Seon Ki Park
Liang Xu *Editors*

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

Springer

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

Seon Ki Park • Liang Xu
Editors

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

Springer

Editors
Seon Ki Park
Atmospheric Science and Engineering
Ewha Womans University
Seoul
Korea

Liang Xu
Naval Research Laboratory
Monterey
California
USA

ISBN 978-3-642-35087-0 ISBN 978-3-642-35088-7 (eBook)
DOI 10.1007/978-3-642-35088-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2008935664

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Yoshikazu Sasaki (*right*) *and his mentor Shigekata Syono working on hydrodynamic theory of vortex motion during Syono's visit to the University of Oklahoma* (*December 1963*). Drawn by John M. Lewis, using pen, brush, and India ink.

To Yoshi K. SASAKI and Roger W. DALEY

Preface

Since the first session for data assimilation (DA) had been organized at the Asia Oceania Geosciences Society (AOGS) Annual Meeting in 2005, we have conducted several successful sessions under the title of "Yoshi K. Sasaki Symposium on Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications." It was to honor Prof. Yoshi K. Sasaki of the University of Oklahoma for his lifelong contributions to DA in geosciences. Yoshi had introduced the variational method to meteorology as early as the 1950s, and since then DA has developed into an utmost important technique in modern numerical prediction in various disciplines of geosciences.

The first volume of this book, under the same title of the Sasaki Symposium, has been published in March 2009 with a collection of notable invited papers along with those selected from previous symposiums up to 2008. Among them, John M. Lewis, one of Yoshi's students, contributed a chapter titled "Sasaki's Pathway to Deterministic Data Assimilation." I. Michael Navon provided a thorough review of variational DA for numerical weather prediction, while Yoshi himself introduced a new theory based on the entropic balance. Milija and Dusanka Zupanski discussed some issues in ensemble DA, and Zhaoxia Pu overviewed the effect of satellite DA to improve forecasts of tropical cyclones. A coastal application of the ocean DA was reviewed by Xiaodong Hong and colleagues, and the variational approach to hydrologic DA was discussed by Francois-Xavier Le Dimet. Rolf H. Reichle and colleagues addressed recent advances in land data assimilation at the NASA/GMAO, and Nasim Alavi and colleagues surveyed assimilation of soil moisture and temperature into land surface models. As demonstrated, the previous volume covered important topics on DA in meteorology, oceanography, and hydrology, by dealing with both theoretical and practical aspects.

It has been more than 3 years since the first volume has been published. Since then we had three successful symposiums - held at Singapore in August 2009, at Hyderabad in July 2010, and at Taipei in August 2011, each with about 30 presentations. Therefore we decided to publish the second volume under the same title, again by collecting both invited papers and selected papers from the three symposiums. This volume includes excellent overviews of estimation theory, nudging and variational methods, and Markov chain Monte Carlo methods. Most prominently, Yoshi has extended his entropy balance theory for tornado DA from the previous volume.

In this volume, theoretical and methodological aspects encompass estimation and entropic balance theory, variational and ensemble methods, nudging and representer methods, Monte Carlo and ensemble adaptive methods, the maximum likelihood ensemble filter, the local ensemble transform Kalman filter, micro-genetic algorithm, etc., with applications to oceanic, meteorological, and hydrologic DA; radar/lidar/satellite assimilation; parameter estimation; adjoint sensitivity; and adaptive (targeting) observations.

This book will be useful to individual researchers as well as graduate students as a reference to the most recent progresses in the field of data assimilation. We appreciate Boon Chua at Naval Research Laboratory and Francois-Xavier Le Dimet, who have served as the co-conveners of the Sasaki Symposium. We are very honored to dedicate this book to Yoshi Sasaki and the late Roger Daley for their significant contributions in data assimilation.

Ewha Womans University, Seoul — Seon Ki Park
Naval Research Laboratory, Monterey — Liang Xu
July 2012

Contents

List of Contributors

Clark Amerault Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Brian C. Ancell Texas Tech University, Department of Geosciences, Lubbock, TX 79409, USA

Hernan G. Arango Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA

Nancy L. Baker Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Craig Bishop Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Carla Cardinali Data Assimilation Section, ECMWF, Shinfield Park, Reading, Berks, RG2 9AX, UK

Matthew Carrier Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Pak Wai Chan Hong Kong Observatory, 134A, Nathan Road, Kowloon, Hong Kong, China

A. Chandrasekar Department of Earth and Space Sciences, Indian Institute of Space Science and Technology, Valiamala, Thiruvananthapuram 695547, India

Boyu Chen National Meteorological Center of China Meteorological Administration, Beijing 100081, China

Boon S. Chua Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

James A. Cummings Oceanography Division, Naval Research Laboratory, Monterey, CA 93943, USA

Dacian N. Daescu Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, Portland, OR 97207, USA

James Doyle Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Patrick Drake Department of Ocean Sciences, University of California at Santa Cruz, CA 95064, USA

Chris Edwards Department of Ocean Sciences, University of California at Santa Cruz, CA 95064, USA

Takeshi Enomoto Disaster Prevention Research Institute, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology, Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan

Selime Gürol ECMWF, Shinfield Park, Reading, Berks, RG2 9AX, UK

Miki Hattori Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushimacho, Yokosuka, Kanagawa 237-0061, Japan

Daniel Hodyss Naval Research Laboratory, Monterey, CA 93943, USA

Xiaodong Hong Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Jun Inoue Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushimacho, Yokosuka, Kanagawa 237-0061, Japan

Bradley M. Isom Advanced Radar Research Center, University of Oklahoma, Norman, OK 73072, USA

Gregg Jacobs Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Marta Janiskova ECMWF, Shinfield Park, Reading, Berks, RG2 9AX, UK

Wei Kang Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA 93943, USA

Nobumasa Komori Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology, Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan

Arthur J. Krener Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA 93943, USA

Matthew R. Kumjian Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK 73072, USA

M. Govindan Kutty Center for Analysis and Prediction of Storms, National Weather Centre, University of Oklahoma, Norman, OK 73072, USA

Akira Kuwano-Yoshida Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology, Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan

S. Lakshmivarahan School of Computer Science, University of Oklahoma, Norman, OK 73019, USA

Rolf H. Langland Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Yong Hee Lee Forecast Research Laboratory, National Institute of Meteorological Research, Korea Meteorological Administration, Seoul 156–720, Republic of Korea

John Lewis National Severe Storms Laboratory and Desert Research Institute, 2215 Raggio Parkway, Reno, NV 89512–1095, USA

Shie-Yui Liong Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore 119227, Singapore

Philippe Lopez ECMWF, Shinfield Park, Reading, Berks, RG2 9AX, UK

Lynn A. McMurdie Dept. of Atmospheric Sciences, University of Washington, Box 351640, Seattle WA 98195–1640, USA

Takemasa Miyoshi RIKEN Advanced Institute for Computational Science, 7-1-26, Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD 20742, USA

Andrew M. Moore Department of Ocean Sciences, University of California at Santa Cruz, CA 95064, USA

Qoosaku Moteki Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushimacho, Yokosuka, Kanagawa 237-0061, Japan

Mu Mu Key Laboratory of Ocean Circulation and Wave, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China

Emilie Neveu Department of Ocean Sciences, University of California at Santa Cruz, CA 95064, USA

Hans Ngodock Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Seon Ki Park Department of Atmospheric Science & Engineering, Ewha Womans University, Seoul 120–750, Republic of Korea

Patricia Pauley Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Derek J. Posselt Atmospheric, Oceanic, and Space Sciences, University of Michigan, 2455 Hayward Street, Ann Arbor, MI 48109–2143, USA

Alfred M. Powell Jr. NOAA/NESDIS/STAR, 5200 Auth Road, WWB, Camp Springs, MD 20746.

