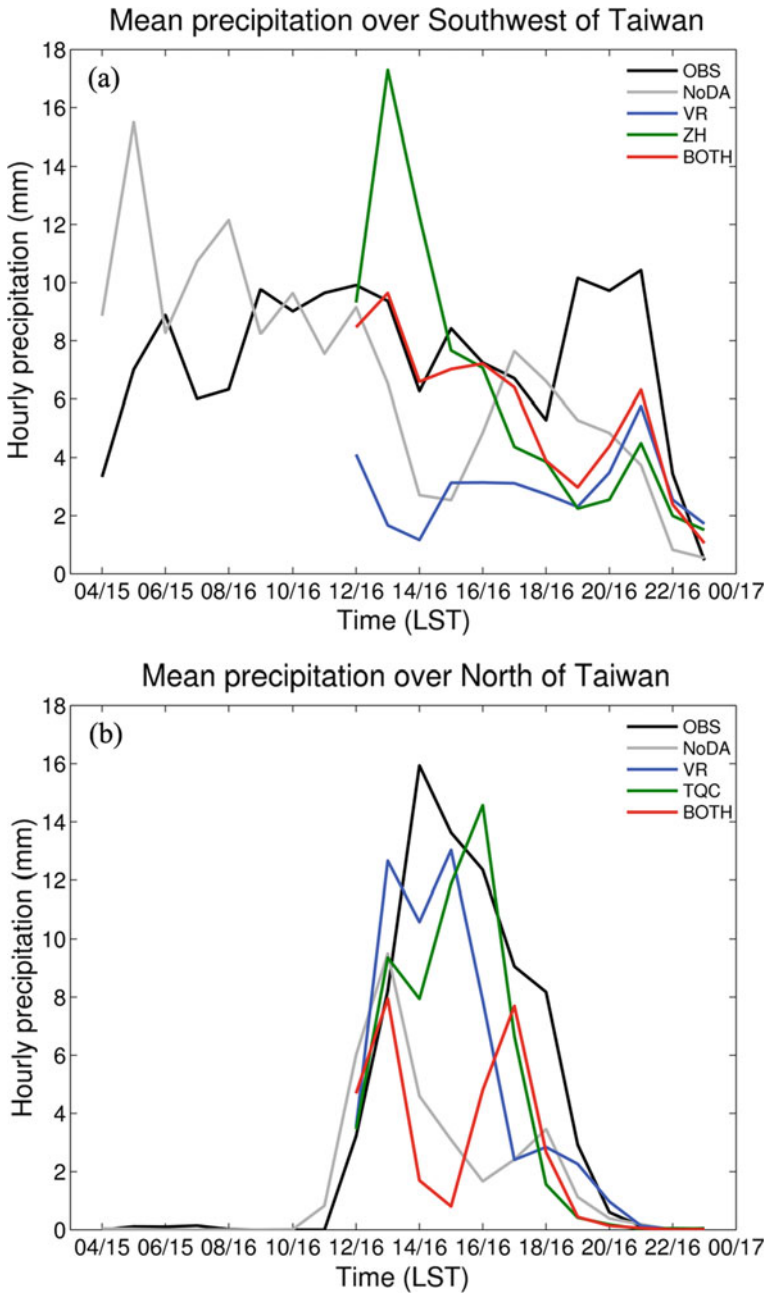

Fig. 7 Time series of the hourly rainfall (mm/h) averaged over the area in **a** southwestern and **b** northern Taiwan (the black boxes in Fig. 6 a)

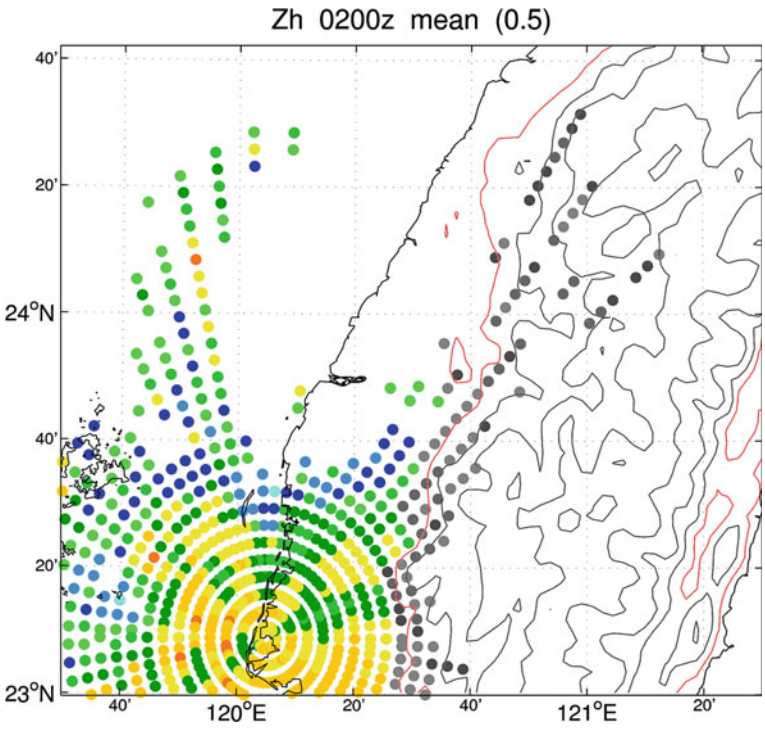

Fig. 8 Superobbed reflectivity (dBZ) at 1000 LST (0200 UTC)16 June 2008. Contours have the same definition as in Fig. 6a. The red contour is used as the threshold of terrain height to remove terrain-related noise

upstream (Fig. 8) does have a significant impact on precipitation prediction (e.g., Fig. 6b, c).

Regarding the rainfall behavior in temporal evolution and spatial distribution, the NoDA forecast has some skill in predicting this event but is less satisfactory. Even though RCCG and RCKT cannot directly observe the convective storms over northern Taiwan, the intensity, location, and temporal variation of the heavy rainfall in this area are significantly improved (Figs. 6c, and 7b) after the model wind field is updated by assimilating the radial velocity upstream. Compared to NoDA, there is no rainfall simulated over the target area in the VR forecast at 1100 LST (Fig. 9(c-1)), and the onshore flow over the upstream area is enhanced with higher T_v (Fig. 10a vs. Fig. 10e). These differences in VR contribute greatly to better predict the formation and development of the afternoon thunderstorm in the downstream, and the location of the rainfall maximum in VR is closer to the observations. In contrast, the BOTH forecast significantly underestimates the amount of precipitation over northern Taiwan and is even weaker than that of the NoDA forecast (Figs. 7b, 9(b-3) vs. 9(d-3)). The main improvement of the precipitation prediction for this afternoon

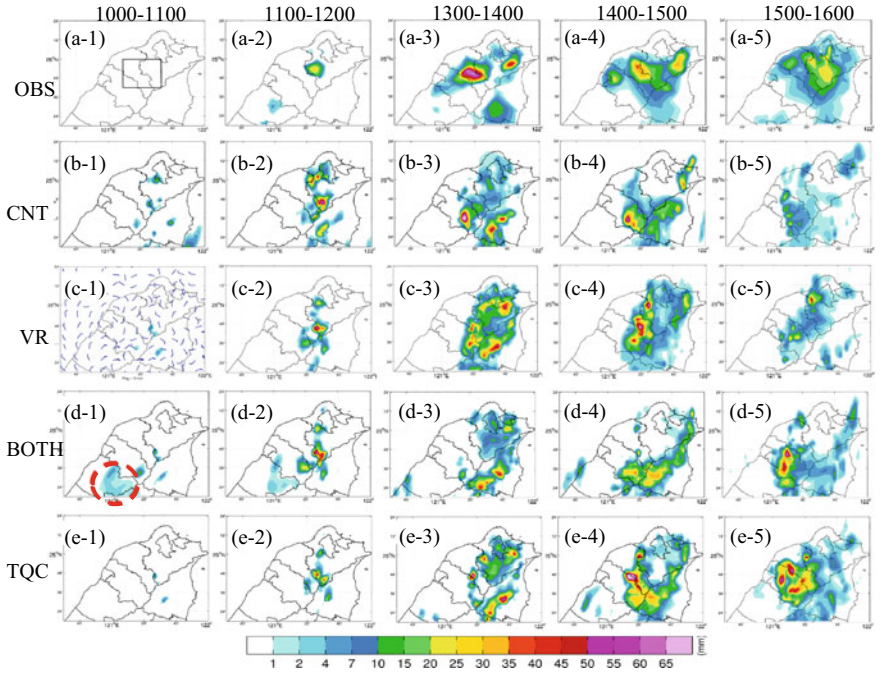

Fig. 9 Hourly rainfall from (1) 1000, (2) 1100, (3) 1300, (4) 1400, and (5) 1500 LST 16 June 2008. From top to bottom rows are observations based on the rain gauge data, the CNTL, VR, BOTH, and TQC forecasts, respectively. The black box indicates the target area, and the red circle in (d-1) highlights the light rain in Miaoli county. This figure is reproduced from Fig. 5 in Cheng et al. (2020)

storm event appears due to the upstream wind adjustment from assimilating the radial velocity, while degradation is obtained from assimilating the reflectivity.

The poorer representation of afternoon thunderstorm in BOTH is attributed to the excessively and unrealistic early precipitation over Miaoli County (red circle in Fig. 9(d-1)) during 1000–1100 LST. The associated evaporation causes a weak cold pool to form near the surface (Fig. 10j) and move northward following the environmental flow. As shown in Fig. 10k, T_v in the BOTH simulation is 3 degrees less over the coastal area in northwestern Taiwan at 1200 LST than in the other simulations. Being located upstream of the afternoon thunderstorm, such a cold pool hinders the high T_v air flowing into northern Taiwan, which leads to a less favorable condition for the development of afternoon convection. This implies that the reflectivity over this area is contaminated by the non-meteorological related signals, but this cannot be filtered out completely by the quality check (QC) criterion, in which the data is recognized as topography-related noise if the wind speed is low and reflectivity is high.

An experiment, named as TQC, that applies an additional terrain-related QC to remove the near-surface reflectivity over the mountainous area (indicated by the

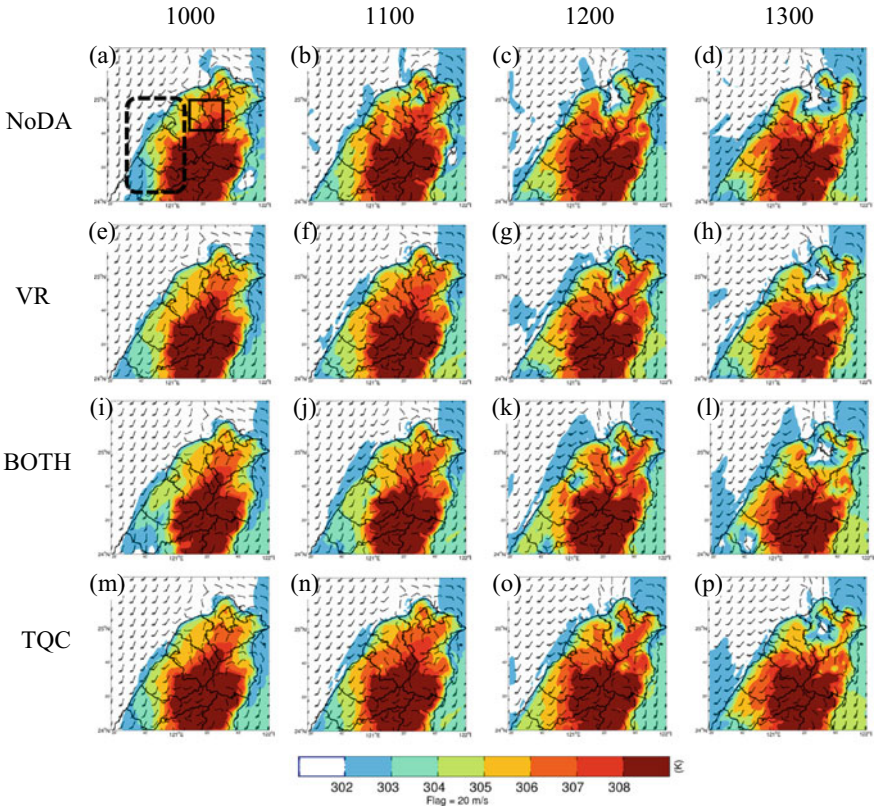

Fig. 10 Virtual potential temperature (K) and wind (ms^{-1}) at the lowest model level from 1000 LST (0200 UTC) to 1300 LST (0500 UTC). From top to bottom rows are results from the NoDA, VR, BOTH, and TQC forecasts, respectively. In (a), the solid box indicates the target area (same as in Fig. 6a), and the dashed rectangle indicates the upstream area. This figure is reproduced from Fig. 4 in Cheng et al. (2020)

gray dots in Fig. 8) shows a great difference in precipitation. Among all experiments using the radar data, TQC is the only experiment that resulted in a very intense precipitation event during 1500–1700 LST and its hourly precipitation rate is the most intense (>14.5 mm h^{-1}), with a longer duration similar to the observations. Therefore, by removing the near-surface reflectivity over the mountainous area, TQC avoids exhibiting the unrealistic early precipitation during 1000–1100 LST, and there is no anomalous cold pool formed near the surface (Fig. 10 i vs. 10 m). Sufficient warm and moist southwesterly flows into northern Taiwan to sustain the development of convective storms (Fig. 9(e-5)).

The accuracy of reflectivity data not only affects the hydrometer fields, but also the thermodynamic fields, which can affect predicting the onset and development of thunderstorms downstream the observing area. It is critical to consider the effect of

the complex terrain with a QC algorithm more sophisticated than using a criterion of low wind speed and high reflectivity.

4 Limitations in WLRAS

As mentioned in the introduction, the performance of WLRAS can be affected by the assimilation setup, such as the ensemble size or localization. Especially, sampling error is a major factor limiting the accuracy of the analysis. This section is an excerpt from Wu et al. (2020) and reviews the characteristics, attributions and impact of the sampling errors induced by using an insufficient ensemble size with WLRAS.

a. Characteristics of sampling errors

In EnRDA, model state variables can be updated directly by radar data through the flow-dependent background error covariance (BECV) estimated by the short-term ensemble forecast. Sampling errors affect the accuracy of the ensemble-estimated BECV, but such impacts are different for the BECV between the observable and unobservable model states in the radar data-based convective-scale EDA. Also, sampling errors can have a larger impact when the error correlation is weak (Houtekamer and Mitchell 1998). Based on the same heavy rainfall event discussed in Sect. 3 a, Wu et al. (2020) identify the attribution of sampling errors in WLRAS by comparing the results of the experiments with small (40 members) and large (256 members) ensembles. The setups of the assimilation experiments are the same as the ones used in Sect. 3, except that the initial and boundary conditions of the ensemble are taken from the NCEP 1° × 1° FNL (Final) Operational Global Analysis data and perturbed according to the background error covariance constructed for the WRF-3DVAR system.

The sensitivity of background error correlation (BECR) to sampling error is quantified by the resemblance between the point BECR derived from the two ensemble sets based on the Standardized Mean Absolute Difference (SMAD) defined as

$$SMAD = \frac{\sum_{i=1}^{N} \left| CORR_i^L - CORR_i^s \right|}{N \times CORR_M}. \tag{7}$$

The SMAD is computed in a 72 × 72 km^2 area centered at the observation point used to calculate the point BECR. In Eq. (7), superscripts L and S denote the large and small ensembles, respectively. N is the total model grid numbers in the calculation area. $CORR_i$ is the point BECR between observation variables at the observation location and model states at a model grid point *i*. $CORR_M$ is the maximum value of absolute BECR with either ensemble set (i.e., $Max(|CORR^L{}_i (i = 1, \ldots, N)|, |CORR^s{}_i (i = 1, \ldots, N)|)$ in the area. A larger SMAD indicates that the BECR derived from the small ensemble shows less similarity to that derived from the large ensemble, i.e., it is more sensitive to sampling error. Figure 11 shows the SMADs computed at different observation points between observation variables (V_r orZ_h) and model

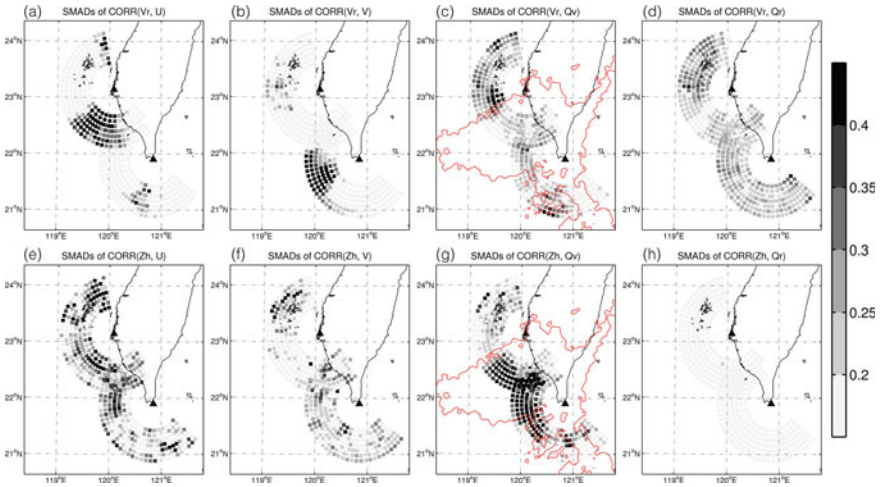

Fig. 11 The SMADs (colored points) of **a** CORR(V_r, U), **b** CORR(V_r, V), **c** CORR(V_r, q_v), **d** CORR(V_r, q_r), **e** CORR(Z_h, U), **f** CORR(Z_h, V), **g** CORR(Z_h, q_v), and **h** CORR(Z_h, q_r) calculated with super-obs points on the first PPI elevation angle (0.5°). The triangles show the location of the RCCG and RCKT radar sites. The red contours in (c) and (g) indicate the 40 dBZ composite reflectivity of PM ensemble mean derived from the small ensemble at 0000 UTC 16 June 2008. Adopted from Fig. 7 in Wu et al. (2020)

variables. The SMAD pattern can be used to illustrate the hot spots of sampling errors associated with this event.

The sampling errors associated with this event are sensitive to the relationships between the simulated observations and model variables. This relationship is divided into direct and indirect depending on whether the information of the model variable will be used in the observation operator. With a direct relationship, CORR($Z_h q_r$) exhibits small SMADs at all observation points while the SMADs of CORR(V_r,q_r) are generally larger (Fig. 11d vs. Fig. 11h). The relationship between V_r and the amount of q_r can be influenced by multiple processes, including but not limited to horizontal/vertical advection, convergence, and condensation. Therefore, sampling errors in these processes can lead to large SMADs in CORR(V_r,q_r). Given the direct relationship between $V_r$and horizontal winds, the SMADs of CORR (V_r, U) and CORR(V_r, V) are generally small; however, they are the only two in Fig. 11 that have much larger values at some specific azimuths of radar beams (Fig. 11a, b). Such a characteristic is related to the fact that radar measurements cannot resolve the wind components tangential to radar beams. When the prevailing southwesterly wind arrives toward the south of RCCG (offshore of southwestern Taiwan), only the southerly component contributes to the radial wind from the RCCG radar. Thus the radial wind has little correlation with the westerly component. Therefore, CORR(V_r, U) over this region can be contaminated easily by sampling error. Similarly, the southerly component does not contribute to the radial velocity west of RCKT, leading to a large SMAD of CORR(V_r, V) in that region. In contrast, the

SMADs of CORR(V_r, V) south of RCKT are smaller since the prevailing wind in this region is predominantly southerly and projects strongly to the radial wind.

Unlike the SMADS for variables with direct relationships, the moisture field has an indirect relationship with either V_r or $Z_h$and their SMADs show more complicated patterns that are sensitive to the intensity of reflectivity. The SMADs of CORR(Z_h,q_v) are generally larger over the intense reflectivity area (enclosed by the red contour in Fig. 11g), while the SMADs of CORR(V_r,q_v) are larger over the area with weaker reflectivity (Fig. 11c). When dynamic forcing dominates, such as moisture transport and convergence in the intense reflectivity area, the wind direction is more consistent among ensemble members, and CORR(V_r,q_v) can be robust and easier to estimate. In comparison, the wind field can be impacted by multiple factors outside the convective cells, such as prevailing wind, offshore flow, and topography; thus, the uncertainty of the wind direction is large there. This makes the estimation of CORR(V_r,q_v) vulnerable to sampling error and leads to the large SMADs of CORR(V_r,q_v) in the area outside of intense reflectivity. On the other hand, q_v and Z_h are implicitly related through the nonlinear phase transition in the microphysics during convective development. The phase transition from q_v to hydrometeors (i.e., Z_h) is related to the dynamical (e.g., vertical motion) and thermodynamical (e.g., temperature) conditions, which have large uncertainties in the region of intense reflectivity. These results imply that if only Z_h is assimilated, sampling error tends to introduce erroneous corrections on the moisture field in the intense reflectivity area, while such a detrimental effect will take place in the weak reflectivity area when only V_r is assimilated. In other words, correcting the moisture field is restrictive by assimilating V_r and Z_h.

b. Impact of covariance localization

Covariance localization (Hamill et al. 2001) is a common strategy applied to remedy the issues associated with sampling error in the EDA algorithms. However, improperly chosen localization scales can degrade the forecast performance. How such an issue affects precipitation prediction is highlighted by the following experiments with different horizontal and vertical localization scales. The number in the experiment name indicates the horizontal and vertical localization scales, and the first letter indicates the small or large ensemble. For instance, S1204 is the experiment in which WLRAS is performed with 40 ensemble members and with the horizontal and vertical localization scale equal to 12 and 4 km, respectively.

Among the experiments with a small ensemble size, only S1204 has a root mean square innovation (RMSI) comparable to the experiments with the large ensemble; this justifies the choice of horizontal localization in TYL14. With large localization, S3612 exhibits the largest RMSI in V_r and Z_h at later cycles. L3612 exhibits the smallest RMSI inZ_h, indicating that larger localization length scale is beneficial when a large enough ensemble is used to better represent the flow-dependent background error covariance.

The differences among these experiments lead to different skills in precipitation prediction (Fig. 12). Both L1204 and S1204 capture the heavy rainfall pattern over

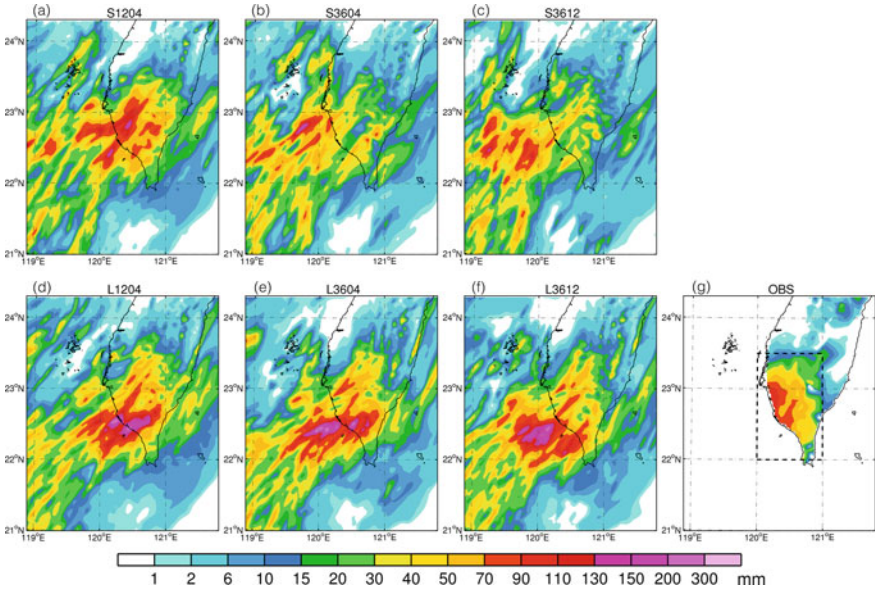

Fig. 12 Accumulated rainfall from 0200 to 0700 UTC of the experiments **a** S1204, **b** S3604, **c** S3612, **d** L1204, **e** L3604, and **f** L3612. **g** The accumulated rainfall from 0200 to 0900 UTC from rain gauge observations interpolated to model grids. The dashed box in **g** shows the area used to calculate verification scores. Adopted from Fig. 11 in Wu et al. (2020)

the coast of southwestern Taiwan, except that the rainfall maximum of L1204 is more intense and extends further south. When an inadequately long horizontal localization is used, sampling error degrades the rainfall prediction performance with a small ensemble size, as S3604 shows a substantial reduction in rainfall intensity compared to S1204. Such precipitation discrepancies can be attributed to the differences of the moisture field and hydrometeors between the analysis means of L3604 and S3604 after a 2-h assimilation (Fig. 13a, c); the differences are much larger than those between the analysis means of L1204 and S1204 (Fig. 13b, d). Therefore, sampling errors can contaminate the distant BECV and adversely affects the moisture and hydrometeor analysis.

Although L3604 has a better forecast skill than S3604 in terms of bias, it has a wet bias at large thresholds (70 mm (5 h)$^{-1}$). Being initialized from the analysis adopting a larger vertical localization scale, the L3612 forecast reduces the overestimation of rainfall intensity over the coastal area of southern Taiwan exhibited in the L3604 forecast (Fig. 12f vs. Fig. 12e). Thus, the choice of the vertical localization affects the vertical adjustment and modifies the vertical instability of the model state. Figure 14 shows the cross-section of the BECV between the simulated Z_h and the thermodynamically-related model variables (θ orq_v) derived from the large ensemble at 0000 UTC. It is evident that there is a strong covariance between the lower and higher levels for both the moisture and temperature fields. Such characteristics reflect

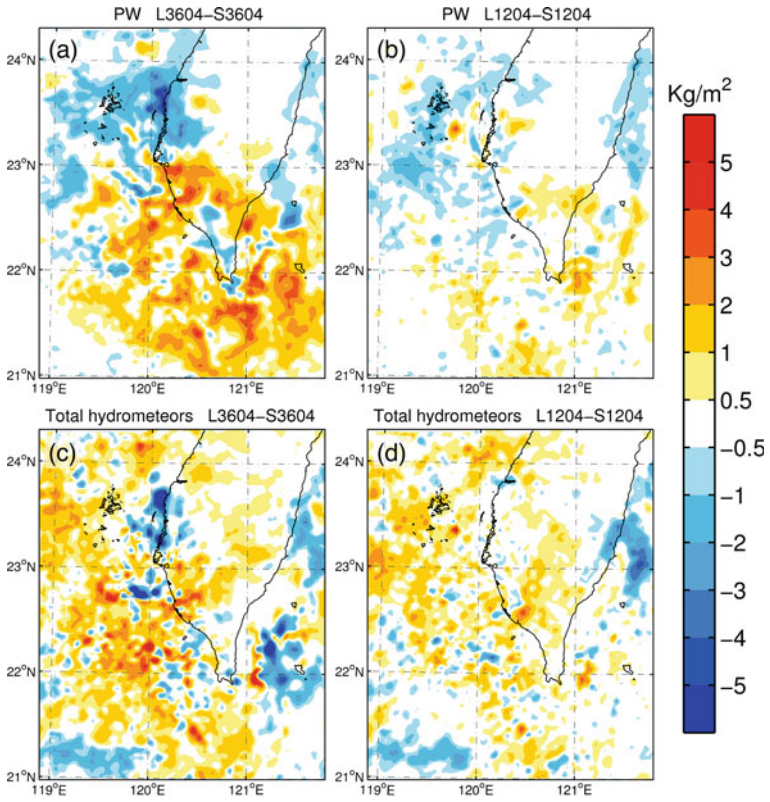

Fig. 13 The difference of **a**, **b** precipitable water and **c**, **d** total hydrometeors between analysis mean of **a**, **c** L3604 and S3604, and **b**, **d** L1204 and S1204 at 0200 UTC 16 June 2008. Adopted from Fig. 12 in Wu et al. (2020)

the vertical scale of convection cells. With a shorter vertical localization (4 km), the correction becomes more locally limited than that for 12 km (as indicated by dashed boxes in Fig. 14), unless the observations are sufficiently dense to capture the vertical structure. Figure 14 also implies that different vertical localizations result in different vertical adjustments, leading to different thermodynamic conditions. Compared to L3604, L3612 has a smaller buoyancy, given that it is cooler and drier below 5 km, but is warmer and moister above 5 km. Thus, the atmospheric condition in L3612 is thermodynamically less unstable than that of L3604, alleviating the excessive rainfall shown in L3604. These results also imply that using a longer vertical localization in convective-scale data assimilation can better address vertical adjustment for convective-scale thermodynamic structures. This proper adjustment in the vertical dimension is essential for correctly representing the strength of the convections and thus precipitation intensity.

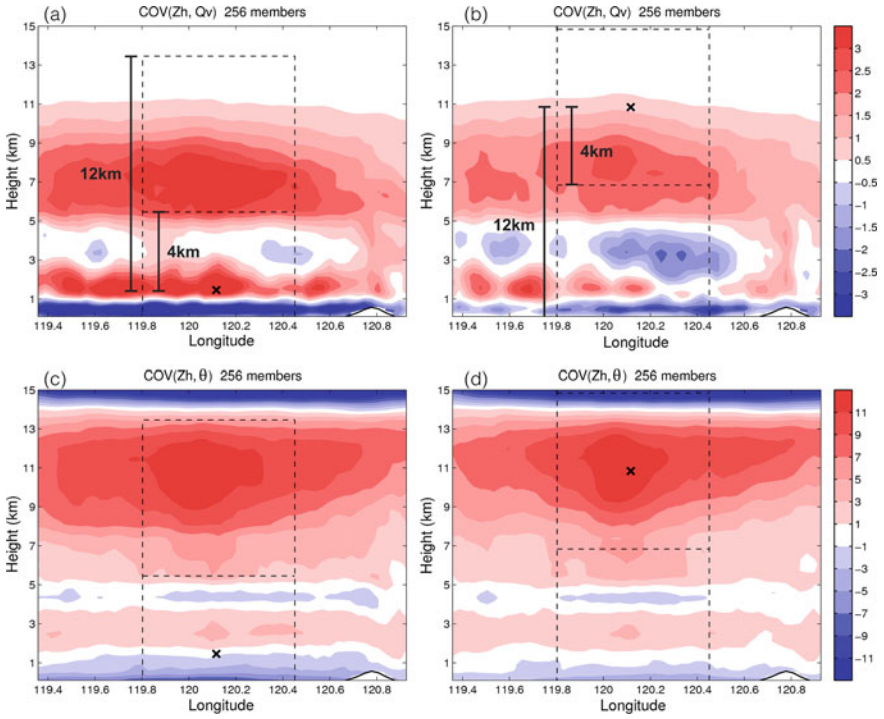

Fig. 14 The background error covariance between Z_h at an observation point (indicated by black x symbol) and model states **a**, **b** q_v, and **c**, **d** θ over a vertical cross-Sect. (22.25 N°) derived from the large ensemble at 0000 UTC 10 June 2008. The observation points are located at 22.25 N° with **a**, **c** 1458 m, and **b**, **d** 10,840 m height. The dashed boxes illustrate the localization range of 36 km for horizontal and 4 or 12 km for vertical. Adopted from Fig. 16 in Wu et al. (2020)

5 Improving Heavy Rainfall Prediction in Taiwan with WLRAS

a. Mitigating the impact of sampling error associated with non-observable tangential wind under the prevailing wind

The unique pattern of sampling errors associated with the projection of U and V wind in the radial direction (Fig. 11a, b) is related to the presence of strong prevailing wind (i.e., the southwesterly), which is a very common feature when heavy rainfall events are associated with fronts or monsoon flow. To alleviate this kind of detrimental impact, Wu et al. (2020) proposed a strategy to lessen the corrections at these specific locations by increasing the observation error of V_r. When updating the V wind with a limited ensemble size, the weighting of $V_r$observation error is tripled if the super-obs is located between azimuth 60° and 120° (east of the radar) or between 240° and 300° (west of the radar). Similarly, the V_r observation error is tripled in the south and

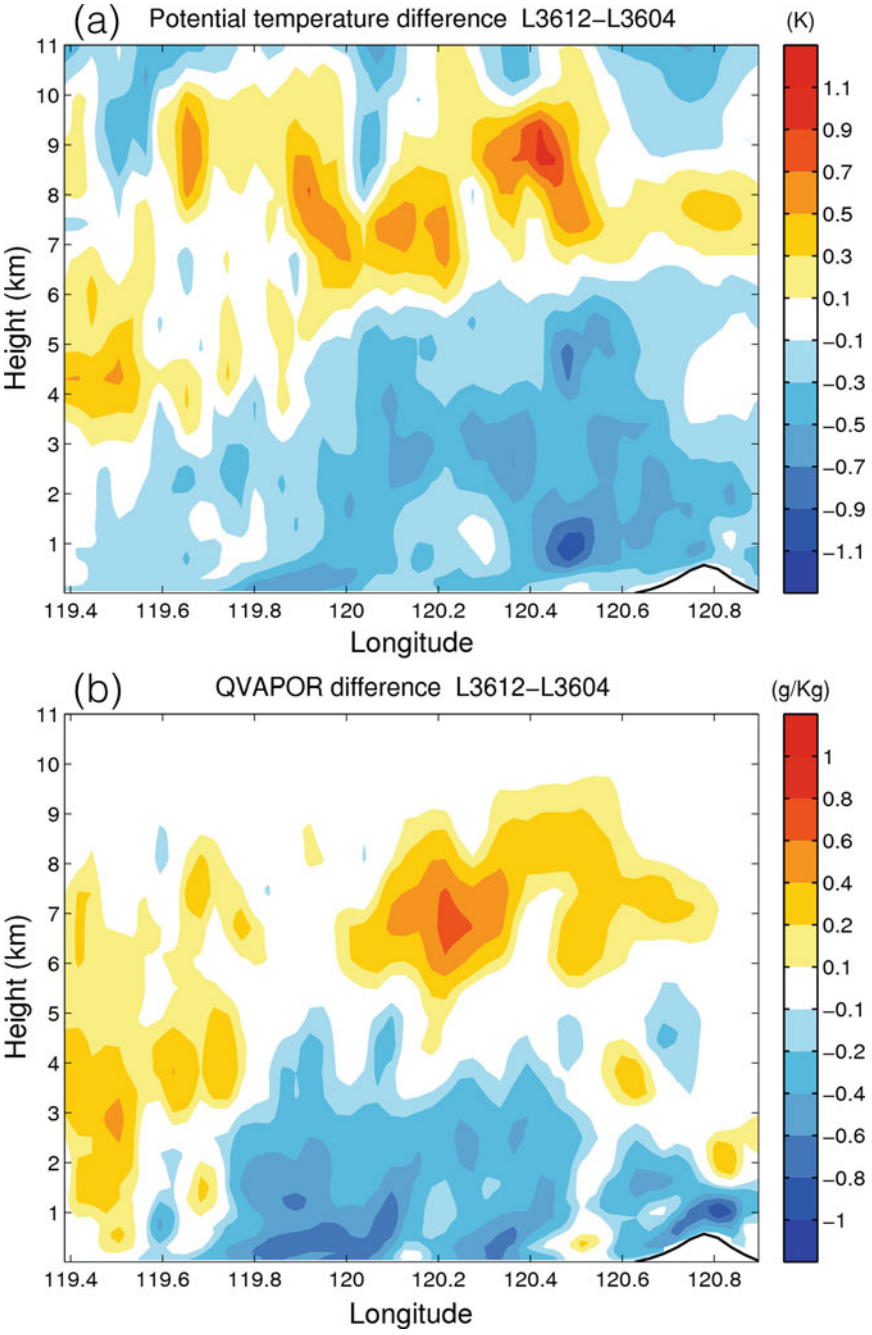

Fig. 15 The vertical cross-section over 22.25 °N of the difference in **a** θ and **d** q_v between analysis means of L3604 and L3612 at 0200 UTC 16 June 2008. Adopted from Fig. 17 in Wu et al. (2020)

north of the radar when updating the U wind. An experiment (named LimVr), using the same setting as S3604 but applying the strategy mentioned above, is conducted to evaluate its relevance.

The results show a substantial difference in the wind fields between the analysis means of the two experiments at 0200 UTC, especially for the zonal wind component (Fig. 16b). The westerly wind difference appears near the coast of southwestern Taiwan, where the intense rainfall was observed (Fig. 12g). By enhancing the westerly wind, LimVr has a larger moisture flux into Taiwan than S3604 (Fig. 16c). Consequently, the amount of precipitation over southwestern Taiwan becomes larger (Fig. 16a), especially over the coastal areas. The result of LimVr suggests that such

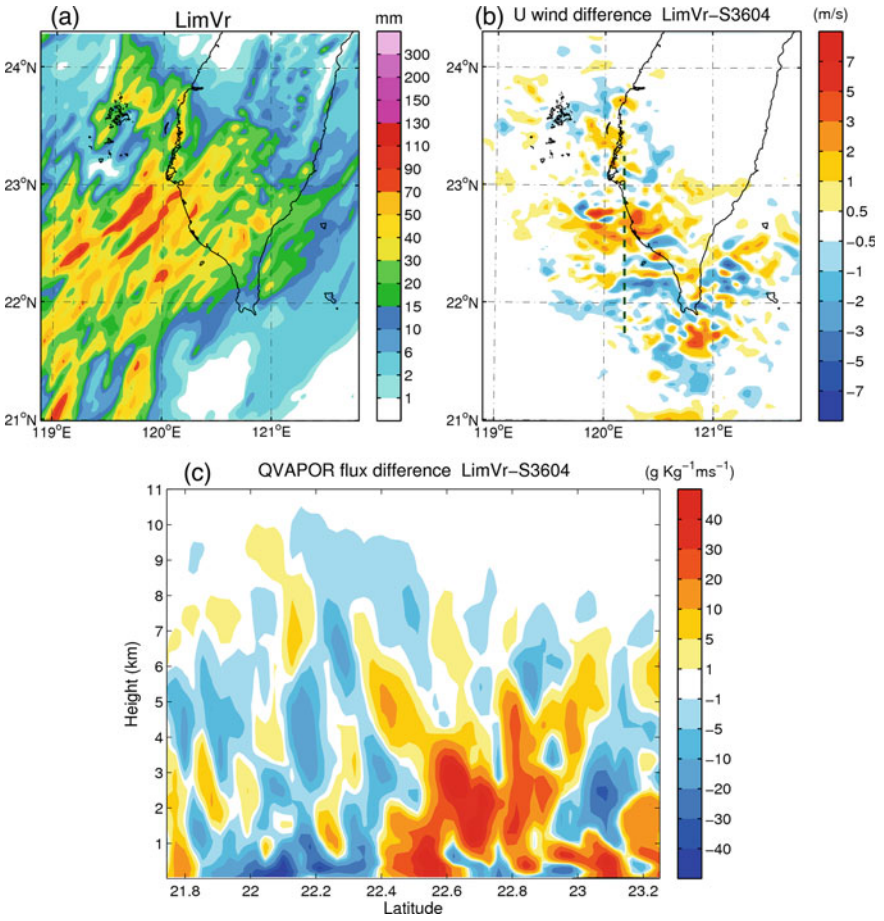

Fig. 16 A The accumulated rainfall from 0200 to 0700 UTC of the experiment LimVr. The difference of **b** U wind near 1-km height and **c** east-westerly q_v flux over the cross-section at 120.2°E between LimVr and S3604's analysis means at 0200 UTC 16 June 2008. The dashed line in (**b**) shows the location of the cross-section in (**c**). Adopted from Fig. 19 in Wu et al. (2020)

$V_r$observation error inflation method has the potential to deal with sampling errors associated with the limitations of radial wind measurements and bring benefits to precipitation prediction. The positive impact of applying this inflation strategy has also been identified for other rainfall events with the prevailing wind in the environment, such as a heavy rainfall episode associated with a stationary front hovering over northern Taiwan.

b. Additional impact from assimilating the GNSS-ZTD data

The accuracy of the moisture field is important for the initialization and development of convections and thus affects precipitation's occurrence, coverage, intensity, and duration. However, it is restrictive to update the moisture field in WLRAS. As discussed in Sect. 4 a, the correlations between the moisture and observed variables are less direct and can be contaminated by sampling errors easily. Besides, moisture correction at the early stage of convection development would be limited since the reflectivity and radial velocity from the conventional radars are available only when the rain starts. Assimilating the moisture observations additionally becomes the direct solution to tackle this issue in WLRAS. Among the remote sensing observing systems that provide moisture information, observations based on the global navigation satellite systems (GNSS) are becoming more important for providing reliable moisture information, such as spaceborne radio occultation (RO) (Ma et al. 2011; Yang et al. 2014; Chen et al. 2020) or ground-based zenith total delay (ZTD) (Bengtsson et al. 2003; Bauer et al. 2015). GNSS-based observations use the fact that the radio rays emitted by GNSS satellites are bent when passing through the atmosphere, causing signaling delays as the rays are received. The degree of bending depends on the change in the density of the atmosphere, which is primarily related to temperature and humidity conditions. One great advantage of GNSS observations is that radio rays are affected little by clouds and are thus sensitive to the changes in temperature and moisture in all weather conditions. A ground-based GNSS receiver measures the delay in the path in receiving a signal from a GNSS satellite (Bevis et al. 1992) and ZTD expresses this delay as the excess path length along the zenith direction. The ZTD data are available at a high temporal frequency (30 min to one hour), filling in the observation gaps between radiosondes and satellite measurements. Recent studies suggest that the ZTD data have an important role in providing valuable rapid moisture information for very short-term forecasts with rapid update analysis cycles (de Haan 2013) and quantitative precipitation estimation (Bauer et al. 2015).

The ZTD can be considered as comprising two components (ZTD = ZHD + ZWD): the zenith hydrostatic delay (ZHD) and zenith wet delay (ZWD). Although the ZWD contributes only less than 10% to the total delay, this component is an important indicator of atmospheric water vapor variations. The total precipitable water (TPW), related to the moisture, can be retrieved based on ZTD (Yeh et al. 2016). Previous studies have assimilated either ground-based ZTD or TPW in the high-resolution data assimilation framework (Shoji et al. 2011; Oigawa et al. 2018; Mahfouf et al. 2015; Yang et al. 2020) to improve the moisture analysis. Directly assimilating ZTD rather than TPW may be more desirable since deriving TPW requires surface pressure

information, which may not be measured at all GNSS ground stations (Cucurull et al. 2004).

The ground-based ZTD observation network has been established in Taiwan, and the data can be processed at near real-time (Yeh et al. 2016). In the following, we review the results of Yang et al. (2020), which investigates the impact of assimilating the ZTD data in addition to the radar data on short-term precipitation prediction over complex terrain based on a prefrontal heavy-rainfall event on 10 June 2010 in Taiwan. Figure 17a shows the locations of the ZTD stations used in Yang et al. (2020). The ZTD observation operator simulates the ground-based ZTD value given a model vertical profile at the observation location. As the first step, the model state is interpolated into a vertical profile of temperature and humidity at a given ground-based GNSS site. Information on the modeled surface pressure and surface altitude at this location is also required. The estimation of ZHD is based on Saastamoinen (1972), which has a very high accuracy compared to more elaborate methods (Vedel et al. 2001). The ZWD is related to the vertical integration of water vapor. Further corrections are required to consider the difference between the model surface height and the altitude of the station. If the altitude of the site is higher than the model terrain, the correction removes the extra ZTD estimation by considering the vertical integration of the refractivity from the model terrain to this site altitude, and vice versa.

For the R-localization in LETKF, the height of the observation is defined at the height of the GNSS station, even though the observation is actually a vertically integrated quantity. This assumption is reasonable since the presence of moisture is usually greatest near the surface. The vertical localization for the ZTD data is described by a Gaussian function with 3 km as the standard deviation (localization scale). To consider the horizontal scale of the moisture distribution, ZTD is assimilated with a horizontal localization scale of 50 km, which is one order of magnitude broader than the one used for assimilating the radar data. In the following results, the assimilation is performed hourly from 0000 to 1200 UTC 10 June 2012 with an analysis grid spacing of 2 km. Results of three experiments that assimilate radar, ZTD, and both data are presented, and they are referred to as RDA, ZDA, and BOTH, respectively.

With the flow-dependent background error covariance, the moisture corrections are sensitive to the characteristics of the background moisture and landcover type. Assimilating the ZTD data has a large and dominant impact on the moisture conditions in Taiwan, and thus the ZTD data can be complementary to the radar data. The impact of assimilating the ZTD data is illustrated by the analysis increment from RDA and BOTH at 1200 UTC (2000 LST). While there is limited radar data available over northwestern Taiwan due to non-precipitating conditions (Fig. 17b), the ZTD data in this area provide significant moisture and southwesterly wind corrections in the BOTH analyses (Fig. 18a). In particular, wind corrections with ZTD data exhibit convergence in the coastal area with positive moisture increments, resulting in broader moisture convergence in BOTH over northwestern Taiwan than that in RDA. In this area, the amount of analysis corrections on the moisture and wind fields from assimilating radar data is limited in the BOTH and RDA analyses (Fig. 18b, c).