Brian S. Powell Department of Oceanography, University of Hawaii at Manoa, Marine Sciences Building, 1000 Pope Road, Honolulu, Hawaii 96822, USA

Xiaohao Qin State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Srivatsan V. Raghavan Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore 119227, Singapore

Alex Reinecke Naval Research Laboratory, Monterey, CA 93943, USA

Tom Rosmond Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Kazuo Saito Meteorological Research Institute, 1–1 Nagamine, Tsukuba, Ibaraki, 305–0052, Japan

Yoshi K. Sasaki School of Meteorology, University of Oklahoma, Norman, OK 73072, USA

Keith Sashegyi Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Hiromu Seko Meteorological Research Institute, 1–1 Nagamine, Tsukuba, Ibaraki, 305–0052, Japan

Ole M. Smedstad QinetiQ North America, Stennis Space Center, MS 39529, USA

Scott Smith Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Tadashi Tsuyuki Meteorological Research Institute, 1–1 Nagamine, Tsukuba, Ibaraki, 305–0052, Japan

Minh Tue Vu Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore 119227, Singapore

Anthony T. Weaver CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France

Wai Kin Wong Hong Kong Observatory, 134A, Nathan Road, Kowloon, Hong Kong, China

Mingqing Xiao Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA

Jianjun Xu Environmental Science and Technological Center, College of Science, George Manson University, Fairfax, VA 22030, USA

Liang Xu Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA

Shozo Yamane Department of Environmental Systems Science, Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe 610-0394 Kyoto, Japan

Max Yaremchuk Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Xing Yu Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore 119227, Singapore

Edward D. Zaron Department of Civil and Environmental Engineering, Portland State University, Portland, OR 97207, USA

Feifan Zhou Laboratory of Cloud-Precipitation Physics and Severe Storms (LACS), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Milija Zupanski Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO 80523–1375, USA

Chapter 1
A Survey of Observers for Nonlinear Dynamical Systems*

Wei Kang, Arthur J. Krener, Mingqing Xiao, and Liang Xu

Abstract The Kalman filter, invented initially for control systems, has been widely used in science and engineering including data assimilation. For the last several decades, the estimation theory for dynamical systems has been actively developed in control theory. In this paper, we survey several observers, including Kalman filters, for nonlinear systems. We also review some fundamental concepts on the observability of systems defined by either differential equations or a numerical model. The hope is that some of these ideas will inspire research that can benefit the area of data assimilation.

Keywords Observers and estimation • Nonlinear systems • Observability

1.1 Introduction

In modern control theory, the term *Observer* has a technical meaning. An observer is a system defined by differential or difference equations and associated computational algorithms which accepts the measured data from another system as input and

*This work was supported in part by NRL and AFOSR

W. Kang (✉) · A.J. Krener
Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, USA
e-mail: wkang@nps.edu; ajkrener@nps.edu

M. Xiao
Department of Mathematics, Southern Illinois University, Carbondale, IL, USA
e-mail: mxiao@siu.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. II)*, DOI 10.1007/978-3-642-35088-7_1,

returns an estimate of the state of the other system. Observers play a critical role in control systems because many feedback controllers depend on the accurate estimate of state variables of the system to be controlled. An accurate estimation of the state in the presence of noise and uncertainties is essential for a controller to achieve high quality performance.

Estimation from data with random noise can be traced to Gauss about 200 years ago who invented the technique of deterministic least-squares for orbit measurements. In the early twentieth century, Fisher introduced maximum likelihood estimation. Then in the middle of the twentieth century Wiener invented his well known optimal filter for stationary processes. Around 1960s, Kalman and Bucy introduced an optimal recursive filter for dynamical systems. This filter, now known as the Kalman filter, is "the very foundation for data mixing in modern multisensor systems (Gelb 1974)." The estimation for systems governed by differential equations has been an active research field in control theory for more than 50 years. In addition to the Kalman filter, which is essentially a recursive solution to the least square problem, estimation processes have been developed for various performance requirements, such as asymptotically stable estimation, H_∞ estimation, and minimum energy estimation. Fundamental theory has been developed to analyze observability, an intrinsic property of systems with sensors that largely determines the invertibility from past measurement to the state of the system.

Data assimilation is an area of estimation theory and an application to systems with extremely high dimensions. Both filtering and smoothing methods are critical to date assimilation. Although we focus on nonlinear filtering methods in this paper, smoothing algorithms can be developed using similar ideas. Approaches such as ensemble Kalman filters and 4D-Var are based on the theory of optimal estimation, especially the Kalman filter and minimum energy estimation. The data assimilation community has done extensive research on these topics for over 30 years. While this book is focused on problems in data assimilation, this article is to provide a survey on some ideas and results that have been actively developed in control theory, but not widely used in data assimilation. The goal is to lay out some related but different concepts and methods. We hope that some of them may inspire different approaches that benefit the area of data assimilation.

1.2 Observability

In this paper, we consider systems defined by differential equations. The sensor measurement is defined by an output function. For example,

$$\begin{aligned} \dot{x} &= Ax \\ y &= Cx \\ x(0) &= x_0 \end{aligned} \tag{1.1}$$

is a linear system in which $x \in \mathbb{R}^n$ is the state variable, $y \in \mathbb{R}^p$ is the output variable whose value can be measured, $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times p}$ are known constant or time varying matrices. Given A, C, and the past sensor information about $y(t)$, the problem is to estimate x or a function of the state variable in the presence of noise and uncertainties. A nonlinear system is defined similarly,

$$\begin{aligned} \dot{x} &= f(x) \\ y &= h(x) \\ x(0) &= x_0 \end{aligned} \tag{1.2}$$

An immediate question to be answered before observer design is whether a system (1.1) or (1.2) admits a convergent estimator. In other words, how to determine that the past values of $y(t)$ contain adequate information to achieve a reliable estimate of $x(t)$. This leads to the concept of observability. Two initial states x_{01} and x_{02} are said to be distinguishable if the outputs $y_1(t)$ and $y_2(t)$ of (1.2) satisfying the initial conditions $x_0 = x_{01}$ and $x_0 = x_{02}$ differ at some time $t \geq 0$. The system is said to be observable if every pair x_{01}, x_{02} are distinguishable. Observability can be easily verified for linear systems. The output of (1.1) and its derivatives at time $t = 0$ are

$$\begin{aligned} y(0) &= Cx_0 \\ \dot{y}(0) &= CAx_0 \\ \ddot{y}(0) &= CA^2x_0 \\ &\vdots \\ y^{(n-1)}(0) &= CA^{n-1}x_0 \end{aligned} \tag{1.3}$$

Obviously, (1.1) is observable if the mapping from x_0 to the derivatives of $y(t)$ is one-to-one. In fact, it can be proved that (1.1) is observable if and only if the following observability matrix has full rank

$$O = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

For nonlinear systems, the output and its derivatives are given by the iterated Lie derivatives

$$\begin{aligned} y(0) &= y(x_0) \\ \dot{y}(0) &= L_f(h)(x_0) = \frac{\partial h}{\partial x}(x_0)f(x_0) \end{aligned}$$

$$\ddot{y}(0) = L_f^2(h)(x_0) = \frac{\partial L_f(h)}{\partial x}(x_0)f(x_0)$$

$$\vdots$$

$$y^{(k-1)}(0) = L_f^{k-1}(h)(x_0) = \frac{\partial L_f^{k-2}(h)}{\partial x}(x_0)f(x_0)$$

for some integer $k > 0$. If the mapping from x_0 to $h, L_f(h), L_f^2(h), \cdots$ distinguishes points then the system is observable. For a real analytic system this is a necessary and sufficient condition for observability. For simplicity of exposition, suppose $p = 1$. Consider the matrix

$$\begin{bmatrix} \frac{\partial h}{\partial x}(x_0) \\ \frac{\partial L_f(h)}{\partial x}(x_0) \\ \vdots \\ \frac{\partial L_f^{n-1}(h)}{\partial x}(x_0) \end{bmatrix}$$

If this matrix is invertible, then the system is locally observable at x_0. This observability matrix is a topic addressed in almost all textbooks of linear and nonlinear control theory, for instance Kailath (1980) for linear systems and Isidori (1995) for nonlinear systems.

For high dimensional systems, it is important to quantitatively define observability. The observability Gramian is a widely used concept for this purpose (Kailath 1980). Consider a linear system (1.1), an arbitrary initial state x_0 of a trajectory

$$x(t) = e^{At}x_0$$

can be uniquely determined from the known function $y(t) = Cx(t)$ if and only if the columns in the matrix

$$Ce^{At}$$

are linearly independent over $[t_0, t_1]$. This is equivalent to say that

$$G = \int_{t_0}^{t_1} e^{A^T t}C^T Ce^{At}dt$$

is nonsingular. This matrix is called the observability Gramian. In fact, the L^2-norm of the output satisfies

$$\int_{t_0}^{t_1} ||y(t)||^2 dt = x_0^T G x_0$$

Therefore, the eigenvalues of G represent the gain from the initial state to the output. If G has a zero eigenvalue, then its eigenvector results in a zero output. The system is unobservable. If G has a very small eigenvalue, then the system is weakly observable, i.e. a small noise in $y(t)$ can cause a large estimation error. Therefore, the smallest eigenvalue of G is used as a quantitative measure of observability.