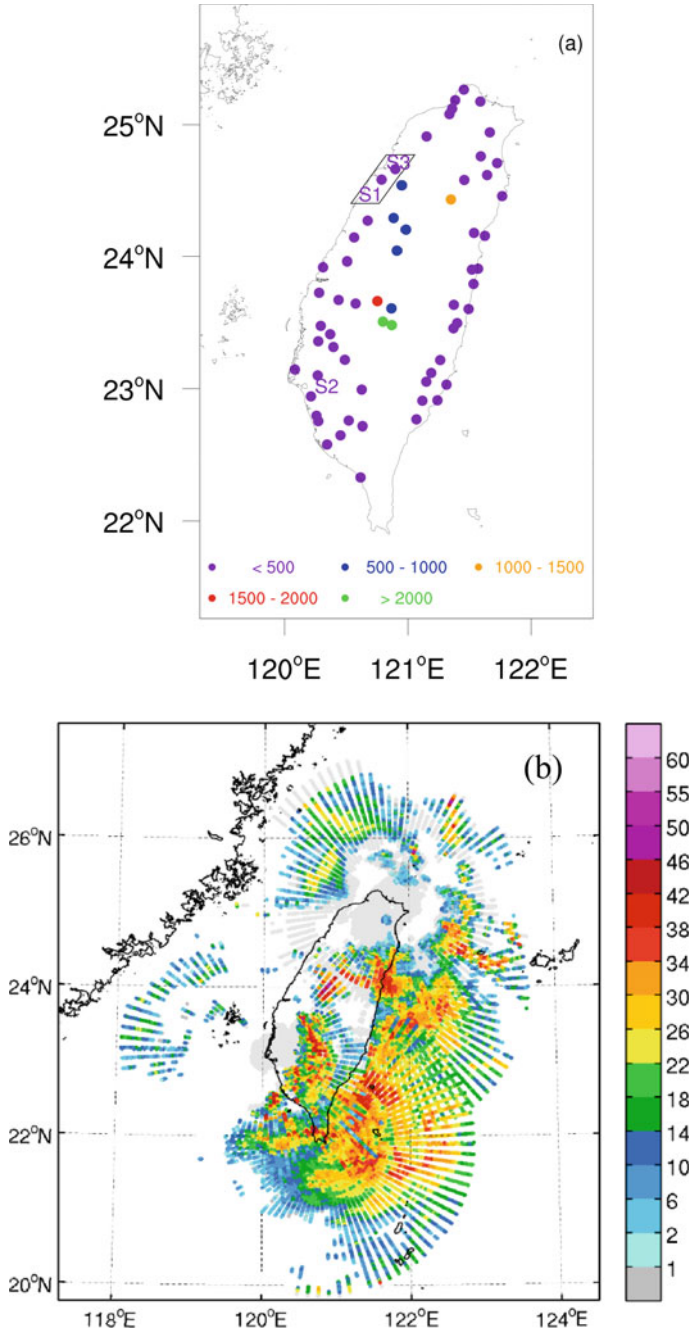

Fig. 17 **a** Observation locations of ZTD stations. The locations are categorized by the altitude (m) of the station. **b** Superobservations of radar data at 1200 UTC (2000 LST) 10 June 2012. Adopted from Fig. 4 in Yang et al. (2020)

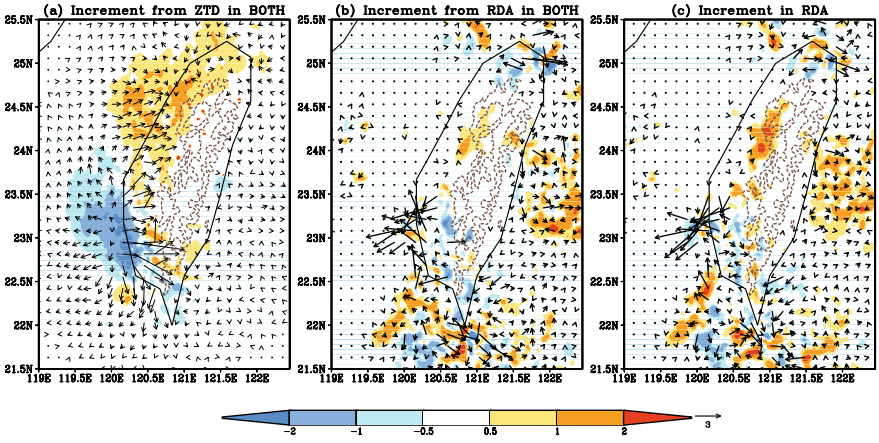

Fig. 18 Analysis increment of TPW (g kg^{-1}) and wind (m s^{-1}) at the level of 950 hPa at 1200 UTC 10 June. **a** is from assimilating ZTD data in BOTH, **b** and **c** are from assimilating radar data in BOTH and RDA, respectively. The dashed contours denote the terrain height at 1 and 2 km. Adopted from Fig. 10 in Yang et al. (2020)

Furthermore, near the coast of southwestern Taiwan, where radar data is abundant and rainfall has persisted, assimilating ZTD data reduces the moisture and enhances the onshore flow. The ZTD data provide a robust moisture correction since these data are directly related to the moisture field, and the data are available regularly at the GNSS stations. The dynamical relationship between moisture and wind allows the ZTD data to provide strong wind corrections as well. It is expected that the ZTD data assimilation modifies the conditions for convection development and thus heavy precipitation. Assimilating radar data also provides strong wind corrections offshore, but the moisture correction is limited.

For this event, RDA generally gives a good precipitation prediction in terms of the precipitation amounts and locations since the synoptic weather conditions are generally well represented in the initial conditions. However, the RDA forecast in areas of heavy rainfall is suboptimal in several aspects but the prediction can be improved by assimilating ZTD data additionally. The heavy precipitation in RDA begins too early and is too strong in southern Taiwan (areas K and P in Fig. 19b, f) but the overpredicted rainfall can be suppressed in BOTH (Fig. 19h) with the moisture reduction provided by the ZTD data (Fig. 18a). Also, the amount of precipitation in RDA in area N is less than the observations during the early forecast hours, but the rainfall is significantly overpredicted at 12 h (Fig. 19n). The precipitating process in this area is spun up in BOTH, even when there are non-precipitating conditions initially. It is noted that ZTD data assimilation alone is also very useful for improving short-term forecasts, and that moisture adjustment can improve the location and intensity of heavy rainfall. The intensity and locations of heavy rainfall generally agrees very well with the observations (Fig. 19o). However, without the direct information of hydrometeors, the simulated rainfall intensity is weaker than

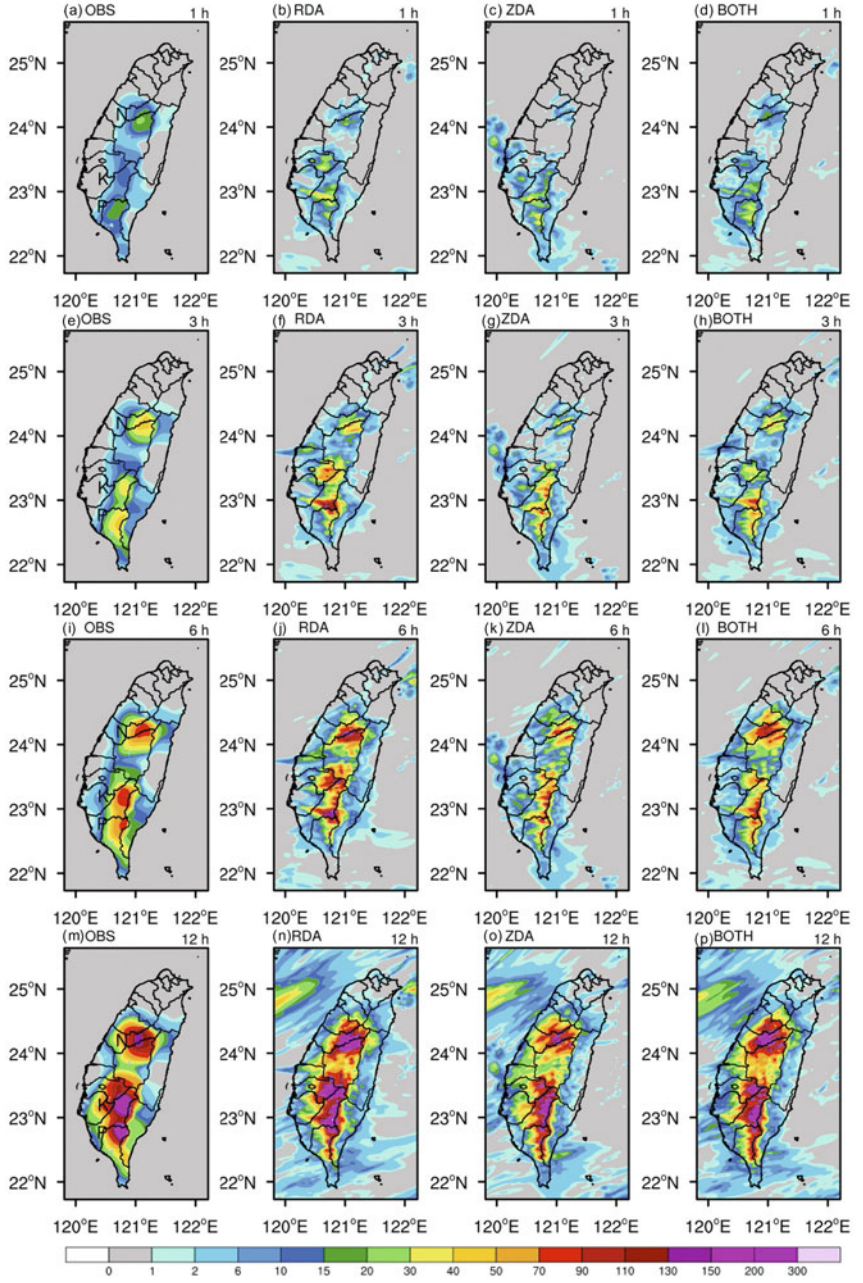

Fig. 19 Rainfall accumulation (mm) from observation (the 1st column) and RDA, ZDA, and BOTH (2nd to 4th columns) forecasts initialized at 1200 UTC (2000 LST) 10 June. Rainfall is accumulated for 1, 3, 6, and 12 h. Adopted from Fig. 11 in Yang et al. (2020)

that from RDA. Furthermore, the difference in the initialization affects the rainfall field within a very short time frame (Fig. 19b–d), indicating that rapid updating of the moisture field also plays an important role in very short-term rainfall predictions.

The moisture adjustment brought by assimilating the ZTD data can also be crucial to capture the initialization of afternoon thunderstorm during the prediction. Figure 20 is the precipitation prediction for an afternoon thunderstorm event over northern Taiwan on 22 July 2019. Torrential rainfall took place south of Taipei at 1720 LST (Fig. 20a). The assimilation is conducted from 1230 LST (0430 UTC) to 1430 LST (0630 UTC), with radar and ZTD data. Unlike the results shown in previous cases, the assimilation of radar data cannot capture the features leading to the heavy rainfall of this event, and the rainfall amount and location are even less accurate than NoDA (Fig. 20c vs. Fig. 20b). In contrast, the ZDA forecast exhibits a much greater amount of rainfall than the RDA forecast. In particular, the general features in the ZDA forecast are similar to the observations, where heavy rainfall occurs over the mountainous area of the New Taipei City, Hsinchu, and Miaoli counties, even though the

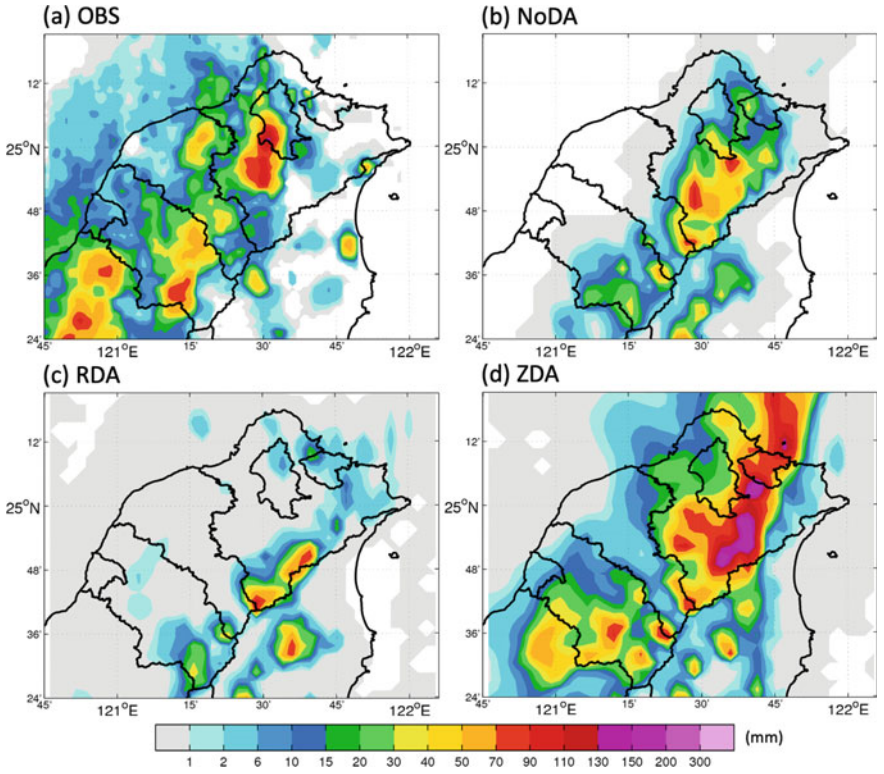

Fig. 20 Rainfall accumulated from 1430 to 2030 LST (0630 to 1230 UTC) 22 July 2019. **a** Observation from the rain gauge, **b** NoDA, **c** RDA, and **d** ZDA forecasts initialized at 0200 UTC 22 July

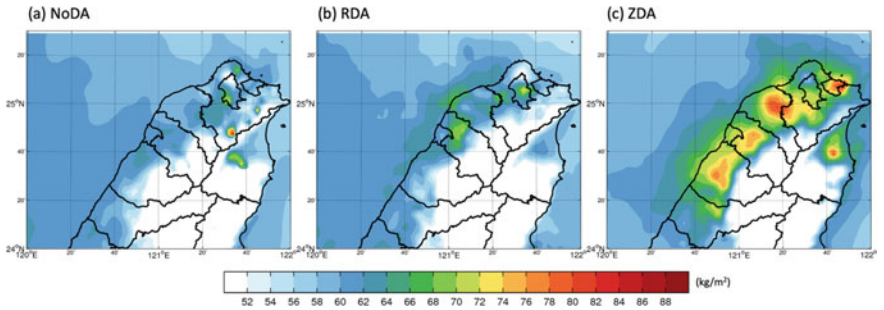

Fig. 21 Total precipitable water (mm) of **a** NoDA, **b** RDA, and **c** ZDA at 0630 UTC

rainfall in New Taipei City is overpredicted. Compared to NoDA, the ZTD assimilation provides large moisture enhancement while the moisture in RDA is very limited (Fig. 21b, c). The moisture corrections are effective in modifying the thermodynamic conditions and lead to great differences in precipitation prediction.

In conclusion, the ground-based GNSS ZTD data can provide rapid moisture information that is crucial in convective-scale data assimilation for representing the convection initialization and development.

6 Summary and Future Work

This chapter reviews the impact and challenges of applying WLRAS for precipitation prediction in Taiwan, which has complex terrain. The results are mainly based on a thorough investigation of a classical heavy rainfall event in Taiwan on 16 June 2008.

For this event, the sustainability of the convergence zone is the key for producing long-lived heavy rainfall events. The convergence zone is established by the warm southwesterly offshore and a cold pool outflow at the coast of southwestern Taiwan, induced by evaporative cooling from precipitation. Assimilating V_r can adjust the dynamical field effectively but is not good enough to generate heavy precipitation at the right location. In comparison, assimilating Z_h has a direct effect on correcting the rainfall location by adjusting the hydrometeor concentrations and indirect effects on the moisture and temperature fields. However, without the support from the updraft, most hydrometeors rain out during the first hour, as a feature of model shock. Assimilating both V_r and Z_h provides good cooperation between the dynamical and thermodynamic conditions at the coast with characteristics of orography effect. This leads to the best prediction in the intensity, location, and even the heavy rainfall's temporal variation.

For the same case, the characteristics, attribution, and impact of sampling error are identified by comparing the results of WLRAS with 40 and 256 members. In general, sampling errors for this event are sensitive to the relationships between the simulated observations and model variables. The ensemble-based estimation of BECR is less

sensitive to the ensemble size when the relationship between variables is direct and robust, such as Z_h and q_r, and V_r and U/V. For sampling errors arising from less-direct relationships between variables, moisture is a particularly difficult variable to update by assimilating V_r or Z_h since BECR between the simulated radar observations and moisture is much more uncertain. In the areas where the radar beams cannot resolve the U or V component, the BECR between the simulated V_r and model horizontal wind is sensitive to sampling error when the projections of the U and V component to the simulated V_r are comparable. In summary, the sampling errors in EnRDA are greatly affected by the dynamic/thermodynamic properties of the weather systems and the limitation of measurement simulated by the model. With a small ensemble, the sampling errors lead to an underprediction of heavy rainfall when the horizontal localization radius is excessively large. When BECR in vertical can be well estimated by a large ensemble, using a long vertical localization is critical to provide necessary adjustments for representing the vertical thermodynamic structure of convection, which further improves precipitation prediction.

On the same day, the prediction of a short-lived, intense afternoon thunderstorm over northern Taiwan is very sensitive to the initial conditions upstream. A QC criterion involving low wind speed and high reflectivity values may not completely filter out the terrain-related noises. Assimilating these problematic reflectivities upstream produces unrealistic outbreaks of light rain, whose evaporative cooling blocks the transport of warm and moist air into the target (downstream) area. As a result, this affects the initialization and development of the downstream thunderstorm. Removing these problematic reflectivities over the mountainous region upstream, a condition of thermodynamic instability can be established. The afternoon thunderstorm in the target area can be triggered during the forecast, with a well-represented location and intensity of this short-lived heavy rainfall event. A sophisticated procedure to examine the data quality over the terrain area is essential for optimizing the impact of radar data on precipitation prediction.

Assimilation of the GNSS ground-based ZTD data in Taiwan could be a remedy to compensate for the limitation of radar data on moisture corrections because the fast variations in the ZTD data are highly related to variations in the moisture field. For the case that assimilating radar data has led to somewhat skillful precipitation prediction, assimilating ZTD additionally can provide significant moisture adjustment in the area where the convection is initialized and when no radar data was available; and in the area where heavy precipitation had persisted and the precipitation is overpredicted with assimilation of radar data. For the case that assimilating radar data does not demonstrate improvement in precipitation prediction due to its limited moisture correction, the ZTD data assimilation can effectively modify the moisture field, leading to a thermodynamically unstable condition favorable for developing strong convection. For both events, the intensity and location of the heavy rainfall are improved. In particular, the ZTD data are expected to capture the meso-alpha scale moisture interacting with topography, such as the moisture transport during the Meiyu season. Therefore, the horizontal covariance localization scale is chosen to be one order of magnitude larger than what is used for radar data. With the advantages of inexpensive instrumentation and easy installation, the ZTD data becomes

an important set of observations complementary to radar data and provides a great potential to improve very short-term predictions over Taiwan.

Several challenges with the convective-scale data assimilation system in Taiwan need to be further tackled to improve heavy rainfall prediction. The near-surface condition is critical for representing the thermodynamic instability over topography. Assimilating surface observations can compensate for the limitation that the radar data cannot observe the atmospheric condition near the surface. The advantage of assimilating high-resolution radar data needs to be optimized by considering the observation error correlation (Yeh et al. 2019). Finally, improving the prediction of heavy precipitation with multi-scale interactions needs to be achieved through better incorporation with global-scale analysis (Hsiao et al. 2015) or data assimilation methodologies that provide multi-scale corrections (Zhang et al. 2009; Miyoshi and Kondo 2013; Yang et al. 2020).

Acknowledgements The authors are sponsored by the National Space Organization and the Ministry of Science and Technology in Taiwan (MOST Grant 109-2625-M-008 -015 and 107-2111-M-008 -003 -MYs). We are very grateful to Central Weather Bureau CWB for providing the ZTD and the QPESUMS data.

References

Aksoy A, Dowell DC, Snyder C (2009) A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: Storm-scale analyses. Mon Wea Rev 137:1805–1824

Bauer H-S, Schwitalla T, Wulfmeyer V, Bakhshaii A, Ehret U, Neuper M, Caumont O (2015) Quantitative precipitation estimation based on high-resolution numerical weather prediction and data assimilation with WRF—a performance test. Tellus A 67:25047

Bevis M, Businger S, Herring T, Rocken C, Anthes R, Ware R (1992) GPS meteorology: remote sensing of atmospheric water vapor using the Global Positioning System. J Geophys Res 97:15787–15820

Bengtsson L et al (2003) The use of GPS measurements for water vapor determination. Bull Am Meteorol Soc 84:1249–1258

Bick T, Coauthors, (2016) Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale. Quart J Roy Meteor Soc. 142:1490–1504. https://doi.org/10.1002/qj.2751

Bormann N, Bonavita M, Dragani R, Eresmaa R, Matricardi M, McNally A (2016) Enhancing the impact of IASI observations through an updated observation-error covariance matrix. Q J R Meteorol Soc 142:1767–1780. https://doi.org/10.1002/qj.2774

Campbell WF, Satterfield EA, Ruston B, Baker NL (2017) Accounting for correlated observation error in a dual-formulation 4D variational data assimilation system. Mon Wea Rev 145:1019–1032. https://doi.org/10.1175/MWR-D-16-0240.1

Carlin JT, Gao J, Snyder JC, Ryzhkov AV (2017) Assimilation of ZDR columns for improving the spinup and forecast of convective storms in storm-scale models: Proof-of-concept experiments. Mon Wea Rev 145:5033–5057. https://doi.org/10.1175/MWR-D-17-0103.1

Chang P-L, Lin P-F, Jou BJ-D, Zhang J (2009) An application of reflectivity climatology in constructing radar hybrid scans over complex terrain. J Atmos Oceanic Technol 26:1315–1327

Chang W, Chung K, Fillion L, Baek S (2014) Radar data assimilation in the Canadian high-resolution ensemble Kalman filter system: performance and verification with real summer cases. Mon Wea Rev 142:2118–2138

Cheng H-W, Yang S-C, Chen C-S (2019) Impact of Doppler radar network in Taiwan on the convective-scale data assimilation and precipitation prediction: the extreme precipitation event during 1–3 June 2017. AOGS 16th Annual Meeting, Singapore, Asia Oceanic GeoScience Society, AS30-A029

Cheng H-W, Yang S-C, Chen C-S, Liou Y-C (2020) An investigation of the sensitivity of a WRF-based convective-scale assimilation system on an afternoon thunderstorm in northern Taiwan. SOLA, (accepted)

Chung K-S, Chang W, Fillion L, Tanguay M (2013) Examination of situation-dependent background error covariances at the convective scale in the context of the ensemble Kalman filter. Mon Wea Rev 141:3369–3387. https://doi.org/10.1175/MWR-D-12-00353.1

Cucurull L, Vandenberghe F, Barker D, Vilaclara E, Rius Z (2004) Three-dimensional variational data assimilation of ground-based GPS ZTD and meteorological observations during the 14 December 2001 storm event over the Western Mediterranean Sea. Mon Wea Rev 132:749–763

de Haan S (2013) Assimilation of GNSS ZTD and radar radial velocity for the benefit of very-short-range regional weather forecasts. Q J R Meteorol Soc 139:2097–2107

Desroziers G, Berre L, Chapnick B, Poli P (2005) Diagnosis of observation, background and analysis-error statistics in observation space. Q J R Meteorol Soc 131:3385–3396. https://doi.org/10.1256/qj.05.108

Dowell DC, Wicker LJ, Snyder C (2011) Ensemble Kalman filter assimilation of radar observations of the 8 May 2003 Oklahoma city supercell: influences of reflectivity observations on storm-scale analyses. Mon Wea Rev 139:272–294. https://doi.org/10.1175/2010MWR3438.1

Fowler AM, Dance SL, Waller JA (2018) On the interaction of observation and prior error correlations in data assimilation. Q J R Meteorol Soc 144:48–62. https://doi.org/10.1002/qj.3183

Gao J, Xue M (2008) An efficient dual-resolution approach for ensemble data assimilation and tests with simulated Doppler radar data. Mon Wea Rev 136:945–963. https://doi.org/10.1175/2007MWR2120.1

Gao J, Stensrud DJ (2012) Assimilation of reflectivity data in a convective-scale, cycled 3DVAR framework with hydrometeor classification. J Atmos Sci 69:1054–1065

Gastaldo T, Poli V, Marsigli C, Alberoni PP, Paccagnella T (2018) Data assimilation of radar reflectivity volumes in a LETKF scheme. Nonlin Processes Geophys 25:747–764

Hamill TM, Whitaker JS, Snyder C (2001) Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Mon Wea Rev 129:2776–2790. https://doi.org/10.1175/1520-0493(2001)129,2776:DDFOBE.2.0.CO;2

Honda T, Miyoshi T, Lien G-Y, Nishizawa S, Yoshida R, Adachi SA, Terasaki K, Okamoto K, Tomita H, Bessho K (2018) Assimilating All-Sky Himawari-8 Satellite infrared radiances: a case of typhoon soudelor (2015). Mon Wea Rev 146:213–229. https://doi.org/10.1175/MWR-D-16-0357.1

Houtekamer PL, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. Mon Wea Rev 126:796–811. https://doi.org/10.1175/1520-0493(1998)126%3c0796:DAUAEK%3e2.0.CO;2.s

Hsiao L, Huang X, Kuo Y, Chen D, Wang H, Tsai C, Yeh T, Hong J, Fong C, Lee C (2015) Blending of global and regional analyses with a spatial filter: application to typhoon prediction over the western north pacific ocean, *Wea.* Forecasting 30:754–770

Huang Y, Wang X, Kerr C, Mahre A, Yu T, Bodine D (2020) Impact of assimilating future clear-air radial velocity observations from phased-array radar on a supercell thunderstorm forecast: an observing system simulation experiment study. Mon Wea Rev 148:3825–3845. https://doi.org/10.1175/MWR-D-19-0391.1

Hunt BR, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126

Jung Y, Xue M, Zhang G (2010) Simultaneous estimation of microphysical parameters and the atmospheric state using simulated polarimetric radar data and an ensemble Kalman filter in the presence of an observation operator error. Mon Wea Rev 138:539–562. https://doi.org/10.1175/2009MWR2748.1

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability, pp 341, Cambridge University Press, New York

Li X, Mecikalski JR, Posselt D (2017) An ice-phase microphysics forward model and preliminary results of polarimetric radar data assimilation. Mon Wea Rev 145:683–708. https://doi.org/10.1175/MWR-D-16-0035.1

Lien G-Y, Miyoshi T, Nishizawa S, Yoshida R, Yashiro H, Adachi SA, Yamaura T, Tomita H (2017) The near-real-time SCALE-LETKF system: a case of the September 2015 Kanto-Tohoku heavy rainfall. SOLA 13:1–6

Lindskog M, Salonen K, Järvinen H, Michelson DB (2004) Doppler radar wind data assimilation with HIRLAM 3DVAR. Mon Wea Rev 132:1081–1092

Liou Y-C, Chang S-F, Sun J (2012) An application of the immersed boundary method for recovering the three-dimensional wind fields over complex terrain using multiple-Doppler radar data. Mon Wea Rev 140:1603–1619

Ma Z, Kuo Y-H, Ralph FM, Neiman PJ, Wick GA, Sukovich E, Wang B (2011) Assimilation of GPS radio occultation data for an intense atmospheric river with the NCEP regional GSI system. Mon Wea Rev 139:2170–2183

Mahfouf J-F, Ahmed F, Moll P, Teferle FN (2015) Assimilation of zenith total delays in the AROME France convective scale model: a recent assessment. Tellus A 67:1–20

Maldonado P, Ruiz J, Saulo C (2020) Parameter sensitivity of the WRF–LETKF system for assimilation of radar observations: imperfect-model observing system simulation experiments. Wea Forecast 35:1345–1362

Ménétrier B, Montmerle T, Berre L, Michel Y (2014) Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles. Q J R Meteorol Soc 140:2050–2061. https://doi.org/10.1002/qj.2267

Miyoshi T, Coauthors, (2016) "Big Data Assimilation" Revolutionizing Severe Weather Prediction. Bull Am Meteorol Soc 97:1347–1354. https://doi.org/10.1175/BAMS-D-15-00144.1

Miyoshi T, Kondo K (2013) A multi-scale localization ap- proach to an ensemble Kalman filter. SOLA 9:170–173. https://doi.org/10.2151/SOLA.2013-038

Miyoshi T, Imamura T (2014) 10240-member ensemble Kalman filtering with an intermediate AGCM. Geophys Res Lett 41:5264–5271. https://doi.org/10.1002/2014GL060863

Oigawa M, Tsuda T, Seko H, Shoji Y, Realini E (2018) Data assimilation experiment of precipitable water vapor observed by a hyper-dense GNSS receiver network using a nested NHM-LETKF system. Earth Planets Space 70:74

Posselt DJ, Bishop CH (2012) Nonlinear parameter estimation: comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. Mon Wea Rev 140:1957–1974. https://doi.org/10.1175/MWR-D-11-00242.1

Posselt DJ, Hodyss D, Bishop CH (2014) Errors in ensemble Kalman smoother estimates of cloud microphysical parameters. Mon Wea Rev 142:1631–1654. https://doi.org/10.1175/MWR-D-13-00290.1

Putnam B, Xue M, Jung Y, Snook N, Zhang G (2019) Ensemble Kalman filter assimilation of polarimetric radar observations for the 20 May 2013 Oklahoma tornadic supercell case. Mon Wea Rev 147:2511–2533. https://doi.org/10.1175/MWR-D-18-0251.1

Rainwater S, Bishop CH, Campbell WF (2015) The benefits of correlated observation errors for small scales. Q J R Meteorol Soc 141:3439–3445. https://doi.org/10.1002/qj.2582

Saastamoinen J (1972) Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In *The Use of Artificial Satellites for Geodesy, Geophysical* Monograph Series, Volume 15. American Geophysical Union

Schraff C, Reich H, Rhodin A, Schomburg A, Stephan K, Perianez A, Potthast R (2016) Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). Q J R Meteoro. Soc. doi: https://doi.org/10.1002/qj.2748

Shao, Y-M (2015) Improving short-term rainfall prediction with an LETKF radar data assimilation system: a case study of SoWMEX IOP8 (in Chinese with an English abstract). M.S. thesis, Dept. of Atmospheric Sciences, National Central University, Taiwan, pp 78. https://hdl.handle.net/11296/ndddn2

Shoji Y, Kunii M, Saito K (2011) Mesoscale data assimilation of Myanmar cyclone Nargis part II: assimilation of GPS-derived precipitable water vapor. J Meteorol Soc Japan 89:67–88

Skamarock WC, Coauthors (2008) A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-4751STR, pp 113. doi:https://doi.org/10.5065/D68S4MVH

Snook N, Xue M, Jung Y (2015) Multiscale EnKF assimilation of radar and conventional observations and ensemble forecasting for a tornadic mesoscale convective system. Mon Wea Rev 143:1035–1057. https://doi.org/10.1175/MWR-D-13-00262.1

Snyder C, Zhang FQ (2003) Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. Mon Wea Rev 131:1663–1677. https://doi.org/10.1175//2555.1

Sobash RA, Stensrud DJ (2013) The impact of covariance localization for radar data on EnKF analyses of a developing MCS: observing system simulation experiments. Mon Wea Rev 141:3691–3709. https://doi.org/10.1175/MWR-D-12-00203.1

Sun JZ, Wang HL (2013) Radar data assimilation with WRF 4D-Var. Part II: comparison with 3D-Var for a squall line over the U.S. Great Plains. Mon Wea Rev 141:2245–2264. https://doi.org/10.1175/MWR-D-12-00169.1

Sobash RA, Wicker LJ (2015) On the impact of additive noise in storm-scale EnKF experiments. Mon Wea Rev 143:3067–3086. https://doi.org/10.1175/MWR-D-14-00323.1

Tao W-K, Coauthors (2003) Microphysics, radiation and surface processes in the Goddard Cumulus Ensemble (GCE) model. Meteor Atmos Phys, 82, 97–137

Terasaki K, Miyoshi T (2014) Data assimilation with error-correlated and non-orthogonal observations: Experiments with the Lorenz-96 model. SOLA 10:210–213. https://doi.org/10.2151/sola.2014-044

Terasaki K, Miyoshi T (2017) Assimilating AMSU-A radiances with the NICAM-LETKF. J Meteorol Soc Japan 95:433–446. https://doi.org/10.2151/jmsj.2017-028

Tong MJ, Xue M (2005) Ensemble Kalman Filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. Mon Wea Rev 133:1789–1807. https://doi.org/10.1175/MWR2898.1

Tsai C-C, Yang S-C, Liou Y-C (2014) Improving quantitative precipitation nowcasting with a local ensemble transform Kalman filter radar data assimilation system: observing system simulation experiments. Tellus 66A:21804. https://doi.org/10.3402/tellusa.v66.21804

Tsai C-C, Yang S-C, Lin C-Y, Liou J-C (2016) The development of WLRAS and its very short-term QPF performance in multiple heavy rainfall events. *The 7th EnKF Data Assimilation Workshop*, State College, PA, U.S.A., Peter Houtekamer and Fuqing Zhang. http://www.adapt.psu.edu/2016EnKFWorkshop/ABSTRACTS/Chih-Chien_Tsai_Poster.html

Vedel H, Mogensen KS, Huang X-Y (2001) Calculation of delays from meteorological data, comparison of NWP model and observed delays. Phys Chem Earth 26:497–502

Vukicevic T, Posselt DJ (2008) Analysis of the impact of model nonlinearities in inverse problem solving. J Atmos Sci 65:2803–2823. https://doi.org/10.1175/2008JAS2534.1

Waller JA, Dance SL, Nichols NK (2017) On diagnosing observation-error statistics with local ensemble data assimilation. Q J R Meteorol Soc 143:2677–2686. https://doi.org/10.1002/qj.3117

Weston PP, Bell W, Eyre JR (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. Q J R Meteorol Soc 140:2420–2429. https://doi.org/10.1002/qj.2306

Wheatley DM, Knopfmeier KH, Jones TA, Creager GJ (2015) Storm-scale data assimilation and ensemble forecasting with the NSSL experimental warn-on-forecast system. Part I: radar data experiments. Wea Forecasting 30:1795–1817

Wu P, Yang S, Tsai C, Cheng H (2020) Convective-scale sampling error and its impact on the ensemble radar data assimilation system: a case study of a heavy rainfall event on 16 June 2008 in Taiwan. Mon Wea Rev 148:3631–3652. https://doi.org/10.1175/MWR-D-19-0319.1

Xu W, Zipser EJ, Chen Y-L, Liu C, Liou Y-C, Lee W-C, Jong-Dao Jou B (2012) An orography-associated extreme rainfall event during TiMREX: initiation, storm evolution, and maintenance. Mon Wea Rev 140:2555–2574. https://doi.org/10.1175/MWR-D-11-00208.1

Yang S-C, Huang Z-M, Huang C-Y, Tsai C-C, Yeh D-K (2020) Convective-scale assimilation with the GNSS-ZTD and radar data and its impact on heavy rainfall prediction in Taiwan. Mon Wea Rev 148:1075–1098. https://doi.org/10.1175/MWR-D-18-0418.1

Yang S-C, Chen S-H, Chen S-Y, Huang C-Y, Chen C-S (2014) Evaluating the impact of the COSMIC-RO bending angle data on predicting the heavy precipitation episode on 16 June 2008 during SoWMEX-IOP8. Mon Wea Rev 142:4139–4163

Yang S-C, Chen S-H, Kondo K, Miyoshi T, Liou Y-C, Deng Y-L, Chang H-L (2017) Multilocalization data assimilation for predicting heavy precipitation associated with a multi-scale weather system. J Adv Model Earth Syst 9:1684–1702

Yano J, Coauthors (2018) Scientific challenges of convective-scale numerical weather prediction. Bull Amer Meteorol Soc 99, 699–710. https://doi.org/10.1175/BAMS-D-17-0125.1

Yeh T-K, Hong J-S, Wang C-S, Chen C-H, Chen K-H, Fong C-T (2016) Determining the precipitable water vapor with ground-based GPS and comparing its yearly variation to rainfall over Taiwan. Adv Space Res 57:2496–2507

Yeh H-L, Cheng H-W, Yang S-C (2019) Diagnosing observation error correlations for doppler radar radial winds based on the WRF-LETKF data assimilation system. 16th Annual Meeting on Asia Oceania Geosciences Society, Suntec, Singapore, AS30-A031

Yokota S, Seko H, Kunii M, Yamauchi H, Sato E (2018) Improving short-term rainfall forecasts by assimilating weather radar reflectivity using additive ensemble perturbations. J Geophys Res 123:9047–9062. https://doi.org/10.1029/2018JD028723

Yussouf N, Kain JS, Clark AJ (2016) Short-term probabilistic forecasts of the 31 May 2013 Oklahoma Tornado and flash flood event using a continuous-update-cycle storm-scale ensemble system. Wea Forecasting 31:957–983. https://doi.org/10.1175/WAF-D-15-0160.1

Zeng Y, Janjić T, de Lozar A, Blahak U, Reich H, Keil C, Seifert A (2018) Representation of model error in convective-scale data assimilation: Additive noise, relaxation methods, and combinations. J. Adv. Model. Earth Syst. 10:2889–2911

Zhang FQ, Weng YH, Sippel JA, Meng ZY, Bishop CH (2009) Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter. Mon Wea Rev 137(7):2105–2125. https://doi.org/10.1175/2009MWR2645.1

Zhang J, Howard K, Langston C, Vasiloff S, Kaney B, Arthur A, Van Cooten S, Kelleher K, Kitzmiller D, Ding F, Seo D-J, Wells E, Dempsey C (2011) National mosaic and multi-sensor (QPE) (NMQ) system: description, results, and future plans. B Am Meteor Soc 92:1321–1338

Interpretation of Forecast Sensitivity Observation Impact in Data Denial Experiments

Nancy L. Baker, Patricia M. Pauley, Rebecca E. Stone, and Rolf H. Langland

Abstract The U.S. Navy's global operational data assimilation and forecast system has significantly greater beneficial impact from the assimilation of global and polar Atmospheric Motion Vectors (AMVs) as compared to that from other Numerical Weather Prediction (NWP) centers. Results from an earlier multi-agency data denial inter-comparison study, presented at the 11th International Winds Working Group meeting (Baker et al. 2012a), demonstrated that this relatively large observation impact for the Navy system could be attributed to the assimilation of AMVs from multiple data providers which provided both a greater number of observations and better spatial and temporal coverage (Merkova et al. 2012). One important conclusion from Baker et al. (2012a) was that the interpretation of Forecast Sensitivity Observation Impact (FSOI; Langland and Baker 2004) for data denial studies can be problematic, particularly when the change to the Global Observing System is substantial (such as denying all satellite AMVs). Typically, such comparisons between two NWP systems for different data assimilation experiments explicitly assume that the quality of the two analyses are similar, and that the FSOI can be computed independently for the control and data denial experiments. However, this assumption may not be valid for data denial experiments with appreciable changes to the observing system. These considerations were further explored in the Baker et al. (2012b) presentation at the Fifth WMO Workshop on the impact of Various Observing Systems on Numerical Weather Prediction. These implications of data denial experiments on the interpretation of FSOI metrics are generally not well recognized. Additionally, the interpretation of FSOI may also be problematic for any set of experiments where the quality of the underlying analyses differ considerably from each other. In this chapter, the previous AMV data denial experimental studies are re-examined within the context of the implications on the interpretation of FSOI for data denial experiments.

N. L. Baker (✉) · P. M. Pauley · R. H. Langland
Marine Meteorology Division, Naval Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943-5502, USA
e-mail: nancy.baker@nrlmry.navy.mil

R. E. Stone
SAIC, 7 Grace Hopper Avenue, Monterey, CA 93943-5502, USA

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_22

1 Introduction

Until recently, Atmospheric Motion Vectors (AMVs) have had a much larger adjoint-based observation impact in the U.S. Navy's global operational numerical weather prediction (NWP) system compared to other global NWP centers. Conversely, the observation impact from microwave and infrared atmospheric sounders has been less. The total observation impact for the Navy and NASA Global Modeling and Assimilation Office (GMAO) global NWP systems for May 2012 is summarized in Fig. 1. The top five (six) observation categories for Navy were: AMV, radiosonde, aircraft (AMDAR and MDCRS), land surface, IASI, (SSMIS integrated water vapor), and AMSUA. The top five categories for GMAO are AMSUA, radiosonde, aircraft, IASI, and AMV. The Navy system uses a moist total energy error norm, while the GMAO system uses a dry total energy norm, and this accounts for large impact from SSMIS integrated water vapor (which GMAO did not assimilate). A more recent

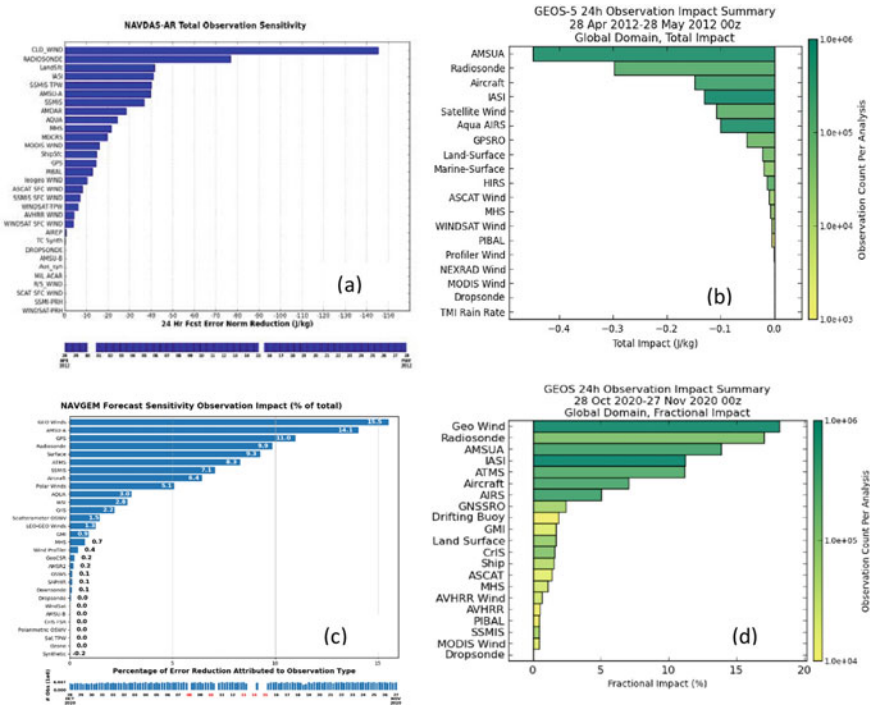

Fig. 1 Total Forecast Sensitivity Observation Impact for the **a** Navy system, and **b** NASA GMAO system (right), for the month ending on May 28 2012, and for percent impact November 2020 for the **c** Navy system, and **d** NASA GMAO system. The NRL observation impact is computed for each assimilation cycle, e.g. four times per day, while the GMAO observation impact is computed only for 00UTC. Note that the sign convention for GMAO was changed between 2012 and 2020 (cf. (d) from https://gmao.gsfc.nasa.gov/forecasts/systems/fp/obs_impact/, last retrieved 30 May 2012.)

comparison from November 2020 indicates that geostationary AMVs now have the largest adjoint-based observation impact for both systems (Fig. 1c, d).

The U.S. Navy's global operational NWP system in 2012 was composed of NAVDAS-AR (NRL Atmospheric Variational Data Assimilation System—Accelerated Representer), a 4D-Var (four-dimensional variational) global data assimilation system in observation space (Xu et al. 2005; Xu et al. 2006; Rosmond and Xu 2006; Chua et al. 2009), and the NOGAPS (Navy Operational Global Atmospheric Prediction System), a global atmospheric model that at that time was run operationally with a resolution of 319 spectral triangular truncation on 42 levels (Hogan and Rosmond 1991; Peng et al. 2004).

As discussed in Pauley et al. (2012), the NRL system differed from other operational centers in the use of geostationary AMVs from multiple operational data providers (NESDIS, EUMETSAT, and JMA), the University of Wisconsin-Madison's Cooperative Institute for Meteorological Satellite Studies (CIMSS), with additional winds provided from AFWA (Air Force Weather Agency). Representative data coverage charts for these AMVs are shown in Fig. 2. Another difference from other operational centers is that the AMVs are used as "superobs", as described Pauley et al. (2012).

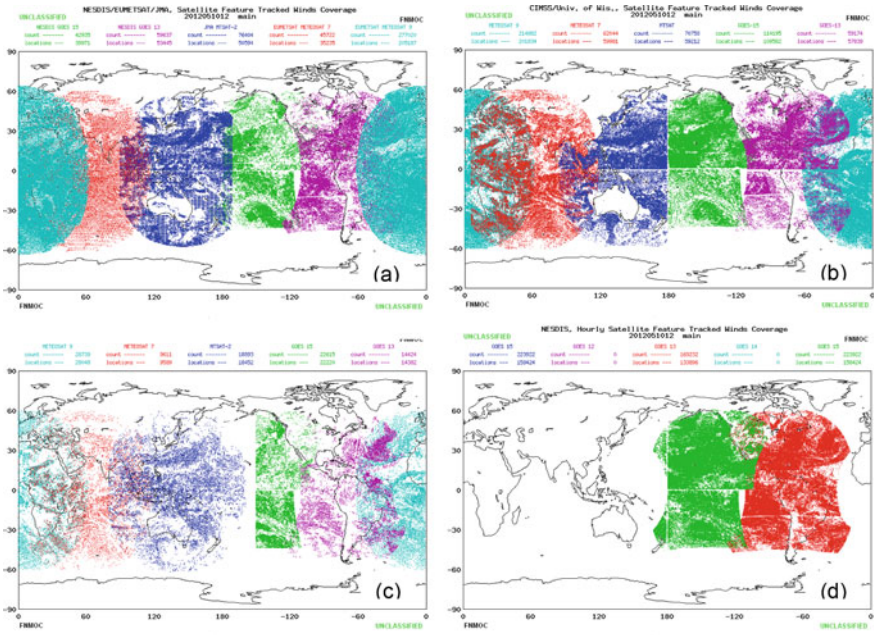

Fig. 2 Atmospheric Motion Vector Data Coverage for May 10 2012, for **a** NESDIS/EUMETSAT/JMA, **b** CIMSS/UW, **c** AFWA, and **d** NESDIS Hourly winds

2 Experiment Design

The experiments used a configuration of NAVDAS-AR and NOGAPS that closely matched the Navy's operational configuration in use in 2012. For the assimilation component, NAVDAS-AR was run with T319 outer loop resolution (approximately 42 km), and a T119 inner loop resolution (approximately 110 km), with 42 vertical levels from the surface to 0.4 hPa (around 70 km). Approximately 1.1 million observations were assimilated every 6 hours for the control run, while the denial run removed all of the geostationary and polar orbiting AMVs (see Table 1). The AMVs were averaged to form superobs, following the procedure outlined in Pauley et al. (2012). Approximately 175,000 superobs were assimilated for each assimilation cycle, representing around 15% of the total observation count. The NOGAPS forecast model used Eulerian differencing, with the Emanuel cumulus scheme. The satellite radiance bias correction method followed an offline two-predictor approach (Harris and Kelly 2001) and used the previous 15 days of radiance innovations to generate the bias coefficients, which were updated with each assimilation cycle.

The data denial experiments discussed in Baker et al. (2012a) were run for both Northern Hemisphere (NH) winter and summer cases. Here, we focus on the NH Summer, as the previous study found that the impact of AMVs is greatest for the summer hemisphere, which corresponds to the Northern Hemisphere tropical cyclone season. The control run was initialized with the 00UTC cycle on 1 August 2010, and spun up through the 18UTC cycle on 14 August 2010. The initial conditions, bias coefficients, and bias statistics for this date and time were archived to provide the starting point for the AMV denial run. The first assimilation cycle with the

Table 1 Sources of AMVs, according to satellite, orbit, and frequency band and data providers. Acronyms are defined in the Appendix

Satellite	Orbit	Frequency Band	Data Providers
GOES-11	Geostationary	IR, SWIR, WV, WVCLD, VIS	NESDIS, CIMSS, AFWA
GOES-13	Geostationary	IR, SWIR, WV, WVCLD, VIS	NESDIS, CIMSS, AFWA
Meteosat-7	Geostationary	IR, WV, WVCLD, VIS	EUMETSAT, CIMSS, AFWA
Meteosat-9	Geostationary	IR, WV, WVCLD, VIS	EUMETSAT, CIMSS, AFWA
MTSAT-2	Geostationary	IR, SWIR, WVCLR, WVCLD, WV, VIS	JMA, CIMSS
MODIS Terra	Polar	IR, WV	CIMSS
MODIS Aqua	Polar	IR, WV	CIMSS
MODIS Terra/Aqua mixed	Polar	IR, WV	CIMSS
AVHRR (NOAA 15,16,18,19), METOP-A	Polar	IR	CIMSS

AMV data denial was for the 00UTC cycle on 15 August 2010. The 5-day forecasts were initialized from the 12UTC analyses, with the final forecast on 30 September 2010. The assimilation cycles were continued for another 5 days to provide verifying analyses for the long forecasts. This specific experiment configuration was specified for the multi-agency data denial inter-comparison study for the 11th International Winds Working Group meeting.

3 Northern Hemisphere Summer Forecast Verification Results

In this section, the forecast verification results from the Northern Hemisphere summer AMV denial experiment are summarized. The 500 hPa geopotential height anomaly correlation die-off curves are displayed in Fig. 3. These figures suggest that the largest impact from AMV winds is in the summer (northern) hemisphere (Fig. 3a), with minimal impact in the winter (southern) hemisphere (Fig. 3b). The 500 hPa geopotential height anomaly correlation scores are analogous to those presented for the NH winter experiments (not shown; see Baker et al. 2012a), in that most of the benefit from the AMV assimilation appears to be in the summer hemisphere, with minimal impact in the winter hemisphere.

The impact of the AMV assimilation on the analyses (initial conditions) were assessed by computing the mean differences between the control and denial runs for the period from 15 August through 30 September 2010. The mean denial minus control wind speed and vector differences at 250 hPa are plotted in Fig. 4. The largest differences are located in the tropics, where the AMV assimilation noticeably alters the analyzed mean wind speed and direction. Other areas with moderate wind

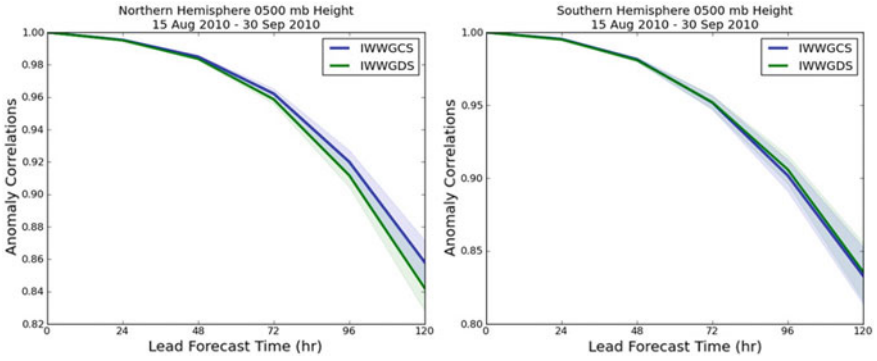

Fig. 3 Anomaly correlation die-off curves for 500 hPa geopotential heights for **a** the Northern Hemisphere and **b** Southern Hemisphere for 15 August 2010 through 30 September 2010. The forecast lead time in hours is given on the abscissa, and the anomaly correlation is given on the ordinate. Shading indicates the 95% confidence level

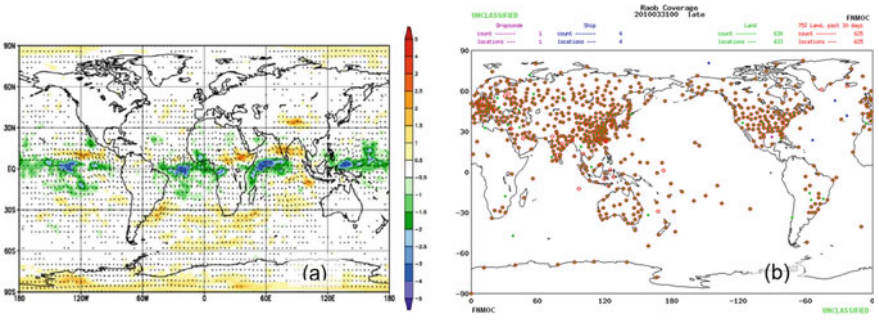

Fig. 4 **a** Averaged wind speed and wind vector differences at analysis time at 250 hPa. The color shading indicates the mean wind speed difference and the arrows give the mean wind vector difference (both in ms^{-1}). **b** Distribution of radiosondes stations for 00UTC on 31 March 2010

differences tend to coincide with regions with few conventional wind observations from radiosondes (Fig. 4b), pilot balloons, or aircraft.

The 24-hr forecast 200 and 850 hPa vector wind RMS errors for the tropics, as verified against radiosondes, are plotted in Fig. 5. Although the differences are not statistically significant (in part due to the limited number of radiosondes stations in the tropics; Fig. 4b, there is a clear trend for lower vector wind RMS errors in the forecast when AMVs are assimilated.

The homogeneous tropical cyclone (TC) track forecast errors for the control and denial experiments are shown in Fig. 6. The storm position in the forecast was determined by the NOGAPS tropical cyclone storm tracker which verified against the TC warning position as opposed to the post-season best-track position. Although there were not many individual tropical cyclones during the test period, these results highlight the importance of AMV assimilation for improving TC track forecasts.

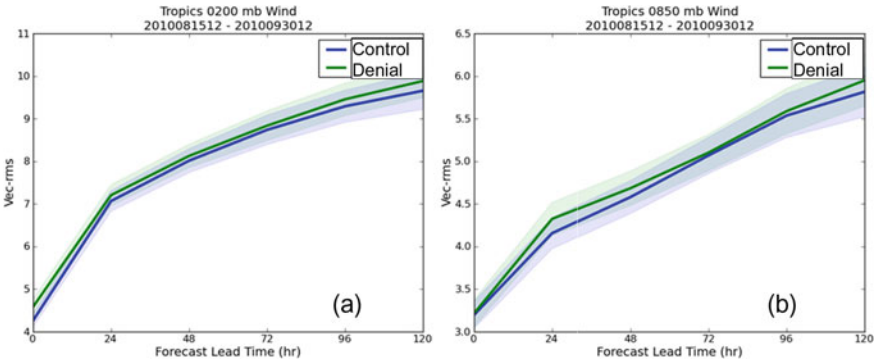

Fig. 5 **a** 200 hPa and **b** 850 hPa vector wind RMS errors in the tropics for the Northern Hemisphere summer AMV control experiment (blue line; Denial) compared to the AMV denial experiment (green line; Control), as verified against radiosondes. The forecast lead time (hours) is given on the abscissa, and the vector wind error ($m\ s^{-1}$) is given on the ordinate

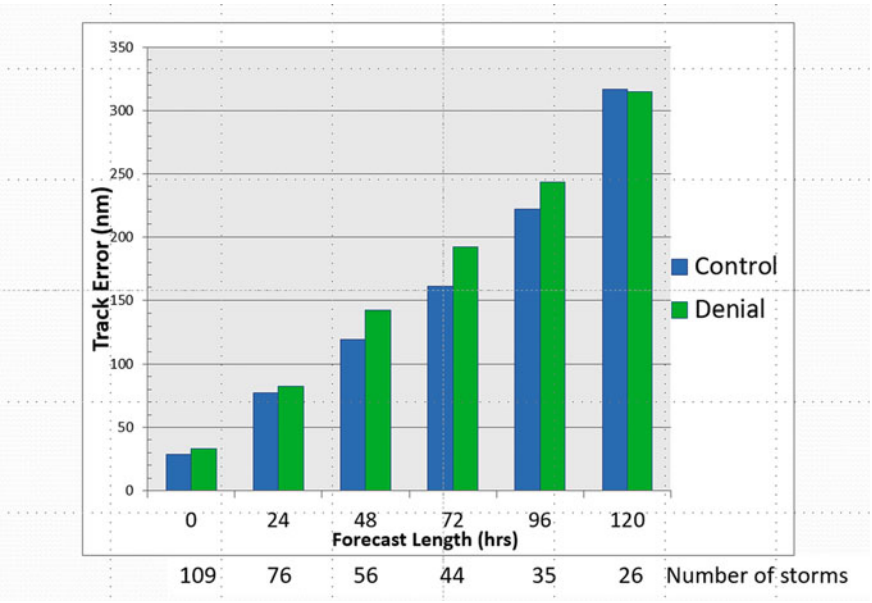

Fig. 6 Tropical cyclone track prediction error in nautical miles (nm; ordinate) as a function of forecast lead time (hrs; abscissa) for 15 August 2010, through 30 September 2010. The control experiment errors are given by blue bars, and the denial experiment errors are given by green bars. The number of verifying storm positions for this homogeneous comparison, for each forecast lead time is given by the number below the graph. The differences are significant for all forecast lengths to t + 72 at the 99.0 to 99.5% confidence levels

4 Northern Hemisphere Summer Forecast Sensitivity Observation Impact

The forecast sensitivity observation impact (FSOI; Langland and Baker 2004) was computed for each 6-hr assimilation cycle, using a moist total energy error norm. The observation impact for the major observing platforms/categories are presented in Fig. 7 for the control experiment and the AMV denial experiments. These results show that, in the absence of AMVs, other observations account for a larger share of the computed FSOI. The largest increases in FSOI in the denial run are for the satellite radiances and for other wind observations. For the control run, the greatest beneficial impact overall (Fig. 8a) is from the IR winds, although on a per-observation basis (Fig. 8b), the VIS winds provide more impact than either IR or WV winds, despite the low observation count (Fig. 8c). The FSOI contributions due to the assimilation of radiances for the various satellite sounders (see Appendix for the acronym list) are presented in Fig. 9 for the control and denial experiments. Interestingly, the largest increases in FSOI occur for the two hyperspectral IR sounders.

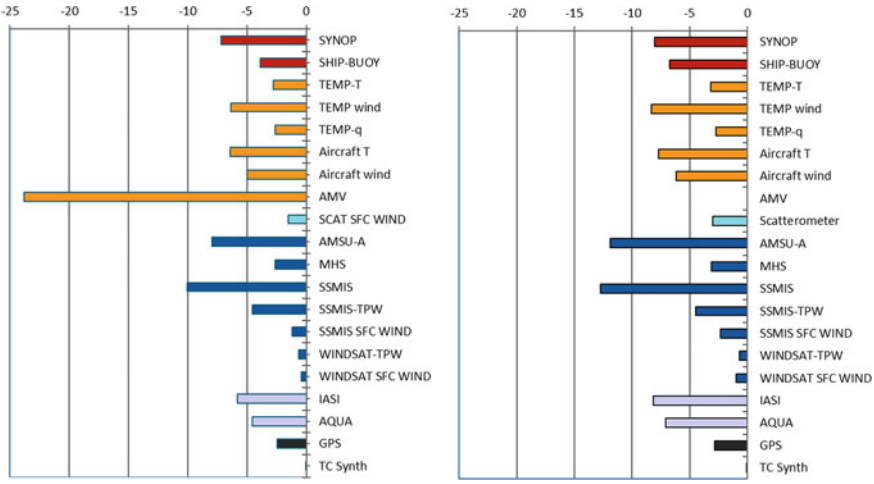

Fig. 7 Percent reduction of the total FSOI for each major observing category for 15 August through 30 September 2010, **a** for the control experiment and **b** the AMV denial experiment. The observation impact is computed every 6 h

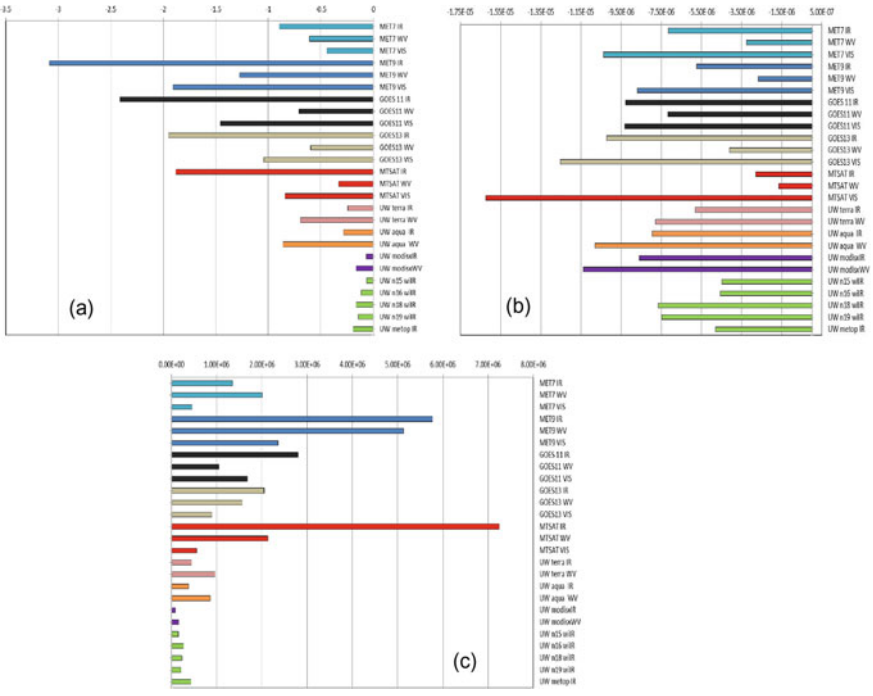

Fig. 8 Statistics for the control run as a function of satellite and type (infrared, water vapor, visible), for **a** the percent reduction of the total FSOI and **b** the FSOI per observation (J kg^{-1}), and **c** the number of superobs; each wind component counts as one observation

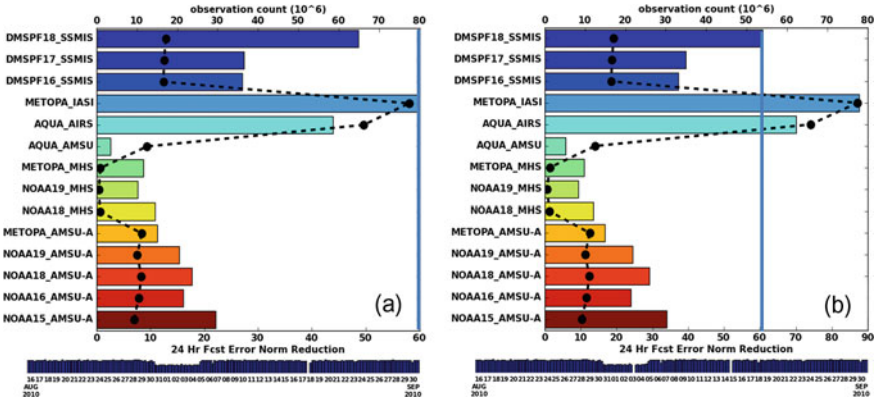

Fig. 9 FSOI for the assimilated microwave and infrared satellite sounders for **a** control and **b** AMV denial runs. The colored bars represent the FSOI, while the black dashed lines and dots give the total observation counts divided by 106. Note that the abscissa scales differ, and the vertical blue line highlights the 60 J kg-1 value. These figures show that the MW and IR satellite sounders have greater impact when AMVs are denied

5 Northern Hemisphere Summer 24-H Moist Total Energy Error Norms

The observation impact results presented in Sect. 4 show that AMV assimilation contributes to a large reduction in the total moist energy error norm. The 24-hr total moist energy error norms for the control run were separated into the four main components of temperature, humidity, vorticity and divergence (neglecting the much smaller terrain pressure error), and averaged starting with the 00UTC cycle on 1 August 2010, and continuing through the 12UTC cycle on 9 September 2010. The average 24-hr forecast error norm components, shown in Fig. 10, are largest in spatial extent and magnitude for humidity followed by vorticity. The corresponding FSOI values were computed for each of the four components of the total moist energy error norm. According to Fig. 7a, the assimilation of the AMVs contribute around 24% to the total FSOI for all observations. Of that 24%, approximately 12% is from the reduction in the humidity error, and 9.5% from the vorticity error, with less than 2% each attributable to temperature and divergence (Fig. 11). To summarize, AMVs provide the greatest percent reduction in the total moist error norm for any observation category. The total moist error norm is dominated by humidity and vorticity components, and the assimilation of AMVs is effective in reducing these two components of the 24-hr total moist energy error norm.

However, reducing the difference in the two error norms (24-hr and 30-hr) is not the same as reducing the total 24-hr error norm, particularly when the denied observations represent a significant proportion of the observing system, or contribution to the overall reduction in forecast error. Under these circumstances, the accuracy of the denial run analysis can be significantly less than for the control run. If the error

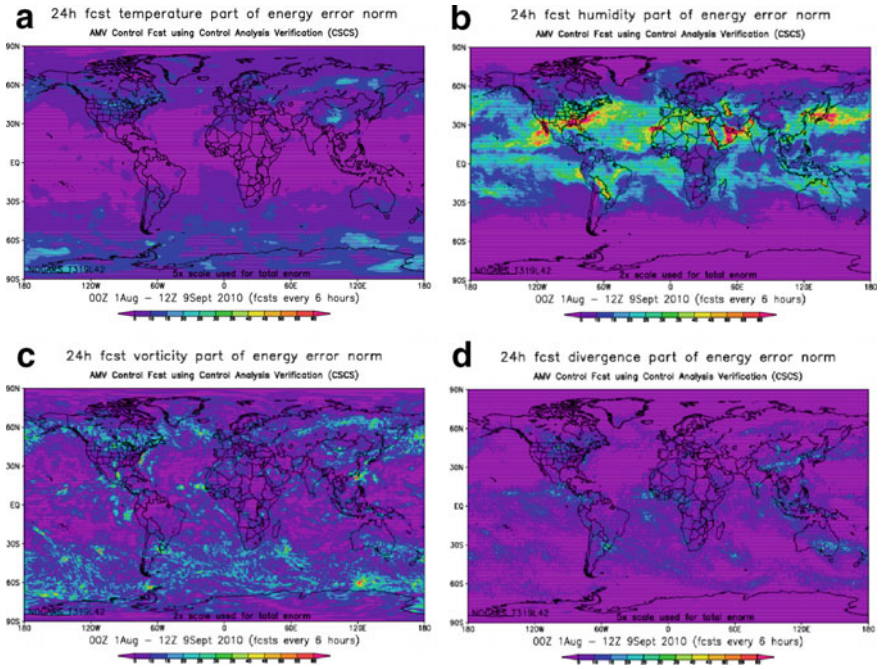

Fig. 10 Components of the averaged 24-h moist energy error norm (J kg^{-1}), for **a** temperature, **b** humidity, **c** vorticity and **d** divergence

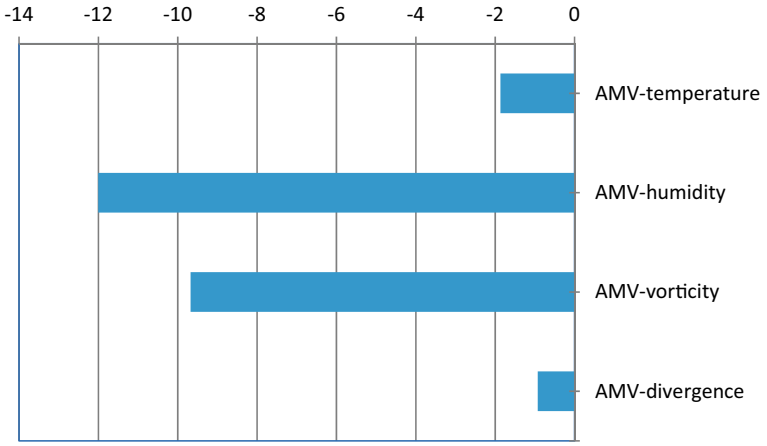

Fig. 11 The percent reduction of the total FSOI for AMVs (from Fig. 7a), but computed for the temperature, humidity, vorticity and divergence components of the total moist energy error norm (Fig. 9)

norms are computed relative to their own (self) analysis, then the control and denial runs may appear to have similar 24-hr moist energy error norms. The associated FSOI values may be similar and highly misleading. For this discussion, we note that denying all satellite AMVs is a large change to the NRL global analysis/forecast system, and explicitly assume that the control analyses (with AMV assimilation) are more accurate than the analyses produced without AMV assimilation.

The 24-hr moist total energy error norms for the control and denial cases, where both runs are verified against their own (self) analysis, are compared in Table 2. The column labeled "Total" contains the global total 24-hr moist error norm values, which are unexpectedly smaller (less error) for the denial run than for the control run. To better understand this apparent discrepancy, the global errors are partitioned into the main components of vorticity, divergence, temperature and humidity (neglecting the much smaller terrain pressure error). The error norms for all components except vorticity are larger for the control run than the denial run, again suggesting that AMV assimilation increases the 24-hr error. These results are contradictory with the forecast verification results in Sect. 3.

The 24-hr error norm values were then further stratified by region, NHEM (20-80N), Tropics (20N-20S) and SHEM (20S-80S). For the Northern (summer) Hemisphere, AMV assimilation decreases the 24-hr error norm for all components, which is consistent with the 500 hPa geopotential height anomaly correction scores. For the Southern (winter) Hemisphere, AMV assimilation primarily reduces the vorticity errors, and slightly increases the humidity errors. In the tropics, however, AMV assimilation apparently increases all components of the 24-hr error norm values.

The 24-hr moist energy error norms' time series are plotted for 15 August through 30 September in Fig. 12. The solid green line represents the 24-hr error norm for the control run as verified against the control analyses, while the cyan line shows the 24-hr error norm for the denial run as verified against the denial analyses. When verified against self-analyses, the control and denial runs have similar 24-hr moist energy error norms. However, when verified against the control analyses, the denial forecasts have much larger 24-hr errors using the total energy error norm, and all components of the error norm (vorticity, divergence, temperature, humidity) are larger when AMVs are excluded from the assimilation.

To further investigate these results, the 24-hr total moist energy error norms were computed using (a) the control run 24-hr forecasts verified against the control analyses, (b) the AMV denial 24-hr forecasts verified against the AMV denial analyses, and (c) the AMV denial 24-hr forecasts verified against the control analyses. The error norms were computed for every 6-hr assimilation cycle and averaged from 15 August 2010 through 26 September 2010. The time-averaged 24-hr total moist energy error norms for the control (Fig. 13a) and denial (Fig. 13b) runs, when computed using their own (self) analyses, are remarkably similar in appearance (structure and magnitude). Such similarities might lead to the erroneous conclusion that AMV assimilation has little impact on either the forecasts or analyses.

However, when the 24-hr total moist energy error norms for the denial run are verified against the control analyses, the resulting plots are strikingly different (Fig. 13c). The time-averaged error norms clearly show that the denial of the AMVs results in

Table 2 Summary of 24-h total moist energy error norms (J kg^{-1}) for the control run with AMV assimilation and the denial run, as verified against self-analysis. The lower (smaller) 24-h moist energy norm values are color-coded, red for control run or blue for the denial run

Avg values	Total		Vorticity		Divergence		Temperature		Humidity	
	Control	Denial	Control	Denial	Control	Denial	Control	Denial	Control	Denial
Global	23.257	22.976	7.469	7.647	1.673	1.604	2.028	2.001	12.024	11.658
NHEM (20–80)	8.587	8.844	2.322	2.509	0.432	0.441	0.713	0.718	5.101	5.157
SHEM (20–80)	6.264	6.377	3.097	3.248	0.530	0.530	0.900	0.902	1.703	1.661
TROPICS (20–20)	8.387	7.678	1.959	1.755	0.706	0.626	0.393	0.352	5.319	4.936

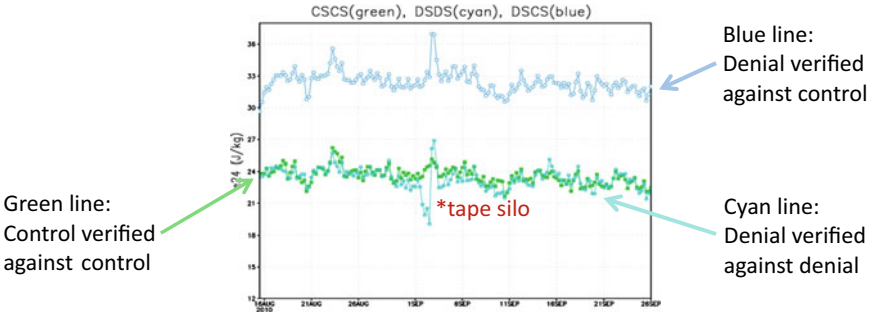

Fig. 12 Time series of 24-h total moist energy error norms (J kg^{-1}) for the control run verified against control (self-analyses) (green line), for the denial run verified against denial self-analyses (cyan line), and the denial run verified against the control analyses (blue line). The spike in the cyan line was due to tape silo issues

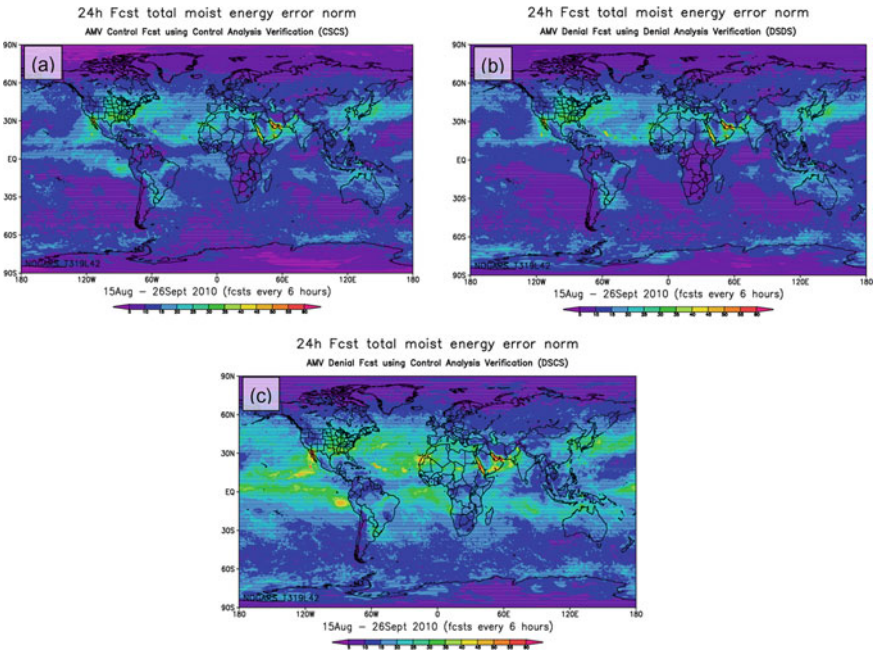

Fig. 13. 24-h total moist energy error norm (J kg^{-1}) computed using **a** the control run 24-h forecast verified against the control analysis, **b** the AMV denial 24-h forecast verified against the AMV denial analysis, and **c** the AMV denial 24-h forecast verified against the control analysis

large departures from the control analyses, not only in the tropics (30S-30N), but also over the Northern Hemisphere mid-latitude ocean regions. This is consistent with the forecast verification and FSOI results where the largest benefit from AMV assimilation was similarly located in the tropics and for the Northern Hemisphere.

6 Summary and Discussion

The Navy's global operational data assimilation and forecast system has long demonstrated significantly greater benefit from the assimilation of Atmospheric Motion Vectors (AMVs) compared to other NWP centers. In particular, AMVs provide the largest FSOI (around 24%) for any observing category. This has been attributed to the assimilation of AMVs from multiple data providers that provide more observations with better spatial and temporal coverage and the use of super-observations. The data denial experiments presented in this chapter show that, in the absence of AMVs, other observations (primarily satellite radiances—especially from IR sounders—and other wind observations) largely compensate for the missing winds with respect to the reduction in the 24-hr error norms and FSOI computations.

Typically, for data denial experiments, the comparisons between two (or more) data assimilation experiments explicitly assume that the quality of the analyses are similar, and that the FSOI can be computed independently for each control and data denial experiment. However, this assumption may not be valid for data denial experiments with appreciable changes to the observing system. Under these circumstances, the accuracy of the both the denial run analyses and forecasts can be significantly less than for the control run, and the resulting FSOI computations may be highly misleading.

This was demonstrated by comparing time series and time-averaged spatial maps computed using the 24-hr total moist energy error norms from the control 24-hr forecasts and verifying analyses, and the AMV denial 24-hr forecasts and verifying analyses. The resulting total moist energy error norms were remarkably similar in both magnitude and appearance for both the control and denial runs. Such similarities might lead to the erroneous conclusion that AMV assimilation has little impact on either the forecasts or analyses. In contrast, when the 24-hr total moist energy error norms for the denial run were computed using the control analyses, the resulting time series and spatial maps were strikingly different in both magnitude and spatial structure. In particular, these maps clearly show that the denial of the AMVs leads to large departures from the control analyses, not only in the tropics (30S-30N), but also over the Northern (summer) Hemisphere mid-latitude ocean regions—results that are consistent with the FSOI and forecast verification metrics.

The fundamental conclusion from this study is that the interpretation of FSOI for data denial studies can be problematic, and the results may be highly misleading. This is especially true when the change to the observing system is substantial (such as denying all satellite AMVs).

Acknowledgements We gratefully acknowledge support from the Naval Research Laboratory under program elements 0601153N and 062435N.

Appendix: Definitions of Acronyms

4D-Var:	4-Dimensional VARiational data assimilation
ACARS:	Aircraft Communications, Addressing, and Reporting System
AFWA:	(U.S.) Air Force Weather Agency
AIREP:	Aircraft Report
AMDAR:	Aircraft Meteorological Data Relay.
AMSR-2:	Advanced Microwave Scanning Radiometer - 2
AMSU-A:	Advanced Microwave Sounding Unit-A
AMSU-B:	Advanced Microwave Sounding Unit-B
AMV:	Atmospheric Motion Vector
AQUA (AIRS):	Atmospheric InfraRed Sounder, flown on the NASA Aqua satellite.
ASCAT:	Advanced Scatterometer, flown on the METOP satellites.
AVHRR:	Advanced Very High Resolution Radiometer
CIMSS/UW:	Cooperative Institute for Meteorological Satellite Studies
CLD_WIND:	AMVs from geostationary satellites (also referred to as GEO WINDS)
CrIS:	Cross-track Infrared Sounder
CrIS FSR:	Cross-track Infrared Sounder, Full Spectral Resolution
EUMETSAT:	European operational satellite agency for monitoring weather, climate and the environment from space.
FSOI:	Forecast Sensitivity to Observation Impact
GeoCSR:	Geostationary satellite Clear Sky Radiance
GMAO:	Global Modeling and Assimilation Office at NASA Goddard.
GMI:	GPM (Global Precipitation Measurement) Microwave Imager
GOES:	(U.S.) Geostationary Operational Environment Satellite
GNSS:	Global Navigation Satellite System (which includes GPS).
GPS:	Global Positioning System
GPS RO:	GPS Radio Occultation observations (also called GNSS RO).
HIRS:	High-resolution Infrared Radiation Sounder
IASI:	Infrared Atmospheric Sounding Interferometer

IR:	Infrared.
JMA:	Japanese Meteorological Agency.
LeoGeo:	CIMSS AMVs determined from composite imagery based on data from both geostationary and polar-orbiting satellites
MDCRS:	Meteorological Data Collection and Reporting System.
Meteosat:	EUMETSAT geostationary satellites, abbreviated as MET7 for Meteosat-7, MET9 for Meteosat-9, etc.
METOP:	METeorological Operational (polar-orbiting) satellites, operated by EUMETSAT.
MHS:	Microwave Humidity Sensor
MODIS:	Moderate Resolution Imaging Spectroradiometer, flown on the NASA Aqua and Terra satellites.
MTSAT:	Multi-functional Transport Satellite, geostationary satellites operated by JMA.
NASA:	(U.S.) National Aeronautics and Space Administration
NAVDAS-AR:	NRL Atmospheric Variational Data Assimilation System—Accelerated Representer.
NESDIS:	(U.S.) National Environmental Satellite and Data Information Service.
NEXRAD:	(U.S.) Next-generation Radar
NH or NHEM:	Northern Hemisphere.
NOAA:	(U.S.) National Oceanic and Atmospheric Administration
NOGAPS:	Navy Operational Global Atmospheric Prediction System
NRL:	(U.S.) Naval Research Laboratory
NWP:	Numerical Weather Prediction
OSWS:	Ocean Surface Wind Speed
OSWV:	Ocean Surface Wind Vector
PIBAL:	Pilot Balloon
SAPHIR:	Sondeur Atmosphérique du Profil d'Humidité Intertropicale par Radiométrie
SH or SHEM:	Southern Hemisphere
SHIP-BUOY:	Observations from fixed and mobile ships and buoys.
SSMIS:	Special Sensor Microwave Imager Sounder
SSMIS TPW:	Total Precipitable Water retrievals from SSMIS.
SSMIS SFC WIND:	Ocean surface wind speed retrievals from SSMIS.
SWIR:	Shortwave IR
SYNOP:	WMO-format surface data, primarily from land-based stations
TC Synth or "Synthetic":	Synthetic observations generated from TC warning messages
TC:	Tropical Cyclone.

TEMP:	WMO-format radiosonde data (including T (temperature), wind, and q (humidity)
TMI:	TRMM (Tropical Rainfall Measuring Mission) Microwave Imager
WINDSAT-TPW:	NRL polarimetric microwave satellite Total Precipitable Water retrievals
WINDSAT SFC WIND:	WindSat wind vector retrievals
VIS:	Visible
WMO:	World Meteorological Organization
WV:	Water Vapor
WVCLD:	Cloud-Top Water Vapor
WVCLR:	Clear-Sky Water Vapor

References

Baker N, Langland R, Pauley P, Xu L, Velden C (2012a) The impact of satellite atmospheric motion vectors in the U.S. Navy global data assimilation system—NWP results. Extended abstract, 11th International Winds Workshop, 20–24 February 2012, University of Auckland, New Zealand. http://cimss.ssec.wisc.edu/iwwg/iww11/talks/Session4_Baker.pdf

Baker N, Langland R, Pauley P, Xu L, Merkova D, Gelaro R (2012b) The impact of satellite atmospheric motion vectors in the U.S. navy global data assimilation system. Presented at the Fifth WMO workshop on the impact of various observing systems on numerical weather prediction. 22–25 May 2012, Sedona, AZ (USA). https://www.wmo.int/pages/prog/www/OSY/Meetings/NWP5_Sedona2012/1a6_Baker.pdf

Chua B, Xu L, Rosmond T, Zaron E (2009) Preconditioning representer-based variational data assimilation systems: application to NAVDAS-AR. Oceanic and Hydrologic Applications, Springer-Verlag, Data Assimilation for Atmospheric, p 493

Harris B, Kelly G (2001) A satellite radiance-bias correction scheme for data assimilation. Q J R Meteorol Soc 127:1453–1468

Hogan TF, Rosmond TE (1991) The description of the navy operational global atmospheric prediction system's spectral forecast model. Mon Wea Rev 119:1786–1815

Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201

Merkova D, Gelaro R, Baker N, Pauley P, Langland R, Xu L (2012) Observation impact of satellite winds in the NASA GEOS-5 forecast system. Presented at the 11th International Winds Workshop, 20–24 February 2012, University of Auckland, New Zealand. http://cimss.ssec.wisc.edu/iwwg/iww11/talks/Session4_Merkova.pdf

Pauley P, Baker N, Langland R, Xu L, Merkova D, Gelaro R, Velden C (2012) The impact of satellite atmospheric motion vectors in the U.S. Navy global data assimilation system—the superob procedure. Extended abstract, 11th International Winds Workshop, 20–24 February 2012, University of Auckland, New Zealand. http://cimss.ssec.wisc.edu/iwwg/iww11/talks/Session4_Pauley.pdf

Peng MS, Ridout JA, Hogan TF (2004) Recent modifications of the Emanuel convective scheme in the naval operational global atmospheric prediction system. Mon Wea Rev 132:1254–1268

Rosmond T, Xu L (2006) Development of NAVDAS-AR: Non-linear formulation and outer loop tests. Tellus 58A:45–58

Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR: Formulation and initial tests of the linear problem. Tellus 57A:546–559

Xu L, Langland R, Baker N, Rosmond T (2006) Development of the NRL 4D-Var data assimilation adjoint system. Geophys Res Abs 8:8773

Modelling the Background Error Covariance Matrix: Applicability Over the Maritime Continent

Joshua Chun Kwang Lee and Xiang-Yu Huang

Abstract The background error covariance matrix is fundamental to any data assimilation system. Since it cannot be explicitly specified, methods have been developed to estimate and model it. These involve certain assumptions which may be invalid over the Maritime Continent. In this chapter, the applicability of the main methods employed to estimate the background covariance matrix and the validity of the main assumptions in modelling it are explored, particularly for the Maritime Continent context. A brief demonstration of the methods over the region, where applicable, is provided to explore possible limitations in their conceptualisation. The manifestation of the main assumptions in the structures of the background error covariance matrix is also demonstrated using pseudo-single observation experiments. Additional comments are included to highlight areas for further work and echo the call for much needed research on modelling the background error covariance matrix for the Maritime Continent.

1 Introduction

One of the most important components of any data assimilation system is the background error covariance matrix. It controls the weightage between the background information and observational information. It also determines how observational information may be spread spatially to other variables. However, the background error covariance matrix is not trivial to specify. First and foremost, it is computationally unfeasible to explicitly specify such a large matrix (Bannister 2008a). Consider a limited-area numerical weather prediction (NWP) system with a square horizontal domain of 500 by 500 gridpoints and 80 vertical levels, along with six prognostic model variables. The background error covariance matrix would then have to be

J. C. K. Lee (✉) · X.-Y. Huang
Centre for Climate Research Singapore, 36 Kim Chuan Road, Singapore 537054, Singapore
e-mail: joshua_lee@nea.gov.sg

X.-Y. Huang
e-mail: xyhuang@ium.cn

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_23

populated with ~10^{16} (10 quadrillion!) elements, which is impractical to store even if the number is halved by taking advantage of matrix symmetry. Additionally, it is also impossible to accurately prescribe the background errors of the NWP system because the "true state" is unknown. Therefore, it must be estimated and modelled (Bannister 2008a).

Most estimation and modelling methods are conceived whilst involving multiple assumptions (Bannister 2008b), some of which may be invalid in the Tropics. This is particularly pertinent for the Maritime Continent, where the interaction of different weather phenomena and substantial topography can further complicate any attempt to accurately estimate and model the background errors. The application of some of the estimation methods may also require a dense observation network, which is absent in this region.

These limitations have induced certain questions: (i) How applicable are the methods employed to estimate the background errors over the Maritime Continent? (ii) How valid are the main assumptions applied in the context of the Maritime Continent? An overview of the main estimation and modelling methods and exploration of these key questions are provided with the aid of a convective-scale NWP data assimilation system over the Maritime Continent (SINGV-DA; Huang et al. 2019; Heng et al. 2020). Possible future research directions and suggestions are included, highlighting the need for research in this specific area.

2 Estimation of the Background Errors

In the initial "calibration step", one may employ different methods to estimate the background error covariance matrix (**B**) of the NWP system. These are often based on either existing available information such as innovation statistics, or a surrogate quantity which have error statistics that are arguably similar to the background errors.

2.1 Innovations Method

Innovation (observation-minus-background) statistics are often used to estimate the background errors without requiring knowledge of the "true state" (Rutherford 1972). This has been implemented in many previous studies (Hollingsworth and Lönnberg 1986; Lönnberg and Hollingsworth 1986; Järvinen 2001; Xu and Wei 2001, 2002; Xu et al. 2001, 2007), often considered the most accurate method for estimating background errors (Xu et al. 2001) but is subject to key assumptions. The essence of this method relies on disentangling observation errors and background errors from the innovation statistics.

To illustrate this, consider background errors and observation errors given by:

$$\boldsymbol{\epsilon_b} = \mathbf{x_b} - \mathbf{x_t} \tag{1}$$

$$\boldsymbol{\epsilon_o} = \mathbf{y_o} - H(\mathbf{x_t}) \tag{2}$$

where $\boldsymbol{\epsilon_b}$ is the background error, $\mathbf{x_b}$ is the background state vector, $\mathbf{x_t}$ is the "true state" which is unknown, $\boldsymbol{\epsilon_o}$ is the observation error (including errors in instrument and representativeness), $\mathbf{y_o}$ is the observation vector, and H is the mapping to observation space. Strictly speaking, $\boldsymbol{\epsilon_b}$ and $\mathbf{x_b}$ can be continuous fields, although they are normally represented discretely depending on the model resolution.

The unknown $\mathbf{x_t}$ is omitted by substituting Eq. 1 into Eq. 2,

$$\boldsymbol{\epsilon_o} = \mathbf{y_o} - H(\mathbf{x_b} - \boldsymbol{\epsilon_b}) \tag{3}$$

Linearising H about $\mathbf{x_b}$ (denoted $\mathbf{H}$),

$$\boldsymbol{\epsilon_o} \approx \mathbf{y_o} - H(\mathbf{x_b}) + \mathbf{H}\boldsymbol{\epsilon_b} \tag{4}$$

Re-arranging to get the innovation vector $\mathbf{d}$,

$$\mathbf{d} = \mathbf{y_o} - H(\mathbf{x_b}) \approx \boldsymbol{\epsilon_o} - \mathbf{H}\boldsymbol{\epsilon_b} \tag{5}$$

demonstrates its relationship with the background and observation errors.

For simplicity, consider an idealised scenario with direct observations of a single prognostic variable geopotential height (h) at some model gridpoints. From Eq. 5, for the model gridpoints with observations,

$$d(i, j, k) \approx \epsilon_{\mathrm{o}}(i, j, k) - \epsilon_{\mathrm{b}}(i, j, k) \tag{6}$$

where i, j, k are indexes in a 3-dimensional model grid (x, y, z). Typically, it is assumed that $\boldsymbol{\epsilon_b}$ and $\boldsymbol{\epsilon_o}$ are unbiased, independent, and $\boldsymbol{\epsilon_o}$ is uncorrelated horizontally in space and time,

$$\langle \boldsymbol{\epsilon_b} \rangle = 0, \quad \langle \boldsymbol{\epsilon_o} \rangle = 0 \tag{7}$$

$$\langle \epsilon_{\mathrm{o}}(i,\ j,\ k)\ \epsilon_{\mathrm{b}}(l, m, n) \rangle = 0 \quad \forall i,\ j,\ k,\ l,\ m,\ n \tag{8}$$

$$\langle \epsilon_{\mathrm{o}}(i,\ j,\ k)\ \epsilon_{\mathrm{o}}(l, m, n) \rangle = 0 \quad \text{for } k = n \text{ and } \forall i,\ j,\ l,\ m \tag{9}$$

where l, m, n also denote indexes in a 3-dimensional model grid; $k = n$ denotes the same vertical level, and $\langle . \rangle$ denotes time-averaging (the assumption of ergodicity).

Using Eqs. 6 to 9, the background error autocovariance (any covariance within the same field) matrices of h for each of the model gridpoints with observations (which are components of the full $\mathbf{B}$) can then be retrieved from the partitioning of

the innovation statistics:

$$\langle d(i,\ j,\ k)\, d(l,\ m, n)\rangle \ \approx\ \langle \epsilon_{\mathrm{o}}(i,\ j,\ k)\epsilon_{\mathrm{o}}(l,\ m,\ n)\rangle \ +\ \langle \epsilon_{\mathrm{b}}(i,\ j,\ k)\epsilon_{\mathrm{b}}(l,\ m,\ n)\rangle \tag{10}$$

On the same vertical level ($k = n$), applying Eq. 9 implies that the first term on the RHS is zero and the background error covariances are given fully by the innovation covariances, therefore allowing for the derivation of the background errors from innovation statistics. Further extension from the autocovariances, assuming the observation errors are also uncorrelated between variables, to the height-wind cross-covariances can be retrieved (Lönnberg and Hollingsworth 1986; Xu and Wei 2002). The cross-covariance with respect to a derived height background error h' at a point $(i,\ j, k)$ can be computed after scaling the derived velocity background errors u' and v' at (l, m, n) by a Coriolis term (see Eq. 2.1 of Xu and Wei 2002) and projecting onto the tangential $t'(l, m, n)$ and radial $r'(l, m, n)$ components relative to the direction from point $(i,\ j, k)$ to(l, m, n). The resulting h'-t' and h'-r' cross-covariances (see Eq. 2.5a, b of Xu and Wei 2002) are then further used to represent the height-wind cross-covariances.

Note that the above and Eq. (10) only determine the covariances between the observation locations, which in this idealised scenario coincide with model gridpoints. Without further assumptions, it cannot be applied for all model gridpoints in the domain. As innovation statistics are limited by the availability of observations, the crude assumptions of isotropic and homogeneous background errors are required for the method to be viable in an operational setting. These assumptions are discussed in Sect. 3.

In the context of the Maritime Continent, such a method would be challenging to apply. In the current operational model at the Meteorological Service Singapore (MSS), SINGV-DA only assimilates around 10 radiosondes from the World Meteorological Organization (WMO) Global Telecommunication System (GTS) consistently throughout the entire domain (Heng et al. 2020). The lack of radiosonde observations in the region presents a significant hurdle since the innovations method requires sufficient *in situ* observations of all quantities in the background state for the statistics to be sufficiently robust. While this method has been discussed, it is not practical to apply and evaluate it in SINGV-DA. Perhaps with increased data sharing efforts and more intensive observation period experiments conducted, such a method could become viable for exploratory studies in the future.