For nonlinear systems, an empirical observability Gramian can be numerically computed (Krener and Ide 2009). Consider (1.2) and a nominal trajectory $x(t)$ with initial state $x(0) = x_0$. Define a mapping

$$\begin{aligned} &\delta x_0 \to h(\hat{x}(t)) - h(x(t)) \\ &\text{subject to} \\ &\dot{\hat{x}}(t) = f(\hat{x}(t)) \\ &\hat{x}(0) = x_0 + \delta x_0 \end{aligned} \tag{1.4}$$

Let $v_1, v_2, \cdots, v_n$ be an orthonormal basis in $\mathbb{R}^n$. Let $\rho > 0$ be a small number. In the direction of ρv_i, the variation of the output can be estimated empirically by

$$\Delta_i(t) = \frac{1}{2\rho}\left(h(x^+(t)) - h(x^-(t))\right), \tag{1.5}$$

where

$$\begin{aligned} \dot{x}^{\pm}(t) &= f(x^{\pm}(t)) \\ \hat{x}^{\pm}(0) &= x_0 \pm \rho v_i, \end{aligned}$$

The mapping, (1.4), from the initial state to the output space can be locally approximated by a linear function

$$\delta x_0 = \sum_{i=0}^{n} \alpha_i v_i \to \sum_{i=0}^{n} \alpha_i \Delta_i(t) \tag{1.6}$$

Therefore, the observability Gramian of the nonlinear system can be approximated by the Gramian associated to (1.6)

$$\begin{aligned} G &= (G_{ij})_{i,j=1}^n \\ G_{ij} &= \int_{t_0}^{t_1} \Delta_i^T(t)\Delta_j(t)dt \end{aligned} \tag{1.7}$$

Locally around the nominal trajectory, the eigenvalues of (1.7) measure the gain from the variation of the initial state to the variation of the output. If G has a small eigenvalue, then $x(t)$ is weakly observable. A small noise in $y(t)$ can result in a large estimation error.

The Gramian or empirical Gramian in Kailath (1980) and Krener and Ide (2009) measures the observability of full initial states. However, for systems with very high dimensions, the problem of full observability is, in many cases, ill-posed. Some discussions on the partial observability, or Z-observability, for complex systems were introduced in Kang and Barbot (2007). Meanwhile, quantitatively measure partial observability has been rapidly developed in a sequence of papers (Kang 2011; Kang and Xu 2009a,b, 2011). For PDEs, the observability is defined and computed for the finite dimensional approximations of the original model. In Kang and Xu (2009a,b), dynamic optimization is used as a tool for the definition.

Definition 1.1. Given a trajectory $x(t)$, $t \in [t_0, t_1]$. Let $W \subseteq \mathbb{R}^n$ be a subspace. Let $\rho > 0$ be a constant. Define ϵ as follows

$$\begin{aligned}
&\epsilon = \min_{\bar{x}(t)} ||h(\bar{x}(t)) - h(x(t))|| \\
&\text{subject to} \\
&\dot{\bar{x}} = f(\bar{x}), \\
&||\bar{x}(0) - x_0|| = \rho \\
&\bar{x}(0) - x_0 \in W
\end{aligned}$$

Then the ratio ρ/ϵ is a measure of observability for the W-component of $x(0)$.

If $\rho \to 0$, the ratio ρ/ϵ can be considered as an extension of the observability Gramian. Consider a linear system (1.1). Suppose $W = \mathbb{R}^n$. Then the observability Gramian, G, satisfies (Kailath 1980; Krener and Ide 2009)

$$||y||_{L^2}^2 = x_0^T P x_0 \tag{1.8}$$

Given $||x_0|| = \rho$, we have

$$\epsilon^2 = \lambda_{min} \rho^2 \tag{1.9}$$

where λ_{min} is the smallest eigenvalue of G. Therefore, the ratio ρ^2/ϵ^2 equals the reciprocal of the smallest eigenvalue of the observability Gramian. In Kang and Xu (2009a,b), the concept of partial observability was applied to more general problems using various types of norms and knowledge of the system. An example of optimal sensor location by maximizing the observability for data assimilations was given in Kang and Xu (2011).

1.3 Asymptotic Observers

Following control theory, asymptotic observers are systems defined by differential or difference equations and associated computational algorithms which accepts the measured data from another system as input and returns an estimate of the state of the other system. In the case of a perfect model without noise and uncertainties, the estimated state should converge to the true state of the system being observed. Also if the initial state of the observer equals the true state, then the estimation error is zero along the entire trajectory. In most observer designs, such as Luenberger observers and Kalman filters, an observer consists of a copy of the original system plus a correction term which is a function of the measured data.

Asymptotic observers are widely used in control systems to achieve stable estimates of state variables. The design emphasizes the stability and simplicity of the estimation process. In general it does not optimize any performance measure. The Luenberger observer for linear systems is a simplest example that illustrates the fundamental idea of asymptotic observers.

1.3.1 Luenberger Observer

Given a dynamical system with an output

$$\begin{aligned} \dot{x} &= Ax \\ y &= Cx \end{aligned} \tag{1.10}$$

where $x \in \mathbb{R}^n$ is the state variable, $y \in \mathbb{R}^p$ is the output which can be measured, $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$ are matrices. We assume that A, C and the output $y(t)$ are known information. The goal is to find an estimate, estimate, denoted by $\hat{x}(t)$, of the state variable so that $\hat{x}(t)$ asymptotically approaches $x(t)$. The observer has the following form

$$\dot{\hat{x}} = A\hat{x} + G(y - C\hat{x}) \tag{1.11}$$

The matrix $G \in \mathbb{R}^{n\times p}$ is called the observer gain, which is used to stabilize the estimation error. Define

$$e = x - \hat{x}$$

then the error dynamics has the following form

$$\begin{aligned} \dot{e} &= Ae - G(y - C\hat{x}) \\ &= (A - GC)e \end{aligned} \tag{1.12}$$

It is obvious that $e(t)$ asymptotically approaches zero if the eigenvalues of $A - GC$ are all located in the left half plane. To estimate $x(t)$, one can use any initial guess $\hat{x}(0)$. Then $\hat{x}(t)$ from (1.11) satisfies

$$\lim_{t\to\infty} e(t) = 0$$

When applying the observer, $y(t)$ is measured online and the (1.11) is numerically propagated in real-time to provide an estimate of $x(t)$.

It can be proved that, for any set of n complex numbers, there always exists an observer gain, G, so that the eigenvalues of the error dynamics (1.12) are placed at these locations, if the pair (A, C) is observable, i.e. the following observability matrix has full rank

$$\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

This result guarantees that one can always find linear observers with stable error dynamics for observable systems. For systems in which (A, C) is not observable, it is still possible to achieve asymptotic stability of (1.12). This depends on the spectrum of A, which can be divided into observable modes and unobservable modes. Details are referred to Kailath (1980). If all the unobservable modes are on the left half plane, then there always exists a G that stabilizes (1.12).

The error dynamics does not include measurement error. If the output is corrupted by noise, the asymptotic stability of the observer guarantees that $\hat{x}(t)$ is stabilized around the true value. There are infinitely many observer gains to stabilize the observer. A high gain observer has fast convergence to the true value of the system, however it is very sensitive to sensor noise. Although asymptotic observers do not guarantee optimal performance in any sense, their advantage lies in the simplicity. For real time applications, each estimate at a given time is simply computed by one step integration of the observer equation, which can be implemented using any numerical algorithm for solving ordinary differential equations (ODEs). Luenberger observers can be found as a standard topic in almost all textbooks on control theory, for instance (Kailath 1980; Khalil 2002).

1.3.2 Observers with Linear Error Dynamics

For nonlinear systems, observer design with a guaranteed asymptotically stable error dynamics is a difficult task (Hermann and Krener 1977). The Luenberger observer works for linear systems because its error dynamics is decoupled from the unknown trajectory being observed. For nonlinear systems, however, this is not true in general. There is a large volume of literature on the construction of nonlinear observers that admit a linear error dynamics. In the pioneering work (Krener and Isidori 1983) a technique called output injection was introduced. In addition, necessary and sufficient conditions are found under which the error dynamics of the nonlinear observer is equivalent to a linear ODE. Consider a nonlinear dynamical system with an output

$$\begin{aligned} \dot{x} &= f(x) \\ y &= h(x) \end{aligned} \tag{1.13}$$

in which $x \in \mathbb{R}^n$ is the state variable, $y \in \mathbb{R}^p$ is the output which can be measured, $f(x)$ and $h(x)$ are vector valued functions with adequate smoothness. In Krener and Isidori (1983), it is propose to find a change of coordinates around a fixed point x_0

$$\begin{aligned} z &= z(x) \\ z(x_0) &= 0 \end{aligned} \tag{1.14}$$

so that (1.13) is transformed into a linear system with a nonlinear output injection

$$\begin{aligned} \dot{z} &= Az + \phi(y) \\ y &= Cz \end{aligned} \tag{1.15}$$

for some matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. If this is the case, then we can easily construct a Luenberger type of observer as follows.

$$\dot{\hat{z}} = A\hat{z} + \phi(y) + G(y - C\hat{z}) \tag{1.16}$$

Let

$$e = z - \hat{z}$$

then the error dynamics is a linear system decoupled from $z(t)$

$$\dot{e} = (A - GC)e$$

If G, the observer gain, is chosen so that the eigenvalues of $(A - GC)$ are all in the left half plane, then

$$\lim_{t\to\infty} e(t) = 0$$

Not all nonlinear systems can be transformed into a linear system with output injection. The existence of the change of coordinates (1.14) can be determined using Lie differentiation. Given a function $h(x)$, let dh represents the 1-form, or the gradient,

$$dh(x) = \left[\frac{\partial h}{\partial x_1}(x) \ \frac{\partial h}{\partial x_2}(x) \cdots \frac{\partial h}{\partial x_n}(x) \right]$$

The Lie derivative is defined as follows

$$\begin{aligned} L_f(h) &= dh \cdot f \\ L_f(dh) &= f^T \frac{\partial^2 h}{\partial x} + dh \frac{\partial f}{\partial x} \end{aligned}$$

The following theorem was proved in Krener and Isidori (1983).

Theorem 1.1. *There exists a local change of coordinates (1.14) that transforms (1.13) into a linear system with output inject (1.15) if and only if*

$$\begin{aligned} f(x_0) &= 0 \\ h(x_0) &= 0 \end{aligned}$$

and $L_f^n(dh)$ is a linear combination of $L_f^k(dh)$ for $k = 0, 1, \cdots, n-1$.

Note that the theorem guarantees the existence of a local change of coordinates around an equilibrium. Therefore, the observers are limited in a local neighborhood of an equilibrium point. Among a large number of publications on the observer design by achieving linearized error dynamics, we would like to bring up (Kazantzis and Kravaris 1998). In this work, the formulation of the observer design problem is realized via a system of singular first-order linear partial differential equations (PDE). The theory is applicable to a larger family of systems than that addressed in Krener and Isidori (1983). In fact, after a nonlinear change of coordinates, the resulting system is not required to have a linear output like in (1.15). Another advantage of the work in Kazantzis and Kravaris (1998) is that the solution to the PDEs is locally analytic and this enables the development of a series solution method, that is programmable using symbolic software packages. In the presence of noise, some types of output injection, such as a y^2 term, may result in a biased estimation because $E[(y+n)^2] = E[y^2] + E[n^2]$, where n is a random noise.

Other related work includes Zeitz's extended Luenberger observer based upon a local linearization technique (Zeitz 1987). Nonlinear coordinate transformations have also been employed to transform the nonlinear system to a suitable observer canonical form, where the observer design problem may be solved (Bestle and Zeitz 1983; Ding et al. 1990; Xia and Gao 1989; Zheng et al. 2007).

1.3.3 Observers Based on Lyapunov Functions

For systems that do not admit a linear error dynamics, nonlinear observers can be derived so that its stability is guaranteed by a Lyapunov function. A widely used approach is based on the high gain observer proved in Gauthier et al. (1992). Once again, consider the nonlinear system (1.13). Using a single output case as an example, consider the mapping, $z = z(x) : \mathbb{R}^n \to \mathbb{R}^n$, defined by

$$z(x) = \begin{bmatrix} h(x) \\ L_f h(x) \\ \vdots \\ L_f^{n-1} h(x) \end{bmatrix} \tag{1.17}$$

We assume that $z = z(x)$ is a diffeomorphism on a region $\Omega \subseteq \mathbb{R}^n$. Under this transformation, the original system is equivalent to the system in the form

$$\begin{aligned} \dot{z} &= \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \\ \phi(z) \end{bmatrix} \\ y &= z_1 \end{aligned} \tag{1.18}$$

Using the following notation

$$\begin{aligned}\bar{f}(z) &= \begin{bmatrix} z_2 \; z_3 \; \cdots \; z_n \; \phi(z) \end{bmatrix}^T \\ \bar{A} &= \begin{bmatrix} 0 \; 1 \; 0 \cdots 0 \; 0 \\ 0 \; 0 \; 1 \cdots 0 \; 0 \\ \vdots \; \vdots \; \vdots \; \cdots \; \vdots \; \vdots \end{bmatrix} \\ \bar{C} &= \begin{bmatrix} 1 \; 0 \cdots 0 \end{bmatrix}\end{aligned}$$

the observer has the form

$$\dot{\hat{z}} = \bar{f}(\hat{z}) - S^{-1}\bar{C}(\bar{C}\hat{z} - y)$$

where S is the solution of the equation

$$-\theta S - \bar{A}^T S - S\bar{A} + \bar{C}^T\bar{C} = 0$$

where θ is a constant. It is proved in Gauthier et al. (1992) that the error of the observer

$$e(t) = z - \hat{z}$$

approaches zero if θ is large enough (thus the name "high gain observer"). While the proof is carried out in the z-space, the observer can be constructed in the original state space

$$\dot{\hat{x}} = f(\hat{x}) - \left(\frac{\partial z}{\partial x}\right)^{-1} S^{-1}\bar{C}(h(\hat{x}) - y)$$

The simplicity in the construction of a high gain observer makes it a convenient tool for nonlinear systems (Gauthier and Kupka 1994). However, in the presence of noise, a high gain observer should be used with caution. It may significantly enlarge the impact of the noise and result in large estimation errors. In addition, the "homomorphism" requirement for (1.17) limits the region in the state space in which the observer is applicable. In Krener and Kang (2003), a nonlinear observer is constructed without a global homomorphism requirement. In addition, the observer gain depends on the state of the system so that it is not constantly high. Global or semi-global observers can also be derived based on Lyapunove functions for systems with a triangular structure or bounded nonlinear terms in its differential equations (Lei et al. 2007; Krener and Kang 2003; Tsinias 1989). Deriving Lyapunov functions for nonlinear systems is always difficult. An alternative is to directly apply convergent numerical algorithms in nonlinear observers. For instance, an Euler-Newton observer is introduced in Kang (2006). Moving horizon observers are also computational based methodologies (Findeisen et al. 2002; Michalska and Mayne 1995).

The sliding mode observer is another Lyapunov function based approach. It has the capability of handling unknown inputs or unknown parameters (Floquet and Barbot 2007). A survey of various types of sliding mode observers can be found in

Spurgeon (2008). It is interesting to point out that an engineering approach for fast estimation is to build an electronic analogue realization of a sliding mode observer (L'Hernault et al. 2008).

1.4 Optimal Filtering

Optimal filtering is a class of observers that achieve optimal performance by minimizing some metrics of the estimation error. Due to the optimality requirement, the online computational load required for optimal filters is usually higher than that needed for asymptotic observers.

1.4.1 Kalman Filters

Consider a system with random noise

$$\begin{aligned} \dot{x} &= f(x) + Gw \\ y &= h(x) + Dv \end{aligned} \tag{1.19}$$

where w and v are standard white Gaussian noises. Suppose the estimated state is $\hat{x}(t)$. If (1.19) is nonlinear, we linearize it around $\hat{x}(t)$

$$\begin{aligned} \dot{x} &= A(t)x + w \\ y &= C(t)x + v \end{aligned} \tag{1.20}$$

where

$$A(t) = \frac{\partial f}{\partial x}(\hat{x}(t)),\ C(t) = \frac{\partial h}{\partial x}(\hat{x}(t))$$

A Kalman filter based upon the linearization of a nonlinear system is called an extended Kalman filter (EKF). It includes the estimates of the state variable, $\hat{x}$, and the estimation error covariance matrix, $P(t) \in \mathbb{R}^n$. More specifically, EKF is an observer with a dynamic gain

$$\begin{aligned} \dot{\hat{x}} &= f(\hat{x}) + K(t)(y - \hat{y}) \\ \dot{P}(t) &= A(t)P(t) + P(t)A^T(t) + Q(t) - P(t)C^T(t)R^{-1}C(t)P(t) \\ \hat{x}(t_0) &= \hat{x}_0,\ P(t_0) = P_0 \\ \hat{y} &= h(\hat{x}) \\ Q(t) &= G(t)G^T(t) \\ R(t) &= D(t)D^T(t) \\ K(t) &= P(t)C^T(t)R^{-1}(t) \end{aligned}$$

The matrices Q(t) and R(t) are the driving noise covariance in the system dynamics and the measurement noise covariance, respectively. The matrix $Q(t)$ must be nonnegative definite and $R(t)$ must be positive definite. The initial state estimate x_0 and its covariance P_0 describe the prior knowledge of the true state at the beginning of the process. The Kalman filter "represents the most widely applied and demonstrable useful result to emerge from the state variable approach of modern control theory" (Sorenson 1985). It can be found in many textbooks on control theory, for instance (Gelb 1974; Brown and Hwang 1997). A drawback of EKF is that the convergence is, in general, not guaranteed. Simple examples can be found in which an EKF estimation process diverges (Krener 2004). Various proofs of its local convergence exist in the literature. Interested readers are referred to Krener (2003b) and references therein.