2.2 *'National Meteorological Center' (NMC) Method*

The NMC method relies on computing forecast differences between forecasts of varying lengths but valid at the same time and using them as surrogates to estimate the background error. Parrish and Derber (1992) first incorporated this in the NMC's spectral statistical-interpolation analysis system, computing the difference between

a 24-h forecast and an initialised analysis (0-h forecast) valid at the same time. The forecast differences were compiled over an extended period of time to produce the climatological sample needed to estimate the background error statistics. This method has proved popular, with many studies and operational centres previously adopting this approach to model background error covariances for their data assimilation systems (Rabier et al. 1998; Derber and Bouttier 1999; Lorenc et al. 2000; Ingleby 2001; Barker et al. 2004, 2012; Rawlins et al. 2007; Huang et al. 2009; Buehner et al. 2010; Heng et al. 2020; Lee and Huang 2020), and also for the generation of reanalyses data (Onogi et al. 2007; Kobayashi et al. 2015).

Mathematically, the estimate of the background error can be illustrated by considering the difference between a 36-h forecast ($\mathbf{x}_{36}$) and a 12-h forecast ($\mathbf{x}_{12}$) which are valid at the same time, initialised from analyses 24 h apart to eliminate the errors in modelling the diurnal cycle. The choice of forecast length is not fixed and other alternatives are viable (Bannister 2008a). We consider the background errors ($\boldsymbol{\epsilon}_{36}$ and $\boldsymbol{\epsilon}_{12}$) associated with $\mathbf{x}_{36}$ and $\mathbf{x}_{12}$ respectively without the effect of model bias, given by:

$$\boldsymbol{\epsilon}_{36} = \mathbf{x}_{36} - \mathbf{x}_{\mathbf{t}} \Rightarrow \mathbf{x}_{36} = \boldsymbol{\epsilon}_{36} + \mathbf{x}_{\mathbf{t}} \tag{11}$$

$$\boldsymbol{\epsilon}_{12} = \mathbf{x}_{12} - \mathbf{x}_{\mathbf{t}} \Rightarrow \mathbf{x}_{12} = \boldsymbol{\epsilon}_{12} + \mathbf{x}_{\mathbf{t}} \tag{12}$$

As before, the unknown $\mathbf{x}_{\mathbf{t}}$ is omitted by considering the difference between the forecasts, using Eqs. 11 to 12,

$$\mathbf{x}_{36} - \mathbf{x}_{12} = \boldsymbol{\epsilon}_{36} - \boldsymbol{\epsilon}_{12} \tag{13}$$

$\mathbf{B}$ can then be computed over many realisations (typically a series of forecast differences collated over an extended period of time). Previous studies have used a sample size of around 30 to 60 to compute the statistics (Parrish and Derber 1992; Derber and Bouttier 1999; Wu et al. 2002; Chen et al. 2013; Ban et al. 2017; Lee and Huang 2020), although as with any statistical estimation, a larger sample is often desirable. Over many realisations with the sample mean removed, the covariances can be computed:

$$\langle (\mathbf{x}_{36} - \mathbf{x}_{12})(\mathbf{x}_{36} - \mathbf{x}_{12})^{\mathrm{T}} \rangle = \langle (\boldsymbol{\epsilon}_{36} - \boldsymbol{\epsilon}_{12})(\boldsymbol{\epsilon}_{36} - \boldsymbol{\epsilon}_{12})^{\mathrm{T}} \rangle \tag{14}$$

Expanding the RHS, this gives:

$$\langle (\mathbf{x}_{36} - \mathbf{x}_{12})(\mathbf{x}_{36} - \mathbf{x}_{12})^{\mathrm{T}} \rangle = \langle \boldsymbol{\epsilon}_{36}\boldsymbol{\epsilon}_{36}^{\mathrm{T}} \rangle - \langle \boldsymbol{\epsilon}_{36}\boldsymbol{\epsilon}_{12}^{\mathrm{T}} \rangle - \langle \boldsymbol{\epsilon}_{12}\boldsymbol{\epsilon}_{36}^{\mathrm{T}} \rangle + \langle \boldsymbol{\epsilon}_{12}\boldsymbol{\epsilon}_{12}^{\mathrm{T}} \rangle \tag{15}$$

Often, one crudely assumes that the background error covariances are similar for $\boldsymbol{\epsilon}_{12}$ and $\boldsymbol{\epsilon}_{36}$, implying that:

$$\langle \boldsymbol{\epsilon}_{36}\boldsymbol{\epsilon}_{36}^{\mathrm{T}} \rangle = \langle \boldsymbol{\epsilon}_{12}\boldsymbol{\epsilon}_{12}^{\mathrm{T}} \rangle = \mathbf{B} \tag{16}$$

Using Eqs. 14 to 16, and rearranging, this gives:

$$\mathbf{B} = \tfrac{1}{2}\{\langle(\mathbf{x}_{36} - \mathbf{x}_{12})(\mathbf{x}_{36} - \mathbf{x}_{12})^{\mathrm{T}}\rangle + \langle\boldsymbol{\epsilon}_{36}\boldsymbol{\epsilon}_{12}^{\mathrm{T}}\rangle + \langle\boldsymbol{\epsilon}_{12}\boldsymbol{\epsilon}_{36}^{\mathrm{T}}\rangle\} \tag{17}$$

Furthermore, it is often assumed that the background errors $\boldsymbol{\epsilon}_{12}$ and $\boldsymbol{\epsilon}_{36}$ are uncorrelated,

$$\langle\boldsymbol{\epsilon}_{36}\boldsymbol{\epsilon}_{12}^{\mathrm{T}}\rangle = 0, \quad \langle\boldsymbol{\epsilon}_{12}\boldsymbol{\epsilon}_{36}^{\mathrm{T}}\rangle = 0 \tag{18}$$

Strictly speaking, this is unlikely to be valid, leading to non-negligible positive contributions as a result of the correlation and therefore a possible underestimation of **B**. A crude scaling factor may then be applied to inflate or tune **B** depending on the system. Even though one is able to prescribe **B** using this method explicitly, it is still computationally impractical to store and further assumptions of isotropy and homogeneity must be made to simplify it. These are discussed in Sect. 3.

Following the development of limited area models (LAMs), an alternative formulation of the NMC method has been developed to handle the errors originating from the lateral boundary conditions (LBCs) accordingly, often termed the "lagged" NMC method as opposed to the standard NMC method described above. This was initially employed in Široká et al. (2003). They found that the exclusion of error sources from the driving model (DM), which provides the lateral boundary conditions, resulted in sharper structures in the analysis increments corresponding to mesoscale features. This was due to sharper correlation structures in the background error covariances, which was deemed as beneficial for mesoscale NWP systems. Further studies have also explored using this method over other LAM domains (Fischer et al. 2005; Sadiki and Fischer 2005; Berre et al. 2006; Bölöni 2006; Montmerle et al. 2006; Stanesic et al. 2019; Lee and Huang 2020).

The approach to deriving the background errors is similar to the standard NMC method; the state is propagated by different non-linear forecast models (LAM and DM) for the two forecasts of varying length. The main difference between the "lagged" and standard NMC method lies in the initial conditions (ICs) and LBCs driving the shorter forecast, as well as the observational information that is introduced by the cycling DM prior to the initialisation of the shorter forecast. As a simple illustration, Fig. 1 shows a 12-hourly cycling DM with the same cycle run (DM_initial) and latest cycle run (DM_latest) providing the ICs and LBCs for the initialisation of the shorter forecast in the "lagged" and standard NMC method respectively. For brevity, the mathematical notations for the full non-linear forecast models are not described.

Despite the wide acceptance of the NMC method, it is surprising that there have been little theoretical justification and formalism on its validity. Bouttier (1996) noted that in its early conception, it was only tested against the innovations method in a few cases. Since then, very few studies have attempted to validate it. Horvath and Bölöni (2004) compared the standard and "lagged" NMC methods with the innovations method in the Hungarian version of the Aire Limitée Adaptation dynamique Développement InterNational (ALADIN/HU) model and found that all three methods

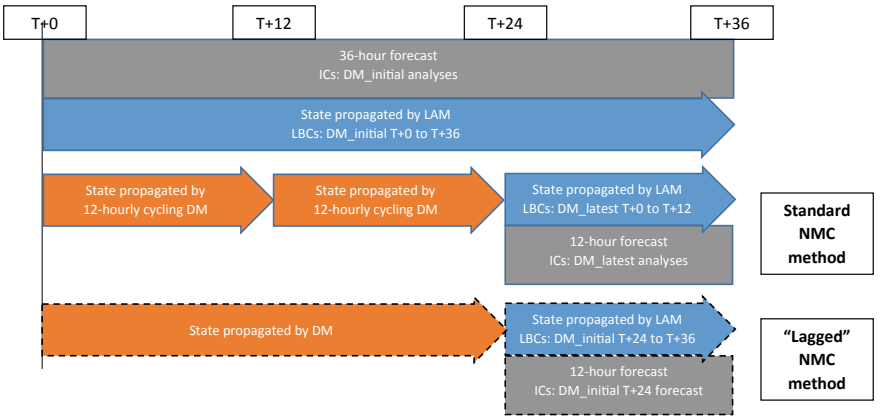

Fig. 1 Simple schematic to show the difference between the standard NMC method and "lagged" NMC method illustrated in an LAM. See text for abbreviations

had fairly comparable temperature autocovariances especially near to the surface, but had large differences in the specific humidity autocovariances near the surface and in the upper troposphere. In particular, the "lagged" NMC method exhibited an underestimation of both the temperature and specific humidity autocovariances compared to the innovations method. This is consistent with Sadiki and Fischer (2005), who hypothesised that the "lagged" NMC method likely underestimates the variances. In a different approach, Errico et al. (2015) used an observation system simulation experiment to validate the variance estimates from the untuned standard NMC method and found that the rescaling of variances and horizontal correlation length-scales depend both on pressure level and latitude. Particularly in the Tropics, the variances for temperature, specific humidity, zonal and meridional wind were all underestimated by the standard NMC method. It is unsurprising that many studies have artificially inflated and deflated **B** or altered the corresponding correlation length-scales derived using the NMC method (Barker et al. 2004; Chapnik et al. 2004; Gu et al. 2005; Kleist et al. 2009; Ha and Lee 2012; Liu et al. 2015). Often, the optimal tuning factors are only empirically determined.

Given the many assumptions in the NMC method, it is possible that some are less valid in the Tropics. Rabier et al. (1998) noted that the NMC method is prone to breaking down in the Tropics where the large-scale flow has little temporal variability. A similar argument is presented in Žagar et al. (2005), who noted that in a data-sparse region such as the Tropics, the two forecasts of varying lengths may tend to be very similar and thus the NMC method may underestimate the variances. Therefore, while it is a useful method to apply in a variational data assimilation system for a data-sparse area such as the Maritime Continent, the derived **B** must be interpreted with caution. Nevertheless, it is worth examining and more discussion is provided in Sect. 4.

2.3 Ensemble Methods

An ensemble of forecasts can be used to generate a set of random samples for the background state probability distribution, with **B** often defined around the ensemble mean as an estimate for the "true state" which is unknown. The approach is essentially a Monte Carlo method that was highlighted by Epstein (1969), who noted that the deterministic trajectories from a sample of discrete initial points in phase space (of all possible state vectors) could be used to compute and estimate the mean and variances of the background probability distribution. Evensen (1994) first implemented this approach, which became commonly known as the ensemble-based Kalman filter or ensemble Kalman filter (EnKF), to estimate the background errors statistics in a two-layer non-linear quasi-geostrophic ocean model. Since then, numerous studies have employed ensemble-based methods. A review of the basic EnKF method is covered in Evensen (2003). Many other modifications of the EnKF method and variations in methods to generate an ensemble have also been developed (see Hamill 2006; Ehrendorfer 2007; and references therein). A comprehensive review of the most recent developments and outstanding challenges, as well as operational implementations are presented in Houtekamer and Zhang (2016).

The ensemble of forecasts can be generated by various approaches, through a separate parallel-run ensemble system using either stochastic EnKF or deterministic filters (such as those in Anderson 2001; Bishop et al. 2001; Whitaker and Hamill 2002; Tippett et al. 2003) or through the introduction of perturbations into an existing forecast system to generate an ensemble of analyses as initial conditions for the forecasts (Houtekamer et al. 1996; Fischer 2003; Berre et al. 2006). Buehner (2005) compares both approaches with the NMC method. There is an added complexity for ensemble-based methods in LAMs (see Meng and Zhang 2011; and references therein) since additional sources of errors from the LBCs also have to be accounted for. A widely used method to generate the ensemble for LAMs is to perform a dynamical downscaling from a global ensemble (Hohenegger et al. 2008; Bölöni et al. 2015; Bojarova and Gustafsson 2019). The errors in the LBCs and analysis in the LAM are naturally derived from the global ensemble after interpolation. While it is simple to implement, this may exclude meaningful smaller scale structures which influence the ensemble spread and thus lead to an underestimation of the forecast ensemble variance. For the estimation of **B**, the main focus is on the usage of the ensemble forecasts, irrespective of the ensemble generation approach.

There are two main strategies to estimate **B** using an ensemble of forecasts with N members. One strategy is to approximate the "true state" using the ensemble mean. The estimate of the background errors and the spread of the errors which define **B** are given by:

$$\boldsymbol{\epsilon}_{\mathbf{b}}^{i} = \mathbf{x}_{\mathbf{b}}^{i} - \mathbf{x}_{\mathbf{t}} \approx \mathbf{x}_{\mathbf{b}}^{i} - \bar{\mathbf{x}}_{\mathbf{b}} \tag{19}$$

$$\mathbf{B} = \frac{1}{N-1} \sum_{i=1}^{N} \boldsymbol{\epsilon}_{\mathbf{b}}^{i} \boldsymbol{\epsilon}_{\mathbf{b}}^{i\,\mathrm{T}} \tag{20}$$

where $\bar{\mathbf{x}}_\mathbf{b} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_\mathbf{b}^i$ is the ensemble mean and index i represents each ensemble member, $\epsilon_\mathbf{b}^i$ is also often referred to as error modes. This strategy was suggested in Evensen (1994) and was widely implemented in many recent studies for both global models and LAMs (e.g. Wang et al. 2008; Storto and Randriamampianina 2010; Clayton et al. 2013; Gustafsson et al. 2014; Montmerle et al. 2018; Caron et al. 2019).

The second strategy involves considering the ensemble perturbations about a central forecast. The approach is similar to the NMC method for computing the forecast differences (Eqs. 11 to 17) but computing the differences between forecasts from each ensemble member and a central forecast instead. For clarity, the estimate of the background errors and **B** derived from the ensemble is thus given by:

$$\boldsymbol{\epsilon}_\mathbf{b}^i \approx \mathbf{x}_\mathbf{b}^i - \mathbf{x}_\mathbf{b}^c \tag{21}$$

$$\mathbf{B} = \frac{1}{2}\left\{\frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{\epsilon}_\mathbf{b}^i\boldsymbol{\epsilon}_\mathbf{b}^{i\,\mathrm{T}}\right\} \tag{22}$$

where $\mathbf{x}_\mathbf{b}^c$ is the central forecast (which is usually the unperturbed ensemble forecast). In this formulation, there is no proxy used for $\mathbf{x}_\mathbf{t}$ and it is assumed that the ensemble perturbations between members are uncorrelated, which is the reason for the additional factor of half in Eq. 22 (Berre et al. 2006). As highlighted by Bölöni et al. (2015), the same central forecast is used for the computation, resulting in $N-1$ pairs. This strategy was also mentioned in Evensen (1994), and has been applied in previous studies (Fischer 2003; Berre et al. 2006; Bölöni et al. 2015). For both strategies, additional time-averaging (as denoted by $\langle . \rangle$) may not necessarily be applied unlike in the NMC method, although it is still possible (e.g. Buehner 2005; Storto and Randriamampianina 2010; Varella et al. 2010; Bölöni et al. 2015).

Evensen (1994) discusses how both strategies may be equally suitable in NWP applications. It was noted that the structures of the central forecast and ensemble mean could be similar, although the ensemble mean may reflect a lower amplitude in the structures than the central forecast. The more appropriate method could depend on the data availability in the region and the size of the ensemble. For operational implementation of ensemble methods, there also exist many variations (see Bannister 2017; and references therein).

The application of the ensemble method is particularly attractive for the Tropics, due to the possibility of incorporating flow-dependence in the estimation of **B** (Sect. 3.3). Previous studies have provided evidence of the importance of flow-dependent covariances for data-sparse regions (Hamill and Snyder 2000; Whitaker et al. 2004, 2008), such as over the Maritime Continent. However, no study has explored the flow-dependence of **B** over the Maritime Continent. A brief illustration of the application of the ensemble method and the resulting flow-dependent structures over the Maritime Continent are thus presented in Sect. 4. It would be useful for future studies to investigate them further.

3 Modelling the Background Error Covariance Matrix

Apart from estimating the background errors, one also needs to consider how to model and prescribe a computationally feasible **B** in a data assimilation system. For variational data assimilation, pre-conditioning of a cost function is required. This involves a series of transforms (see Bannister 2008b) which determine the square root of **B**. The design of these transforms may invoke further assumptions which reduce the elements required to specify such a large matrix.

The structures of **B** can be illustrated using pseudo-single observation experiments (Derber and Bouttier 1999; Huang et al. 2009), which provide a visual demonstration of how the main assumptions manifest themselves in the transforms. For reference, the solution for the pseudo-single observation (Gustafsson et al. 2012) can be written as:

$$\delta \mathbf{x} = \mathbf{B}_i (\mathbf{B}_{ii} + \mathbf{R}_{ii})^{-1} \mathbf{d}_i = \mathbf{B}_i \left(\sigma_\mathrm{b}^2 + \sigma_\mathrm{o}^2\right)^{-1} d \tag{23}$$

where $\delta \mathbf{x}$ is the analysis increments, **R** is the observation error covariance matrix, σ_b, σ_o and d are the background error standard deviation, observation error and innovation at gridpoint i respectively.

3.1 Homogeneity

The assumption of homogeneous background error statistics is often applied for the innovations, NMC and sometimes the ensemble method. Typically, to retrieve the horizontal autocorrelations for each model level, one must consider the correlation between the entire two-dimensional model grid to a specific gridpoint, repeated for each gridpoint in the domain. For the innovations method, one can only compute the correlations between observation locations. However, invoking this assumption, one now only needs to consider how (on average) the same correlation structure can be used for each gridpoint or observation location in the entire domain.

The definition of homogeneity (Gaspari and Cohn 1999) applied on a two-dimensional model domain D is as follows. Consider a correlation function C applied on two gridpoints ($\mathbf{x}$) and ($\mathbf{y}$) in D, if $\forall T$ (a horizontal translation function) of D,

$$C(T(\mathbf{x}), T(\mathbf{y})) = C(\mathbf{x}, \mathbf{y}) \tag{24}$$

holds, then the correlation function is homogeneous throughout D. Often, the domain-averaged statistics are computed during the "calibration step" to be applied for the entire domain. One may pose the question: "What is the correlation function that best represents that for the entire domain?"

To briefly illustrate homogeneity, pseudo-single pressure observations 100 Pa above the background ($d = 100$) are inserted near the surface at model level 3 with an observation error of 20 Pa ($\sigma_o = 20$) at four different points in the SINGV-DA domain and the resulting pressure increments for the four experiments are presented (Fig. 2). The classification of the points correspond to coastal lowland, coastal ocean, open ocean as defined by Birch et al. (2016) and coastal highland, located along the Sumatran mountain range. It is clear that the components of the correlation function do not exhibit spatial variation and are thus homogeneous on each model level (although only shown for one model level here). Other interesting features in Fig. 2, such as wind increments, are discussed in Sect. 4.1.

Wu et al. (2002) demonstrated the importance of accounting for the latitudinal dependence of the background error statistics, especially in the Tropics. Chen et al. (2014) explored its impacts on typhoon forecasts. Apart from simply computing the domain-averaged statistics, one may instead choose to split the domain into latitudinal bands and perform zonal averaging within each band. The homogeneity would then be confined over each latitudinal band. However, this would still be a crude assumption, which is likely invalid over the Maritime Continent given the spatial variation in orography and land-sea contrast from the cornucopia of small islands. For convective-scales, the interaction of weather phenomena of different spatial scales over the region also facilitates the inclusion of substantial inhomogeneity in the background error statistics. For example, the error structures which are physically consistent with a land-sea breeze may not be homogeneous in the domain but are aligned parallel to the coasts instead. The error structures associated with monsoon surges over open ocean are likely to be different from the error structures over coastal lowland where speed convergence may occur. These expectations are consistent with Michel and Auligné (2010), who showed that the inhomogeneity in the background error statistics over Antarctica can be related to orography, as well as land-sea contrasts.

Also, while the discussion has been mainly on the horizontal correlations, perhaps another important consideration is that the vertical correlations are often also subject to homogeneity; the vertical transforms do not vary with position. However, in reality, the vertical covariances of the background errors are strongly dependent on position (Bannister 2008b). Vetra-Carvalho et al. (2012) demonstrated how hydrostatic balance in the vertical covariances can break down in regions of convection. Montmerle and Berre (2010) also showed how the vertical covariances may vary between precipitating and non-precipitating regions. Given that convection, and thus precipitation, is abundant in the Maritime Continent, appropriate methods for modelling **B** must therefore allow for inhomogeneity.

3.2 Isotropy

The assumption of isotropy is closely linked to the assumption of homogeneity and is also often applied for the innovations, NMC and sometimes the ensemble method. The question to be posed is: "How should one specify the general correlation

function?" While there are various methods one may construct, isotropy is often assumed for simplicity and convenience.

The definition of isotropy (Gaspari and Cohn 1999) applied on a two-dimensional model domain D is as follows. Consider a correlation function C applied on two gridpoints ($\mathbf{x}$) and ($\mathbf{y}$) in D, if $\forall V$ (orthogonal horizontal transformation function) of D,

$$C(V(\mathbf{x}), V(\mathbf{y})) = C(\mathbf{x}, \mathbf{y}) \tag{25}$$

holds, then the correlation function is isotropic on D. Simply put, if the correlation function does not change due to rotation and reflection (and thus is solely a function of the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$), then it is isotropic.

A simple isotropic (Gaussian) correlation function that has been used frequently (Daley 1991) is defined as:

$$C(r) = e^{-\frac{r^2}{2L^2}} \tag{26}$$

where r is the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$, and L is a constant length-scale which may be computed during an offline "calibration step". Other families of correlation functions may also be used, such as those derived from stochastic autoregressive process representations (see Thiébaux et al. 1986). For example, the default option in the Met Office setup uses the second order autoregressive (SOAR) function (Lorenc et al. 2000) following Thiébaux et al. (1986), given by:

$$C(r) = (1 + \tfrac{|r|}{s})e^{-\frac{|r|}{L}} \tag{27}$$

in the offline "calibration step". The SOAR is used to fit the distance-binned covariances computed during the "calibration step" to determine the optimal values of L to be eventually used in $\mathbf{B}$ (for the horizontal transform). Strictly speaking, a spectral representation of SOAR is used in the horizontal transform. This is different from the application of recursive filters (Purser et al. 2003), such as in Barker et al. (2003) which allows for the inclusion of anisotropy (Barker et al. 2004). An example is shown in Lorenc (1997). Figure 2 illustrates the isotropy prescribed using SOAR. The circular shape of the pressure increments related to the horizontal background error autocorrelations are solely a function of distance from the point of the pseudo-observation. Thus, the increments are invariant with respect to orientation.

Other possible methods for modelling the correlations include the application of spectral convolutions (Courtier et al. 1998) and diffusion operators which are derived from the integral solution of diffusion equations (Weaver and Courtier 2001). The latter is more relevant for application in ocean models as it can account for boundary conditions along the coast which may be difficult to treat using other methods. Apart from these, the use of wavelets to represent local correlations has also been proposed

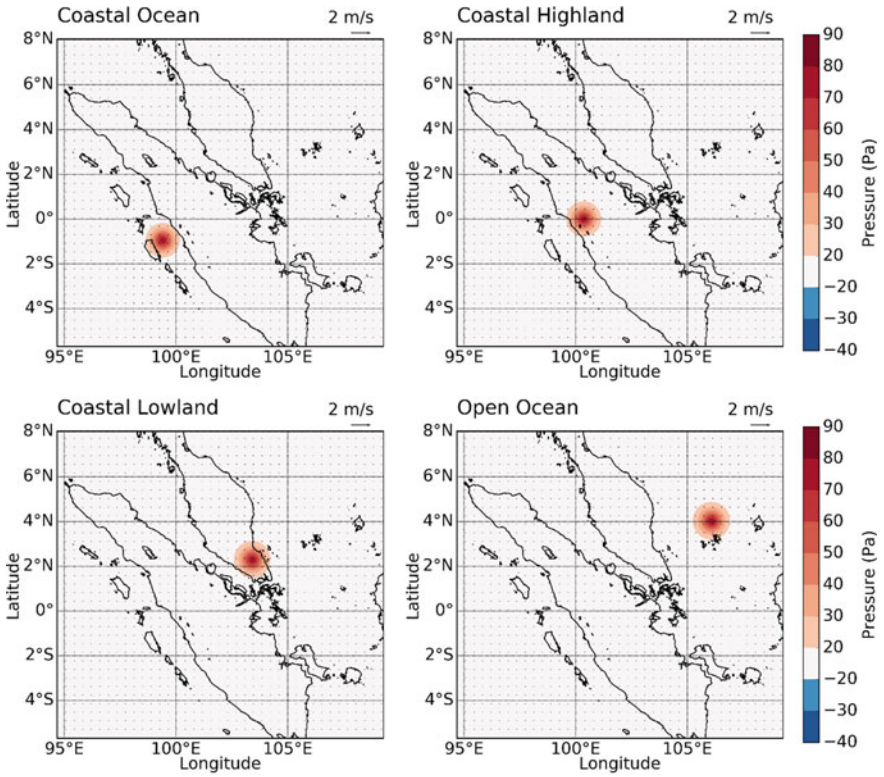

Fig. 2 Pressure and horizontal wind increments as a result of a pseudo-single observation of pressure with respect to four classification points in the domain, related to the associated modelled background error autocovariances and cross-covariances (see Eq. 23) using the "lagged" NMC method. Training data comprise 56 pairs of forecast differences in September 2018

(Fisher and Andersson 2001). Wavelets possess information on both geographical position and scale; a scheme can be carefully designed to allow for the inclusion of inhomogeneity and anisotropy. This has recently been applied in both global models and LAMs (Fisher 2003; Deckmyn and Berre 2005; Bannister 2007; Varella et al. 2011; Berre et al. 2015). There exist some variety in the choice of wavelet functions and its applications. The conventional method involves the use of non-orthogonal wavelets to capture the spatial and spectral variation in the correlations, as demonstrated in Fisher (2003) and Bannister (2007). Bannister (2007) further proposed another wavelet-based method as an approximation to the conventional method. Deckmyn and Berre (2005) explored the use of orthogonal wavelets instead. They also discuss the possible representation of local anisotropies with the wavelet formulation.

The extent to which isotropy holds in the Maritime Continent requires further investigation. The length-scales calibrated for the correlation functions need to be appropriate for the scale of the phenomena, and hence background errors, expected

over the region. One might argue that for the convective-scale, such length-scales should be short so that the increments are relatively localised around the observations. However, with shorter length-scales, larger scale information due to the passage of equatorially-trapped waves or the Madden–Julian Oscillation (MJO) could be omitted simply because the observations in the region are too sparse. With elongated coastlines and oblong-shaped islands in the domain, prescribing isotropic correlations in the background error statistics also appears particularly unrealistic. Sattler and Huang (2002) previously highlighted how the presence of a land-sea contrast can influence the validity of the isotropic assumption in the background error statistics over Western Europe. This could be relevant in the Maritime Continent as well.

3.3 Time Stationarity

In variational data assimilation, time stationarity is another assumption usually applied in the modelling of**B**. One may be interested in the question: "How does **B** change over time?" **B** should ideally be updated constantly in order to represent these changes. This may be subject to computational constraints. As highlighted in Sect. 2, the assumption of ergodicity is often invoked; the time-averaged (climatological) statistics are often used to estimate the real background errors using the innovations, NMC, and sometimes ensemble methods (e.g. Buehner 2005; Storto and Randriamampianina 2010; Varella et al. 2011; Bölöni et al. 2015).

By definition, time stationarity simply means that the estimated background error statistics do not change with time. The same **B** is used for each cycle in a data assimilation system, often because it may be unfeasible to update such a large matrix at every cycle. This is mainly applicable for traditional 3D-Var algorithms. Even for 4D-Var algorithms, **B** is evolved by the linear model within the assimilation window, but is reset at the beginning of the next window. Only in ensemble methods (e.g. EnKF) can this assumption be relaxed, since it is possible to update the background errors using the analysis errors from previous windows or from the ensemble state trajectories.

Ideally, the background errors should be dynamically consistent with the flow conditions in the system, which is unlikely to be constant, giving rise to the term "flow-dependence". For the Maritime Continent, apart from the possible presence of inhomogeneity and anisotropy in the real background errors discussed in Sect. 3.1 and 3.2, additional complexities arise due to their time dependence. On the seasonal timescale, the intra-annual monsoon migration could play a dominant role. For example, the error structures during the Southwest monsoon (around June to September for Singapore) are likely to be different from the Northeast monsoon (around November to February for Singapore). The prior is influenced by warmer and drier air originating from Australia and more frequent squall lines while the latter is influenced by moister and cooler air originating from the South China Sea and possible cold surges. Even within the Maritime Continent, the diagnosed inter-monsoon periods differ depending on latitude due to the intra-annual migration of

the Inter Tropical Convergence Zone. On the sub-seasonal timescale, the passage of equatorial waves and the MJO also determine the flow conditions in the system and thus can further conflate the resulting error structures. On longer timescales, the Quasi-Biennial Oscillation can also influence the tropical background errors (Žagar et al. 2007). It is thus unsurprising that the assumption of time stationarity is often invalid when applied to the Maritime Continent, especially on convective-scales.

At present, the most advanced operational systems have incorporated flow-dependence in the background error statistics and have yielded promising results (Bonavita et al. 2011; Clayton et al. 2013; Montmerle et al. 2018; Caron et al. 2019).

3.4 Multivariate Balance Relationships

In variational data assimilation, observational information from one variable (e.g. wind) may be spread by the cross-covariances of **B** to other variables (e.g. mass). It is assumed that there exist some statistical form of balance in the estimated background errors of the prognostic variables, which originate from balance in the real atmosphere. Therefore, one might consider how balance constraints (e.g. geostrophic balance) can be exploited to increase the value of observational information while ensuring dynamically consistent increments to the system (e.g. as illustrated in Derber and Bouttier 1999; Kalnay 2003; Bannister 2008a). Often, a transformation (variable transform) to a set of variables (also referred to as control variables) which implicitly captures the multivariate aspects of the background errors is chosen. This transformation is designed based on prescribed balances between prognostic variables, to control variables which are assumed to be uncorrelated to improve the conditioning of the cost function. Regardless of the transformation, it is important that the resulting implied cross-covariances of the prognostic variables should attempt to represent the cross-covariances of the real background errors, which is not known.

Common approaches adopted by most operational systems include the use of linear regression and the linear balance equation as mass-wind balance relations in the variable transform (see Table 1 of Bannister 2008b for a detailed breakdown). However, linear balance breaks down near the Equator where geostrophy is vanishing. Even in the mid-latitudes, geostrophy becomes less important approaching convective-scales (Berre 2000). These may have implications on the implied mass-wind coupling and covariance structures. Daley (1996) demonstrated how any mass-wind coupling could be a fallacious result of using the linear balance equation for multivariate covariances in the Tropics. One would not expect the assimilation of a mass observation to yield information about the wind field. However, given that the large-scale variability in the dynamical fields over the Tropics may be explained by convectively coupled equatorial waves (Wheeler et al. 2000), it is possible that the covariance structures may project on and resemble these dominant modes which exhibit mass-wind coupling.

Instead of applying a variable transform, Žagar et al. (2004) explored the use of tropical eigenmodes which describe balanced tropical motion to model **B**. They

demonstrated the importance of Kelvin waves and mixed Rossby-gravity waves for the horizontal structure of increments and their effect on mass-wind coupling in the Tropics. Žagar et al. (2005) further showed that the mass-wind coupling in the ECMWF short-range forecast errors statistics projected strongly onto the equatorial Kelvin wave structure. Ingleby (2001) found that the mass-wind coupling in the Tropics was negligible using the Met Office variational data assimilation system which applies the linear balance equation in the variable transform. Using a convective-scale tropical configuration of the same system, Lee and Huang (2020) found that the mass-wind coupling could depend on the training data used to generate different covariances. Some covariances had negligible mass-wind coupling while others had structures that resembled an equatorial Kelvin wave. Using the Weather Research and Forecasting data assimilation (WRFDA) system which applies linear regression in the variable transform, Chen et al. (2013) found that the mass-wind coupling yielded a collocation of the divergent wind field with an inserted mass perturbation. This was also noted in Derber and Bouttier (1999) using the early ECMWF system which also applies linear regression in the variable transform, although the magnitude of the divergent wind vectors appear negligible (~0.06 m s^{-1} resulting from a geopotential height d of 10 m). As the real background errors and their covariances cannot be known, it is difficult to validate the implied mass-wind coupling structures from the application of a variable transform or other methods. However, it is worth noting that these error structures are likely determined by the dominant evolving modes in the Tropics.

From the diagnosis of linear regression coefficients, Chen et al. (2013) provided evidence that the (non-negligible) contribution to the balanced mass field came mainly from the divergent part of the wind field in the Tropics, thereby highlighting a possible physically meaningful multivariate relationship. It is thus surprising that even with mass-wind coupling prescribed in terms of equatorial waves in the background error covariance model, Žagar et al. (2008) showed that the analysis was still essentially univariate. As for the linear balance equation, a key criticism is that it excludes the divergent part of the wind field, and thus barely relates the mass and wind field close to the equator. The background error covariance model becomes effectively univariate; the mass control variables (e.g. unbalanced pressure in the Met Office system) yields little information about the wind field and only recover the prognostic mass field. Worse still, any little wind information yielded could contaminate the analysis, following Daley (1996). Hence, while a possible mass-wind relationship may exist, it is unsurprising that some operational centres have treated the analysis univariately for the Tropics, such as using the linear balance operator or as described in Courtier et al. (1998).

Recently, studies have gone one step further and explored taking an entirely univariate approach for modelling the background errors; the variable transform is set as the identity matrix and control variables are identical to the prognostic variables (apart from the moisture control variable). These have purportedly resulted in a higher quality analyses in convective-scale models (Li et al. 2016; Sun et al. 2016). A somewhat radical suggestion would be to test this approach over the Tropics and compare verification scores with existing formulations. However, perhaps most

importantly, effort must be placed on identifying a dominant balance relationship for the Tropics which can be exploited. In this regard, one potential candidate would be to derive a "weak temperature gradient balance" (personal communication) for mass and wind variables. This originates from the application of the weak temperature gradient approximation (Sobel et al. 2001) on the shallow water equations to retrieve an elliptic equation to solve for a mass perturbation. Key control variables would then be zonal and meridional wind, and unbalanced mass (personal communication), although the extent of their correlation has not been investigated.

As the above discussion relates broadly to the Tropics, it is directly applicable to the Maritime Continent context. The choice of control variables and the design of the variable transform with appropriate mass-wind balance relations should be carefully considered, yet it is not thoroughly explored. Further studies are much needed in this aspect.

4 Comparison of Methods and Validity of Assumptions

Apart from the discussion related to the questions raised in Sect. 1, it is perhaps more important to demonstrate the methods, where applicable, over the Maritime Continent using SINGV-DA, and comment on the validity of the assumptions raised in relation to the results.

For a start, it is impossible to know the real background errors because the "true state" is unknown. It is also impractical to apply the innovations method in SINGV-DA because of the sparsity of observations, as discussed previously. However, it is possible to compute the raw covariances of the training data used in the "lagged" NMC method and the error modes used in the ensemble method and compare them with their respective modelled **B**. For the following section, the examples presented are for single case events. Further research needs to be undertaken to generalise the results or comments.

4.1 Modelled Background Error Covariance Matrix Using the "Lagged" NMC Method

The original formulation of Lorenc et al. (2000) is applied, including the removal of the time-mean from the forecast differences. The training data comprise 56 pairs of forecast differences from the period of September 2018 which undergo a series of offline "calibration steps". Further details of the modelling process for SINGV-DA are described in Lee and Huang (2020). The resulting covariance structures are illustrated in Fig. 2, to demonstrate the assumptions of homogeneity, isotropy and time stationarity. Additionally, one may also note that for this particular covariance, the mass-wind coupling appears to be weak. Although not noticeable, the largest wind

vectors have a magnitude of around 0.1 m s^{-1}. Other covariances exhibiting stronger mass-wind coupling (Lee and Huang 2020) have wind vectors with magnitude of up to 0.3 m s^{-1}. These large wind vectors are collocated with the pseudo-single observation.

4.2 *Raw Background Error Covariance Matrix Using the "Lagged" NMC Method*

For the computation of the raw **B** using the same training data, Eq. 17 is applied with the assumption of Eq. 18. For easier comparison of the raw and modelled **B**, the raw **B** is transformed to analysis increments space (Fig. 3) through direct scaling of the

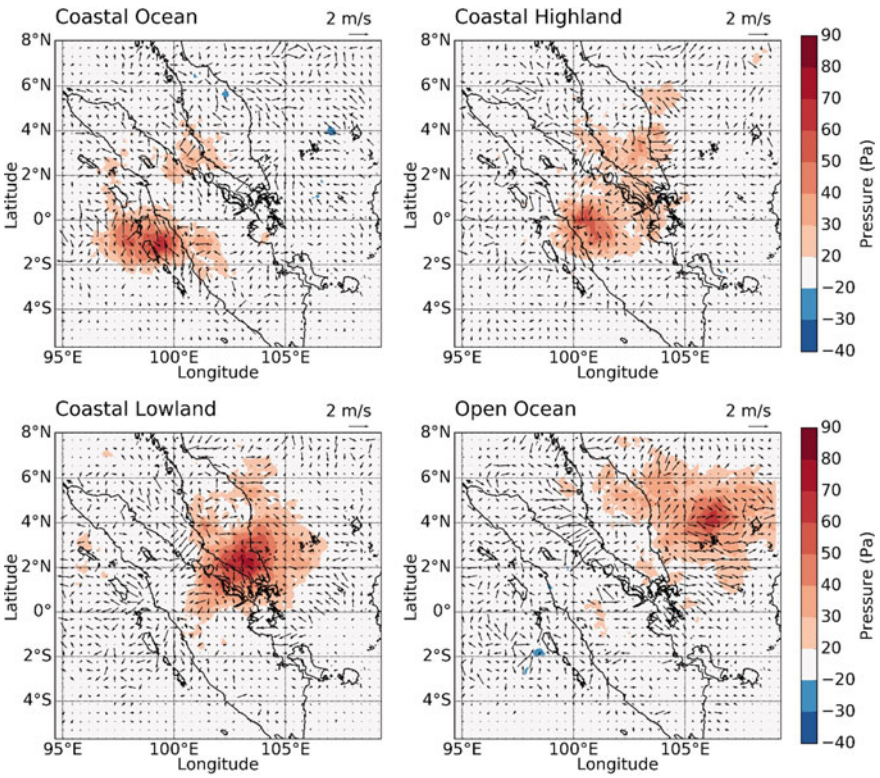

Fig. 3 Raw background error autocovariances and cross-covariances using the "lagged" NMC method with respect to four classification points in the domain, transformed to associated pressure and horizontal wind increments (see Eq. 23) for comparison with other figures (see text). Training data comprise 56 pairs of forecast differences in September 2018

associated horizontal covariance field $\mathbf{B}_i$ with the same d and σ_o (see Eq. 23) used for Fig. 2 and the computed σ_b at each of the 4 classified points.

Notably, the pressure autocovariance fields of the raw **B** include substantial inhomogeneity, in contrast to Fig. 2. Over coastal lowland, the horizontal length-scales also appear slightly longer compared to over coastal highland, possibly because of the lack of orography. Isotropy appears to be valid to a first approximation over coastal lowland and open ocean, but appears more elongated over coastal highland along the coast. One might expect the inhomogeneity and anisotropy to be larger closer to the surface (model level 3 as shown; Deckmyn and Berre 2005). It should be noted that this elongated pattern over coastal highland extends vertically up to model level 20 (~2 km), beyond which the horizontal length-scales decrease substantially and autocovariances become more isotropic. With regards to the multivariate aspects, the cross-covariances between horizontal wind and pressure for the raw **B**

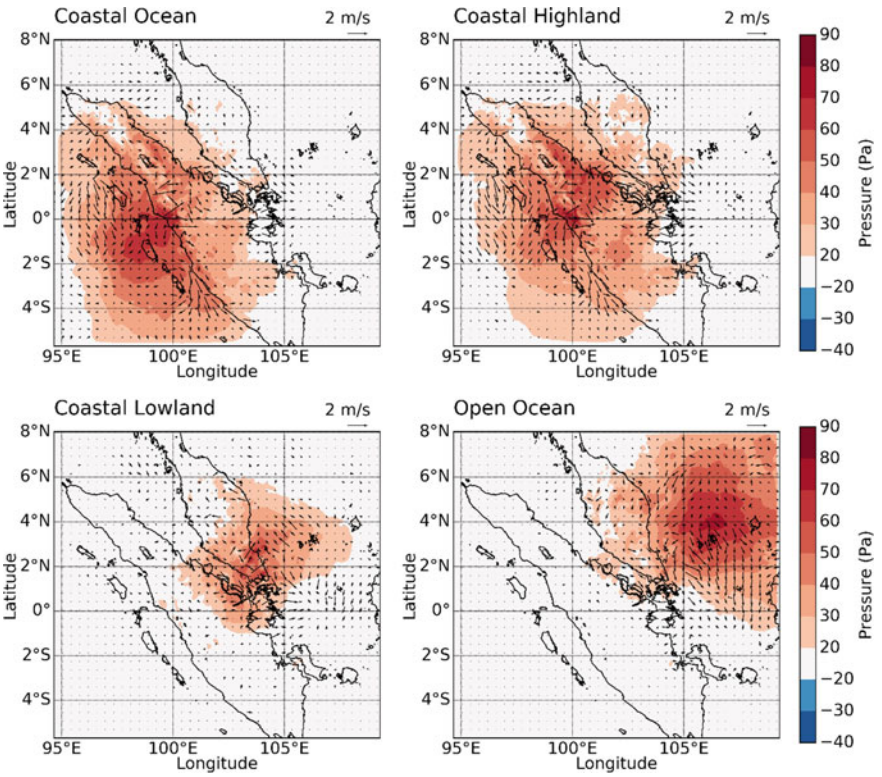

Fig. 4 Raw background error autocovariances and cross-covariances using the ensemble method with respect to four classification points in the domain, transformed to associated pressure and horizontal wind increments (see Eq. 23) for comparison with other figures (see text). Error modes from 50 ensemble 6-h forecasts initialised on 10 September 2018 0000UTC are used. A Gaussian filter with a localisation length-scale of 500 km is applied

appears to be larger than in the modelled **B**, particularly in the vicinity of the observation. The magnitude of the wind vectors range from 1–3 m s^{-1}. This suggests that there could exist some meaningful mass-wind coupling in the estimated background errors that is perhaps smoothed by spatial averaging (application of homogeneity) during the "calibration step". One should also note the noise in the autocovariance and cross-covariance fields throughout the domain. This can be potentially attributed to sampling noise since one would not expect spurious long-range covariances in the real background error covariances. This indicates that even with 56 pairs of forecast differences, it may still be insufficient to address any sampling noise associated with the rank deficiency when explicitly computing **B**.

4.3 Raw Background Error Covariance Matrix Using the Ensemble Method

A widely used ensemble method for LAMs involves generating the ensemble using dynamical downscaling from a global ensemble. As a demonstration of concept, 50 ensemble forecasts are generated for one cycle (10 September 2018 0000UTC), downscaled from European Centre for Medium-Range Weather Forecasts (ECMWF) forecasts. This case study is for demonstrative purposes and not related to any targeted flow conditions. The computation of the raw ensemble covariance ($\mathbf{B_e}$) is performed applying Eq. 20 on 6-hour LAM forecasts (or ensemble trajectories). As before for easier comparison, the raw $\mathbf{B_e}$ is transformed to analysis increments space (Fig. 4). A simple Gaussian filter with a horizontal length-scale of 500 km is applied in prognostic variable space to remove any potential spurious long-range covariances and for easier comparison with the following sub-sections. This choice of length-scale is estimated based on a Gaussian curve-fit to the distance-binned domain correlations to the four classification points.

The raw $\mathbf{B_e}$ generated using the ensemble method contain error structures which are largest at the forecast validity time. It appears that the pressure autocovariance field has generally longer horizontal length-scales compared to Fig. 3, especially over open ocean, and over the coastal ocean and highland points along the Sumatran coast. This is possibly an inherent weakness of the downscaling concept; the background error structures at forecast validity time are inherited from the large-scale analysis error structures of the coarse resolution global ensemble (Bölöni et al. 2015). However, the 6-hour LAM forecast (valid at 0600UTC) may still allow for the spin-up of some mesoscale structures, which appear in the computation of the raw $\mathbf{B_e}$. One can easily distinguish the strong positive pressure autocovariances with respect to coastal ocean and highland points aligned parallel to the Sumatran coast. Following the cross-covariances, a negative surface pressure perturbation would imply surface convergence along the coastal regions, as expected during the typical development of a sea breeze at 0600UTC (1300LT). It is therefore physically reasonable that there is a larger uncertainty along the coastal regions at 0600UTC.

It is also apparent that there exists substantial inhomogeneity and anisotropy in the covariance fields. The large-scale covariance fields differ between classification points and strong positive values are often aligned along the coasts instead of being described solely as a function of distance from each classification point. It also appears that the mass-wind coupling is not negligible in the background errors estimated using the ensemble method, although the influence of sampling noise again has to be considered.

4.4 Modelled Background Error Covariance Matrix Using the Ensemble Method

It is possible to pass the same training data (error modes) used to compute the raw $\mathbf{B_e}$ through a variational data assimilation system in a so-called "EnVar" approach. This is the same as hybrid EnVar methods (Bannister 2017) except that full weight is given to the implied ensemble background error covariance (modelled $\mathbf{B_e}$). A localisation Schur (or Hadamard) product (Lorenc 2003) is applied on the modelled $\mathbf{B_e}$ with a horizontal length-scale of 500 km. A Gaussian filter similar to the one described in Sect. 4.3 is applied, except that localisation is performed in control variable space. Vertical localisation is also applied (see Clayton et al. 2013), although it is not primarily relevant for this illustration.