An EKF requires the linearization of system models, which may not be easily available during real-time operations. In addition, the linearization is changed if the model is modified or updated. A different approach is to use the Unscented Kalman Filter (UKF) which does not require the online computation of the linearization. Following Julier and Uhlmann (2004), consider a discrete time nonlinear system

$$\begin{aligned} x_k &= f(x_{k-1}, w_{k-1}) \\ y_k &= h(x_{k-1}, v_{k-1}) \end{aligned}$$

where x, y, v and w are the state, measurement, process noise and measurement noise respectively. The UKF is "founded on the intuition that it is easier to approximate a probability distribution than it is to approximate an arbitrary nonlinear function or transformation" (Julier and Uhlmann 2004). The UKF assumes that at every sampling instance, the state x is always a normally distributed variable. The mean and the covariance information of this random variable can be stored in a set of specially chosen points called sigma points. One simple choice of such sigma points is given below (Julier and Uhlmann 2004)

$$\sigma^i = E(x) \pm \sqrt{nP},\ i = 1, 2, \ldots, n$$

where $E(x)$ is the mean of the random variable x, P is the covariance matrix and n is the dimension of x. It can be shown that the nonlinear transformation of the sigma points preserves statistics up to second order in a Taylor serious expansion (Julier and Uhlmann 2004). Based on this fact, a prediction of the state and the covariance matrices in the filter algorithm can be carried out as follows:

- Based on the previous-step estimation of the state, $\hat{x}_{k-1}$, and the covariance matrix, $\hat{P}^{xx}_{k-1}$, calculate a set of sigma points as

$$\sigma^i = \hat{x}_{k-1} \pm \sqrt{n\hat{P}^{xx}_{k-1}},\ i = 1, 2, \ldots, n;$$

- Propagate all the sigma points through the nonlinear dynamic and the output equations,

$$\begin{aligned} z^i &= f(\sigma^i, 0) \\ g^i &= h(\sigma^i, 0),\ i = 1, 2, \ldots, n; \end{aligned}$$

- Calculate the mean (prediction) of the state and output,

$$\begin{aligned} \tilde{x}_k &= \tfrac{1}{2n} \textstyle\sum_{i=1}^{2n} z^i \\ \tilde{y}_k &= \tfrac{1}{2n} \textstyle\sum_{i=1}^{2n} g^i; \end{aligned}$$

- The prediction of the covariance matrices are given by,

$$\begin{aligned} \tilde{P}_k^{xx} &= \tfrac{1}{2n} \textstyle\sum_{i=1}^{2n} (z^i - \tilde{x}_k)(z^i - \tilde{x}_k)^T \\ \tilde{P}_k^{yy} &= \tfrac{1}{2n} \textstyle\sum_{i=1}^{2n} (g^i - \tilde{y}_k)(g^i - \tilde{y}_k)^T \\ \tilde{P}_k^{xy} &= \tfrac{1}{2n} \textstyle\sum_{i=1}^{2n} (z^i - \tilde{x}_k)(g^i - \tilde{y}_k)^T \end{aligned}$$

Once the prediction of $\tilde{x}_k$, $\tilde{P}_k^{xx}$, $\tilde{P}_k^{yy}$ and $\tilde{P}_k^{xy}$ are available, the update is given by

$$\hat{x}_k = \tilde{x}_k + K(y_k - \tilde{y}_k)$$

where

$$\begin{aligned} K &= \tilde{P}_k^{xy} [\tilde{P}_k^{yy}]^{-1} \\ \hat{P}_k^{xx} &= \tilde{P}_k^{xx} - K \tilde{P}_k^{xy} K^T. \end{aligned}$$

While UKF avoids the computation of linearization, it requires the integration of $2n$ trajectories. For nonlinear systems with a moderate dimension, the UKF is an reliable and efficient filter for real-time estimation. However, it is not clear if the idea is applicable to large scale systems with tens of thousands or even millions of dimensions. For systems with very high dimensions, such as the models for numerical weather forecast, currently popular approaches include the ensemble Kalman filter (EnKF) and 4D-Var estimation and prediction (Anderson 2003; Evensen 2007, 1994; Houtekamer and Mitchell 1998; Chua and Bennett 2001; Courtier et al. 1994; Rabier et al. 2000; Xu et al. 2005). These methods are primarily developed and widely used in the data assimilation community. They are extensively addressed in the other chapters of this book. Therefore, we skip the details on EnKF and 4D-Var methods.

1.4.2 H_∞ Filter

The Kalman filter is optimal in a stochastic sense. However, the probability model of disturbances may not be available for a given system. In this case, one may assume that the noises are not stochastic but unknown L^2 functions. The goal of H_∞ filters is to estimate the state variables in such a way that the gain from noise to estimation error is as small as possible. Following Krener (2004), consider a system

$$\begin{aligned}\dot{x} &= f(x) + g(x)w \\ y &= h(x) + v \\ x(0) &= x^0 + \tilde{x}_0\end{aligned} \tag{1.21}$$

where disturbances w and v are unknown L_2 functions, $\tilde{x}_0$ is an unknown error in the initial condition. The total "energy" of the disturbances is formulated using L^2 norms

$$||\tilde{x}_0||^2 + \int_0^t ||w(\tau)||^2 + ||v(\tau)||^2 d\tau$$

For some $\gamma > 0$, if a filter satisfies

$$\int_0^t ||x(\tau) - \hat{x}(\tau)||^2 d\tau \le \gamma^2 \left(||\tilde{x}_0||^2 + \int_0^t ||w(\tau)||^2 + ||v(\tau)||^2 d\tau \right)$$

for arbitrary $\tilde{x}_0$, w, and v, then we say that the gain from the disturbance to the estimation error is bounded by γ. We seek an estimator based on worst case scenarios. Define

$$\begin{aligned}Q(x,t) = \inf_{\tilde{x}_0, w(\cdot)} \Bigg(&\frac{\gamma^2}{2} ||\tilde{x}_0||^2 + \frac{\gamma^2}{2} \int_0^t ||w(\tau)||^2 + ||y(\tau) - h(x(\tau))||^2 d\tau \\ &- \frac{1}{2} \int_0^t ||x(\tau) - \hat{x}(\tau)||^2 d\tau \Bigg)\end{aligned}$$

where $x(\cdot)$ is subject to (1.21) and $x(t) = x$. If $Q(x,t) \ge 0$, then it is guaranteed that the gain from the disturbance to the estimation error is bounded by γ. From dynamic programming, $Q(x,t)$ satisfies the following partial differential equation

$$\begin{aligned}0 = &\frac{\partial Q}{\partial t}(x,t) + \sum_{i=1}^{n} \frac{\partial Q}{\partial x_i}(x,t) f_i(x) + \frac{1}{2\gamma^2} \sum_{i,j=1}^{n} \frac{\partial Q}{\partial x_i}(x,t) a_{ij}(x) \frac{\partial Q}{\partial x_j}(x,t) \\ &- \frac{\gamma^2}{2} ||y(t) - h(x)||^2 + \frac{1}{2} ||x - \hat{x}||^2\end{aligned} \tag{1.22}$$

If the equation has a solution, then the optimal estimate is given by

$$\hat{x}(t) = \text{argmin}_x Q(x,t) \tag{1.23}$$

It is of Hamilton-Jacobi type, first order, nonlinear PDE driven by the observations. It is very difficult, if not impossible, to compute an accurate solution in real time. Moreover it may not admit a smooth solution so the (1.22) must be interpreted in the viscosity sense. This is an infinite dimensional observer with state $Q(\cdot, t)$ evolving according to (1.22) with state estimate given by (1.23). Hence it is of limited practical use.