The structures of the modelled $\mathbf{B_e}$ (Fig. 5) are generally similar to the raw $\mathbf{B_e}$, since they are closely related by definition, except that the pressure autocovariances are smaller in magnitude. The structures are inhomogeneous and largely anisotropic. Isotropy is only valid to a first order approximation over coastal lowland and open ocean. As noted from the cross-covariances of raw $\mathbf{B_e}$, the mass-wind coupling in the modelled $\mathbf{B_e}$ is also not negligible.

From Sect. 3.3, it is possible to apply the ensemble method in an ergodic fashion (through the accumulation of error modes throughout a period, time-shifted or time-lagged) to increase the sample size and possibly reduce spurious long-range covariances. However, this effectively results in a static covariance, much like in the innovations or NMC method. The ensemble method can be exploited to further investigate the validity of time stationarity. To illustrate the flow-dependence of the background error statistics, 50 ensemble forecasts are generated for another cycle (14 July 2019 0000 UTC), downscaled from ECMWF forecasts. The same localisation and modelling steps are applied on the error modes as done previously to get another modelled $\mathbf{B_e}$ (hereafter referred to as alternate $\mathbf{B_e}$ for convenience).

There are some distinct differences in the structures of the alternate $\mathbf{B_e}$ (Fig. 6) compared to the previous modelled $\mathbf{B_e}$. The pressure autocovariances with respect to coastal ocean and highland points are distinctly different; the positive autocovariances are less widespread and negative autocovariances appear to the west of the classification point. The horizontal wind cross-covariances are also in general different throughout the domain. However, there are some noticeable similarities

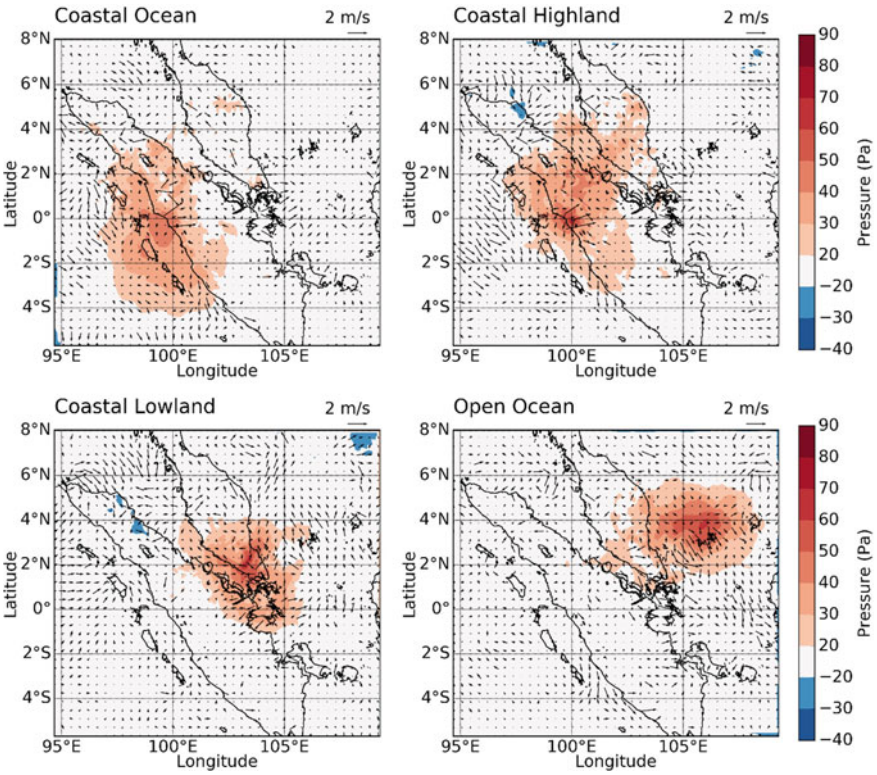

Fig. 5 Pressure and horizontal wind increments as a result of a pseudo-single observation of pressure with respect to four classification points in the domain, related to the associated modelled background error autocovariances and cross-covariances (see Eq. 23) using the ensemble method. Error modes from 50 ensemble 6-h forecasts initialised on 10 September 2018 0000UTC are used. See text for details of the localisation applied

between the alternate and previously modelled $\mathbf{B_e}$. Their pressure autocovariances with respect to coastal ocean are both generally aligned along the Sumatran coast, as opposed to being generally isotropic over coastal lowland and open ocean. Their horizontal wind cross-covariances with respect to the coastal highland point both imply the adherence to a sea breeze pattern along the Sumatran coast, as described in Sect. 4.3. This appears to be a robust structure along the coast, although further evidence is required to determine if there exists any intra-annual variation.

While time-stationarity is a crude assumption, there may be instances where robust structures can exist within the background error statistics. These may vary on different timescales so a clearer understanding on the physical processes underpinning these error structures is required, especially for the Maritime Continent.

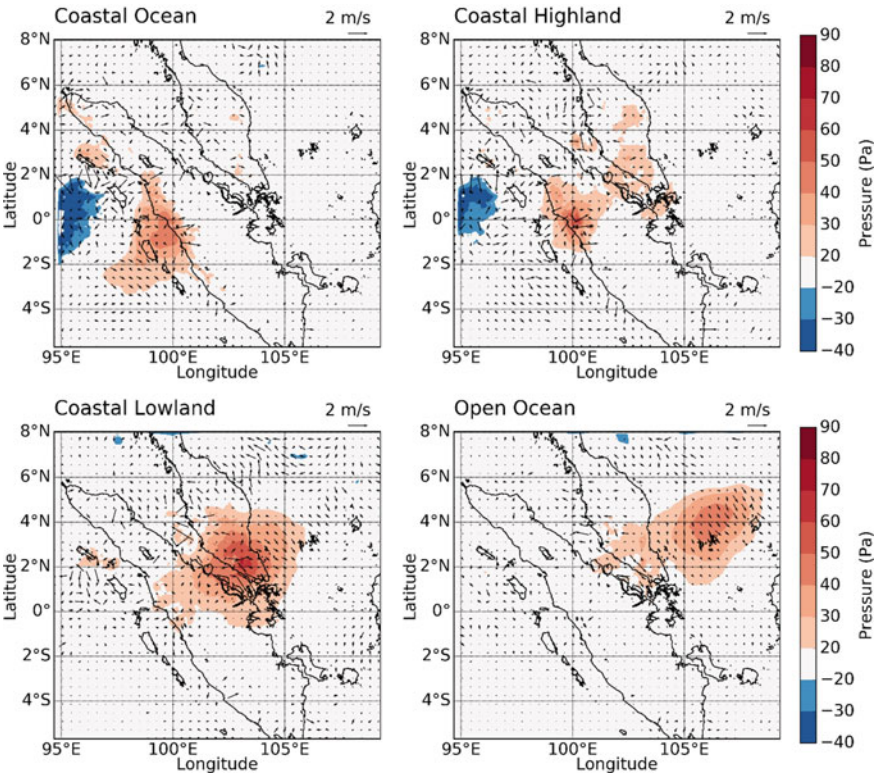

Fig. 6 Pressure and horizontal wind increments as a result of a pseudo-single observation of pressure with respect to four classification points in the domain, related to the associated modelled background error autocovariances and cross-covariances (see Eq. 23) using the ensemble method. Error modes from 50 ensemble 6-h forecasts initialised on 14 July 2019 0000UTC are used. See text for details of the localisation applied

5 Concluding Remarks

In this chapter, an overview of the applicability of the main methods employed to estimate the background covariance matrix and the validity of the main assumptions in modelling it is discussed, particularly for the Maritime Continent context.

A brief demonstration of the methods, where applicable, is provided to illustrate how different methods involve various quantities as surrogates for the real background errors. These may not necessarily accurately represent the structures of the real background errors due to inherent limitations and deficiencies in the conceptualisation. In modelling the background error covariance matrix, certain assumptions applied may also be invalid because of the complexity of the background error structures over the Maritime Continent.

Additional comments are provided in the discussion to highlight possible areas for further research. These include the need for more evidence to establish the validity of the main assumptions of homogeneity, isotropy and time stationarity. It is emphasised that effort should be placed on developing innovative methods to relax these assumptions and to test them over the Maritime Continent. In particular, a possible design of a model for the multivariate component of the background error covariance matrix appropriate for the Tropics is mentioned. As hybrid ensemble-variational methods gain traction, perhaps implementation of one or more of these variants in operational systems and assessment of their impacts would also shed insights on the intricacies of modelling the background error covariance matrix for the Maritime Continent.

References

Anderson JL (2001) An ensemble adjustment Kalman filter for data assimilation. Mon Weather Rev 129(12):2884–2903

Ban J, Liu Z, Zhang X, Huang XY, Wang H (2017) Precipitation data assimilation in WRFDA 4D-Var: implementation and application to convection-permitting forecasts over United States. Tellus A 69(1):1368310

Bannister RN (2007) Can wavelets improve the representation of forecast error covariances in variational data assimilation? Mon Weather Rev 135(2):387–408

Bannister RN (2008a) A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. Q J R Meteorol Soc 134(637):1951–1970

Bannister RN (2008b) A review of forecast error covariance statistics in atmospheric variational data assimilation. II: modelling the forecast error covariance statistics. Q J R Meteorol Soc 134(637):1971–1996

Bannister RN (2017) A review of operational methods of variational and ensemble-variational data assimilation. Q J R Meteorol Soc 143(703):607–633

Barker DM, Huang W, Guo YR, Bourgeois A (2003) A three-dimensional variational (3DVAR) data assimilation system for use with MM5. NCAR Tech Note 68

Barker DM, Huang W, Guo YR, Bourgeois AJ, Xiao QN (2004) A three-dimensional variational data assimilation system for MM5: Implementation and initial results. Mon Weather Rev 132(4):897–914

Barker D, Huang XY, Liu Z, Auligné T, Zhang X, Rugg S et al (2012) The weather research and forecasting model's community variational/ensemble data assimilation system: WRFDA. Bull Am Meteorol Soc 93(6):831–843

Berre L (2000) Estimation of synoptic and mesoscale forecast error covariances in a limited-area model. Mon Weather Rev 128(3):644–667

Berre L, Ştefaănescu SE, Pereira MB (2006) The representation of the analysis effect in three error simulation techniques. Tellus A 58(2):196–209

Berre L, Varella H, Desroziers G (2015) Modelling of flow-dependent ensemble-based background-error correlations using a wavelet formulation in 4D-Var at Météo-France. Q J R Meteorol Soc 141(692):2803–2812

Birch CE, Webster S, Peatman SC, Parker DJ, Matthews AJ, Li Y et al (2016) Scale interactions between the MJO and the western maritime continent. J Clim 29(7):2471–2492

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. Mon Weather Rev 129(3):420–436

Bojarova J, Gustafsson N (2019) Relevance of climatological background error statistics for mesoscale data assimilation. Tellus A 71(1):1615168
Bölöni G (2006) Development of a variational data assimilation system for a limited area model at the Hungarian Meteorological Service. Q J Hung Meteorol Serv 110(3–4):309–327
Bölöni G, Berre L, Adamcsek E (2015) Comparison of static mesoscale background-error covariances estimated by three different ensemble data assimilation techniques. Q J R Meteorol Soc 141(687):413–425
Bonavita M, Raynaud L, Isaksen L (2011) Estimating background-error variances with the ECMWF Ensemble of Data Assimilations system: some effects of ensemble size and day-to-day variability. Q J R Meteorol Soc 137(655):423–434
Bouttier F (1996) Application of Kalman filtering to numerical weather prediction. In Proceeding 1996 ECMWF Seminar on data assimilation and workshop on non-linear aspects of data assimilation. ECMWF, Reading, UK, pp 61–90
Buehner M (2005) Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. Q J R Meteorol Soc 131(607):1013–1043
Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B (2010) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. Mon Weather Rev 138(5):1550–1566
Caron JF, Michel Y, Montmerle T, Arbogast É (2019) Improving background error covariances in a 3D ensemble–variational data assimilation system for regional NWP. Mon Weather Rev 147(1):135–151
Chapnik B, Desroziers G, Rabier F, Talagrand O (2004) Properties and first application of an error-statistics tuning method in variational assimilation. Q J R Meteorol Soc 130(601):2253–2275
Chen Y, Rizvi SR, Huang XY, Min J, Zhang X (2013) Balance characteristics of multivariate background error covariances and their impact on analyses and forecasts in tropical and Arctic regions. Meteorol Atmos Phys 121(1–2):79–98
Chen Y, Zeng L, Huang XY, Wang H, Rizvi SR (2014) A study of latitude dependent background error covariance and its impact. J Trop Meteorol 30(4):654–662
Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. Q J R Meteorol Soc 139(675):1445–1461
Courtier P, Andersson E, Heckley W, Vasiljevic D, Hamrud M, Hollingsworth A et al. (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Q J R Meteorol Soc 124(550):1783–1807
Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge, UK
Daley R (1996) Generation of global multivariate error covariances by singular-value decomposition of the linear balance equation. Mon Weather Rev 124(11):2574–2587
Deckmyn A, Berre L (2005) A wavelet approach to representing background error covariances in a limited-area model. Mon Weather Rev 133(5):1279–1294
Derber J, Bouttier F (1999) A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus A 51(2):195–221
Ehrendorfer M (2007) A review of issues in ensemble-based Kalman filtering. Meteorol Z 16(6):795–818
Epstein ES (1969) Stochastic dynamic prediction. Tellus 21(6):739–759
Errico RM, Privé NC, Gu W (2015) Use of an OSSE to evaluate background-error covariances estimated by the NMC method. Q J R Meteorol Soc 141(687):611–618
Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res: Oceans 99(C5):10143–10162
Evensen G (2003) The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean Dyn 53(4):343–367
Fisher M (2003) Background error covariance modelling. In Seminar on recent development in data assimilation for atmosphere and ocean. ECMWF, Reading, UK, pp 45–63
Fisher M, Andersson E (2001) Developments in 4D-Var and Kalman filtering. ECMWF Tech Memo 347, pp 38

Fischer C, Montmerle T, Berre L, Auger L, Ştefănescu SE (2005) An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. Q J R Meteorol Soc 131(613):3477–3492

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc 125(554):723–757

Gu J, Xiao Q, Kuo YH, Barker DM, Jishan X, Xiaoxing MA (2005) Assimilation and simulation of typhoon Rusa (2002) using the WRF system. Adv Atmos Sci 22(3):415–427

Gustafsson N, Bojarova J, Vignes O (2014) A hybrid variational ensemble data assimilation for the HIgh Resolution Limited Area Model (HIRLAM). Nonlinear Process Geophys 21(1):303–323

Gustafsson N, Huang XY, Yang X, Mogensen K, Lindskog M, Vignes O et al (2012) Four-dimensional variational data assimilation for a limited area model. Tellus A 64(1):14985

Ha JH, Lee DK (2012) Effect of length scale tuning of background error in WRF-3DVAR system on assimilation of high-resolution surface data for heavy rainfall simulation. Adv Atmos Sci 29(6):1142–1158

Hamill TM (2006) Ensemble-based atmospheric data assimilation. In: Palmer T, Hagedorn R (eds) Predictability of weather and climate. Cambridge University Press, Cambridge, pp 124–156

Hamill TM, Snyder C (2000) A hybrid ensemble Kalman filter–3D variational analysis scheme. Mon Weather Rev 128(8):2905–2919

Heng BP, Tubbs R, Huang XY, Macpherson B, Barker DM, Boyd DF et al (2020) SINGV-DA: a data assimilation system for convective-scale numerical weather prediction over Singapore. Q J R Meteorol Soc 146(729):1923–1938

Hohenegger C, Walser A, Langhans W, Schär C (2008) Cloud-resolving ensemble simulations of the August 2005 Alpine flood. Q J R Meteorol Soc 134(633):889–904

Hollingsworth A, Lönnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. Tellus A 38(2):111–136

Horvath K, Gergely B (2004) Comparison of NMC and Lönnberg-Hollingsworth type of background error statistics of the ALADIN/HU model. RC LACE internal report, pp 18

Houtekamer PL, Lefaivre L, Derome J, Ritchie H, Mitchell HL (1996) A system simulation approach to ensemble prediction. Mon Weather Rev 124(6):1225–1242

Houtekamer PL, Zhang F (2016) Review of the ensemble Kalman filter for atmospheric data assimilation. Mon Weather Rev 144(12):4489–4532

Huang XY, Barker D, Webster S, Dipankar A, Lock A, Mittermaier M et al (2019) SINGV–the convective-scale numerical weather prediction system for Singapore. ASEAN J Sci Technol Dev 36(3):81–90

Huang XY, Xiao Q, Barker DM, Zhang X, Michalakes J, Huang W et al (2009) Four-dimensional variational data assimilation for WRF: formulation and preliminary results. Mon Weather Rev 137(1):299–314

Ingleby NB (2001) The statistical structure of forecast errors and its representation in The Met. Office global 3-D variational data assimilation scheme. Q J R Meteorol Soc 127(571):209–231

Järvinen H (2001) Temporal evolution of innovation and residual statistics in the ECMWF variational data assimilation systems. Tellus A 53(3):333–347

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, UK

Kleist DT, Parrish DF, Derber JC, Treadon R, Wu WS, Lord S (2009) Introduction of the GSI into the NCEP global data assimilation system. Weather Forecasting 24(6):1691–1705

Kobayashi S, Ota Y, Harada Y, Ebita A, Moriya M, Onoda H et al (2015) The JRA-55 reanalysis: general specifications and basic characteristics. J Meteorol Soc Jpn Ser II 93(1):5–48

Lee JCK, Huang XY (2020) Background error statistics in the tropics: structures and impact in a convective-scale numerical weather prediction system. Q J R Meteorol Soc 146(730):2154–2173

Li X, Zeng M, Wang Y, Wang W, Wu H, Mei H (2016) Evaluation of two momentum control variable schemes and their impact on the variational assimilation of radarwind data: case study of a squall line. Adv Atmos Sci 33(10):1143–1157

Liu YA, Huang HL, Gao W, Lim AH, Liu C, Shi R (2015) Tuning of background error statistics through sensitivity experiments and its impact on typhoon forecast. J Appl Remote Sens 9(1):096051

Lönnberg P, Hollingsworth A (1986) The statistical structure of short-range forecast errors as determined from radiosonde data Part II: the covariance of height and wind errors. Tellus A 38(2):137–161

Lorenc AC (1997) Development of an operational variational assimilation scheme. J Meteorol Soc Jpn Ser II 75(1B):339–346

Lorenc AC (2003) The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. Q J R Meteorol Soc 129(595):3183–3203

Lorenc AC, Ballard SP, Bell RS, Ingleby NB, Andrews PL, Barker DM et al. (2000) The Met. office global three-dimensional variational data assimilation scheme. Q J R Meteorol Soc 126(570):2991–3012

Meng Z, Zhang F (2011) Limited-area ensemble-based data assimilation. Mon Weather Rev 139(7):2025–2045

Michel Y, Auligné T (2010) Inhomogeneous background error modeling and estimation over Antarctica. Mon Weather Rev 138(6):2229–2252

Montmerle T, Berre L (2010) Diagnosis and formulation of heterogeneous background-error covariances at the mesoscale. Q J R Meteorol Soc 136(651):1408–1420

Montmerle T, Lafore JP, Berre L, Fischer C (2006) Limited-area model error statistics over Western Africa: comparisons with midlatitude results. Q J R Meteorol Soc 132(614):213–230

Montmerle T, Michel Y, Arbogast E, Ménétrier B, Brousseau P (2018) A 3D ensemble variational data assimilation scheme for the limited-area AROME model: formulation and preliminary results. Q J R Meteorol Soc 144(716):2196–2215

Onogi K, Tsutsui J, Koide H, Sakamoto M, Kobayashi S, Hatsushika H et al (2007) The JRA-25 reanalysis. J Meteorol Soc Jpn Ser II 85(3):369–432

Parrish DF, Derber JC (1992) The national meteorological center's spectral statistical-interpolation analysis system. Mon Weather Rev 120(8):1747–1763

Purser RJ, Wu WS, Parrish DF, Roberts NM (2003) Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. Mon Weather Rev 131(8):1524–35

Rabier F, McNally A, Andersson E, Courtier P, Unden P, Eyre J et al. (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). II: Structure functions. Q J R Meteorol Soc 124(550):1809–1829

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW et al (2007) The met office global four-dimensional variational data assimilation scheme. Q J R Meteorol Soc 133(623):347–362

Rutherford ID (1972) Data assimilation by statistical interpolation of forecast error fields. J Atmos Sci 29(5):809–815

Sadiki W, Fischer C (2005) A posteriori validation applied to the 3D-VAR Arpège and Aladin data assimilation systems. Tellus A 57(1):21–34

Sattler K, Huang XY (2002) Structure function characteristics for 2 metre temperature and relative humidity in different horizontal resolutions. Tellus A 54(1):14–33

Široká M, Fischer C, Cassé V, Brožková R, Geleyn JF (2003) The definition of mesoscale selective forecast error covariances for a limited area variational analysis. Meteorol Atmos Phys 82(1–4):227–244

Sobel AH, Nilsson J, Polvani LM (2001) The weak temperature gradient approximation and balanced tropical moisture waves. J Atmos Sci 58(23):3650–3665

Stanesic A, Horvath K, Keresturi E (2019) Comparison of NMC and ensemble-based climatological background-error covariances in an operational limited-area data assimilation system. Atmos 10(10):570

Storto A, Randriamampianina R (2010) Ensemble variational assimilation for the representation of background error covariances in a high-latitude regional model. J Geophys Res: Atmos 115(D17)

Sun J, Wang H, Tong W, Zhang Y, Lin CY, Xu D (2016) Comparison of the impacts of momentum control variables on high-resolution variational data assimilation and precipitation forecasting. Mon Weather Rev 144(1):149–169

Thiébaux HJ, Mitchell HL, Shantz DW (1986) Horizontal structure of hemispheric forecast error correlations for geopotential and temperature. Mon Weather Rev 114(6):1048–1066

Tippett MK, Anderson JL, Bishop CH, Hamill TM, Whitaker JS (2003) Ensemble square root filters. Mon Weather Rev 131(7):1485–1490

Varella H, Berre L, Desroziers G (2011) Diagnostic and impact studies of a wavelet formulation of background-error correlations in a global model. Q J R Meteorol Soc 137(658):1369–1379

Vetra-Carvalho S, Dixon M, Migliorini S, Nichols NK, Ballard SP (2012) Breakdown of hydrostatic balance at convective scales in the forecast errors in the met office unified model. Q J R Meteorol Soc 138(668):1709–1720

Wang X, Barker DM, Snyder C, Hamill TM (2008) A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. Mon Weather Rev 136(12):5116–5131

Weaver A, Courtier P (2001) Correlation modelling on the sphere using a generalized diffusion equation. Q J R Meteorol Soc 127(575):1815–1846

Wheeler M, Kiladis GN, Webster PJ (2000) Large-scale dynamical fields associated with convectively coupled equatorial waves. J Atmos Sci 57(5):613–640

Whitaker JS, Compo GP, Wei X, Hamill TM (2004) Reanalysis without radiosondes using ensemble data assimilation. Mon Weather Rev 132(5):1190–1200

Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 130(7):1913–1924

Whitaker JS, Hamill TM, Wei X, Song Y, Toth Z (2008) Ensemble data assimilation with the NCEP global forecast system. Mon Weather Rev 136(2):463–482

Wu WS, Purser RJ, Parrish DF (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. Mon Weather Rev 130(12):2905–2916

Xu Q, Nai K, Wei L (2007) An innovation method for estimating radar radial-velocity observation error and background wind error covariances. Q J R Meteorol Soc 133(623):407–415

Xu Q, Wei L (2001) Estimation of three-dimensional error covariances. Part II: analysis of wind innovation vectors. Mon Weather Rev 129(12):2939–2954

Xu Q, Wei L (2002) Estimation of three-dimensional error covariances. Part III: Height–wind forecast error correlation and related geostrophy. Mon Weather Rev 130(4):1052–1062

Xu Q, Wei L, Van Tuyl A, Barker EH (2001) Estimation of three-dimensional error covariances. Part I: analysis of height innovation vectors. Mon Weather Rev 129(8):2126–2135

Žagar N, Andersson E, Fisher M (2005) Balanced tropical data assimilation based on a study of equatorial waves in ECMWF short-range forecast errors. Q J R Meteorol Soc 131(607):987–1011

Žagar N, Andersson E, Fisher M, Untch A (2007) Influence of the Quasi-Biennial Oscillation on the ECMWF model short-range-forecast errors in the tropical stratosphere. Q J R Meteorol Soc 133(628):1843–1853

Žagar N, Gustafsson N, Källén E (2004) Variational data assimilation in the tropics: the impact of a background-error constraint. Q J R Meteorol Soc 130(596):103–125

Žagar N, Stoffelen A, Marseille GJ, Accadia C, Schlüssel P (2008) Impact assessment of simulated Doppler wind lidars with a multivariate variational assimilation in the tropics. Mon Weather Rev 136(7):2443–2460

Operational Assimilation of Radar Data from the European EUMETNET Programme OPERA in the Météo-France Convective-Scale Model AROME

Maud Martet, Pierre Brousseau, Eric Wattrelot, Frank Guillaume, and Jean-François Mahfouf

Abstract AROME-France is the convective-scale numerical weather prediction system which has been running operationally at Météo-France since the end of 2008. It uses a 3D-Var assimilation scheme to determine its initial conditions, in which radar data have a significant impact, due to their high temporal and spatial resolutions. Météo-France radar data are assimilated since 2008 for radial velocities and 2010 for reflectivities. Within the framework of the OPERA program, EUMETNET provides radar data from 25 countries with a common data format and a common quality index that allows uniform use independent of country of origin. Pre-processing of radar observations in the AROME-France data assimilation system has been updated in order to correctly take them into account, particularly regarding to the common quality index provided by OPERA. Monitoring experiments of these observations versus AROME-France backgrounds show that OPERA radar data are coherent with other radar data from the Météo-France ARAMIS network. Their assimilation has shown positive impacts on AROME-France forecasts over a two-month trial period and also on a specific heavy precipitation case in December 2019. Since January 2020, OPERA data from 62 radars among 9 French neighbouring countries are operationally assimilated in the AROME-France 3D-Var system.

1 Introduction

National Weather Services need to issue accurate forecasts of high impact weather at small scale (severe thunderstorms, wind gusts, fog, ...). AROME-France is a convective-scale numerical weather prediction system which has been running operationally at Météo-France since the end of 2008 (Brousseau et al. 2016). In order to determine its initial conditions, radar observations (radial winds and reflectivities) are used in the 3D-Var assimilation system, in addition to conventional and satellite observations. Due to their high temporal and spatial resolutions, they have a significant impact on rain forecast performances.

M. Martet (✉) · P. Brousseau · E. Wattrelot · F. Guillaume · J.-F. Mahfouf
CNRM, Météo-France, CNRS, Universit de Toulouse, Toulouse, France
e-mail: maud.martet@meteo.fr

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_24

Since the end of 2008 for radial velocities (Montmerle and Faccani 2009) and 2010 for reflectivities (Wattrelot et al. 2014), radar data from the French network ARAMIS (Application Radar la Météorologie Infra-Synoptique) are assimilated. A specific product has been set up, including Plan Position Indicator (PPI) of radial winds, reflectivity, attenuation by precipitation and an echo type discrimination using dual-polarisation and fuzzy-logic, as described in Tabary (2007) and Tabary et al. (2007).

In addition to the French network, assimilation of European radar data is a challenge. In the framework of the Operational Program for Exchange of Weather Radar Information (OPERA) within the European Meteorological Services Network (EUMETNET), data from 164 radars originating from 25 countries are provided in real time for numerical weather prediction (NWP) purposes (Saltikoff et al. 2019). Unlike the Next Generation Weather Radar (NEXRAD) network over the US territory, the European radar network is extremely heterogeneous in installation date, manufacturers, scanning strategy, signal processing and product generation. Two challenges are addressed within OPERA: unifying data format and dealing with data heterogeneity. The OPERA Data Information Model (ODIM) is the common format used within the European network. This information model describes, in general terms, radar volume data, radar products, and the associated metadata. With this technical uniformity, data providers send equivalent radar products, in particular reflectivity and radial velocity. A centralised pre-processing adds a quality index in order to provide homogeneous information for each individual radar.

The goal of this article is to describe how to efficiently assimilate OPERA radar data in the AROME-France 3D-Var data assimilation system. A first section is dedicated to OPERA radar data: these observations have specificities in terms of format, using ODIM specifications, and also in terms of quality fields added in a centralised processing performed by OPERA operational service. A good understanding of these data is required to properly use them in an assimilation system. Monitoring of the data is also important in order to verify their quality before they are actually assimilated.

Section 3 is dedicated to impacts of assimilating OPERA radar data on the performances of the AROME-France data assimilation/forecast system. Scores over a two month period are first examined in order to estimate the mean impact of OPERA radar data assimilation in AROME-France. Then, results are compared on a case study where the assimilation over Spain and part of the Mediterranean sea has impacted heavy rain over South-East of France.

2 OPERA Radar Data

2.1 Description of the Data

In addition to the 31 French radars from the Météo-France network, 62 OPERA radars, from 9 countries, provide data over the AROME-France geographical domain. As required by the OPERA program, the data providers send three different types

of observations: radial velocity, filtered reflectivity and raw reflectivity (only noise thresholded). A centralised OPERA processing adds homogeneous quality indices, and more specifically a "total quality index" that gathers all quality information from three modules applied to the PPIs (Saltikoff et al. 2019). This index is to be used in the pre-processing of the data to discriminate observations suitable for assimilation purposes.

The comparison of raw and filtered reflectivities provides reliable information on non rainy areas. For OPERA radar data, a common definition of dry pixels was needed, leading to the following definitions: "nodata" indicates that the pixel is out of range or in a blanked sector, "undetect" means that the received radar signal is at or below noise level. A problem occurs when ground clutters are removed and reflectivity value is set to "undetect". In this case, there is a risk to assimilate data as "no rain" whereas the absence of rain is not certain. In order to discriminate dry areas from ground clutters, the comparison between raw and filtered is performed: if the filtered reflectivity DBZH value is "undetect", when the raw reflectivity is also "undetect", we are sure it is a non rainy pixel, otherwise, the reflectivity is removed since there is a doubt whether the pixel is rainy or not.

2.2 Pre-processing of the Data

In AROME-France, a 1D+3D-Var assimilation method is implemented for reflectivity radar data. A 1D Bayesian retrieval of relative humidity columns from reflectivity columns has been developed upstream of the 3D-Var as explained in Wattrelot et al. (2014). However, in OPERA data, no specifications are given on horizontal or vertical resolutions. Each radar can have different elevations (numbers and angles) and each elevation can have a different number of azimuths or gates. So, in order to be able to consider vertical profiles of reflectivity required by the Bayesian method, a choice has been made in the observation preprocessing software: for each radar, the elevations with the number of azimuths with the two most "populated" are used, the others are discarded.

A qualitative study has been performed on the quality index. This study consists of a visual comparison of PPIs where reflectivity has been filtered with different thresholds on quality index. Three values have been considered: 0.6, 0.7 and 0.8. Figure 1 gives an example of filtered reflectivity for these thresholds. Radars from different countries and with various performances have been studied. In some areas, a satellite filter deletes "true" rain in the data. Using a threshold on quality index increases this phenomenon but gives more confidence in the fact that the reflectivity kept is really rain. On this particular set of values, 0.6 and 0.7 seem to be acceptable thresholds but as we want to be sure not to use non-meteorological echoes, the highest one is preferred. After such a qualitative study, the threshold value has been set to 0.7.

In order to assimilate "non-rainy" observations, it is important to know the minimum detectable signal observable by each radar. When this information is not available in the metadata, an approximation is made using the minimum value of

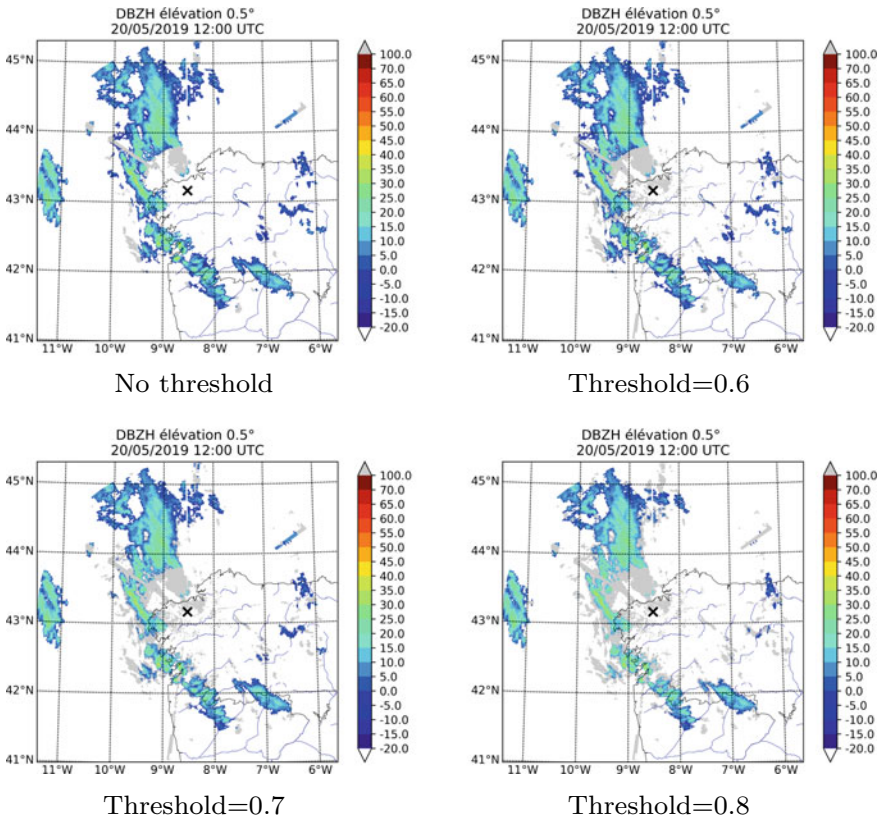

Fig. 1 PPIs from Cerceda (Spain) radar, 20/05/2019 12h00 UTC, elevation 0.5°. Reflectivities have been filtered with different thresholds on the quality index

reflectivity observed in the PPI. If this value is available, the sensitivity threshold is calculated, as a function of the distance, in order to assimilate the "non-rainy" observations and to be able to dry the model where no rain is observed.

Concerning radial velocity, another constraint is that no quality index is available on this parameter as is on reflectivity. In order to be able to discriminate non-meteorological echoes, the quality index calculated for reflectivity is also used for radial velocity, when it is possible. This is the case when reflectivity and radial velocity are observed during the same PPI or when a common elevation is used for reflectivity and radial velocity (even if the time of observation is not strictly equal). Even though this approximation is not ideal, it is the best compromise to avoid the assimilation of radial velocities from non-meteorological echoes.

Moreover, we chose not to consider radial velocity when the Nyquist velocity (NI) is under a threshold set to 30 m/s. This value is a compromise between avoiding aliased radial velocities and keeping data from some radars (many radial velocity

PPIs have smaller Nyquist velocities). If the Nyquist velocity is not present in the file metadata, the radial velocity is not considered for assimilation.

This pre-processing is specific to OPERA radar data. For Météo-France network observations, a discrimination of ground clutters and clear sky echoes is made at radar level and echo type is associated with the reflectivity and radial wind observations. Moreover, Nyquist velocity of Météo-France observations is high enough (60 m/s) to be considered in the assimilation system.

2.3 Monitoring

The first step before active assimilation of OPERA radar data in AROME-France is to check their quality, compared to model background (a 1 h forecast from the previous atmospheric analysis time). Thus, an experiment without radar data assimilation has been performed. This experiment allowed the consideration of model background fields that are not influenced by any radar data (Météo-France and OPERA radars). A 5-day experiment has been performed between 15 and 20 October 2019, chosen because of a favourable rainy context. The background are compared with radar data in observation space.

Figure 2 represents the number of active (i.e. that could be assimilated) observations for each parameter (relative humidity and radial winds) for both groups of radars (Météo-France on one hand, OPERA on the other). The number of observations is dependent of the quantity of rain in particular for radial winds given that radar can't measure radial velocity without rain. Concerning relative humidity, the use of "non-rainy" observations implies less variability in the number of observations retained. When comparing the two groups of radars, the number of observations is equivalent for relative humidity but much less for radial wind, OPERA radar data representing only 20% of Météo-France radar data. This is due to chosen pre-processing (described previously) that leads to the rejection of numerous data.

Figure 3 represents distributions of first guess innovations for the two parameters (radial wind and retrieved relative humidity), compared between Météo-France and OPERA radar data. Only active data are taken into account during this 5-day experiment. The four histograms show unbiased datasets and that innovation distributions follow a Gaussian law.

This monitoring shows that the OPERA radar data quality is suitable for assimilation purposes. The pre-processing developed for these observations allows quality observations to be retained. These good quality observations will contribute to the improvement of the AROME-France data assimilation system.

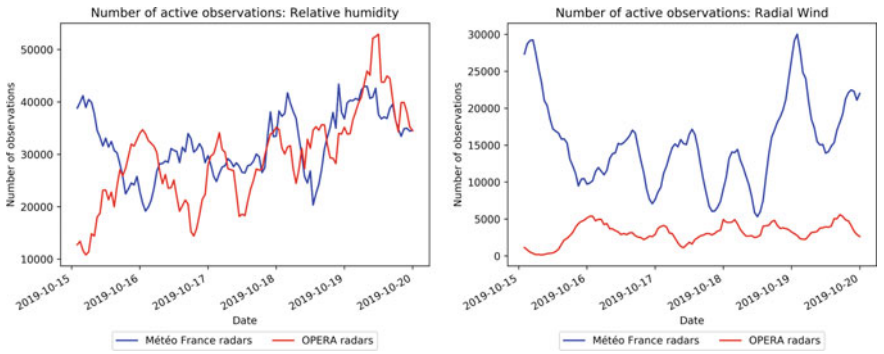

Fig. 2 Number of active observations for relative humidity (left) and radial winds (right), Météo-France radars in blue and OPERA radars in red, between 15 and 20 October 2019

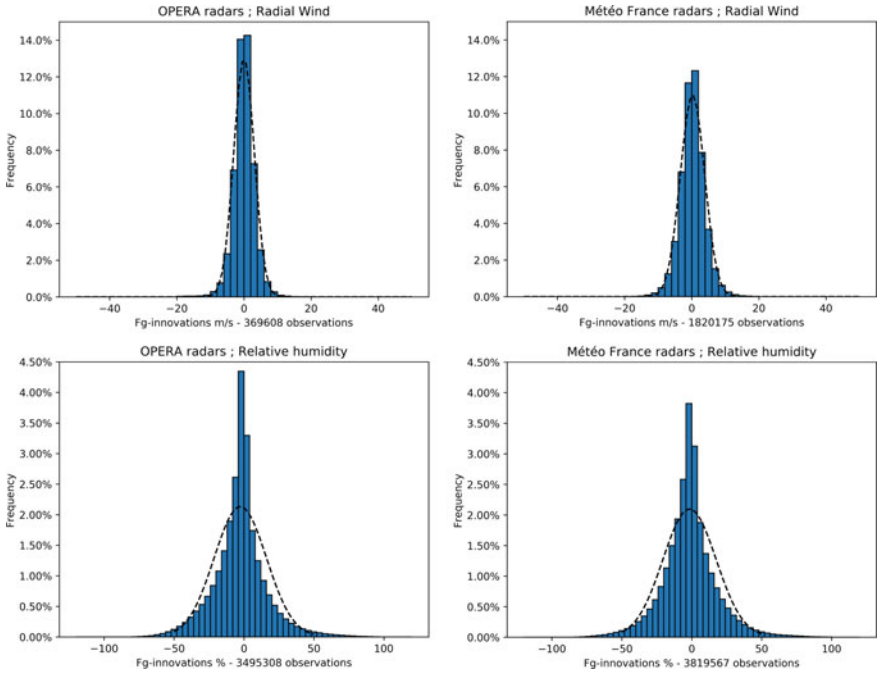

Fig. 3 Histogram of innovations for radial winds (top) and relative humidity (bottom) for OPERA radar data (left) and Météo-France radar data (right) between 15 and 20 October 2019

3 Assimilation in AROME-France

3.1 AROME-France Assimilation System

The AROME–France data assimilation system is configured as a sequential 3D-Var assimilation at the 1.3 km model horizontal resolution in an hourly cycle. It takes full advantage of the high spatial and, to some extent, temporal density of observing systems such as automated surface stations, ground-based GNSS delays and radars. In AROME–France, each 3D-Var step analyzes the two components of horizontal wind, temperature, specific water-vapor humidity and surface-pressure fields on the model grid at full resolution. The other prognostic model fields (turbulent kinetic energy, pressure departure from hydrostatism, vertical divergence and specific content of five condensed water species) are not updated by the analyses but copied from the background. In other words, they are left to adjust to the updated model fields during the forecast steps of the assimilation cycle (Seity et al. 2011).

In this assimilation process, radar data are well suited to provide high-resolution information on wind and precipitation (Brousseau et al. 2014). Moreover, volumes of radar reflectivities play a key role as they provide information on the three-dimensional structure of precipitating systems useful for the initialization of high-resolution models (Gustafsson et al. 2018). Concerning radial velocities, an observation operator, which allows the simulation of radial winds from model variables, is included in the 3D-Var system. Concerning reflectivity, a one-dimensional (1D) Bayesian retrieval of relative humidity profiles followed their assimilation in the 3D-Var is adopted. Since the early 2010s, radar data from Météo-France observation network are operationally assimilated.

Since January 2020, 62 OPERA radars are also assimilated operationally in the AROME-France system. Figure 4 shows the location of these radars and what quantities are used by country (no data, reflectivity or reflectivity and radial wind). Adding these observations in the assimilation system may lead to improved forecast quality, not only over areas where radar observations are available but also over French territory, by enhancing the representation of upstream meteorological phenomena.

3.2 Performances of the Assimilation in the Observation Space

Impacts of OPERA radar data assimilation in the AROME-France 3D-Var system can be assessed via statistics of observation minus background and observation minus analysis quantities. A 10-day period is considered between 21 November and 1st December 2019 to compare forecast performances. In the reference experiment, only Météo-France radar data are assimilated whereas OPERA radar data are added in the other experiment.

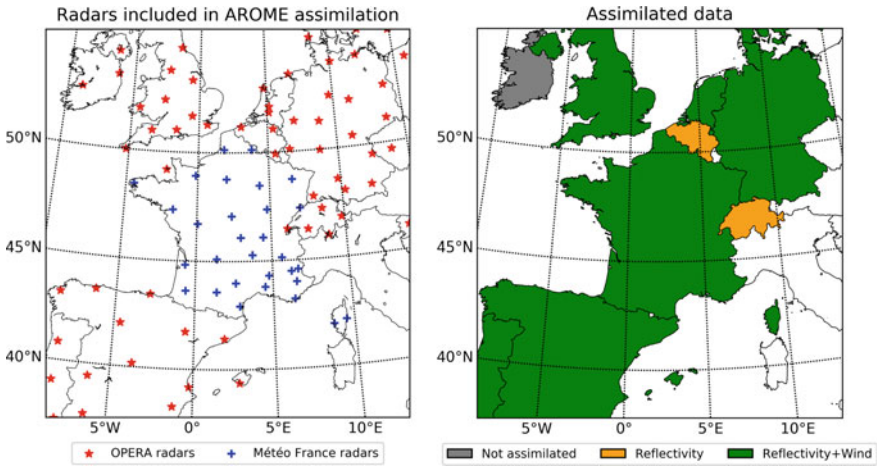

Fig. 4 Location of radars used in the AROME-France assimilation system (left) and assimilated radar data (right)

Figure 5 represents root mean square error (RMS) and bias for relative humidity and radial winds of observation minus background and observation minus analysis quantities. The differences in the number of assimilated observations is coherent with the monitoring experiment (see above). The vast majority of OPERA radar data are collected at elevations lower than 4°, corresponding to exploitation modes chosen by the meteorological services in order to operate their radar network.

Adding OPERA radar data in the data assimilation system has a noticeable impact on relative humidity. In terms of bias, the background has larger values at lower elevations, linked with the added observations. The quality of the observations can vary from a country to another, particularly for low elevations that can be affected by ground clutter and beam blocking, leading to an increase of the background bias. This bias decreases in the analysis when OPERA radar data are used. This reduction underlines the benefit of assimilating these data. In terms of RMS, the assimilation of OPERA radar data improves the quality of both background and analysis.

Concerning radial winds, bias is reduced for background, RMS and bias are reduced both for background and analysis. This result also confirms that the data selection in the pre-processing step is efficient to select only valuable data.

Comparison with other ground-based GNSS data or satellite channels sensitive to humidity shows that background fields are less coherent with these observations (not shown). This result is expected given the large number of radar data that are added in our new configuration, radar observations represent about 75% of the assimilated observations in rainy situations and consequently, can be slightly overfitted.