For linear systems, (1.22) reduces to a Riccati differential equation. Consider

$$\begin{aligned} \dot{x} &= Ax + Bw \\ y &= Cx + Dv \\ x(0) &= 0 \end{aligned}$$

Its H_∞ filter has the following form

$$\begin{aligned} &\dot{\hat{x}}(t) = A\hat{x} + K(t)(y(t) - C\hat{x}(t)) \\ &K(t) = Q(t)C^T \end{aligned}$$

where $Q(t)$ is a solution of the Riccati differential equation

$$\begin{aligned} &\dot{Q}(t) = AQ(t) + Q(t)A^T + BB^T - Q(t)(C^TC - \gamma^{-2}I)Q(t) \\ &Q(0) = 0 \end{aligned}$$

This is an observer similar to Kalman filter. However, this filter can be modified to estimate a linear combination of the state variables, $z = Lx$. There are many books and papers on H_∞ filters, for instance (Green and Limebeer 1995).

1.4.3 Minimum Energy Estimation

Consider a system model (1.21) in which the noises are unknown L^2 functions. We seek the initial state error $\tilde{x}_0$ and the noises $w(t)$ and $v(t)$ of "minimum energy"

$$\frac{1}{2}||\tilde{x}_0||^2 + \frac{1}{2}\int_0^t ||w(\tau)||^2 + ||v(\tau)||^2 d\tau$$

where $v(t)$ is consistent with the observation. The quality of estimation is defined by an optimal control problem

$$Q(x,t) = \inf_{\tilde{x}_0, w(\cdot)} \left(\frac{1}{2}||\tilde{x}_0||^2 + \frac{1}{2}\int_0^t ||w(\tau)||^2 + ||y(\tau) - h(x(\tau))||^2 d\tau \right)$$

in which $x(\tau)$ is subject to (1.21) and $x(t) = x$. The optimal estimation is given by the one that minimizes $Q(x,t)$,

$$\hat{x}(t) = \text{argmin}_x Q(x,t)$$

This approach is similar to and predates a H_∞ estimation, except that it does not require the searching for gain γ. The dynamic programming approach yields a partial differential equation for $Q(x,t)$

$$
\begin{aligned}
0 = \frac{\partial Q}{\partial t}(x,t) + \sum_{i=1}^{n} \frac{\partial Q}{\partial x_i}(x,t) f_i(x) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial Q}{\partial x_i}(x,t) a_{ij}(x) \frac{\partial Q}{\partial x_j}(x,t) \\
-\frac{1}{2} ||y(t) - h(x)||^2
\end{aligned}
$$

Similar to the H_∞ filter, this equation is very difficult to solve, if not impossible, either analytically or numerically. For linear systems, the partial differential equation is reduced to a linear Riccati equation, which is numerically solvable for systems with a moderate dimension. For systems with extremely high dimensions, such as the models used for numerical weather forecast, special treatment must be applied in the optimization process. In principle, 4D-Var is a discrete minimum energy filter using a weighted norm. Some matrices of extremely large size exceed the capacity of computational facility. The way to get around these difficulties is to use tangent linear model and adjoint model in the computation (Liang, some references here). More information on the general idea of minimum energy estimation methods is referred to Hijab (1980), Krener (2003a), and Mortensen (1968) and references therein.

1.5 Observer Construction for PDE Systems

1.5.1 Linear Case

We consider a PDE system written in an abstract form

$$
\begin{aligned}
\dot{x}(t) &= Ax(t), \quad x(0) = x_0, \quad t \geq 0 \\
y(t) &= Cx(t), \quad t \geq 0,
\end{aligned}
\tag{1.24}
$$

on a Hilbert space X, where A is the infinitesimal generator of the strongly continuous semigroup e^{At} on X and C is a bounded operator from X to a second Hilbert space Y (Curtain and Zwart 1995).

Similar to the finite dimensional case, the observability map of (1.24) on $[0, T]$ is a bounded linear operator $\mathcal{C}_T : X \to L_2([0, T]; Y)$ defined as follows

$$
\mathcal{C}_T(x)(t) = Ce^{A(T-t)}x. \tag{1.25}
$$

A widely adopted definition of observability is based on the property that the knowledge about the output y over a finite time interval uniquely determines the initial state. The following definition is essentially the same as the one following (1.3) for finite dimensional systems:

Definition 1.2. System (1.24) is exactly observable on $[0, T]$ (for some $T > 0$) if $\mathcal{C}_T$ is injective and its inverse is bounded on the range of $\mathcal{C}_T$.

In other words, (C, A) is exactly observable on $[0, T]$ if $\mathrm{Ker}(\mathcal{C}_T)$ is $\{0\}$ and $\mathcal{C}_T$ has a closed range. Although the exact observability is consistent with the one for finite dimensional systems, there is no general observability test for infinite dimensional systems. The observability is equivalent to the following inequality

$$\int_0^T \|(\mathcal{C}_T x)(s)\|_Y^2 ds = \int_0^T \|Ce^{A(T-s)}x\|_Y^2 ds \geq \gamma \|x\|_X^2 \tag{1.26}$$

where $\gamma > 0$ is a constant which may depend on T. However, in many cases the inverse of $\mathcal{C}_T$ may not be bounded. Thus this leads to the following definition of weak observability:

Definition 1.3. System (1.24) is approximately observable on $[0, T]$ (for some $T > 0$) if

$$\ker(\mathcal{C}_T) = \{0\}.$$

In other words, (A, C) is approximately observable on $[0, T]$ if $\mathcal{C}_T$ is injective. This definition has a drawback. Some observable systems can be ill-posed, i.e. the inverse mapping from the output variable to the estimated state is extremely sensitive to noise. In this case, studying partial observability makes more sense. In fact, in Kang (2011) it was proved that Definition 1.1 can be applied to PDEs to quantitatively measure the observability of a finite dimensional subspace of the state variables.

A Luenberger observer for (1.24) is an abstract system in the form of

$$\begin{aligned}\dot{\hat{x}}(t) &= A\hat{x} + L(\hat{y}(t) - y(t))\\ \hat{y}(t) &= C\hat{x}(t)\end{aligned} \tag{1.27}$$

where $L : Y \to X$ is a linear operator. Unlike the finite dimensional case, even if (1.24) is exactly observable on some interval $[0, T]$, we may not have a convergent observer (1.27) (see Curtain and Zwart 1995 and references therein). If we define the error $e(t) = x(t) - \hat{x}(t)$, then $e(t)$ approaches zero exponentially as t increases provided that (A, C) is exponentially detectable, which means that there exists a linear operator $L : Y \to X$ such that $A + LC$ generates an exponentially stable C_0-semigroup $e^{(A+LC)t}$.

When C is a compact operator, a necessary condition for (A, C) to be detectable is that the unstable part of the spectrum of A consists only of eigenvalues (Curtain and Zwart 1995). In infinite dimensions, it is impossible to achieve arbitrary eigenvalue assignment, but some interesting results on partial assignment can be found in Clarke and Williamson (1981), Curtain and Zwart (1995), Russell (1968), Sun (1981), and Rebarber (1999). In the following, we present a relatively complete result when C has a finite rank, i.e. $\mathrm{Rang}(C)$ is finite dimensional, a typical case in engineering problems.

For a given real number α, the spectrum of A can be decomposed into two parts in the complex plane

$$\begin{aligned}\sigma_\alpha^+ &= \sigma(A) \cap \{\lambda \in \mathbb{C} \mid Re(\lambda) \geq \alpha\}\\ \sigma_\alpha^- &= \sigma(A) \cap \{\lambda \in \mathbb{C} \mid Re(\lambda) < \alpha\}.\end{aligned} \tag{1.28}$$

An operator A is said to satisfy the *spectrum decomposition assumption at* α if σ_α^+ is bounded and separated from σ_α^-, i.e. the boundaries of σ_α^+ and σ_α^- have no intersection. Under this assumption, we can define the following spectral projection

$$P_\alpha x = \frac{1}{2\pi j} \int_{\Gamma_\alpha} (\lambda I - A)^{-1} x d\lambda,$$

where $j^2 = -1$ and Γ_α is a curve traversed once in the positive direction (counterclockwise) to enclose an open set containing σ_α^+ in its interior and σ_α^- in its exterior. The projection induces a decomposition of the state space X:

$$X = X_\alpha^+ \oplus X_\alpha^-, \quad \text{where } X_\alpha^+ = P_\alpha X \text{ and } X_\alpha^- = (I - P_\alpha)X.$$

Next let us denote

$$\begin{aligned} A_\alpha^+ &= P_\alpha A, & A_\alpha^- &= (I - P_\alpha)A, \\ C_\alpha^+ &= C P_\alpha, & C_\alpha^- &= C(I - P_\alpha). \end{aligned} \tag{1.29}$$

Assume C has finite rank. We say that (A, C) is detectable with stability margin greater than or equal to $-\alpha$ if there exists $L \in \mathcal{L}(Y, X)$ such that the C_0-semigroup $e^{(A+LC)t}$ generated by $A + LC$ satisfies

$$\|e^{(A+LC)t}\| \leq M e^{-\beta t}, \quad M \geq 1 \tag{1.30}$$

holds for any $\beta < \alpha$. The pair (A, C) is detectable if and only if

- A satisfies spectrum decomposition assumption at α;
- X_α^+ is finite dimensional;
- (A_α^+, C_α^+) is observable;
- $e^{A_\alpha^- t}$ is exponentially stable with a stability margin that is least $-\alpha$.

Sufficient conditions for the exponential detectability were obtained in 1975 by Triggiani (1975) (also see surveys by Pritchard and Zabczyk (1981) and by Russell (1978)). In 1985 Desch and Schappacher (1985) show that these conditions are also necessary for finite-rank inputs. These conditions can be simplified for systems of either the Riesz-spectral type or the retarded delay type (Bhat 1986; Curtain and Pritchard 1974; Curtain and Zwart 1995).

The observer for (1.24) has been studied by many authors (see Orner and Foster 1971; Kitamura et al. 1972; Sakawa and Matsushita 1975; Balas 1980; Gressang and Lamont 1975; Fuji 1980) under the framework of distributed parameter systems. However, due to its infinite-dimensional feature, in general, it is not implementable in applications. Thus designs of finite dimensional observers (in the context of compensators) were proposed based on eigenfunction projections or direct state space projection (Bernstein and Hyland 1986; Curtain 1982, 1993; Kaman et al. 1985; Sakawa 1984; Schumacher 1983; Xiao and Başar 1999). These projection approximations usually lead to high dimensional observers in order to achieve accurate estimation. For extensions to systems with unbounded input and output operators see Curtain (1984) and Curtain and Salamon (1986). Some recent approaches can be found in Smyshlyaev and Krstic (2008, 2009) and Li et al. (2012) and references therein.

1.5.2 Nonlinear Case

Observer design for systems governed by nonlinear PDEs is very challenging. There are very few results available. Instead of trying to cover a broad class of issues, here we introduce a new idea that may directly connect to finite dimension observer design, presented in Sect. 1.2.

A mathematical description of the long-term behavior of a dynamical system ultimately is to determine its attractor. However, the global attractor can be quite complicated geometrically and can attracts solutions at algebraic rate. It has been found that in many cases the global attractor can be embedded into exponentially attractive finite dimensional manifolds (Chueshov 2002; Chow et al. 1992; Demengel and Ghidaglia 1991; Foias and Temam 1977; Garcia-archilla et al. 1999; Marion 1989; Temam 1997). It turns out that inertial manifolds are an appropriate mathematical tool which has been used in the study of the long-term behavior of dynamical systems. These are finite-dimensional Lipschitz manifolds, which attract all the orbits at an exponential rate. Inertial manifolds are positively invariant under the state dynamics and thus contain the global attractors.

If a system possesses an inertial manifold, the long-time dynamics of the system can be captured by the finite-dimensional dynamical flow on the manifold because the inertial manifold exponentially attracts all the orbits of the system. Hence, inertial manifolds can be used for the reduction of a PDE to a finite dimensional ODE in which the ω-limit set of the solution of the PDE coincides with the ω-limit set of a system of ODEs. In a way, the long-time dynamics of a PDE with an inertial manifold is *completely* determined by the solutions of a system of ODEs in finite dimensions, and one can use the well-established ODE theory for the qualitative analysis in an infinite dimensional setting. Many infinite-dimensional systems, including the well-known Navier-Stokes equations, actually possess inertial manifolds.

Consider an abstract evolution equation of the form

$$\frac{du}{dt} + Au = f(u), \qquad u(0) = u_0, \tag{1.31}$$

in a Banach space X, where f is assumed to be continuous from a Banach space E into another Banach space F, with

$$E \subset F \subset X;$$

The injections are continuous and each space is dense in the following one. Typically, $u(t)$ is a function, for instance $U(t, x)$ with a space variable x and a norm in a Banach space. For PDEs, A is a differential operator with respect to x, for instance $A(u)(t) = U_x(t, x)$. Under a Lipschitz condition for $f(u)$ and some assumptions about A and its spectrum, there exists a finite dimensional subspace $V \subset E$ that is invariant under e^{-tA}. The subspace is generated by a set of eigenvectors of A. Over this subspace is an invariant manifold of (1.31), denoted by $\mathcal{M}$, which can be defined as a graph

$$\mathcal{M} = \{(p, p + \Phi(p)) \mid p \in V\}$$

where Φ is a mapping from V to its orthogonal complement in E. It can be proved that (1.31) induces an ODE in V,

$$\frac{dp}{dt} + Ap = \bar{f}(p + \Phi(p)), \quad \bar{p}(0) = \bar{p}_0 \in V. \tag{1.32}$$

for some function $\bar{f}$ derived from the original PDE. The most important property of $\mathcal{M}$ is that it is *exponentially attractive*, that is, for any solution $u(t)$ of (1.31), there exists an induced trajectory

$$\bar{u}(t) = p(t) + \Phi(p(t)) \in \mathcal{M}$$

such that $u(t)$ approaches $\bar{u}(t)$ exponentially. Thus, the finite dimensional dynamics (1.32) determines the long-term behavior of the original system (1.31). An observer designed for (1.32) can be used to estimate $u(t)$.

For example, let us consider the system of reaction-diffusion equations

$$\frac{\partial u}{\partial t} = \nu \Delta u + f(u, \nabla u), \quad \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = 0, \tag{1.33}$$

in a bounded domain $\Omega \subset \mathbb{R}^d$. Here $u = (u_1, \ldots, u_m)$. The function $f(u, w)$ satisfies the global Lipschitz condition:

$$|f(u, w_1) - f(v, w_2)| \leq L\sqrt{|u - v|^2 + |w_1 - w_2|^2}, \tag{1.34}$$

where $u, v \in \mathbb{R}^m$, $w_1, w_2 \in \mathbb{R}^{md}$, and $L > 0$. It can be verified that the system satisfies all the conditions to guarantee the existence of an inertial manifold (Chueshov 2002). In fact, in this example we have $\Phi(p) = 0$. The manifold is the same as the invariant space V. The induced ODE in V has the following form.

$$\frac{d\bar{u}}{dt} = f(\bar{u}, 0), \qquad \bar{u}(t) \in V. \tag{1.35}$$

Thus for any PDE solution u to (1.33), there exists an ODE solution $\bar{u}$ to (1.35) such that

$$\|u(t) - \bar{u}(t)\|_1 \leq Ce^{-\gamma t}, \qquad t \geq 0,$$

where the constant $\gamma > 0$ does not depend on $u(t)$ and $\| \cdot \|_1$ is the Sobolev norm of the first order. Therefore, the observer design for the PDE system (1.33) boils down to the observer design for ODE (1.35), and the methods in previous sections are applicable.

1.6 Conclusions

The Kalman filter, invented initially for control systems, has been widely used in science and engineering including data assimilation. For the last several decades, the estimation theory for dynamical systems have been actively developed in control theory. We have surveyed some but not all of the ways of observers for a nonlinear system. Some approaches have been applied to engineering problems for many years, and some others are relatively new. It is not clear which of them is scalable for systems with extremely high dimensions, like atmospheric models or ocean dynamics. However, we certainly hope that some of these ideas will benefit the data assimilation community. The high gain observer is a theoretical finite dimensional solution to a broad class of systems with small noise. The minimum energy and H_∞ observers are theoretical infinite dimensional solutions to broad classes of noisy problems. However, it is not trivial to implement them for nonlinear systems. The linearization techniques give local and sometimes only approximate solutions for narrower classes of problems. The extended Kalman filter is probably still the most robust and practical approach for most problems. If there are substantial nonlinearities, e.g., multiple stable equilibria and/or stable limit cycles then the use of multiple extended Kalman filters is probably the preferred approach. However, a disadvantage for the extended Kalman Filter is the requirement of linearization in real-time. This is why the unscented Kalman filter is getting increasingly popular in engineering applications, although it suffers the requirement of doubling the dimension of the system. To summarize, there is no best estimation method for general nonlinear systems. Observers should be designed to fit the specific behavior and form of a system and its model.