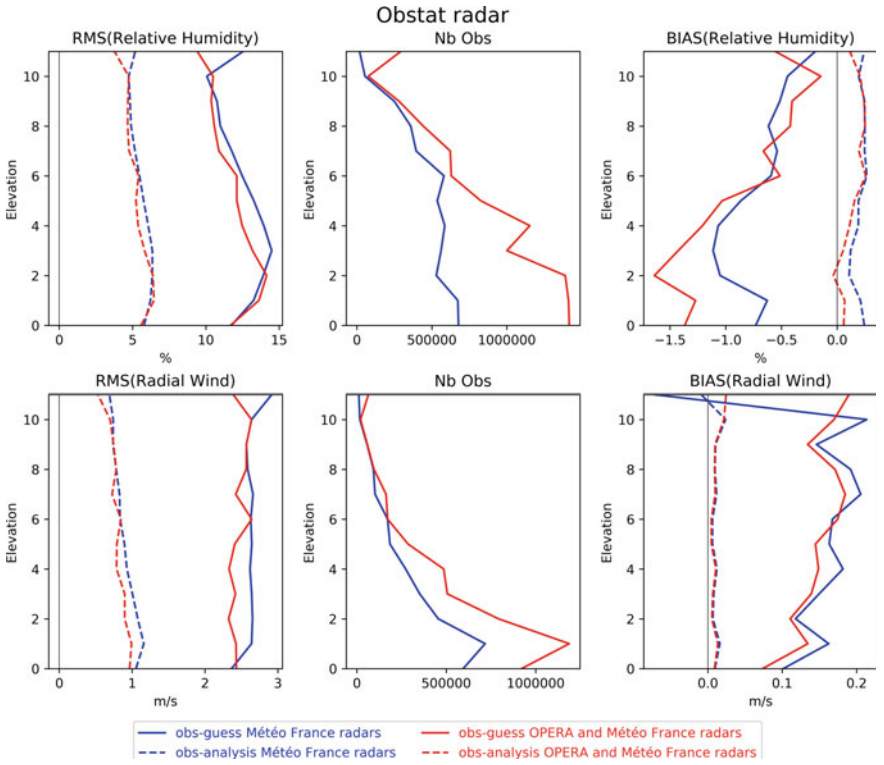

Fig. 5 Statistics of observation minus background (line) and observation minus analysis (dotted line) quantities for relative humidity (top) and Doppler wind (bottom) between 21 November and 1st December 2019. Reference (blue) assimilates only Météo-France radar data, experiment (red) assimilates both OPERA and Météo-France radar data

3.3 Mean Scores

The two configurations of AROME-France, with and without assimilation of OPERA radar data, have been compared over two months, from 21th November 2019 to 15th January 2020. Accumulated precipitation can be compared to the ANTILOPE (ANalyse par spaTIaLisation hOraire des PrEcipitations) product: a quantitative precipitation estimation (QPE) combining radar and rain gauge data whose goal is to provide a real-time rainfall analysis all over France with a 1 km horizontal resolution (Laurantin 2013). It is based on an algorithm of convective cells detection applied on each 5-min radar image in order to separate the large and small scale components of the precipitation field (corresponding respectively to stratiform rainfalls and showers or thunderstorms). Advanced kriging techniques (including automatic non-parametric estimation of 2D correlograms of rainfall fields) are used to combine radar and gauges data.

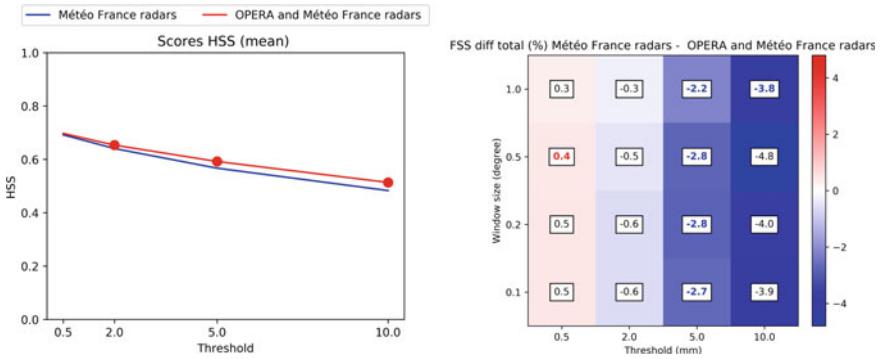

Fig. 6 Heidke Skill Score (HSS, left) and normalised difference of Fractions skill score (FSS, right) of 6 h accumulated precipitation for various threshold (0.5–10 mm). Base time: from 21st November 2019 to 15th January 2020, forecast range 6h. ANTILOPE accumulated precipitation is the reference

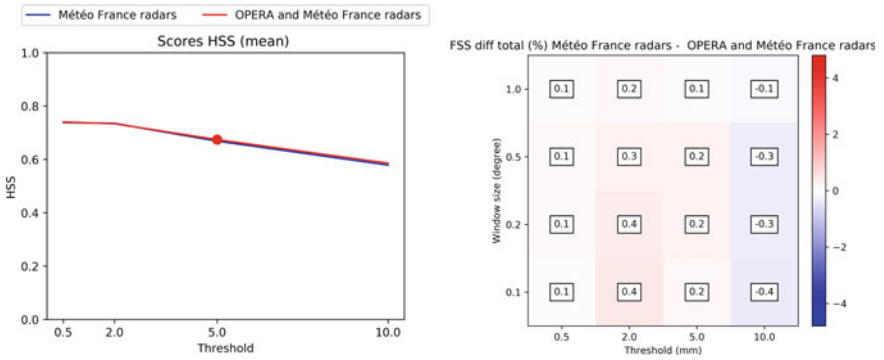

Fig. 7 Heidke Skill Score (HSS, left) and normalised difference of Fractions skill score (FSS, right) of 24 h accumulated precipitation for various threshold (0.5–10 mm). Base time: from 21st November 2019 to 15th January 2020, forecast range 24 h. ANTILOPE accumulated precipitation is the reference

If we focus on the first 6 h of forecast range, the impact on precipitation performance is noticeable. Figure 6 represents the Heidke Skill Score (HSS) and Fractions Skill Score (FSS) computed using ANTILOPE data as reference for 6 h accumulated precipitation at forecast range 6. Red circles on HSS scores highlight significant difference, associated with a bootstrap test. Improvement is more visible for higher thresholds (2 to 10 mm). FSS normalised differences are quite neutral for 0.5 and 2 mm thresholds but improvement is noticeable for 5 and 10 mm thresholds.

Figure 7 represents the same scores for 24 h accumulated precipitation at forecast range 24. The impact of OPERA radar data assimilation is quite neutral at this forecast range. This neutrality can be explained by the fact that the impact of assimilation is visible at the first forecast ranges but decreases with time.

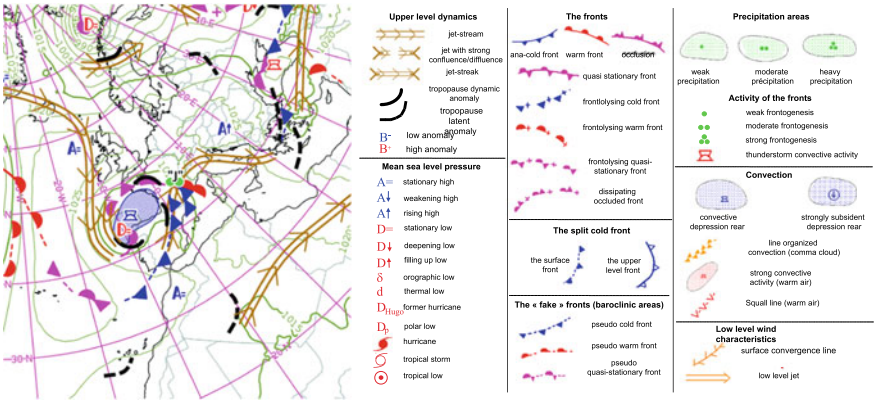

Fig. 8 "ANASYG/PRESYG" describing the meteorological situation on 1st December 2019 (left) and associated legend of the symbols (right)

In order to monitor the evolution of the quality of the AROME-France model over time, a composite index has been developed at Météo-France (Amodei et al. 2015). The "score indicateur" is a mean (on thresholds and neighbourhoods) of Brier Skill Scores against persistence on various parameters: accumulated 6 h rainfall *RR6*, averaged wind gust in six hours between the six and twenty-four hours forecast ranges *FXI6*. For rainfall, thresholds considered are 0.5, 2 and 5 mm; for wind gusts, thresholds considered are 40, 60 and 80 km/h. Scores on SEVIRI brightness temperature (channel 10.8 μm) *BTP6* are also calculated. Observations are from the French ground stations network. From this score (not shown), the impact of the assimilation of the OPERA radar data is quite neutral but small improvements are noticed for the highest precipitation thresholds (2 and 5 mm).

Both scores confirm that the assimilation of OPERA radar data has positive impact on AROME-France forecast performances, specially on the first forecast ranges.

3.4 Case Study

On specific situations, the impact of OPERA radar data assimilation can be noticeable at longer forecast range. Our case study is a low pressure system that occurred on 1st December 2019, named J on the "ANASYG/PRESYG" map, near the shaded blue area (Fig. 8), as described in the Météo-France graphical summary of the synoptic situation (Santurette and Joly 2002). The system enters West of French territory before affecting the South-East of France. High precipitation amounts have been observed during this event that lasts about 24 h: 201 mm in Cannes (Alpes Maritimes) with 103 mm in 3 h and 46 mm in 1 h, 155 mm in Vidauban (Var).

In our study, this meteorological situation is particularly interesting in order to verify if the assimilation of radar data from other European countries, and more

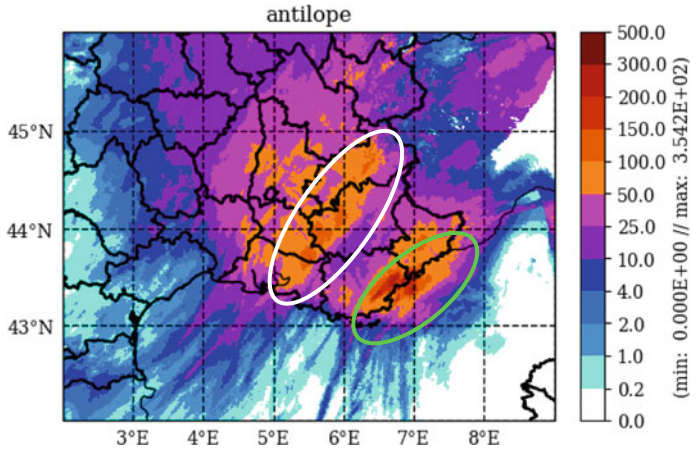

Fig. 9 ANTILOPE accumulated precipitation over 24 h from 1st December 2019 0 h UTC to 2nd December 2019 0 h UTC

specifically from Spanish radars in this case, can have positive impacts on AROME-France forecasts over the French territory. ANTILOPE data (described previously) is be used hereafter as reference for verification of accumulated precipitation.

Figure 9 represents the ANTILOPE accumulated precipitation over 24 h from 1st December 2019 0h UTC to 2nd December 2019 0h UTC. Two areas of heavy rain are observed, identified in Fig. 9 as area "Var" (green ellipse) and area "Hautes-Alpes" (white ellipse).

Figure 10 represents 24-h accumulated precipitation for two AROME-France configurations: without OPERA data data assimilation (REF) and with these additional data (EXPE), base time for the model experiments is 1st December 2019 0h UTC on this illustration. The maximum rain intensity in the "Var" area is underestimated in the EXPE forecast. Nevertheless, the geographical intensity of the maximum accumulated precipitation is not properly located for both simulations. Performances of the EXPE forecast are clearly improved over the "Hautes-Alpes" area. Indeed, both localisation and intensity are better described in the EXPE forecast.

To confirm this qualitative comparison, Heidke Skill Score (HSS) and Fractions Skill Score (FSS) have been computed, using ANTILOPE data as reference (Fig. 11). Hereafter, normalised difference of FSS are represented and negative difference are related to improvement of the forecast quality using OPERA radar data in the assimilation system. These scores are calculated for rain thresholds varying from 2 to 50 mm. HSS is clearly improved in EXPE. Using the same thresholds and considering window sizes from 0.1° to 1° (about 10 to 100 km), FSS also confirms that assimilating OPERA radar data improves forecast quality.

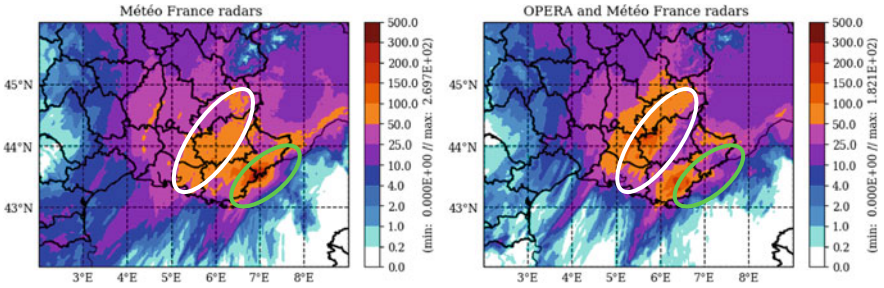

Fig. 10 24 hour accumulated precipitation in AROME-France forecast without (left) and with (right) assimilation of OPERA radar data. Forecast time 24 h, base time 1st December 2019 0 h UTC

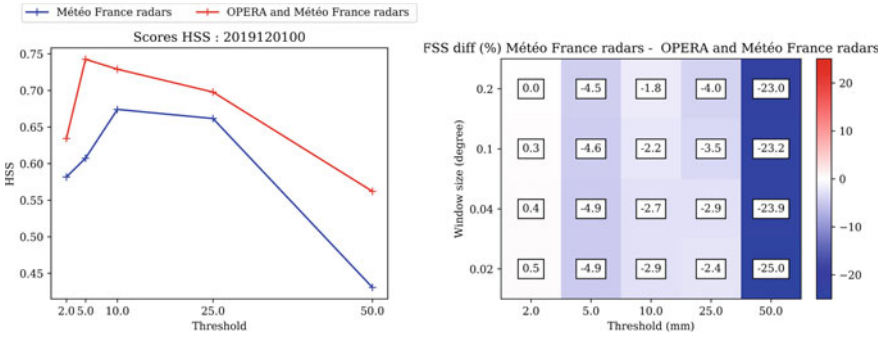

Fig. 11 Heidke Skill Score (HSS, left) and normalised difference of Fractions skill score (FSS, right) of 24 h accumulated precipitation for various thresholds (2–50 mm). Base time: 1st December 2019 0h UTC, forecast range 24 h. ANTILOPE accumulated precipitation is the reference

Another indication of the forecast quality is the variability from one run to the other. Comparison on the same validity period (1st and 2nd December 2019 0h UTC) of three model runs with different base time is displayed in Fig. 12. The results of the three following forecasts in the EXPE configuration show a quite consistent description of the two distinct areas of heavy precipitation. From one run to the other, the underestimation in the "Var" area is less pronounced with the most recent forecasts. On the contrary, a degradation is observed in the REF experiment with more recent forecast. This result indicates that taking into account radar data observations over the Mediterranean sea, in particular from Spain, in the assimilation system allows to better describe rainy areas over South-East of French territory, specially during a Mediterranean event such as the one described here.

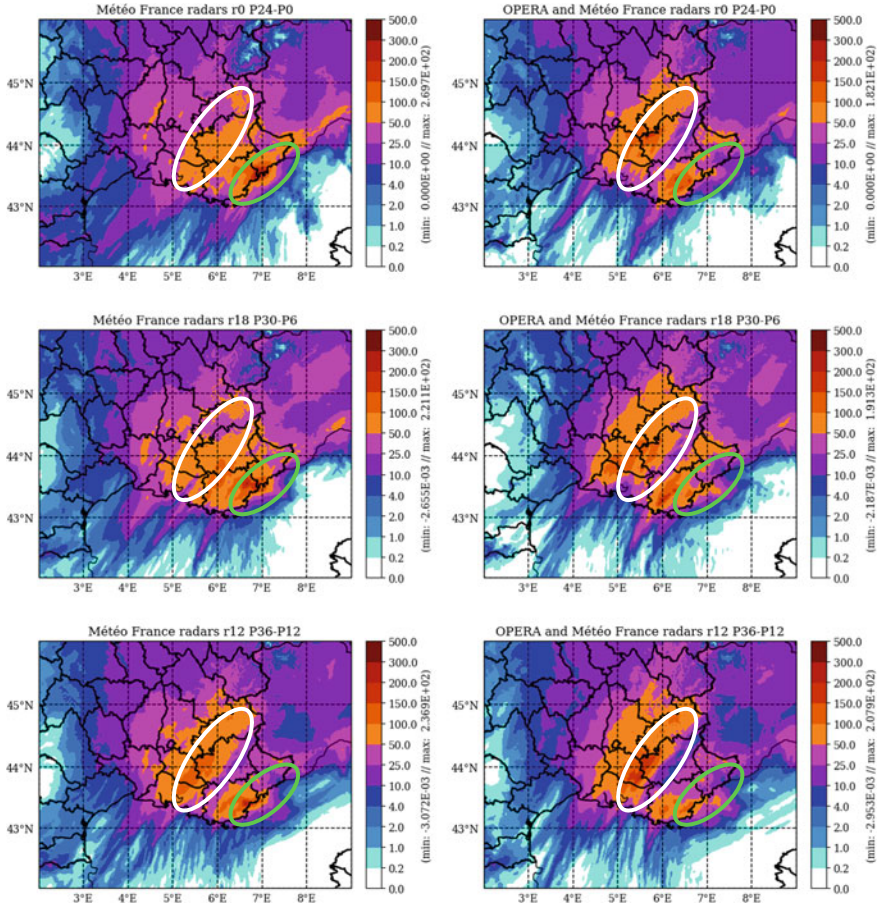

Fig. 12 24 hour accumulated precipitation in AROME-France forecast without (left) and with (right) assimilation of OPERA radar data between 1st and 2nd December 2019 0 h UTC. Base hour and forecast terms vary from top to bottom: 1st December 2019 0 h UTC, forecast 24 h (top), 30th November 2019 18 h UTC, forecast 30 h minus 6 h (middle), 30th November 2019 12 h UTC, forecast 36 h minus 12 h (bottom)

4 Conclusion

In an operational convective-scale numerical weather prediction system such as AROME-France, the assimilation of high temporal and spatial resolution observation such as radar data has significant impact on forecast performances. Since 2008, Météo-France radar data are used in the AROME-France 3D-Var assimilation system. Both radial winds and reflectivities are taken into account with two different methods (3D-Var for radial winds and 1D+3D-Var for reflectivities). Additional radar data from neighbouring countries in Europe (62 radars from 9 countries) has

been a challenge because of the heterogeneity in the national radar networks. In this framework, the OPERA program aims to unify data not only in terms of format but also through the production of a common quality index. This homogeneity allows OPERA radar data to be used uniformly, whatever the originating country.

Firstly, the OPERA radar data have been described: what data are received from OPERA, in particular what processing are used at OPERA level and what quality indices are produced. Then, a selection of the data has been developed in the framework of AROME-France 3D-Var data assimilation system in order to keep data according to the quality index defined at OPERA level. A monitoring experiment of these data during 5 days has shown that OPERA radar data quality is suitable for assimilation purposes. The pre-processing developed for these observations has allowed quality observations to be retained. These good quality observations will contribute the improvement of AROME-France data assimilation system.

Secondly, active assimilation of OPERA radar data in the AROME-France 3D-Var system has been evaluated. Comparison of experiments with and without assimilation of OPERA radar data has been studied in order to quantify the impact of adding these specific observations in the AROME-France system. Two-month experiments with and without OPERA radar data assimilation have been compared. The impact of adding these specific observations in AROME-France 3D-Var assimilation system is globally neutral but performances on rain accumulation forecasts are improved in short time range, when data assimilation is more efficient. In addition to this two-month experiments, a Mediterranean case associated with heavy rain rates over South-East of French territory has been studied in order to evaluate the impact of OPERA radar data assimilation. In this specific meteorological case, Spanish radars appear to have a significant impact over the French territory. For this situation, considering OPERA radar data leads to a better description of the geographic extent of the heavy rain areas and a better quantification of rain accumulations, in particular for higher thresholds. Moreover, the variability between model runs is reduced. These results confirm that adding OPERA radar data in AROME-France improves forecast quality.

Since January 2020, OPERA radar data are part of the Météo-France operational AROME-France data assimilation system.

References

Amodei M, Sanchez I, Stein J (2015) Verification of the French operational high-resolution model AROME with the regional Brier probability score. Meteorol Appl 22(4):731–745

Brousseau P, Desroziers G, Bouttier F, Chapnik B (2014) A posteriori diagnostics of the impact of observations on the AROME-France convective-scale data assimilation system. Q J R Meteorol Soc 140(680):982–994

Brousseau P, Seity Y, Ricard D, Léger J (2016) Improvement of the forecast of convective activity from the AROME-France system. Q J R Meteorol Soc 142(699):2231–2243

Gustafsson N, Janjić T, Schraff C, Leuenberger D, Weissmann M, Reich H, Brousseau P, Montmerle T, Wattrelot E, Bučánek A et al (2018) Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. Q J R Meteorol Soc 144(713):1218–1256

Laurantin O (2013) ANTILOPE: hourly rainfall analysis over France merging radar and rain gauge data. In: 11th international precipitation conference, Ede-Wageningen, The Netherlands

Montmerle T, Faccani C (2009) Mesoscale assimilation of radial velocities from Doppler radars in a preoperational framework. Mon Weather Rev 137(6):1939–1953

Saltikoff E, Haase G, Delobbe L, Gaussiat N, Martet M, Idziorek D, Leijnse H, Novák P, Lukach M, Stephan K (2019) OPERA the radar project. Atmosphere 10(6):320

Santurette P, Joly A (2002) ANASYG/PRESYG, Météo-France's new graphical summary of the synoptic situation. Meteorol Appl 9(2):129–154

Seity Y, Brousseau P, Malardel S, Hello G, Bénard P, Bouttier F, Lac C, Masson V (2011) The AROME-France convective-scale operational model. Mon Weather Rev 139(3):976–991

Tabary P (2007) The new French operational radar rainfall product. Part I: methodology. Weather Forecast 22(3):393–408

Tabary P, Desplat J, Do Khac K, Eideliman F, Gueguen C, Heinrich J (2007) The new French operational radar rainfall product. Part II: validation. Weather Forecast 22(3):409–427

Wattrelot E, Caumont O, Mahfouf J-F (2014) Operational implementation of the 1D+3D-Var assimilation method of radar reflectivity data in the AROME model. Mon Weather Rev 142(5):1852–1873

The 2020 Global Operational NWP Data Assimilation System at Météo-France

F. Bouyssel, L. Berre, H. Bénichou, P. Chambon, N. Girardot, V. Guidard, C. Loo, Jean-François Mahfouf, P. Moll, C. Payan, and D. Raspaud

Abstract The main features of the 2020 version of global Numerical Weather Prediction (NWP) model ARPEGE run operationally at Météo-France are described. This spectral model, developed in collaboration with the Integrated Forecasting System (IFS) of ECMWF, has a tilted and rotated horizontal grid that allows to reach a resolution of 5 km over Europe. The initial conditions are provided by an incremental 4D-Var data assimilation system with a 6-hour time window. Two inner-loops are performed respectively at 100 and 40 km. A comprehensive set of observations is assimilated with a dominance of satellite data representing 90% of them. However in terms of information content, conventional observations reach a fractional value of 20%. A 50-member ensemble data assimilation system based on low resolution 4D-Var is used to estimate daily background error covariances. The most recent improvements on this system regarding model resolutions, ensemble size and observation usage, that took place between mid-2019 and mid-2020, are presented with a selection of evaluations in terms of analysis and forecast skill scores.

1 Introduction

Weather forecasts at various ranges (from few hours to several months) issued by National Meteorological Services rely on outputs from Numerical Weather Prediction (NWP) models that need to be accurately initialized by Data Assimilation (DA) systems. The steadily gain in accuracy of NWP models is the result of improvements on four major aspects: the increase in spatial (horizontal and vertical) model resolutions, the more accurate description of physical processes (with frequently additional prognostic variables), the increase in number of assimilated observations, the improvement of DA algorithms to extract more efficiently information contained

F. Bouyssel · L. Berre · P. Chambon · N. Girardot · V. Guidard · C. Loo · J.-F. Mahfouf (✉) · P. Moll · C. Payan · D. Raspaud
CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
e-mail: jean-francois.mahfouf@meteo.fr

H. Bénichou
Direction des opérations, Météo-France, Toulouse, France

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_25

in observations (particularly through the use of ensembles). Each of these aspects requires additional computational resources (CPU and memory) from High Performance Computers (HPCs) with also increased data storage and improved database management. Therefore major changes that take place on operational NWP systems are very often associated with HPC upgrades.

This chapter illustrates recent changes of the global NWP DA system that took place at Météo-France between mid-2019 and mid-2020, thereby corresponding to a large extent to the configuration that has been migrated from the *HPC Bull DLC B710* to a new *HPC Atos Sequana XH2000*. The main features of the global NWP model are summarized in Sect. 2 . The DA system is described in Sect. 3 for the deterministic high resolution 4D-Var system that is coupled to an Ensemble DA (EDA) system providing background error statistics. An overview of the observational usage is also given in this section, with a summary of recent evolutions. Then, in Sect. 4 a number of results compare the performance of the current operational NWP system against the previous one. Finally, conclusions and planned evolutions are summarized in the last section.

2 Main Features of the Global NWP Model

The global NWP model run operationally at Météo-France since 1992 and named ARPEGE (Action de Recherche Petite Echelle Grande Echelle) has been jointly developed with the Integrated Forecasting System (IFS) of the European Centre for Medium range Weather Forecasts (ECMWF) to foster developments on variational data assimilation in both centers (Courtier et al. 1991). This is a spectral model in the horizontal direction (field decomposition in spherical harmonics with a triangular truncation) with an original grid structure allowing an increased resolution over a specific region of the globe, by a tilted and stretched conformal mapping described in Courtier and Geleyn (1988). The vertical is discretized in finite elements with a staggered grid and a hybrid pressure terrain-following η coordinate system (Simmons and Burridge 1981). The temporal integration is performed with a two-time level Semi-Implicit Semi-Lagrangian (SISL) scheme (Ritchie et al. 1995). The collocation grid of the spectral transforms is a linear reduced Gaussian grid as described in Hortal and Simmons (1991). At the beginning of the project, the prognostic equations of the ARPEGE/IFS model were written with the hydrostatic assumption. Since then, they have been generalized to the compressible Euler equations by Bubnová et al. (1995) using the pressure coordinate change proposed by Laprise (1992). Even though not yet adopted for the global model, it is used for convective-scale versions of the ARPEGE/IFS code adapted to the limited area geometry as described in Termonia et al. (2018).

The horizontal and vertical resolutions of the ARPEGE model have steadily increased with time starting in 1992 with a resolution around 150 km over France (region with highest resolution) and 21 vertical levels. In this chapter we focus on the current highest horizontal resolution of ARPEGE (since July 2019) over Europe

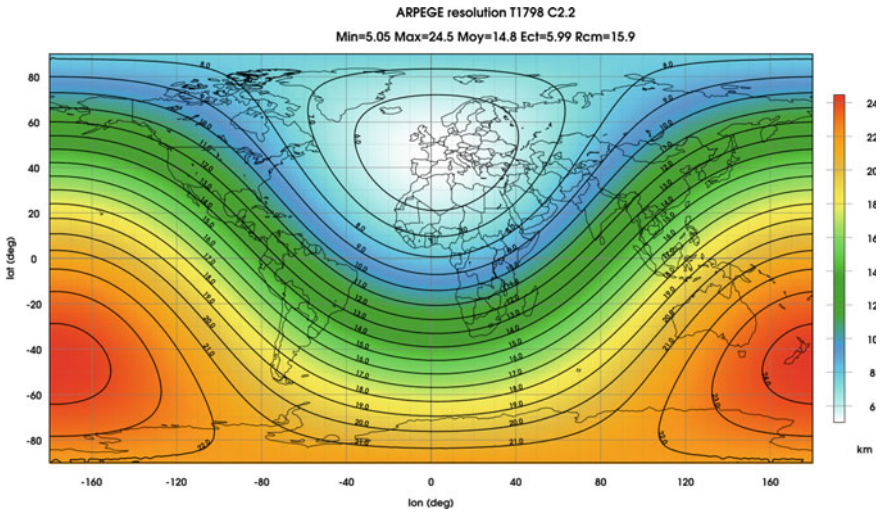

Fig. 1 Horizontal resolution (in km) of the global model ARPEGE resulting from the tilted and stretched grid (2020 configuration)

which is around 5 km and 105 vertical levels. The corresponding truncation is T_L1798 with a stretching factor of c = 2.2 leading to a resolution around 24 km at the antipodes of the numerical pole (New Zealand) as displayed in Fig. 1. This high resolution will be compared to a coarser one (T_L1198c2.2) that was operational at Météo-France between 2015 and 2019 (the highest and lowest resolutions being respectively 7.5 km and 35 km).

The ARPEGE model computes the time evolution of a set of ten prognostic variables: logarithm of surface pressure, temperature, horizontal wind vorticity and divergence, specific humidities for water vapor, cloud liquid water, cloud ice water, rain water, snow water and Turbulent Kinetic Energy (TKE). A set of parametrization schemes describes the effects of sub-grid physical processes on resolved variables: longwave radiation from RRTM (Mlawer et al. 1997), shortwave radiation from an improved version of Fouquart and Bonnel (1980), deep moist convection by a mass-flux scheme with a Kuo-type closure (Bougeault 1985), shallow convection by a mass-flux approach (Bechtold et al. 2001), vertical diffusion by a TKE scheme (Cuxart et al. 2000) with the mixing length formulation of Bougeault and Lacarrre (1989), stratiform clouds and precipitation by the Kessler-type prognostic microphysical scheme of Lopez (2002), gravity wave drag and low level blocking effects (Catry et al. 2008), surface processes from the SURFEX platform (Masson et al. 2013). The deep moist convection scheme, initially designed for large-scale models where cloud fraction within the grid box can be neglected, has been somewhat adapted to the 5 km grid corresponding. to the *grey zone of convection* In practice, convective tendencies are modified in regions where grid-scale vertical motions are large (removal of moisture convergence) to prevent the triggering of spurious grid-

point storms (Marquet et al. 2019). The SISL scheme allows a model time step of 240 s leading to a 4-day forecast within 70 min on 30 nodes (40 CPU Intel Broadwell cores per node) of the *HPC Bull DLC B710*. This rather long model time step, compared to the scale of some microphysical processes, has required to consider precipitation sedimentation in a statistical way (Bouteloup et al. 2011).

3 Description of the Data Assimilation System

3.1 High Resolution 4D-Var System

The ARPEGE model has a dedicated DA system to produce its initial conditions. This is a four dimensional variational (4D-Var) system that has been developed together with the numerical code of the model since the tangent linear and adjoint versions were also required to solve efficiently the minimization problem (Le Dimet and Talagrand 1986). The 4D-Var system formulation is *multi-incremental* as it has been proven to be the only affordable solution of this expensive numerical problem for operational NWP applications (Courtier et al. 1994). The minimisation is thus performed in terms of increments (leading to a quadratic cost function) at low resolution with two-outer loops. Even though the geometry of the non-linear model is tilted and stretched, it has been chosen to keep a uniform grid for the minimisations. The first minimisation is performed at truncation T_L224 (around 100 km) whereas the second one is done at higher resolution T_L499 (around 40 km). A set of 40 iterations is chosen for each minimisation (compromise between the computing time and the convergence of the cost function). The convergence of the numerical algorithm (conjugate gradient) is improved by a pre-conditioning using eigenvectors from the previous minimisation (Lanczos algorithm). The first minimisation is undertaken with almost adiabatic tangent linear and adjoint models (simplified vertical diffusion) whereas a set of improved linearized physical parameterizations is used in the second minimization for vertical diffusion, gravity wave drag and large scale condensation (Janisková et al. 1999). This configuration will be compared to a similar one but with coarser resolutions of the inner loops: T_L149 (around 150 km) and T_L399 (around 60 km) corresponding to the operational settings between 2015 and 2019. The control of fast gravity waves is done with a Digital Filter Incremental (DFI) constraint proposed by Gauthier and Thépaut (2001) and an external digital filtering.

An operational version of the 4D-Var DA system is cycled every 6 h: this is the *assimilation suite*. For this suite the observation cut-off is long enough to collect the vast majority of available measurements before starting the *screening* process (data selection by a systematic comparison of observations with their model counterpart) that is between 7 and 8 h after analysis time. Another suite called the *production suite* is run with a much shorter cut-off (between 2 and 3 h) in order to provide early delivery forecasts that are also used as boundary conditions for the convective scale model

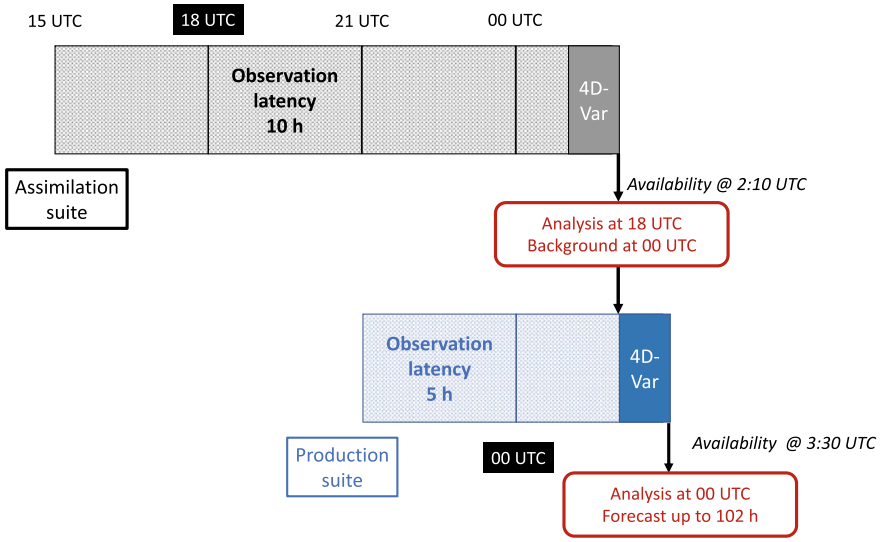

Fig. 2 Structure of the operational production and assimilation suites of the ARPEGE model. The shaded areas correspond to the observation latency before launching the 4D-Var assimilation

AROME (Seity et al. 2011). The disadvantage is that not all observations are available in the database when starting the *screening*, resulting in less accurate analyses and subsequent forecasts. Therefore the production suite is not cycled meaning that the background is taken from a 6-h forecast of the *assimilation suite*. The assimilation window is set to 6 h (+/- 3 h around analysis time). Observations are gathered by 30 min time slots except at the beginning and at the end of the window with 15 min time slots. These two suites and their interactions are displayed in Fig. 2. Forecasts from the *production suite* range between 60 and 114 h depending on analysis time.

Every 6 h an independent surface analysis is performed based on Optimal Interpolation (OI). Over oceans, a Sea Surface Temperature (SST) analysis uses BUOY and SHIP surface reports as observations, the previous SST analysis as a background (i.e. persistence) and a relaxation towards the daily SST analysis product OSTIA (Donlon et al. 2012). Over continents, screen-level temperature and relative humidity measurements from SYNOP reports are interpolated on the model grid using an OI scheme and 6-h forecasts of these parameters as background field. Then, the analysis increments are converted into soil temperature and soil moisture corrections using a method described in Giard and Bazile (2000) with a number of recent changes available in Mahfouf et al. (2009). This simple method prevents spurious positive feedbacks between soil moisture, surface evapotranspiration, and precipitation.

3.2 Ensemble Data Assimilation System

An Ensemble Data Assimilation (EDA) system named AEARP (Assimilation d'Ensemble ARPEGE) is coupled to the 4D-Var system in order to provide flow dependent background error covariances. The method, initially proposed by Fisher (2003), relies on an error simulation approach: observation errors are mimicked by adding observation perturbations, which are random draws of the specified observation error covariance matrix; model errors are simulated through a multiplicative inflation of forecast perturbations (Raynaud et al. 2012). The AEARP ensemble is made of 50 members, using lower resolution and a simplified 4D-Var configuration, compared to the deterministic run. To reduce the computational cost of each member, the non-linear ARPEGE model is run at low resolution $T_L 499$ with a regular grid and the 4D-Var incremental formulation is performed at $T_L 224$ with only one outer-loop (40 iterations). The covariance matrix of background errors is decomposed in a sequence of sparse operators (flow-dependent balance equations and multi-variate scale-dependent regressions, error standard deviations and spatial correlations) as formulated by Derber and Bouttier (1999). The set of perturbed short-range forecasts from the EDA allows to estimate local background error standard deviations which need to be spatially filtered from noise induced by sampling errors (Berre and Desroziers 2010). Moreover, to represent heterogenous 3D correlation functions, a wavelet block-diagonal formulation of the correlation matrix is employed (Fisher 2003). Correlations functions are estimated from the ensemble states over the three most recent cycles (akin to a moving time average), in order to increase sample size (Berre et al. 2015). The previous operational configuration was made of only 25 members at lower resolution ($T_L 479$ for the trajectory and $T_L 149$ for the minimization). Correlation functions were estimated through a time average over the six most recent cycles. In order to maintain a similar execution time for each 4D-Var task on a limited number of nodes of the *HPC Bull DLC 710*, the set of assimilated observations has been reduced (the three most recent infrared hyperspectral sounders out of six). On top of providing background error covariances to the deterministic 4D-Var, the AEARP perturbed states are also used to define the initial conditions of the 35 member ARPEGE Ensemble Prediction System (Descamps et al. 2015) and to provide the lateral boundary conditions of the AROME EDA system (Montmerle et al. 2018).

3.3 The Observation Usage

3.3.1 General Features

The availability of a 4D-Var DA system at Météo-France since 2000 has allowed a progressive increase of satellite data usage. This can be seen in Fig. 3 showing the evolution of the number of monthly observations assimilated in the ARPEGE model.

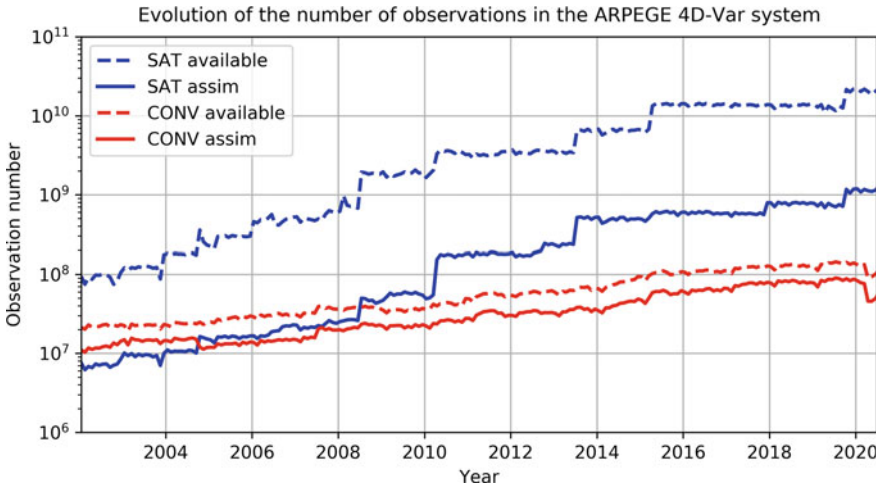

Fig. 3 Evolution of the number of observations assimilated and monitored (i.e. available) monthly in the ARPEGE model over the period 2002–2020 for conventional (CONV: red curves) and satellite (SAT: blue curves) datasets

Over almost 20 years the amount of satellite data has increased by two orders of magnitude whereas during the same period the amount of non-satellite (conventional) data has only increased by a factor of 10 (mostly from aircraft reports). As a result, in the current operational 4D-Var, 93% of the observations come from spaceborne instruments. As also shown in Fig. 3 the amount of satellite data monitored (i.e. available) in the ARPEGE 4D-Var is considerably larger than the one effectively assimilated by about a factor of ten (this is not a feature specific to the Météo-France system). Therefore this one to nine ratio between conventional and satellite observations could even be smaller if satellite data were to be used more efficiently. On the other hand, the fraction of assimilated conventional observations with respect to those monitored is considerably larger. Interestingly, the significant decrease of conventional observations initiated in March 2020 comes from very limited aircraft measurements induced by the Covid-19 worldwide pandemic.

It is known that before assimilation, data have to be carefully checked through a number of quality controls. For example, satellite radiances are systematically compared to a model equivalent in clear sky conditions using the radiative transfer model RTTOV version 11 (Saunders et al. 2018). When the difference between observed and simulated radiances is too large (background check) the observation is discarded and assumed to be cloudy. For hyperspectral infra-red sounders, it is possible to identify a cloud top and not discard the full pixel by keeping channels having most of their weighting function above cloud top (McNally and Watts 2003). Moreover, a CO_2 slicing method allows an estimation of cloud top pressure and effective emissivity to assimilate a number of infra-red radiances affected by clouds (Pangaud et al. 2009). Another important aspect of observation usage concerns the spatial sampling

also named *data thinning*. This pre-processing step is necessary since horizontal correlations of observation errors are not accounted for (both because it simplifies the 4D-Var formulation and also because their modeling is rather complex due to the unstructured nature of observing systems). For satellite data the spatial sampling varies between 100 and 125 km depending upon instruments. The specification of observation errors is based on the knowledge of instrumental errors, and also from background departure statistics. A posteriori diagnostics, as defined by Desroziers et al. (2005a), are used for some observation types. It is also based on pragmatic "trial and error" experiments where the improvement or the degradation of the forecast scores is assessed (since an observation error inflation can partly compensate for neglecting spatial correlations). Satellite radiances are generally affected by biases depending on many factors (scan angle, satellite position, air mass properties, ...), thus they need to be corrected before assimilation. With the advent of hyperspectral sounders having thousands of channels the development of adaptive bias correction methods became necessary to replace rather tedious static methods based on past trial periods. The ARPEGE 4D-Var system uses the variational bias correction scheme developed by Auligné et al. (2007) where the control vector is extended to a set of coefficients depending on predictors proposed by Harris and Kelly (2001). Conventional observations also need to be corrected from known biases. This is the case for radiosondes where the humidity bias is corrected using the proposal of Agust-Panareda et al. (2009) and the temperature bias at high altitudes due to sunlight radiative effects is removed by an empirical formulation depending upon solar elevation and sonde type.

3.3.2 Recent Developments

Table 1 summarizes the major observing systems assimilated nowadays in the ARPEGE 4D-Var. Changes to the operational NWP system usually take place every 18 months and they include changes to the observation usage (new datasets or improved usage of existing ones). However, during a given 18-month period, additional changes can take place to the global observing system that need to be considered before the next version of the NWP suite. A number of new observational datasets were introduced since the implementation of the high resolution system in July 2019, as described more precisely in the following.

In July 2019, the most important changes on observations were:

- Assimilation of a number of instruments from the EUMETSAT/ESA Metop-C polar orbiting satellite: microwave sounders AMSU-A and MHS, GNSS-RO bending angles from the receiver GRAS
- Assimilation of the microwave sounder ATMS onboard the US NOAA-20 polar orbiting satellite
- Assimilation of ocean surface winds from the Indian Ku-band scatterometer OSCAT onboard ScatSat-1

Table 1 Summary of the observing systems assimilated in the 4D-Var system of the ARPEGE model in May 2020. Instruments from Metop-C and NOAA-20 have been introduced progressively between July 2019 and May 2020. The maximum number of satellite channels per instrument is given in the last column with their sensitivity to temperature (T), water vapor (WV) and ozone (O_3). Similarly, the spectral bands (VIS, IR, WV) used for the derivation of atmospheric wind vectors are given. In-situ sensors measure surface pressure (P_s), temperature (T), relative humidity (RH) and winds. The ground based GNSS (GB-GNSS) receivers provide Zenith Total Delay (ZTD) measurements informative on integrated water vapor

Observation type	Instruments / Platform	Comments
LEO IR radiances	IASI (Metop-A/B/C)	129 channels (T, WV, O_3)
	CrIS (S-NPP, NOAA-20)	68 channels (T, WV)
	AIRS (Aqua)	72 channels (T, WV)
GEO IR radiances	SEVIRI (Meteosat-8/11)	6 channels (T, WV)
	AHI (Himawari-8)	5 channels (T, WV)
LEO MW radiances	AMSU-A (NOAA-15/18/19, Aqua, Metop-A/B/C)	9 channels (T)
	ATMS (S-NPP, NOAA-20)	14 channels (T, WV)
	MHS (NOAA-19, Metop-A/B/C)	3 channels (WV)
	MWHS-2 (FY-3C)	3 channels (WV)
	SAPHIR (Megha-Tropiques)	6 channels (WV)
	SSMI/S (DMSP F-17/18)	14 channels (T, WV)
	GMI (GPM-Core)	2 channels (WV)
GNSS-RO bending angles	GRAS (Metop-A/B/C)	above 8 km
	IGOR (COSMIC-1)	–
	IGOR (TerraSAR-X)	–
	IGOR (TanDEM-X)	–
Scatterometer surface winds	C-band ASCAT (Metop-A/B/C)	neutral 10-m winds
	Ku-band OSCAT (ScatSat-1)	neutral 10-m winds
AMVs	SEVIRI (Meteosat-8/11)	(WV, IR, VIS)
	ABI (GOES-16/17)	(WV, IR, VIS)
	AHI (Himawari-8)	(WV, IR, VIS)
	MODIS (Terra, Aqua)	(WV, IR)
	AVHRR (NOAA-15, 18, 19)	(IR)
Aircrafts	AIREP, AMDAR	(T, winds)
Sondes	PILOT, TEMP, Profilers	(T, RH, winds)
Surface	BUOY, SHIP, SYNOP, GB-GNSS	(P_s, T, RH, winds, ZTD)

- Assimilation of Atmospheric Motion Vectors (AMVs) from the US GOES 16 geostationary satellite imagery
- Assimilation of low peaking channels from IASI over continents
- Introduction of interchannel error correlations for infra-red hyperspectral sounders IASI and CrIS
- Use of a variational bias correction scheme for ground based GNSS observations of Zenith Total Delays (ZTD).

In October 2019, additional observations were introduced:

- Assimilation of infra-red hyperspectral sounders IASI from Metop-C and CrIS from NOAA-20

In January 2020, a remaining instrument from Metop-C was added:

- Assimilation of ocean surface winds from the C-band scatterometer ASCAT

Finally in May 2020:

- Assimilation of AMVs from the US GOES 17 geostationary satellite imagery

We examine the proportion of observations assimilated in the ARPEGE 4D-Var in December 2019 together with the information content quantified by the *Degree of Freedom for Signal* (DFS). This diagnostic is regularly used at Météo-France to quantify the impact of observations on ARPEGE analyses. It measures the relative error reduction induced by the observations on the background error covariance matrix $\mathbf{B}$. If $\mathbf{A}$ is the analysis error covariance matrix, the DFS is written as:

$$DFS = Tr[(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}] \tag{1}$$

Since the analysis covariance matrix can be expressed as:

$$\mathbf{A} = (\mathbf{I} - \mathbf{KH})\mathbf{B} \tag{2}$$

where $\mathbf{K}$ is the Kalman gain of the analysis equation (link between analysis increments and background departures) and $\mathbf{H}$ is the tangent-linear version of the observation operator that projects the model state into the observation space, then the DFS is simply $Tr(\mathbf{KH}) = Tr(\mathbf{HK})$. Using a randomization technique, it is possible to obtain the DFS from a set of 4D-Var assimilations, each one being run with randomly perturbed observations following the observation error covariance matrix $\mathbf{R}$ as proposed by Desroziers et al. (2005b). Indeed if $\mathbf{x}_a$ is an unperturbed analysis obtained with an observation vector $\mathbf{y}_o$ and $\mathbf{x}_a^*$ is a perturbed analysis obtained with an observation vector $\mathbf{y}_o^*$, the DFS can be approximated as:

$$DFS \simeq [\mathbf{y}_o^* - \mathbf{y}_o]^T \mathbf{R}^{-1} [H(\mathbf{x}_a^*) - H(\mathbf{x}_a)] \tag{3}$$

since $[H(\mathbf{x}_a^*) - H(\mathbf{x}_a)] \simeq \mathbf{HK}(\mathbf{y}_o^* - \mathbf{y}_o)$ and $Tr(\mathbf{A}) \simeq \zeta^T \mathbf{A} \zeta$ if ζ is a Gaussian random vector (Girard 1989).

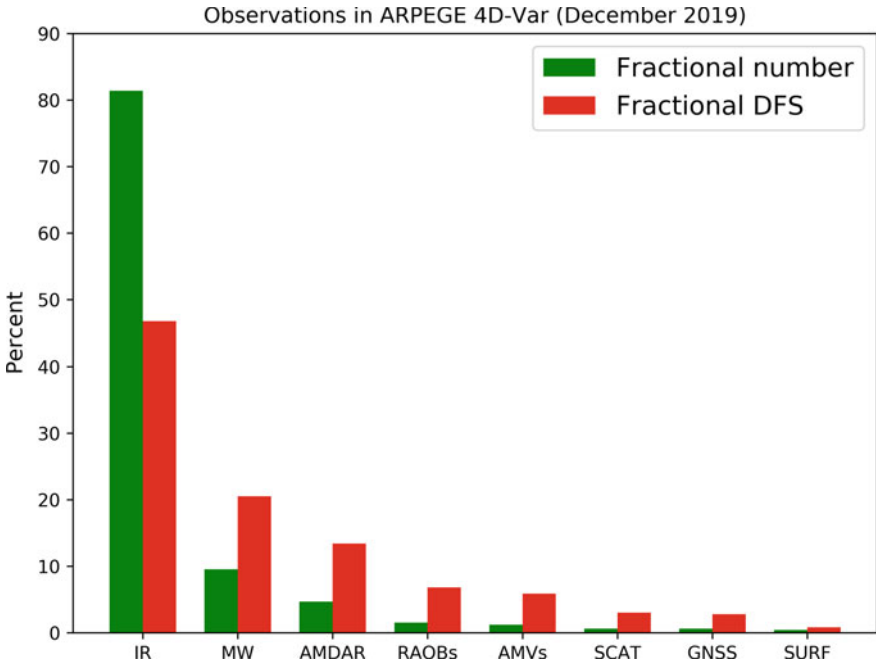

Fig. 4 Proportion of observations assimilated in the ARPEGE model for a 24-h period together with their information content measured by the fractional *DFS* (12 December 2019). Datasets are ranked by decreasing *DFS* values

In practice, a set of six 4D-Var assimilations from the deterministic high resolution system is run over a one day period (4 analysis cycles) in order to produce robust global diagnostics that are regularly updated on the Météo-France NWP monitoring web site.[1]

Figure 4 reveals that the observing system is dominated by infrared radiances (IR) since six hyperspectral sounders are assimilated in ARPEGE from October 2020. Indeed they represent 80% of the total observations. The remaining satellite observations come mostly from microwave radiances (9%). Winds from scatterometers, satellite imagery, and GNSS signals reach only 3%. In terms of conventional observations, aircrafts reports (AMDAR) and radiosoundings (RAOBs) represent 5% of the total number of observations (around 42 millions per day). When examining DFS values for the same period, this rather unbalanced observing system in terms of numbers reveals that the information content is more evenly distributed among the various observation types. The infrared sounders represent 45% of the total DFS that is quite comparable to the contribution of all other satellite instruments (36% but being dominated by microwave radiances (20%)). Despite their small percentage in number, conventional observations (AMDAR, RAOBs, SURF) contribute to

[1] http://www.meteo.fr/special/minisites/monitoring/menu.html.

Table 2 Main features of the two versions of the ARPEGE NWP model with its 4D-Var deterministic and ensemble systems compared in Sect. 3

	Operational version before July 2019 (**OLD**)	Operational version since July 2019 (**NEW**)
ARPEGE Deterministic	T_L1198c2.2L105 (7.5 km Western Europe) 4D-Var inner loops T_L149c1L105 + T_L399c1L105 5 forecasts per day up to 114 h	T_L1798c2.2L105 (5 km Western Europe) 4D-Var inner loops T_L224c1L105 + T_L499c1L105 5 forecasts per day up to 114 h
AEARP (EDA based on ARPEGE)	T_L479c1L105; 25 members 4D-Var: T_L149c1L105 Background error correlations averaged on 1.5 day and updated every 6 h	T_L499c1L105; 50 members 4D-Var: T_L224c1L105 Background error correlations averaged on 0.5 day and updated every 6 h

20% of the total DFS. The small value of the total DFS for a one day period (0.8 million) indicates that the observations do not contain enough information to control all degrees of freedom present in the model state vector (having a size about 100 millions). The measure of the DFS only provides information on the accuracy of the analysis (in theory even without estimating the analysis state). However, in practice it has been noticed that the ranking of the observing systems by DFS values is rather similar to the one obtained from *Forecast Sensitivity to Observation Impact* ($FSOI$) studies which examine the impact of observations on short-range forecast accuracy, particularly when a fair contribution of humidity forecast errors is considered (Marquet et al. 2020).

4 Behavior of the Current Operational Configuration

The high resolution ARPEGE model with significant revisions to its assimilation system (higher resolution of minimisations and upgraded EDA) has been tested against the Météo-France operational system over several months before becoming the new operational system on 2 July 2019. The main differences between this system (named NEW hereafter) and the previous one (named OLD hereafter) are summarized in Table 2. A selection of evaluations undertaken to assess the overall quality of this new NWP DA and forecasting systems is now presented by a comparison of results obtained with NEW and OLD.

A first assessment of the DA system is usually done by examining statistics of observations against model counterparts in terms of short-range forecasts (background departures: OmB) and analyses (analysis residuals: OmA). This provides a first sanity check where (OmA) values should be smaller than (OmB) ones, since the DA system is expected to bring the model state closer to the observations assimilated. Then, if the (OmB) values are smaller for the experiment than for the reference, it

reflects smaller errors of short-range forecasts towards independent observations (not yet assimilated). Contrary to long range forecasts that need large samples to produce statistically significant results, the global (OmB) and (OmA) statistics are often very robust on a small sample (e.g. one week). Over a one month period, the experiment (NEW) shows for almost all observing systems, reduced values of standard deviations of (OmA) and (OmB) with respect to the reference (OLD). The systematic reduction of (OmA) values is a consequence of the increased resolution of the analysis (T_L499 vs. T_L399) and of the introduction of correlated observation errors for IASI and CrIS instruments that represented 70% of the total number of observations in the OLD system. With correlated observation errors, the model state can slightly reduce its fit to those observations at the benefit of other observing systems. Indeed, variances of observation errors have not been modified when introducing interchannel error correlations. This is shown in Fig. 5 for radiosounding profiles (for wind, temperature and specific humidity) and for the microwave instrument ATMS, where the impact is more pronounced for water vapor channels than for temperature ones.

An objective evaluation of the forecasts is performed against meteorological fields that should represent as much as possible the true state of the atmosphere. At Météo-France the comparison is usually performed against ECMWF analyses[2] (providing a uniform coverage) and radiosounding measurements (closer to the truth but rather sparse over tropical regions and in the Southern Hemisphere). Similar signals with both references increase the confidence given to the results. The root-mean square errors (RMSE) of the OLD system are compared to the NEW system in relative percentage (differences divided by the OLD values). Positive values indicate smaller errors with the new (NEW) system. A statistical test allows to estimate the significance of the differences. Such results are displayed in Fig. 6 where for temperature and vector wind fields according to the various forecast ranges and pressure levels, the new system performs better. A slight degradation for temperature is nevertheless noticed in the stratosphere (Northern Hemisphere and Tropics) around 20 hPa. It comes from the fact that the rather coarse vertical resolution at those levels has not been modified with the model horizontal resolution increase leading there to a spurious stratospheric cooling (Lindzen and Fox-Rabinovitz 1989). This weakness has somewhat been reduced by increasing the horizontal diffusion on wind variables and adding an iteration in the search of the origin point of the Semi-Lagrangian scheme.

Since the geometry of the ARPEGE model has been defined in order to provide better forecasts over Europe for the short-range (up to 4 days), a specific NWP skill index has been set-up to evaluate such performances. This NWP index named *IP18* considers three upper air parameters: 500 hPa geopotential (Z500), 850 hPa temperature (T850) and 250 hPa wind (W250) at two forecast ranges (48 and 72 h) issued from the 00 UTC analyses. For each parameter, the RMSE is computed against radiosoundings over Europe. It is then compared and normalized by its value in 2008 as $100 \times (RMSE - RMSE_{2008})/RMSE_{2008}$. The global NWP skill index *IP18* is obtained by an arithmetic average of the six scores. The *IP18* values are displayed in Fig. 7 together with their six components for the current operational system (NEW)

[2] They are averaged on a 0.5° x 0.5° lat/lon grid.

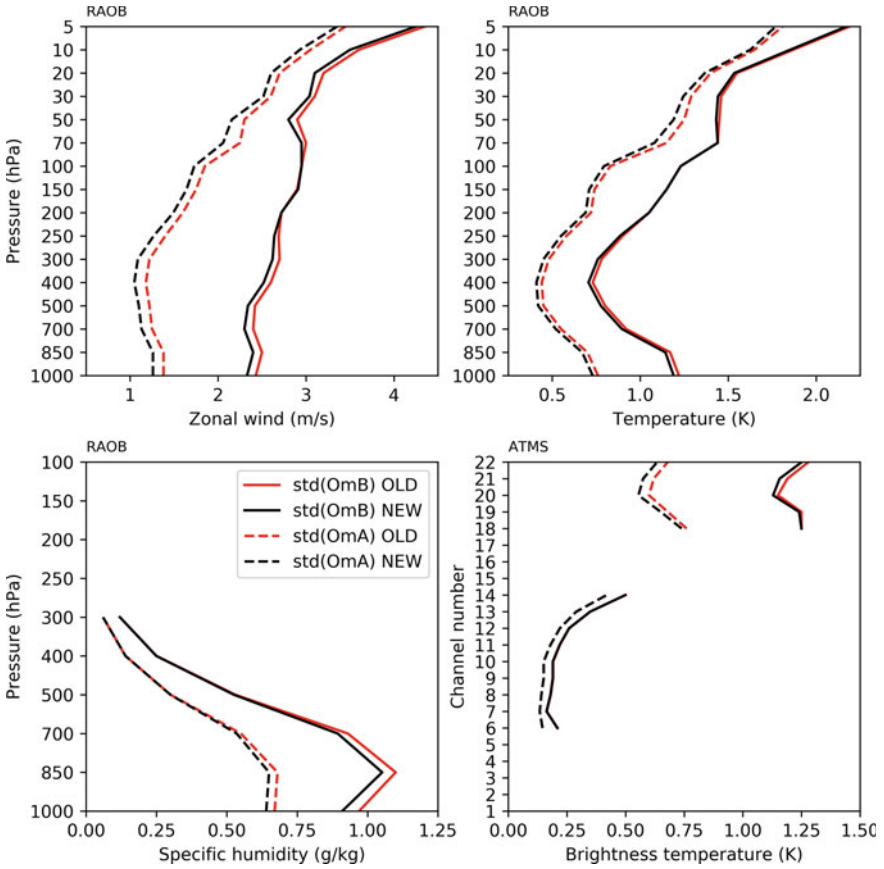

Fig. 5 Global standard deviations of background departures (OmB: solid curves) and analysis residuals (OmA: dashed curves) for zonal wind component, temperature and specific humidity from radiosoundings (RAOB), and radiances from the microwave sounder ATMS (6 to 14 : temperature channels and 18 to 22: water vapor channels) over one month in November 2018. The red curves correspond to the previous operational system (OLD) and the black curves correspond to the new operational system (NEW)

and the previous one (OLD). Positive values indicate improvements with respect to the NWP system in 2008. These values are averaged over a 5-month trial period before going operational. The new system (NEW) that went operational in July 2019 has an *IP18* value of 11% leading to a 3% increase with respect to the previous operational system (OLD). All components of this NWP skill index are improved but the low level temperature at 72 h and the upper level winds show a larger increase in a relative sense.

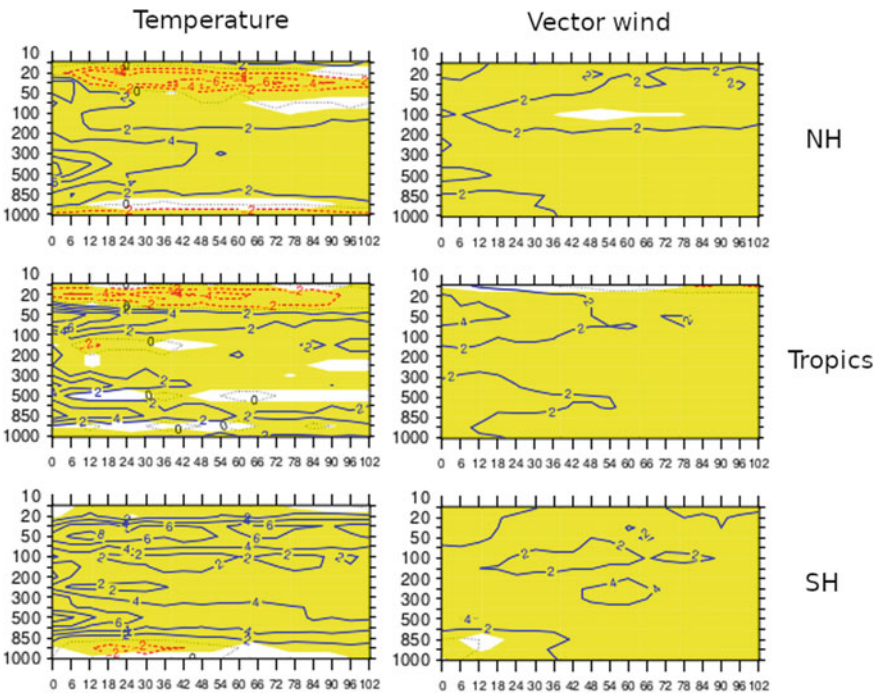

Fig. 6 Normalized RMSE values (in percent) of temperature (left panels) and vector wind (right panels) against ECMWF analyses between the previous operational system (OLD) and the new one (NEW) for forecast ranges up to 102 h. Positive values correspond to an improvement by the new system (reduced errors) and negative values correspond to a degradation by the new system (increased errors). The top row is for the Northern Hemisphere (NH), the middle row for the Tropics and the bottom row is for the Southern Hemisphere (SH). Yellow areas indicate where the differences are significant to 95% confidence level. The period ranges from January to May 2019

5 Conclusions and Planned Evolutions

The 2020 configuration of the operational global DA system run at Météo-France with the ARPEGE model using an incremental 4D-Var has been described and its performances assessed in terms of analysis diagnostics and forecasts scores. The major improvement that took place in July 2019 was an increased horizontal resolution of the tilted and stretched grid of the non-linear model leading to a physical grid size about 5 km over France. Additional changes were made to the DA system with increased resolutions of the increments both in the deterministic 4D-Var and in the EDA providing flow dependent background error statistics. Regarding this last system, the number of members has been increased from 25 to 50 members to provide more reliable statistics by reducing sampling noise. The observational data usage has been significantly modified with the addition of new instruments in particular those from Metop-C and NOAA-20 satellites and by a better usage of

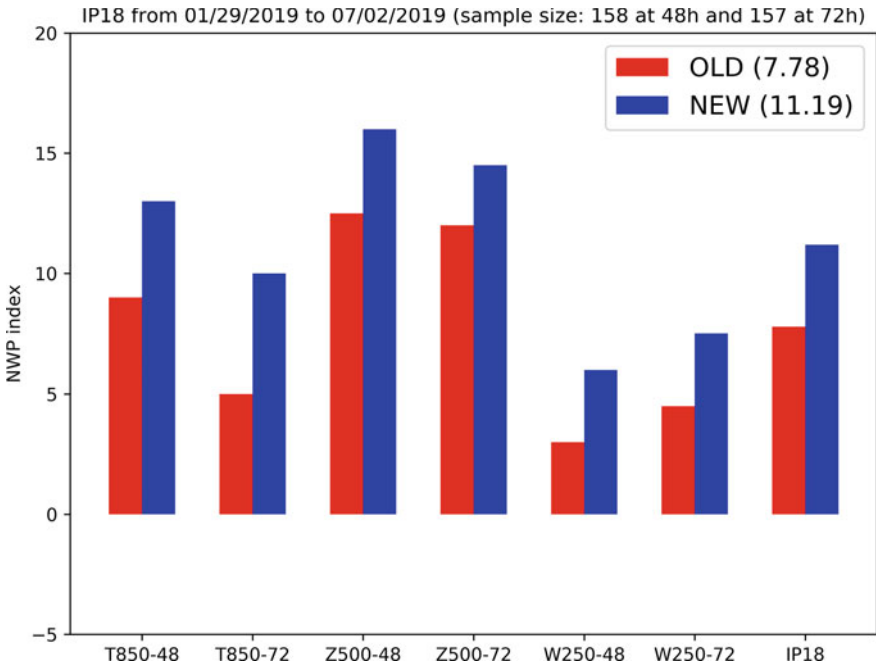

Fig. 7 Combined forecast skill scores of ARPEGE (*IP18* index) over Europe for a 5 month period in 2019 for the OLD and NEW NWP systems

already available observations (variational bias correction of ground based GNSS receivers, correlated observation errors for IASI and CrIS infrared sounders, low peaking channels of IASI over continental surfaces). Since then, new observations have been included such as two hyperspectral sounders (IASI on Metop-C and CrIS on NOAA-20). All these changes led to improved forecast scores as shown by the RMS error of the geopotential at 500 hPa at 72 h forecast range over Europe for a 8 year period compared against other global NWP models (Fig. 8). The quality of the ARPEGE model has steadily improved since mid-2019.

A number of evolutions to this global NWP system are ongoing. Regarding the observing systems the wind profiles from the spaceborne lidar of the satellite mission Aeolus and a set of additional GNSS-RO receivers (in particular those from the equatorial constellation COSMIC-2A) have been recently assimilated (June 2020). Developments on all-sky microwave radiances using a Bayesian inversion method as described in (Duruisseau et al. 2019) are now in a pre-operational stage with improved analyses of humidity in clouds and more accurate forecasts of severe weather events such as tropical hurricane trajectories. The new satellite programmes from EUMETSAT for both polar and geostationary orbits (EPS-SG[3] and MTG[4]) to start in 2023

[3] EUMETSAT Polar System -Second Generation.

[4] Meteosat Third Generation.

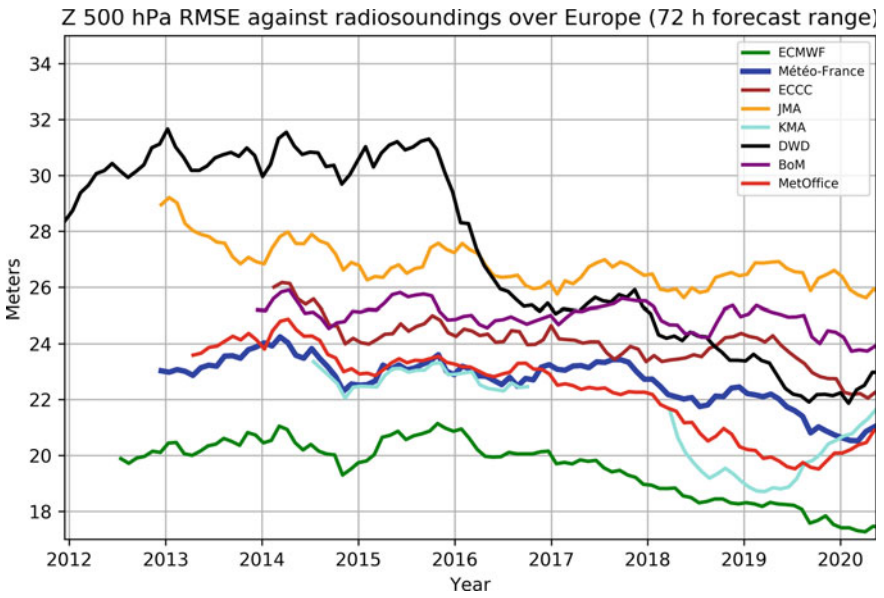

Fig. 8 Root-mean square error of the geopotential at 500 hPa at 72 h forecast range over Europe against radiosoundings (annual running mean between 2012 and 2020) for eight global NWP models including the ARPEGE model (Météo-France)

will be carefully prepared. Indeed they will raise new challenges for exploiting their data efficiently in the DA assimilation context due to the huge amount of additional observations they will provide and also from the novelty of some instruments (e.g. geostationary infrared hyperspectral sounder, lightning imager, sub-millimeter radiometer). Regarding the evolution of the DA algorithm, the 4D-Var system shall evolve towards a 4D-Ensemble Variational (4D-EnVar) system (Desroziers et al. 2014) where the minimization is performed on a 4D state (allowing to relax the perfect model strong constraint) and evolved background error covariances are deduced from the ensemble. This requires specific localisations of spatial correlations including advection within the assimilation window (Desroziers et al. 2016), and variable transforms (Berre et al. 2017) to account for scale-dependent and multivariate features. With the 4D-EnVar system it will also be possible to consider new variables in the control vector such as hydrometeors for clouds and precipitation and to avoid the use of linearized physical parameterization schemes that have been difficult to maintain and develop at Mto-France over the last 15 years. With such evolution all-sky satellite radiances should be assimilated more efficiently.

Acknowledgements We would like to acknowledge the invaluable contribution of the following colleagues in the development of important numerical and physical aspects of the new high resolution version of the ARPEGE model: L. Auger, Y. Bouteloup, L. Descamps, and P. Marquet.

References

Agust-Panareda A, Vasiljevic D, Beljaars A, Bock O, Guichard F, Nuret M, Garcia Mendez A, Andersson E, Bechtold P, Fink A, Hersbach H, Lafore J-P, Ngamini J-B, Parker DJ, Redelsperger J-L, Tompkins AM (2009) Radiosonde humidity bias correction over the west African region for the special AMMA reanalysis at ECMWF. Q J R Meteorol Soc 135(640):595–617

Auligné T, McNally A, Dee D (2007) Adaptive bias correction for satellite data in numerical weather prediction. Q J R Meteorol Soc 133:631–642

Bechtold P, Bazile E, Guichard F, Mascart P, Richard E (2001) A mass-flux convection scheme for regional and global models. Q J R Meteorol Soc 127(573):869–886

Berre L, Arbogast E, Ménétrier B, Desroziers G (2017) Change of variable applied to mass and wind fields for covariance localization, 1–07. WMO CAS/JSC, WGNE Blue Book. http://bluebook.meteoinfo.ru/uploads/2017/docs/01_Berre_Loik_ChageVariableLocalisation.pdf

Berre L, Desroziers G (2010) Filtering of background error variances and correlations by local spatial averaging: a review. Mon Weather Rev 138(10):3693–3720

Berre L, Varella H, Desroziers G (2015) Modelling of flow-dependent ensemble-based background-error correlations using a wavelet formulation in 4D-Var at Météo-France. Q J R Meteorol Soc 141(692):2803–2812

Bougeault P (1985) A simple parameterization of the large-scale effects of cumulus convection. Mon Weather Rev 113(12):2108–2121

Bougeault P, Lacarrère P (1989) Parameterization of orography-induced turbulence in a mesobeta-scale model. Mon Weather Rev 117(8):1872–1890

Bouteloup Y, Seity Y, Bazile E (2011) Description of the sedimentation scheme used operationally in all Météo-France NWP models. Tellus A 63(2):300–311

Bubnová R, Hello G, Bénard P, Geleyn J-F (1995) Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. Mon Weather Rev 123(2):515–535

Catry B, Geleyn J-F, Bouyssel F, Cedilnik J, Brozková R, Derková M, Mladek R (2008) A new sub-grid scale lift formulation in a mountain drag parameterisation scheme. Meteorol Z 17(2):193–208

Courtier P, Freydier C, Geleyn JF, Rabier F, Rochas M (1991) The Arpege project at Mto-France. In: Seminar on numerical methods in atmospheric models, vol II. Shinfield Park, Reading, ECMWF, pp 193–232, 9–13

Courtier P, Geleyn J-F (1988) A global numerical weather prediction model with variable resolution: application to the shallow-water equations at Mto-France. Q J R Meteorol Soc 114(483):1321–1346

Courtier P, Thépaut J-N, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. Q J R Meteorol Soc 120(519):1367–1387

Cuxart J, Bougeault P, Redelsperger J-L (2000) A turbulence scheme allowing for mesoscale and large-eddy simulations. Q J R Meteorol Soc 126(562):1–30

Derber J, Bouttier F (1999) A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus A 51(2):195–221

Descamps L, Labadie C, Joly A, Bazile E, Arbogast P, Cébron P (2015) PEARP, the Mto-France short-range ensemble prediction system. Q J R Meteorol Soc 141(690):1671–1685

Desroziers G, Arbogast E, Berre L (2016) Improving spatial localization in 4DEnVar. Q J R Meteorol Soc 142(701):3171–3185

Desroziers G, Berre L, Chapnik B, Poli P (2005a) Diagnosis of observation, background and analysis-error statistics in observation space. Q J R Meteorol Soc 131(613):3385–3396

Desroziers G, Brousseau P, Chapnik B (2005b) Use of randomization to diagnose the impact of observations on analyses and forecasts. Q J R Meteorol Soc 131(611):2821–2837

Desroziers G, Camino J-T, Berre L (2014) 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. Q J R Meteorol Soc 140(684):2097–2110

Donlon CJ, Martin M, Stark J, Roberts-Jones J, Fiedler E, Wimmer W (2012) The operational sea surface temperature and sea ice analysis (OSTIA) system. Remote Sens Environ 116:140–158. Advanced Along Track Scanning Radiometer (AATSR) Special Issue

Duruisseau F, Chambon P, Wattrelot E, Barreyat M, Mahfouf J-F (2019) Assimilating cloudy and rainy microwave observations from SAPHIR on board Megha Tropiques within the ARPEGE global model. Q J R Meteorol Soc 145(719):620–641

Fisher M (2003) Background error covariance modelling. In: Seminar on recent developments in data assimilation for atmosphere and ocean, 8–12 Sept 2003. Shinfield Park, Reading, ECMWF, pp 45–64

Fouquart Y, Bonnel B (1980) Computations of solar heating of the Earth's atmosphere: a new parameterization. Beitr Phys Atmosph 53:35–62

Gauthier P, Thépaut J-N (2001) Impact of the digital filter as a weak constraint in the preoperational 4D-Var assimilation system of Météo-France. Mon Weather Rev 129(8):2089–2102

Giard D, Bazile E (2000) Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. Mon Weather Rev 128(4):997–1015

Girard D (1989) A fast Monte Carlo cross-validation procedure for large least-squares problems with noisy data. Numer Math 56:1–23

Harris BA, Kelly G (2001) A satellite radiance-bias correction scheme for data assimilation. Q J R Meteorol Soc 127(574):1453–1468

Hortal M, Simmons AJ (1991) Use of reduced gaussian grids in spectral models. Mon Weather Rev 119(4):1057–1074

Janisková M, Veersé F, Thépaut J-N, Desroziers G, Pouponneau B (1999) Impact of a simplified physical package in 4D-Var analyses of FASTEX situations. Q J R Meteorol Soc 125:2465–2485

Laprise R (1992) The Euler equations of motion with hydrostatic pressure as an independent variable. Mon Weather Rev 120(1):197–207

Le Dimet F-X, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A 38A(2):97–110

Lindzen RS, Fox-Rabinovitz M (1989) Consistent vertical and horizontal resolution. Mon Weather Rev 117(11):2575–2583

Lopez P (2002) Implementation and validation of a new prognostic large-scale cloud and precipitation scheme for climate and data-assimilation purposes. Q J R Meteorol Soc 128(579):229–257

Mahfouf J-F, Bergaoui K, Draper C, Bouyssel F, Taillefer F, Taseva L (2009) A comparison of two off-line soil analysis schemes for assimilation of screen level observations. J Geophys Res Atmos 114(D8)

Marquet P, Descamps L, Bouyssel F (2019) A new grid point storm control scheme in the ARPEGE NWP model, vol 49, pp 4–11. WMO CAS/JSC, WGNE Blue Book. http://bluebook.meteoinfo.ru/uploads/2017/docs/04_MARQUET_Pascal_GridPointStormControl.pdf

Marquet P, Mahfouf J-F, Holdaway D (2020) Definition of the moist-air exergy norm: a comparison with existing moist energy norms. Mon Weather Rev 148(3):907–928

Masson V, Le Moigne P, Martin E, Faroux S, Alias A, Alkama R, Belamari S, Barbu A, Boone A, Bouyssel F, Brousseau P, Brun E, Calvet JC, Carrer D, Decharme B, Delire C, Donier S, Essaouini K, Gibelin AL, Giordani H, Habets F, Jidane M, Kerdraon G, Kourzeneva E, Lafaysse M, Lafont S, Lebeaupin Brossier C, Lemonsu A, Mahfouf JF, Marguinaud P, Mokhtari M, Morin S, Pigeon G, Salgado R, Seity Y, Taillefer F, Tanguy G, Tulet P, Vincendon B, Vionnet V, Voldoire A (2013) The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. Geosci Model Develop 6(4):929–960

McNally AP, Watts PD (2003) A cloud detection algorithm for high-spectral-resolution infrared sounders. Q J R Meteorol Soc 129(595):3411–3423

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA (1997) Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. J Geophys Res Atmos 102(D14):16663–16682

Montmerle T, Michel Y, Arbogast E, Ménétrier B, Brousseau P (2018) A 3D ensemble variational data assimilation scheme for the limited-area AROME model: formulation and preliminary results. Q J R Meteorol Soc 144(716):2196–2215

Pangaud T, Fourrié N, Guidard V, Dahoui M, Rabier F (2009) Assimilation of AIRS radiances affected by mid- to low-level clouds. Mon Weather Rev 137(12):4276–4292

Raynaud L, Berre L, Desroziers G (2012) Accounting for model error in the Météo-France ensemble data assimilation system. Q J R Meteorol Soc 138(662):249–262

Ritchie H, Temperton C, Simmons A, Hortal M, Davies T, Dent D, Hamrud M (1995) Implementation of the semi-lagrangian method in a high-resolution version of the ECMWF forecast model. Mon Weather Rev 123(2):489–514

Saunders R, Hocking J, Turner E, Rayer P, Rundle D, Brunel P, Vidot J, Roquet P, Matricardi M, Geer A, Bormann N, Lupu C (2018) An update on the RTTOV fast radiative transfer model (currently at version 12). Geosci Model Develop 11(7):2717–2737

Seity Y, Brousseau P, Malardel S, Hello G, Bénard P, Bouttier F, Lac C, Masson V (2011) The AROME-France convective-scale operational model. Mon Weather Rev 139(3):976–991

Simmons AJ, Burridge DM (1981) An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. Mon Weather Rev 109(4):758–766

Termonia P, Fischer C, Bazile E, Bouyssel F, Brožková R, Bénard P, Bochenek B, Degrauwe D, Derková M, El Khatib R, Hamdi R, Mašek J, Pottier P, Pristov N, Seity Y, Smolíková P, Paniel O, Tudor M, Wang Y, Wittmann C, Joly A (2018) The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1. Geosci Model Develop 11(1):257–281

An Overview of KMA's Operational NWP Data Assimilation Systems

Hyun-Cheol Shin, Ji-Hyun Ha, Kwang Deuk Ahn, Eun Hee Lee, Chang Hwan Kim, Yong Hee Lee, and Adam Clayton

Abstract Currently, the Korea Meteorological Administration (KMA) is operating two global models. One is the Unified Model (UM) developed by the United Kingdom Met Office, and the other is the Korean Integrated Model (KIM), which was developed by the Korea Institute of Atmospheric Prediction Systems (KIAPS) during a 9-year project funded by KMA. The UM global model has been operation since 2010, and the KIM global model since April 2020. The plan is that KIM will be the sole operational global model of KMA after retirement of the UM at the end of 2022. In this chapter, we introduce the data assimilation systems used with the two global models, and with the limited area models that are nested inside the global UM. The UM global model uses a hybrid-4DVar data assimilation system, while the KIM global model uses a hybrid-4DEnVar system. UM-based 1.5 km local and very short range forecast models cycled with 3DVar are also being operated. KMA also runs another very short range forecasting system with a 5 km grid spacing named the Korea Local Analysis and Prediction System (KLAPS) which is also cycled with 3DVar.

1 Introduction

The UM-based numerical weather prediction (NWP) models and data assimilation systems are introduced in Sect. 2. The main characteristics of the UM global, local, and ensemble models are explained, along with the main steps in the global data assimilation system. We also show the observational data used in the UM data assimilation system, and introduce the hybrid-4DVar scheme that was made operational in 2013, and the variational bias correction scheme that was made operational in 2017.

H.-C. Shin (✉) · J.-H. Ha · K. D. Ahn · E. H. Lee
Korea Meteorological Administration, Seoul 07062, Korea
e-mail: sinhyo@korea.kr

C. H. Kim · Y. H. Lee · A. Clayton
Korea Institute of Atmospheric Prediction Systems, Seoul, South Korea

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7_26

In Sect. 3, we introduce the new global model, the Korean Integrated Model (KIM), and its hybrid-4DEnVar data assimilation system, including details of the data assimilation algorithm, the observation types assimilated, and the observation processing system.

The main characteristics of the very short range forecast model, the Korea Local Analysis and Prediction System (KLAPS), are summarized in Sect. 4.

2 UM Data Assimilation

2.1 History

The accuracy of the NWP system has significantly improved since the Korea Meteorological Administration (KMA) introduced the Met Office's Unified Model (UM) in 2010. After initial implementation, KMA has regularly updated the model infrastructure, data assimilation, and physics schemes to keep up with the operational NWP systems running at the Met Office, but normally running a one or two year older version. KMA's current (January 2021) UM-based operational systems are illustrated in Fig. 1, and the history of the main operational changes (model and data assimilation) is summarized in Table 1.

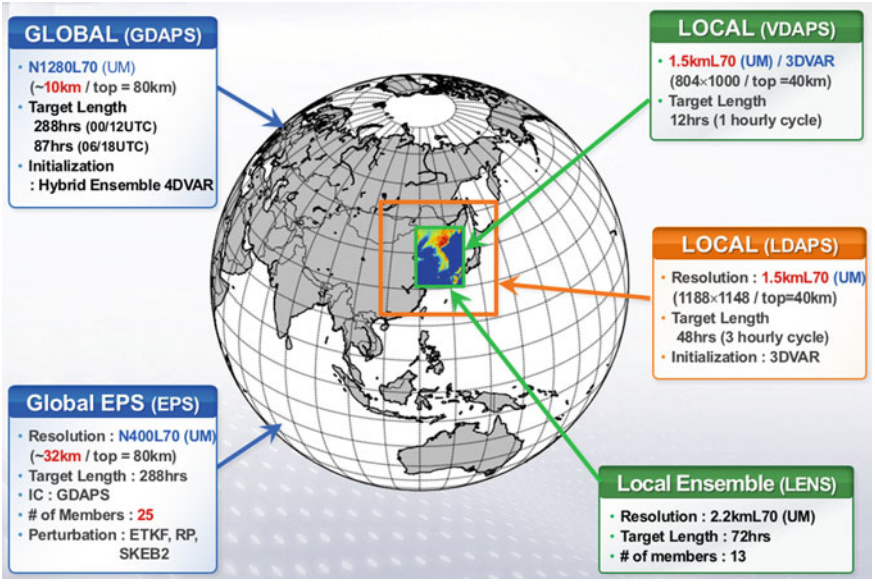

Fig. 1 KMA NWP Operational Models (UM based)

Table 1 Main operational changes of the UM-based KMA NWP systems (2010 ~ 2019)

Year	Operational changes
2010	Global model "GDAPS" introduced (40 km—L50, Top = 65 km) Regional model "RDAPS" introduced (12 km—L38, Top = 45 km/cold run)
2011	Resolution change of global model (→ 25 km—L70, Top = 80 km) Global ensemble "EPSG" introduced (40 km—L70, Top = 80 km, 24 members) Introduction of 4DVar into the regional model Vertical resolution change of regional model (→ L70, Top = 80 km)
2012	Local model "LDAPS" introduced (1.5 km—L70, Top = 40 km)
2013	Hybrid-4DVar for the global model
2015	Local ensemble model "LENS" introduced (3 km—L70, Top = 40 km, 13 members)
2016	Horizontal resolution change of global model (25 km → 17 km) Upgrade of global ensemble (40 km → 32 km, 25 → 49 members)
2017	Introduction of VarBC (Variational Bias Correction) into the global model Expansion of local model domain (Korean Peninsula → East Asia) Very short range forecast model "VDAPS" introduced (1.5 km—L70, Top = 40 km)
2018	Horizontal resolution change of global model (17 km → 10 km) Horizontal resolution change of local ensemble model (3 km → 2.2 km)
2019	Retirement of 12 km regional model

GDAPS: Global Data Assimilation and Prediction System; RDAPS: Regional Data Assimilation and Prediction System; EPSG: Ensemble Prediction System - Global; LDAPS: Local Data Assimilation and Prediction System; LENS: Local ENSemble; VDAPS: Very short range Data Assimilation and Prediction System

2.2 Data Assimilation for the Global Model (GDAPS-UM)

The current version of the KMA UM-based global NWP system was made operational at KMA in June 2018. The global data assimilation system consists of four parts: (1) preparation of the observations database, (2) observation processing (OPS), (3) surface analysis (SURF), and (4) the main variational analysis (VAR).

(1) Observations database. Various observations in ASCII, binary, or BUFR (Binary Universal Form Representation) format are received through the GTS or a dedicated ftp network, and ingested into a database in the ODB (Observation Data Base) format developed by ECMWF.

(2) OPS: The OPS (Observation Processing System) software developed by the Met Office is used to quality control and thin observations read from the ODB database, and prepare them for use in the data assimilation system. In total 15 types of observations are currently managed by the OPS (Table 5). The quality control step removes any observations that have gross errors, fail "buddy" checks, or deviate too far from the background forecast. Many satellite data are then removed through data thinning (Table 2). The biases of satellite radiance data are corrected using parameters updated on the previous data assimilation

Table 2 Thinning ranges (km) of the satellite data used in UM global model

Satellite data	AIRS ATMS CrIS	AMICSR AHICSR SEVIRICSR GOESCSR	AMSR2 FY-3C	ATOVS	IASI	Ground-GNSS	Satwind
30°S ~ 30°N	154	120	80	154	154	100	200
90°S ~ 30°S 30°N ~ 90°N	125	120	80	100	80	100	200

cycle during the late-cutoff analysis, which makes use of the VarBC (Variational Bias Correction) scheme described in Sect. 2.6.

(3) SURF: Surface analysis process. This process produces analyses of soil moisture, SST (Sea Surface Temperature), snow and sea ice. ASCAT superficial soil wetness index data is assimilated using a simplified Extended Kalman Filter (EKF) (Gómez et al. 2020, De Rosnay et al. 2013). Model snow depth is updated based on IMS[1] data, and model SST and sea ice are corrected based on OSTIA[2] (Donlon et al. 2012; Martin et al. 2019) data.

(4) VAR: Variational analysis process. The atmospheric analysis is produced here using a hybrid-4DVar scheme (Rawlins et al. 2007; Clayton et al. 2013). To reduce both computational costs and total run time, the minimization of the cost function starts at lower resolution (88 km) and then proceeds to a higher resolution (40 km). A linear "Perturbation Forecast" (PF) model and its adjoint are included to propagate information through the data assimilation window. Ensemble information is incorporated into the background error covariance matrix as part of a hybrid-4DVar scheme, as described in Sect. 2.6. The control variables are stream function, velocity potential, unbalanced pressure and the humidity variable described in Ingleby et al. (2013). The conjugate-gradient method is used for the minimization of an exactly quadratic cost function, but with relinearization of the observation operators every ten iterations to allow for some weak nonlinearity. The observation window is -3 to $+3$ h, with the FGAT[3] fields for the observation operators obtained by linearly interpolating in time between fields valid at -3, 0 and $+3$ h relative to the nominal analysis time. Increments to the background fields are taken from the nearest PF model timestep.

The global data assimilation cycle is run 4 times (00,06,12,18 UTC) a day. In order to provide model results to forecasters in time, the global model run is started before all observational data within the data assimilation window is received. This run is called the 'early run'. The early run starts 2 h and 40 min after the nominal analysis time in the global model. And the global model is rerun later after a more complete set of

[1] IMS: Interactive Multisensor Snow and Ice Mapping System.

[2] OSTIA: Operational Sea Surface Temperature and Sea Ice Analysis

[3] FGAT: First Guess at Appropriate Time.

observational data has been received to provide a better set of background fields for the next run. This is called the 'late run'. The late run starts 6 h and 40 min after the nominal analysis time in the global model. The early run uses approximately 80% of the observational data used by the late run. A general overview of the UM-based global model and data assimilation systems is provided in Table 3. Synop, Sonde, Aircraft and 13 satellite observation types are assimilated in the global model, as shown in Table 5, and illustrated in Fig. 2 for a particular 6 h period.

2.3 Data Assimilation in Local Models

KMA is also operating a local model (LDAPS) and a very short range model (VDAPS) to forecast convective-scale severe weather events. These two models are cycled with 3DVar. LDAPS has been operational since 2012 and VDAPS has been operational since 2017. As shown in Tables 6 and 7, Synop, Sonde, Aircraft and 1~5 satellite observation types are assimilated in LDAPS and VDAPS. Data from domestic weather radars is also assimilated. The radar radial velocities are assimilated in 3DVar, while the radar surface rainfall rates are assimilated separately using a Latent Heat Nudging (LHN) method (Jones and Macpherson 1997) in which latent heating rate profiles are nudged towards target profiles by adding increments to potential temperature during the model integration (while preserving relative humidity). VDAPS also assimilates visibility observations from automatic weather stations, using the method of Clark et al. (2008). Note that the high nonlinearity of the visibility operator requires a different minimization method—a quasi-Newton method rather than the conjugate-gradient based method that is used with the global and LDAPS variational data assimilation systems. General overviews of the LDAPS and VDAPS data assimilation systems are displayed in Table 3.

2.4 Ensemble Prediction Systems (EPS)

KMA is operating global and local ensemble prediction systems. The global EPS "EPSG" is based on the "MOGREPS-G" system (Bowler et al. 2008) that was operational at the Met Office at the time of the last major NWP upgrade at KMA (June 2018). The ensemble perturbations are updated every 6 h using a localized Ensemble Transform Kalman Filter (ETKF) algorithm (Bowler et al. 2009). These perturbations are added to the early-cutoff global deterministic analysis together with random perturbations to the Sea Surface Temperatures (SSTs), soil moisture and soil temperature to create the initial states for the ensemble forecasts. To account for model uncertainties, the forecasts include stochastic perturbations to key parameters in the physical parametrization schemes, and a Stochastic Kinetic Energy Backscatter (SKEB) scheme to account for the effect of unresolved scales in the dynamics (Tennant et al. 2011). To account for other system deficiencies—particularly in the ETKF-based

Table 3 General overview of UM-based KMA deterministic models and data assimilation systems

	Global model (GDAPS)	Local area model (LDAPS)	Very short range forecast model (VDAPS)
Model resolution	10 km—L70	1.5 km—L70	1.5 km—L70
Model top	80 km	40 km	40 km
Forecast lengths	288 h	48 h	12 h
Model domain	Global domain	East Asia	Korean Peninsula
DA method	Hybrid-4DVar	3DVar	3DVar
DA resolution	88 km → 40 km	2.2 km	2.2 km
DA cycle	6 hourly	3 hourly	1 hourly
Observation window	−3 to +3 h	−1 h 30 min to +1 h 29 min	−49 min to +10 min
Frequency of FGAT fields	Every 3 h	Every 10 min	Every 10 min
Cut-off time (early run)	+2 h and 40 min	+3 h	10 min

(continued)

Table 3 (continued)

	Global model (GDAPS)	Local area model (LDAPS)	Very short range forecast model (VDAPS)
Cut-off time (late run)	+6 h and 40 min	–	–
Lateral boundaries	–	UM global model	UM global model
Minimization method	Conjugate-gradient		Quasi-Newton
Control variables	Stream function, velocity potential, unbalanced pressure, humidity		Same, plus $\log_{10}$(aerosol)

DA: Data Assimilation

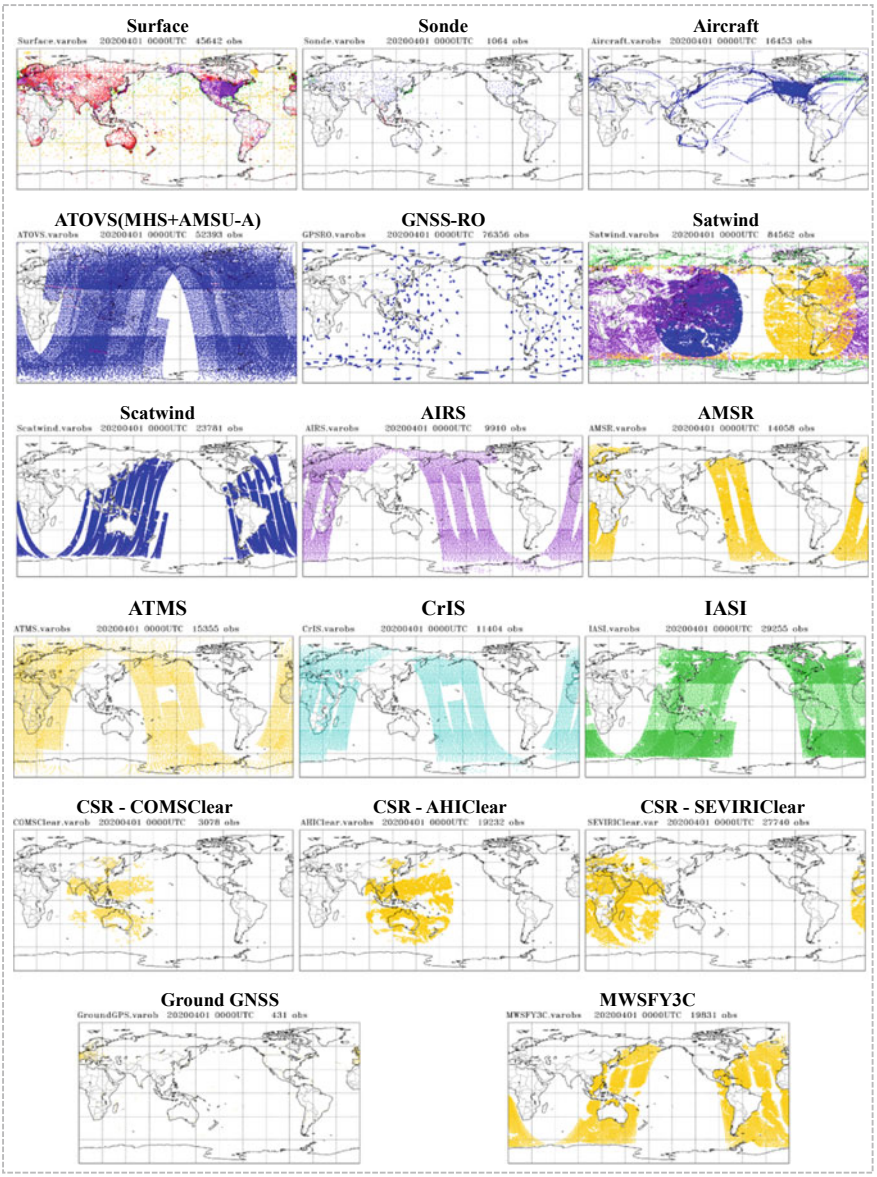

Fig. 2 Coverage of observations used in the 00Z late-cutoff global analysis on 1st April 2020. (The length of the data assimilation window is 6 h.)

Table 4 General overview of KMA global and local ensemble prediction systems

	Global ensemble (EPSG)	Local ensemble (LENS)
Model resolution	32 km—L70	2.2 km—L70
Model top	80 km	40 km
Members	25 for forecast /48 for data assimilation	13 (for forecast only)
Forecast cycles	00Z, 06Z, 12Z, 18Z	00Z, 12Z
Forecast lengths	288 h	72 h
Model domain	Global domain	Korea Peninsula

perturbation update—an adaptive multiplicative inflation scheme is used to approximately match the ensemble spread to the RMSE of the ensemble mean forecast (Flowerdew and Bowler 2012). Tag-lag method was applied to EPSG in 2019 to increase the ensemble spread and use the computational resources more efficiently. 25 ensemble members of 00 (12) UTC are generated from 13 members (including a control member) of 00 (12) UTC and 12 members of 18 (06) UTC.

The local EPS "LENS" simply downscales a subset of the global ensemble members but, as in the global ensemble, the forecasts include stochastic physics and SKEB to help account for model uncertainties. Further basic details of the global and local EPS systems are given in Table 4.

2.5 Observational Data Used in the Assimilation

The amount of observational data used for the assimilation is an important factor for the quality of the model initial fields. Tables 5, 6 and 7 display the various observational data which is used in the global, local area and very short range forecast models, respectively. The distribution of observations assimilated into the global model during a typical late-cutoff cycle is shown in Fig. 2

2.6 Hybrid-4DVAR and VarBC

The global hybrid-4DVar system, which combines 4DVar with the global ensemble forecast system, was introduced at KMA in 2013, following the original implementation at the Met Office in 2011 (Clayton et al. 2013). This first implementation led to significant forecast improvements in all major regions, as illustrated in Fig. 3.

In hybrid-4DVar, the background error covariance matrix $\mathbf{B}$ is a weighted sum of a static "climatological" background error covariance matrix $\mathbf{B_c}$ and an ensemble-based background error covariance matrix $\mathbf{B_e}$ as follows.

Table 5 The observations used in the global model

Observation type	Instrument/Platform
Synop	Surface, Ship, METAR, Buoy
Sonde	TEMP, PILOT, Dropsonde, Wind profiler
Aircraft	AIREPS, AMDARS
GNSS-RO	Metop-A/B, Tandem-X, TerraSAR-X
Satwind	GK-2A, GOES-16, MSG-1/4, Himawari-8, Metop-A/B, Terra, Suomi NPP
Ground GNSS	Ground based receivers
Scatwind	ASCAT/Metop-A/B
Satellite radiances (Microwave)	AMSR2 /GCOM-W1
	MWHS2/FY-3C
	ATMS/Suomi NPP
	AMSU-A/Metop-A/B, NOAA-15/18/19
	MHS/Metop-A/B, NOAA-18/19
Satellite radiances (IR)	CrIS/Suomi NPP
	AIRS/Aqua
	IASI/Metop-A/B
Satellite radiances (CSR)	Himawari-8, GK-2A, MSG-1

GK-2A: Korea's second geostationary satellite launched in 2018; NPP: National Polar-orbiting Partnership; GNSS: Global Navigation Satellite System; AMSR: Advanced Microwave Scanning Radiometer; GCOM-W1: the Global Change Observation Mission-Water 1; MWHS2: MicroWave Humidity Sounder; ATMS: Advanced Technology Microwave Sounder; AMSU: Advanced Microwave Sounding Unit; CrIS: Cross-track Infrared Sounder; AIRS: Atmospheric InfraRed Sounder; IASI: Infrared Atmospheric Sounding Interferometer; Clear Sky Radiance; MSG: Meteosat Second Generation Satellite

Table 6 The observations used in LDAPS

Observation type	Instrument/Platform
Synop	Surface, Ship, METAR, Buoy
Sonde	TEMP, PILOT, Dropsonde, Wind profiler
Aircraft	AIREPS, AMDARS
Satellite radiances (Microwave)	MHS/Metop-A/B
Satwind	Himawari-8, COMS
Scatwind	ASCAT/Metop-A/B
Satellite radiances (CSR)	COMS
Ground GNSS	Ground based receivers
Radar	5 South Korean weather radars (S band; dual polarization)

COMS: Communication, Ocean and Meteorological Satellite. Korea's geostationary satellite launched in 2010

Table 7 The observations used in VDAPS

Observation type	Instrument/Platform
Synop	Surface, Ship, METAR, Buoy
Sonde	TEMP, PILOT, Dropsonde, Wind profiler
Aircraft	AIREPS, AMDARS
Ground GNSS	Ground based receivers
Radar	8 South Korean weather radars (S band; dual polarization)
Visibility	200 ~ 300 Automatic Weather Stations in South Korea

Table 8 β_c^2 and β_e^2 values used in the KMA global hybrid-4DVar system

Vertical level	Below 16 km	Between 16 and 21 km	Above 21 km
β_c^2	0.7	Changed gradually from 0.7 to 0.63	0.63
β_e^2	0.3	Changed gradually from 0.3 to 0.27	0.27

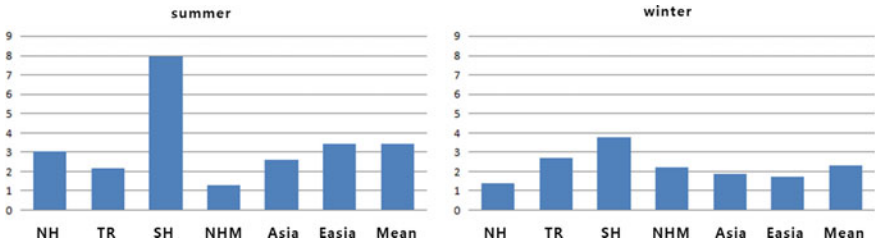

Fig. 3 Percentage RMSE improvements of 5 day 500 hPa geopotential height forecasts against ECMWF analyses due to the change from standard 4DVar to the original 2013 hybrid-4DVar system. The percentage RMSE improvement is defined as 100 * (RMSE_cntl—RMSE_test)/RMSE_cntl, where RMSE_cntl and RMSE_test are the average RMS errors from the control and test trials respectively. The abbreviations used in the plots are as follows. NH: Northern Hemisphere (20°N~ 90°N), TR: Tropics (20°S ~ 20°N), SH: Southern Hemisphere (20°S ~ 90°S), NHM: Northern Hemisphere Mid–latitudes (20°N~ 60°N), Easia: East Asia (20 °N~ 55°N, 100°E ~ 150°E), Asia (25°N ~ 65°N, 60°E ~ 145°E). Mean: the global mean. Summer: June ~ July 2012, Winter: December 2012

$$\mathbf{B} = \beta_c^2 \mathbf{B_c} + \beta_e^2 \mathbf{B_e} \tag{1}$$

where β_c^2 and β_e^2 are the weights given to $\mathbf{B_c}$ and $\mathbf{B_e}$, respectively. The weights are allowed to vary with respect to latitude and vertical level, but in the current system made operational in June 2018 vary only in the vertical, as shown in Table 8.

The details of the current climatological ($\mathbf{B}_c$) and ensemble ($\mathbf{B}_e$) background covariances are as follows.

$\mathbf{B}_c$: This matrix is modelled via a series of spatial and parameter transforms to implement a covariance matrix in which the four control variable fields (stream function, velocity potential, unbalanced pressure and humidity) are uncorrelated. For each of these fields, independent vertical covariance matrices are used for each total horizontal wavenumber, giving spatial covariances that are horizontally homogeneous and isotropic. Geographic variations are introduced via the operator that is used to define the balanced pressure. This is based on an analytic linear balance equation, but includes a subsequent vertical regression procedure to further reduce average correlations between full and unbalanced pressure (Sect. 6.2 of Inverarity et al. 2018). The statistical parameters were obtained from ensemble forecasts produced during pre-operational trials of the 4DEnVar-based global EPS (Bowler et al. 2017) that was eventually made operational at the Met Office in December 2019. (Some postprocessing was also required to resolve issues with the global-mean unbalanced pressure variances, as discussed in Sect. 6.4 of Inverarity et al. 2018).

$\mathbf{B}_e$: This matrix is constructed from fields at T + 3 h from each of the 48 perturbed global ensemble members. To help preserve geostrophic and hydrostatic balances, localization is performed after transforming the ensemble perturbations to the space of the four standard control variables, and splitting the fields into a set of four horizontal wavebands in order to apply the scale-dependent localization technique of Buehner (2012). In this method, the wavebands are treated as independent from each other, and localized with different scales. However, for each waveband we use common localization fields for the four control variable fields, so that inter-variable correlations are kept within each waveband. Vertical localization is also applied, using a common vertical localization matrix for all wavebands and variables. This localization matrix was derived from the same database of ensemble perturbations used to calibrate the climatological covariance matrix, using the objective method of Ménétrier et al. (2015), as explained in Sect. 4 of Inverarity et al. (2018). The horizontal and vertical localization details are illustrated in Fig. 4.

In 2017, VarBC (Variational Bias Correction) for satellite data was introduced to use satellite data more effectively (Cameron and Bell 2018). In VarBC, the bias correction parameters for satellite data (radiances) are updated every cycle during the late-cutoff global hybrid-4DVar analysis. Figure 5 shows how VarBC can reduce the observation increment and bias of satellite data (IASI). VarBC contributed to improving the global model performance (Fig. 6).

3 KIM Data Assimilation

3.1 History

KMA launched a nine-year project (2011–2019) to develop a new global model—the so called Korean Integrated Model (KIM). The purpose of the project was to improve

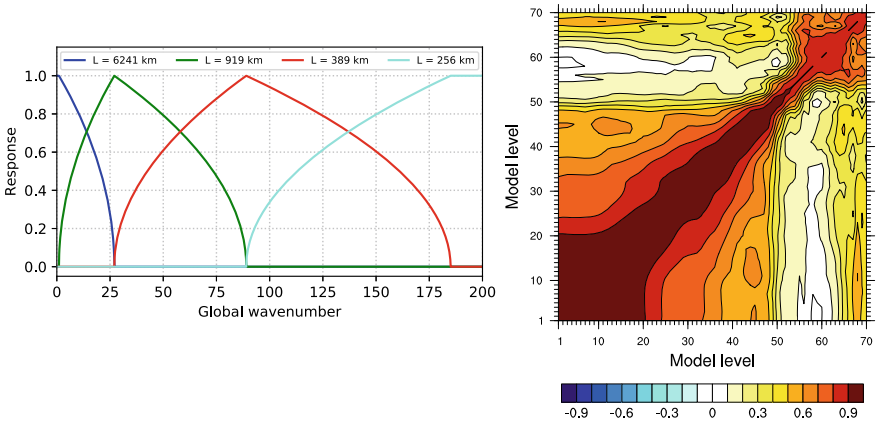

Fig. 4 Horizontal and vertical ensemble localization details for KMA's UM-based hybrid-4DVar system. The left-hand plot shows the horizontal wavebands and corresponding length scales L used for horizontal localization. The horizontal localization functions are Gaussian functions of the form $\exp(-z^2/(2L^2))$, where z is the horizontal separation. The right-hand plot shows the matrix used for vertical localization

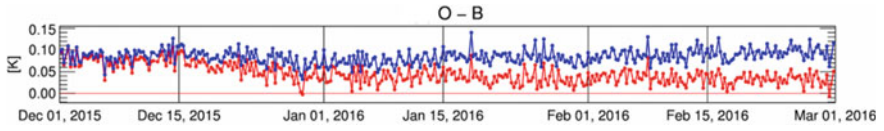

Fig. 5 Background departure time series of satellite data (13.51μm channel from Metop-B IASI) over a 3-month period (01 December 2015 to 01 March 2016). Red curve shows the results with VarBC and blue curve shows the results from the traditional bias correction with fixed bias correction parameters. Blue curve is derived from the operational cycle which did not use VarBC. The initial background and bias correction parameters of VarBC experiment (red curve) are from the operational cycle

the prediction accuracy of worldwide weather phenomena with a particular focus on the Korean peninsula.

The KIM (Choi et al. 2014; Choi and Hong 2016; Choi 2018; Hong et al. 2018; Park and Choi 2020) has been an operational global NWP model for medium-range forecasts up to 12 days (288 h) since April 2020.

3.2 Data Assimilation for the Global Model (GDAPS-KIM)

KIM adopts the hybrid four-dimensional ensemble-variational data assimilation scheme (hybrid-4DEnVar) for its 6 h analyses (Song et al. 2017a; Song et al. 2017b; Song et al. 2017c; Song et al. 2018; Kwon et al. 2018). Technically, the main difference between hybrid-4DEnVar and hybrid-4DVar (as used with the UM-based global

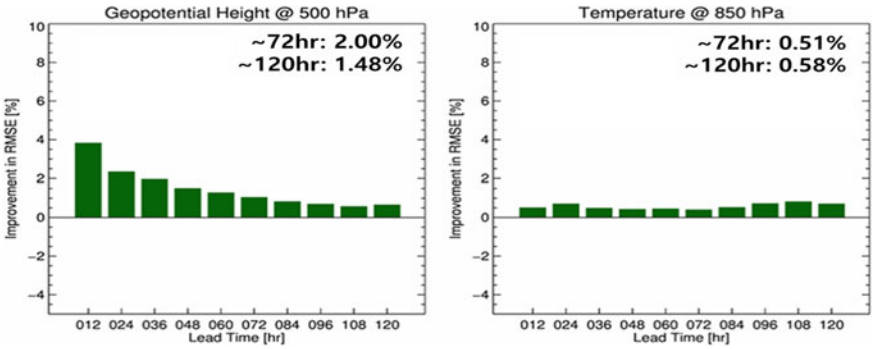

Fig. 6 Percentage RMSE improvement (as defined in the caption of Fig. 3) of the global model with VarBC. Positive values indicate that the global model with VarBC gives improved forecast results relative to the use of traditional bias corrections. The RMSE are calculated based on 5 day forecasts of 500 hPa geopotential height (Left) and 850 hPa temperature (Right) over Northern Hemisphere. The model forecasts were compared against radiosonde observations for August 2016

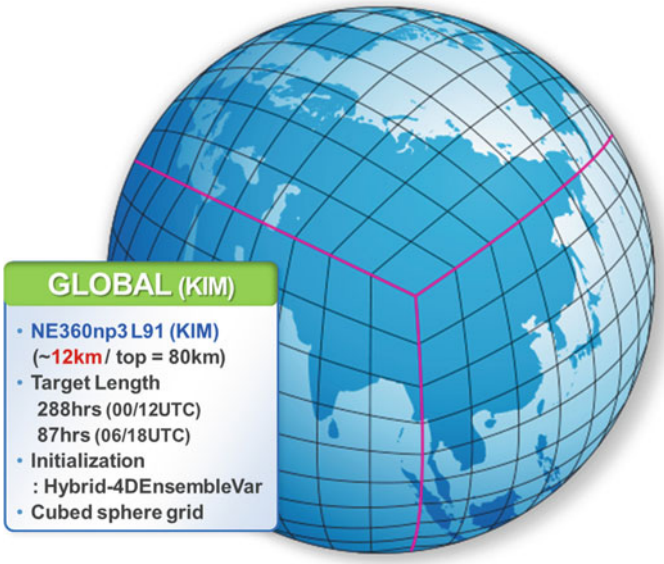

Fig. 7 Main features of the operational KIM based on a cubed-sphere grid system

model) is its use of a four-dimensional ensemble-based background error covariance to directly provide ensemble-based time correlations within the assimilation window, rather than using a linear model and its adjoint to implicitly propagate the full covariance through the window as is done in hybrid-4DVar. In general, hybrid-4DEnVar systems give poorer analyses than produced by a corresponding

hybrid-4DVar system, partly because the climatological covariance remains essentially time-independent, as in a 3DVar system (Lorenc et al. 2015). However, by avoiding the linear and adjoint model integrations used in 4DVar, which have to be run in serial, computational costs are significantly reduced, potentially freeing up resources to run the analysis with higher-resolution ensemble data. Development and maintenance of linear and adjoint models is also costly in terms of human resources, so hybrid-4DEnVar was a natural choice for the first version of the KIM global DA system.

The KIM global DA system consists of the KIM Package for Observation Processing (KPOP; Kang et al. 2018) and a hybrid-4DEnVar system developed for the cubed-sphere grid, referred to here as H4DEV.

(1) KPOP: KPOP prepares quality controlled observations for H4DEV in the desired format. KPOP obtains the observation data with BUFR format and performs quality control, cloud screening, bias corrections, and thinning. And then, it converts quality controlled observation to netcdf format. For bias correction of radiance data, the adaptive bias correction (BC) method that calculates BC coefficients with background at the analysis time rather than using static BC coefficients is used (Kwon et al. 2018). The difference between adaptive BC and VarBC is explained in Auligné et al. (2007). The system manages observation data within the 6 h assimilation window (±3 h around the data assimilation cycle times of 00, 06, 12 and 18 UTC). The list of observations assimilated in the KIM DA system is given in Table 9 (Lee and Song 2018, Ha et al. 2018).

(2) H4DEV: The H4DEV blends the 'climatological' background error with day-to-day varying flow dependent background errors. The climatological background error covariance (BEC) matrix is generated by the NMC method, which is based on differences between pairs of forecasts with the same validity time, but different initial times. The ensemble BEC matrix is generated using 50 ensemble forecasts produced by an Ensemble Prediction System (EPS) based on the Local Ensemble Transform Kalman Filter (LETKF; Hunt et al. 2007, Shin et al. 2016, Shin et al. 2018). A weighted sum of the static and ensemble BEC matrices gives the total BEC matrix as mentioned in Eq. (1) in Sect. 2.6. The operational weights for the climatological and ensemble BEC in H4DEV are set to 0.3 and 0.7, respectively, but vary according to latitude and height. The weight of the ensemble BEC matrix is gradually decreased above 100 hPa

Table 9 Observation types assimilated in the KIM Global Data Assimilation System

Observation type	Instrument/Platform
Surface	Synop, METAR, Ship, Buoy
Sonde	TEMP, PILOT, Windprofiler
Aircraft	AMDARS, AIREPS
Satellite	AMSUA, MHS, ATMS, CrIS, IASI, MWHS2, AMSR2, CSR, AMV, Scatwind, GNSS-RO

and varies with latitude according to a Gaussian function that is 0.7 at the poles and 0.3 at the equator. Figure 8 shows that the impact of the ensemble BEC matrix (in H3DEV—i.e. hybrid-3DEnVar—rather than H4DEV) on the winds (u and v), temperature, and specific humidity in the tropics and subtropics. The model RMSE is generally reduced after introducing the ensemble BEC matrix.

The control variables of the KIM DA system are stream function, unbalanced velocity potential, unbalanced temperature, specific humidity and unbalanced surface pressure, and the H4DEV performs multi-resolution four loops with 42, 85, 170, and 170 total wave numbers (Song et al. 2018). Between the loops, the nonlinear observation operators are re-linearized, but there is no rerun of the nonlinear model, so these are not full outer loops (Kwon et al. 2018). For the minimization of the cost function, preconditioning by transforming the BEC matrix to an identity matrix using its square root form, and a conjugate gradient method is chosen (Song and

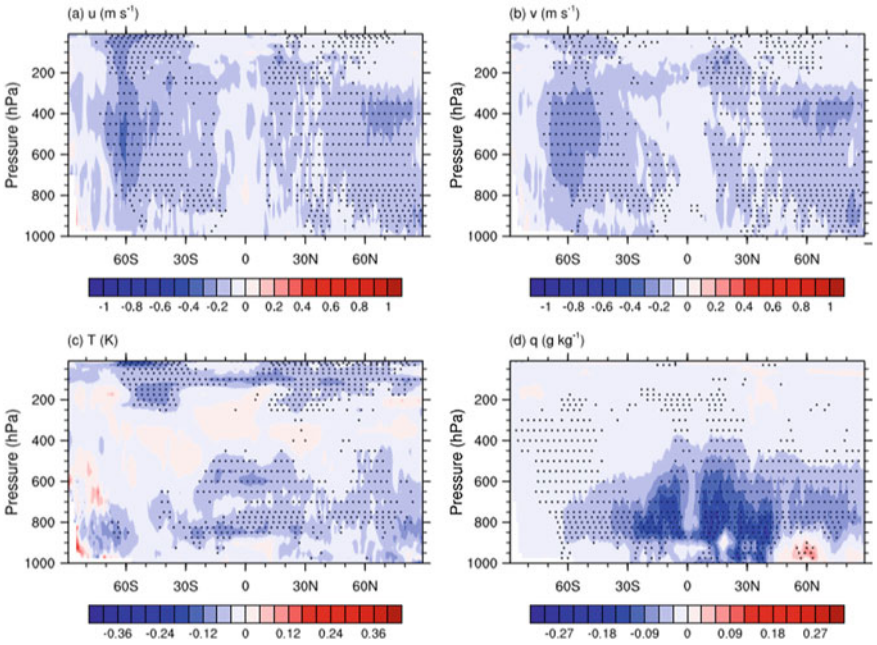

Fig. 8 Composite of analysis RMSE differences of H3DEV relative to 3DVar for (**a**) zonal (u) and (**b**) meridional (v) winds (ms^{-1}), (**c**) temperature (T; K), and (**d**) specific humidity (q;$g\ kg^{-1}$). H3DEV is similar to H4DEV except only one ensemble BEC matrix is given in the observation window. So, H3DEV provides 3 dimensional ensemble BEC matrix, H4DEV provides 4 dimensional ensemble BEC matrix, and 3DVAR does not provide any ensemble BEC matrix. The RMSEs were calculated along longitudes for the same latitude and pressure level against ERA-Interim data. The RMSE differences are inferred from the average of analysis RMSEs from 0000 UTC on 6 July to 1200 UTC on 31 July 2015 in 12 h intervals. The small dots show the variables and regions in which there are RMSE differences with 95% statistical significance, as evaluated with a two-tailed t test. (Song et al. 2017c)

Kwon 2015). After the minimization of the H4DEV, the ensemble mean analysis is re-centered on the H4DEV analysis after the LETKF process. In both the H4DEV and LETKF systems, an analysis increment at the center of the time window (from −3 h to +3 h) is added to (or subtracted from in the case of the LETKF) the background state using an Incremental Analysis Update (IAU) scheme (Bloom et al. 1996) so that high-frequency gravity-inertia waves introduced by the increments are filtered out from the forecasts.

The operational H4DEV system is cycled with a high-resolution deterministic model having approximately 12 km grid spacing. The horizontal resolutions of the H4DEV analysis, the ensemble model, and the LETKF analysis are 50 km, 25 km and 25 km, respectively. There are 91 levels, with the top level at 0.1 hPa. In order to improve the initial fields of the KIM, KMA has a plan to assimilate more satellite observation data. As an illustration along those lines, Fig. 9 shows the impact of CSR observational data from the GK-2A satellite. The performance of KIM is generally improved after assimilating GK-2A CSR data. The impact is particularly clear in the Northern Hemisphere.

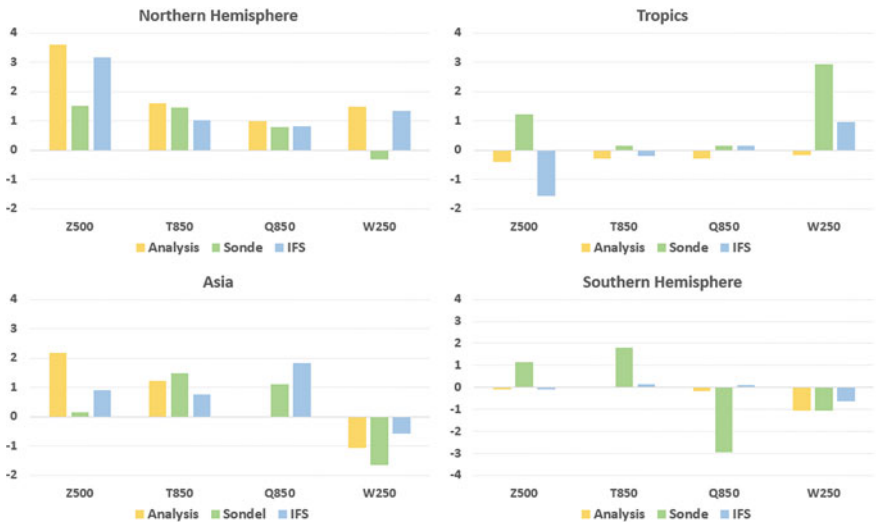

Fig. 9 Percentage RMSE improvement of KIM forecasts after assimilating GK-2A CSR data. Positive value indicates improvement. It is based on the averaged RMSE of 1–5 day forecast. Analysis: RMSE against KIM analysis field, Sonde: RMSE against radiosondes, IFS: RMSE against ECMWF analysis field

4 KLAPS (Very Short Range Forecast Model and Data Assimilation)

KMA is operating another very short range forecast system in addition to VDAPS. It is called the Korea Local Analysis and Prediction System (KLAPS), and runs every 10 min with a 5 km horizontal grid spacing (Albers 1995, Albers et al. 1996, Hiemstra et al. 2006).

Its background field is generated by running the model with the initial and boundary fields from UM global cycle as illustrated in Fig. 10. KLAPS is different from VDAPS in that it is more focused on nowcasting and rapid update cycle. All processes from data assimilation to model run and post processing, are executed on a High Performance Computer in 10 min.

KLAPS has a cloud analysis package and a warm-start diabatic initialization scheme for the quick spin-up of micro physical species through a sequential 3DVAR (Albers et al. 1996) and shares the physics package with KIM. KLAPS produces analyses for 8 variables (temperature, sky condition, precipitation amount, relative humidity, wind direction, wind speed, precipitation types, and lightning) and forecasts for 4 variables (precipitation amount, sky condition, precipitation type, probability of lightning) at the nearest grid points from the observation stations for the digital forecast (detailed forecast for village units), named "Dongrae forecast".

KLAPS assimilates the various types of observation (Table 10). The fast data assimilation system (Sequential-3DVAR (He et al. 2008)) is chosen to speed up the cycle. Only recent observations are assimilated – specifically, those valid at or after the analysis time that have been received by the observation cutoff time, which is 6 min after the analysis time. Thus, the observation window is +00 ~ +06 min.

In order to improve the precipitation forecast, KLAPS results are merged (as a post-processing) with results from the algorithm MAPLE (McGill Algorithm for

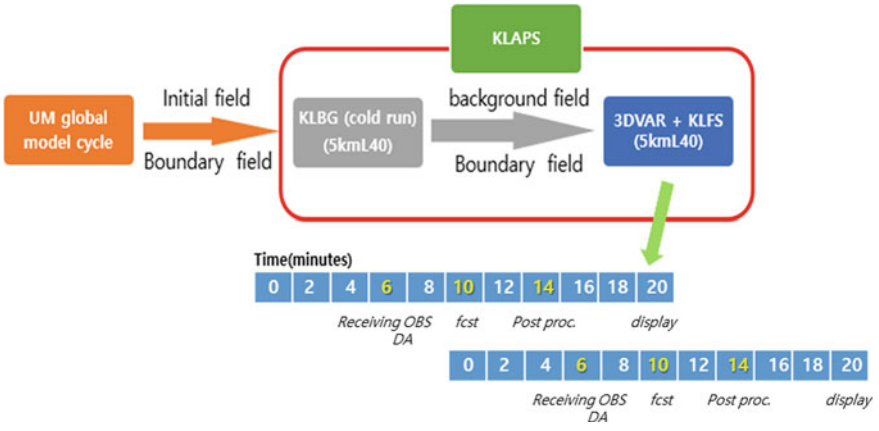

Fig. 10 The structure of KLAPS cycle run which consists of KLBG* and KLFS**. *KLBG: KLAPS run to generate the background/**KLFS: KLAPS run to generate the 12 h forecast

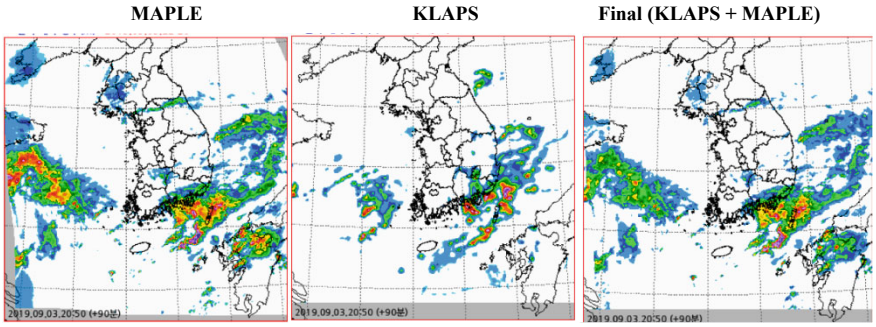

Fig. 11 Comparison between MAPLE, KLAPS, and the final precipitation forecast (90 min forecast issued at 20:50 KST on 3rd September 2019). The final precipitation forecast results (Right) are obtained by merging the KLAPS forecast (Middle) with the MAPLE forecast (Left) which is produced by extrapolating the radar echo image

Table 10 Observation types assimilated in the KLAPS

Observation type	Instrument/Platform
Surface	Buoy, METAR, Automatic Weather Station (AWS)
Upper	TEMP, PILOT, Wind profiler
Aircraft	AMDARS
Satellite	GK-2A, Ground GNSS
Radar	10 South Korean weather radars (S band; dual polarization) (Radial velocity, Reflectivity)
Others	Lightning*, Ceilometer, Visibility meter

*The position of lightning is used for cloud analysis

Precipitation nowcasting by Lagrangian Extrapolation, Bellon et al. 2010, Lee et al. 2010), which produces precipitation forecasts by extrapolating the radar echo image. KLAPS is merged with MAPLE through a weighted average with weights using a hyperbolic tangent function. The weight of KLAPS against MAPLE is zero at analysis time and is increased as forecast hour is changed from 1 to 6 h. Finally at 6 h forecast the weight of MAPLE becomes zero.

5 Conclusion

KMA is running UM-based global and local ensemble systems. The global ensemble provides the ensemble error covariances (from 48 members) for hybrid-4DVar, and produces 12 d forecasts with 25 members. The UM local ensemble which has 2.2 km grid spacing has 13 members and produces 72 h forecast.

Table 11 Operational configuration of KLAPS as of 2019

	KLAPS Run for forecast (KLFS)
Horizontal resolution	5 km
Vertical layers/Model Top	40 sigma layers/Top 50 hPa (18 km)
Number of grid-points	235 × 283
Data assimilation	Sequential-3DVAR + Physical initialization
Domain distance (km)	1,170 × 1,410
Time step for integration	20 s
Cumulus parameterization	Simplified Arakawa–Schubert Scheme (KSAS) (Kwon and Hong 2017)
Boundary layer	Shin–Hong PBL (Shin and Hong 2013)
Microphysics	WDM7 (Bae et al. 2019)
Radiation	RRTMG (Iacono et al. 2008)
Land-surface	Noah LSM (Mitchell 2005)
Forecast length (Interval)	12 h (144 times/day) 10 min interval
Observation window	+00 min ~ +06 min
Cut-off time	6 min
Initial field	Analysis
Background /Boundary field	KLAPS Run given every 10 min (KLBG)

KSAS: A modified GFS Simplified Arakawa-Schubert scheme that enables NSAS to work in various model grids across gray-zone resolutions

KMA started operating the KIM NWP system developed by KMA/KIAPS in April 2020. Thus, KMA is currently operating both the KIM global model and the UM global model. The KIM global model is using hybrid-4DEnVar (H4DEV) as a data assimilation system and the UM global model is cycled with hybrid-4DVar. KIM will soon become the main global model of KMA, but the UM will continue to be run in parallel until at least 2022. The 1.5 km UM-based local model and the very short-range forecast model are cycled with 3DVar systems, producing forecasts for small-scale severe weather events. A KIM regional model will be developed by 2022 and should be operational by 2023. KLAPS is the another very short range forecast system, based on a 10-min interval RUC (Rapid Update Cycle).

The KIM global ensemble system, which provides the ensemble background error covariance data (from 50 members) for the KIM H4DEV and produces 12 d forecast from 14 members, is being tested and will be operational with additional members (from 14 to 26) in 2021. A KIM-based local ensemble will be completed by 2022, and operationally launched in 2023.

KMA also runs a UM-based long range forecast system (named Glosea-5) which has 60 km grid spacing and 85 vertical levels and is coupled to a global ocean

model, producing 8 week sub-seasonal forecasts and 6 month seasonal forecasts (not mentioned here).

References

Albers SC (1995) The LAPS wind analysis. Wea Forecasting 10(2):342–352. https://doi.org/10.1175/1520-0434(1995)010%3c0342:TLWA%3e2.0.CO;2

Albers SC, McGinley JA, Birkenheuer DL, Smart JR (1996) The Local Analysis and Prediction System (LAPS): analyses of clouds, precipitation, and temperature. Wea Forecasting 11(3):273–287. https://doi.org/10.1175/1520-0434(1996)011%3c0273:TLAAPS%3e2.0.CO;2

Auligné T, McNally AP, Dee DP (2007) Adaptive bias correction for satellite data in a numerical weather prediction system. Q J R Meteorol Soc 133(624):631–642. https://doi.org/10.1002/qj.56

Bae SY, Hong S-Y, Tao W-K (2019) Development of a single-moment cloud microphysics scheme with prognostic hail for the Weather Research and Forecasting (WRF) model. Asia-Pac J Atmos Sci 55:233–245. https://doi.org/10.1007/s13143-018-0066-3

Bellon A, Zawadzki I, Kilambi A, Lee HC, Lee YH, Lee G (2010) McGill algorithm for precipitation nowcasting by lagrangian extrapolation (MAPLE) applied to the South Korean radar network. Part I: sensitivity studies of the Variational Echo Tracking (VET) technique. Asia-Pac J Atmos Sci 46:369–381. https://doi.org/10.1007/s13143-010-1008-x

Bloom SC, Tacks LL, Silva Da AM, Ledvina D (1996) Data assimilation using incremental analysis updates. Mon Wea Rev 124(6):1256–1271. https://doi.org/10.1007/s13143-018-0066-3

Bowler NB, Arribas A, Mylne KR, Robertson KB, Beare SE (2008) The MOGREPS short-range ensemble prediction system. Quart J Roy Meteor Soc 134:703–722. https://doi.org/10.1002/qj.234

Bowler NB, Arribas A, Beare SE, Mylne KR, Shutts GJ (2009) The local ETKF and SKEB: upgrades to the MOGREPS short-range ensemble prediction. Quart J Roy Meteor Soc 135:767–776. https://doi.org/10.1002/qj.394

Bowler NB, Clayton AM, Jardak M, Lee E, Lorenc AC, Piccolo C, Pring SR, Wlasak MA, Barker DM, Inverarity GW, Swinbank R (2017) Inflation and localization tests in the development of 4D-ensemble variational assimilations. Quart J Roy Meteor Soc 143: 1280–1302. https://doi.org/10.1002/qj.3004

Buehner M (2012) Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. Mon Wea Rev 143:1622–1643. https://doi.org/10.1175/MWR-D-14-00157

Cameron J, Bell W (2018) The testing and implementation of variational bias correction (VarBC) in the Met Office global NWP system. Wea Sci Tech Rep 631. https://digital.nmla.metoffice.gov.uk/IO_b5adb22c-d5f1-4edb-90ac-a88f471fae77

Choi S-J, Giraldo FX, Kim J, Shin S (2014) Verification of a non-hydrostatic dynamical core using the horizontal spectral element method and vertical finite difference method. Geosci Model Dev 7:2717–2731. https://doi.org/10.5194/gmd-7-2717-2014

Choi S-J, Hong S-Y (2016) A global non-hydrostatic dynamical core using the spectral element method on a cubed-sphere grid. Asia-Pac J Atmos Sci 52:291–307. https://doi.org/10.1007/s13143-016-0005-0

Choi S-J (2018) Structure of eigenvalues in the advection-diffusion equation by the spectral element method on a cubed-sphere grid. Asia-Pac J Atmos Sci 54:293–301. https://doi.org/10.1007/s13143-018-0020-4

Clark PA, Harcourt SA, Macpherson B, Mathison CT, Cusack S, Naylor M (2008) Prediction of visibility and aerosol within the operational Met Office Unified Model. I: model formulation and variational assimilation. Quart J Roy Meteor Soc 134:1801–1816. https://doi.org/10.1002/qj.318

Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. Quart J Roy Meteor Soc 139:1445–1461. https://doi.org/10.1002/qj.2054

De Rosnay P, Drusch M, Vasiljevic D, Balsamo G, Albergel C, Isaksen L (2013) A simplified extended Kalman filter for the global operational soil moisture analysis at ECMWF. Quart J Roy Meteor Soc 139:1199–1213.https://doi.org/10.1002/qj.2023

Donlon CJ, Martin M, Stark J, Jones JR, Fiedler E, Wimmer W (2012) The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. Remote Sens Environ 116:140–158. https://doi.org/10.1016/j.rse.2010.10.017

Flowerdew J, Bowler NE (2012) Online calibration of the vertical distribution of ensemble spread. Quart J Roy Meteor Soc 139:1863–1874. https://doi.org/10.1002/qj.2072

Gómez B, Charlton-Pérez CL, Lewis H, Candy B (2020) The Met Office operational soil moisture analysis system. Remote Sens 12:3691–3711. https://doi.org/10.3390/rs12223691

Ha J-H, Kang J-H, Choi S-J (2018) The impact of vertical resolution in the assimilation of GPS radio occultation data. Wea Forecasting 33(4):1033–1044. https://doi.org/10.1175/WAF-D-17-0061.1

He Z, Xie Y, Li W, Li D, Han G, Liu K, Ma J (2008) Application of the sequential three-dimensional variational method to assimilating SST in a global ocean model. Wea Forecasting 25(6):1018–1033. https://doi.org/10.1175/2007JTECHO540.1

Hiemstra CA, Liston GE, Pielke RA Sr, Birkenheuer DL, Albers SC (2006) Comparing Local Analysis and Prediction System (LAPS) assimilations with independent observations. Wea Forecasting 21(6):1024–1040. https://doi.org/10.1175/WAF961.1

Hong S-Y, Kwon YC, Kim TH, Kim JE, Choi SJ, Kwon IH, Kim J, Lee EH, Park RS, Kim DI (2018) The Korean Integrated Model (KIM) system for global weather forecasting. Asia-Pac J Atmos Sci 54:267–292. https://doi.org/10.1007/s13143-018-0028-9

Hunt B, Kostelich E, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126. https://doi.org/10.1016/j.physd.2006.11.008

Iacono MJ, Delamere JS, Mlawer EJ, Shephard MW, Clough SA, Collins WD (2008) Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models. J Geophys Res 113:D13103. https://doi.org/10.1029/2008JD009944

Ingleby NB, Lorenc AC, Ngan K, Rawlins FR, Jackson DR (2013) Improved variational analyses using a nonlinear humidity control. Quart J Roy Meteor Soc 139:1875–1887. https://doi.org/10.1002/qj.2073

Inverarity GW, Wlasak MA, Jardak M, Lorenc AC (2018) Deterministic data assimilation developments—July 2017. Forecasting Res Tech Rep 625. https://digital.nmla.metoffice.gov.uk/IO_7f284258-0571-43bb-89a0-aad25b992118

Jones CD, Macpherson B (1997) A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. Meteor Appl 4:269–277. https://doi.org/10.1017/S1350482797000522

Kang J, Chun HW, Lee S, Ha JH, Song HJ, Kwon IH, Han HJ, Jeong H, Kwon HN, Kim TH (2018) Development of an observation processing package for data assimilation in KIAPS. Asia-Pac J Atmos Sci 54:303–318. https://doi.org/10.1007/s13143-018-0030-2

Kwon I-H, Song HJ, Ha JH, Chun HW, Kang JH, Lee S, Lim S, Jo Y, Han HJ, Jeong H, Kwon HN, Shin S, Kim TH (2018) Development of an operational hybrid data assimilation system at KIAPS. Asia-Pacific J Atmos Sci 54:319–335. https://doi.org/10.1007/s13143-018-0029-8

Kwon Y-C, Hong S-Y (2017) A mass-flux cumulus parameterization scheme across grayzone resolutions. Mon Wea Rev 145:583–598. https://doi.org/10.1175/MWR-D-16-0034.1

Lee HC, Lee YH, Ha JC, Chang DE, Bellon A, Zawadzki I, Lee G (2010) McGill algorithm for precipitation nowcasting by lagrangian extrapolation (MAPLE) applied to the South Korean radar network. Part II: real-time verification for the summer season. Asia-Pac J Atmos Sci 46:383–391. https://doi.org/10.1007/s13143-010-1009-9

Lee S, Song H-J (2018) Impacts of the LEOGEO atmospheric motion vectors on the East Asian weather forecast. Quart J Roy Meteor Soc 144. https://doi.org/10.1002/qj.3392

Lorenc AC, Bowler NE, Clayton AM, Pring SR, Fairbairn D (2015) Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. Mon Wea Rev 143:212–229. https://doi.org/10.1175/MWR-D-14-00195.1

Martin M, Fiedler E, Roberts-Jones J, Blockley E, McLaren A, Good S (2019) For OSTIA near real time level 4 SST products over the global ocean (product user manual). COPERNICUS

Ménétrier B, Montmerle T, Michel Y, Berre L (2015) Linear filtering of sample covariances for ensemble-based data assimilation. Part I: optimality criteria and application to variance filtering and covariance localization. Mon Wea Rev 143:1622–1643. https://doi.org/10.1175/MWR-D-14-00157.1

Mitchell K (2005) The community Noah Land-surface model (LSM), user' guide, recovered from ftp://ftp.emc.ncep.noaa.gov/mmb/gcp/ldas/noahlsm/ver_2.7.1. National centers for environmental prediction/environmental modelling center

Park J-R, Choi S-J (2020) Verification of the time-split method for higher-order diffusion in the spectral element method model on a cubed-sphere grid. Asia-Pac J Atmos Sci 56:173–184. https://doi.org/10.1007/s13143-019-00137-6

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW, Lorenc AC, Payne TJ (2007) The Met Office global four-dimensional variational data assimilation scheme. Quart J Roy Meteor Soc 133:347–362. https://doi.org/10.1002/qj.32

Shin HH, Hong S-Y (2013) Analysis of resolved and parameterized vertical transports in convective boundary layers at gray-zone resolutions. J Atmos Sci 70(10):3248–3261. https://doi.org/10.1007/s13143-019-00137-6

Shin S, Kang JS, Jo Y (2016) The local ensemble transform Kalman filter (LETKF) with a global NWP model on the cubed sphere. Pure Appl Geophys 173:2555–2570. https://doi.org/10.1007/s00024-016-1269-0

Shin S, Kang JH, Chun HW, Lee S, Sung K, Cho K, Jo Y, Kim JE, Kwon IH, Lim S, Kang JS (2018) Real data assimilation using the Local Ensemble Transform Kalman Filter (LETKF) system for a global non-hydrostatic NWP model on the cubed-sphere. Asia-Pacific J Atmos Sci 54:351–360. https://doi.org/10.1007/s13143-018-0022-2

Song H-J, Kwon I-H (2015) Spectral transformation using a cubed-sphere grid for a three-dimensional variational data assimilation system. Mon Wea Rev 143:2581–2599. https://doi.org/10.1175/MWR-D-14-00089.1

Song H-J, Kwon I-H, Kim J (2017a) Characteristics of a spectral inverse of the Laplacian using spherical harmonic functions on a cubed-sphere grid for background error covariance modeling. Mon Wea Rev 145:307–322. https://doi.org/10.1175/MWR-D-16-0134.1

Song H-J, Kwun J, Kwon I-H, Ha J-H, Kang J-H, Lee S, Chun H-W, Lim S (2017b) The impact of the nonlinear balance equation on a 3D-Var cycle during an Australian-winter month as compared with the regressed wind-mass balance. Quart J Roy Meteor Soc 143:2036–2049. https://doi.org/10.1002/qj.3065

Song H-J, Shin S, Ha J-H, Lim S (2017c) The advantages of hybrid 4DEnVar in the context of the forecast sensitivity to initial conditions. J Geophys Res 122(12):226–244. https://doi.org/10.1002/2017JD027598

Song H-J, Ha J-H, Kwon I-H, Kim J, Kwun J (2018) Multi-resolution Hybrid Data Assimilation core on a cubed-sphere grid (HybDA). Asia-Pacific J Atmos Sci 54:337–350. https://doi.org/10.1007/s13143-018-0018-y

Tennant W, Shutts G, Arribas A, Thompson S (2011) Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. Mon Wea Rev 139:1190–1206. https://doi.org/10.1175/2010MWR3430.1

Index

S. K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, https://doi.org/10.1007/978-3-030-77722-7

D

H

S