All these methods rely on the observability of the system to insure convergence. The concept of observability has the potential to benefit data assimilations in several ways, including optimal sensor network design, data thinning, targeted sensing, etc. For these applications, numerically computing the observability for large systems is a challenge that needs further research.

References

Anderson JL (2003) A local least squares framework for ensemble filtering. Mon Weather Rev 131:634–642

Balas M (1980) Towards a (more) practical control theory for distributed parameter systems, control and dynamic systems. In: Leondes CT (ed) Advances in theory and applications, vol 18. Academic, New York

Bernstein DS, Hyland DC (1986) The optimal projection equations for finite-dimensional fixed-order dynamics compensation of infinite-dimensional systems. SIAM J Control Optim 24:122–151

Bestle D, Zeitz M (1983) Canonical form observer design for nonlinear time variable systems. Int J Control 38:419

Bhat KPM (1986) Regulator theory for evolution systems. Ph.D. thesis, University of Toroto

Brown RG, Hwang PYC (1997) Introduction to random signals and applied Kalman filtering, 3rd edn. Wiley, New York

Chow SN, Lu K, Sell GR (1992) Smoothness of inertial manifolds. J Math Anal Appl 169:283–321

Chua BS, Bennett AF (2001) An inverse ocean modeling system. Ocean Model 3:137–165

Chueshov ID (2002) Introduction to the theory of infinite-dimensional dissipative systems. ACTA Scientific Publishing House, Kharkiv

Clarke BMN, Williamson D (1981) Control canonical forms and eigenvalue assignment by feedback for a class of linear hyperbolic systems. SIAM J Control Optim 19:711–729

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 120:1367–1387

Curtain RF (1982) Stabilization of boundary control distributed systems via integral dynamics output feedback of a finite-dimensional compensator. In: Bensoussan A, Lions JL (eds) Analysis and optimization of systems. Lecture notes in control and information, vol 44. Springer, Berlin/New York, pp 761–776

Curtain RF (1984) Finite-dimensional compensators for parabolic distributed systems with unbounded control and observation. SIAM J Control Optim 22:255–276

Curtain RF (1993) A comparison of finite-dimensional controller designs for distributed parameter systems. Control-Theory Adv Technol 9:609–629

Curtain RF, Pritchard AJ (1974) The infinite dimensional Riccati equation. J Math Anal Appl 47:43–57

Curtain RF, Salamon D (1986) Finite dimensional compensators for infinite dimensional systems with unbounded input operators. SIAM J Control Optim 24:797–816

Curtain RF, Zwart HJ (1995) An introduction to infinite-dimensional linear systems theory. Spring, New York

Demengel F, Ghidaglia J-M (1991) Inertial manifolds for partial differential equations under time-discretization: existence, convergence, and applications. J Math Anal Appl 155:177–225

Desch W, Schappacher W (1985) Spectral properties of finite-dimensional perturbed linear semigroup. J Differ Equ 59:80–102

Ding X, Frank P, Guo L (1990) Nonlinear observer design via an extended observer canonical form. Syst Control Lett 15:313

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 (2007) Data assimilation: the ensemble Kalman filter. Springer, Berlin

Findeisen R, Diehl M, Burner T, Allgower F, Bock HG, Schloder JP (2002) Efficient output feedback nonlinear model predictive control. In: Proceedings of the American control conference Anchorage, AK, pp. 4752–4757

Floquet T, Barbot JP (2007) Super twisting algorithm based step-by-step sliding mode observers for nonlinear systems with unknown inputs. Special Issue of IJSS on Advances in Sliding Mode Observation and Estimation 38(10):803–815

Foias C, Temam R (1977) Structure of the set of stationary solutions of the Navier-Stokes equations. Commun Pure Appl Math 30:149–164

Fuji N (1980) Feedback stabilization of distributed parameter systems by a functional observed. SIAM J Control Optim 18:108–121

Garcia-archilla B, Novo J, Titi ES (1999) An approximate inertial manifolds approach to postrocessing the Galerkin method for the Navier-Stoke equations. Math Comput 68:893–911

Gauthier JP, Kupka IAK (1994) Observability and observers for nonlinear systems. SIAM J Control Optim 32:975–994

Gauthier JP, Hammouri H, Othman S (1992) A simple observer for nonlinear systems with applications to bioreactors. IEEE Trans Autom Control 37:875–880

Gelb A (1974) Applied optimal estimation. MIT, Cambridge

Green M, Limebeer DJN (1995) Linear robust control. Prentice Hall, Englewood Cliffs

Gressang R, Lamont G (1975) Observers for systems characterized by semigroups. IEEE Trans Autom Control AC-20:523–528

Hermann R, Krener A (1977) Nonlinear controllability and observability. IEEE Trans Autom Control 22:728–740

Hijab O (1980) Minimum energy estimation. Ph.D thesis, University of California, Berkeley CA

Houtekamer P, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. Mon Weather Rev 126:796–811

Isidori A (1995) Nonlinear control systems. Springer, London

Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc IEEE 92(3):401–422

Kailath T (1980) Linear systems. Prentice-Hall, Englewood Cliffs

Kaman EW, Khargonekar PP, Tannenbaum A (1985) Stabilization of time-delay systems using finite-dimensional compensators. IEEE Trans Autom Control AC-30:75–78

Kang W (2011) The Consistency of Partial Observability for PDEs, arXiv:1111.5846v1, November

Kang W (2006) Moving horizon numerical observers of nonlinear control systems. IEEE Trans Autom Control 51(2):344–350

Kang W, Barbot J-P (2007) Discussions on observability and invertibility. In: NOLCOS 2007, Pretoria, South Africa

Kang W, Xu L (2011) Observability and optimal sensor placement. Int J Sens Wirel Commun Control (to appear) 1(2):93–101

Kang W, Xu L (2009a) A quantitative measure of observability and controllability. In: Proceedings of the IEEE conference on decision and control, Shanghai, China

Kang W, Xu L (2009b) Computational analysis of control systems using dynamic optimization. arXiv:0906.0215v2

Kazantzis M, Kravaris C (1998) Nonlinear observer design using Lyapunov's auxiliary theorem. Syst Control Lett 34:241–147

Khalil HK (2002) Nonlinear systems, 3rd edn. Prentice Hall, Upper Saddle River

Kitamura S, Sakairi H, Mishimura M (1972) Observers for distributed parameter systems. Electr Eng Jpn 92:142–149

Krener AJ (2003a) The convergence of the minimum energy estimator. In: Kang W, Xiao M, Borges C (eds) New trends in nonlinear dynamics and control, and their applications. Springer, Heidelberg, pp 187–208

Krener AJ (2003b) The convergence of the extended Kalman filter. In: Rantzer A, Byrnes CI (eds) Directions in mathematics systems theory and optimization. Springer, Berlin

Krener AJ (2004) Nonlinear observers. In: Unbehauen H (ed) Control systems, robotics and automation. Encyclopedia of life support systems (EOLSS), Developed under the auspices of the UNESCO. Eolss Publishers, Oxford

Krener AJ, Ide K (2009) Measures of unobservability. In: Proceedings of the IEEE conference on decision and control, Shanghai, China

Krener AJ, Isidori A (1983) Linearization by output injection and nonlinear observers. Syst Control Lett 3:47–52

Krener AJ, Kang W (2003) Locally convergent nonlinear observers. SIAM J Control Optim 42(1):155–177

Lei H, Wei JF, Lin W A global observer for autonomous systems with bounded trajectories. Int J Robust Nonlinear Control 17:1088–1105 (2007).

L'Hernault M, Barbot J-P, Ouslimani A (2008) Feasibility of analogue realization of a sliding mode observer: application to data transmission. IEEE Trans Circuit Syst – I 55(2) pp 614–624

Li L, Huang Y, Xiao M (2012) Observer design for wave equations with van der Pol type boundary conditions. SIAM J Control Optim (accepted to appear). Proceedings of the 10th world congress on intelligent control and automation, Beijing, July 2012, pp 1471–1476

Marion M (1989) Approximate inertial manifolds for reaction diffusion equations in high space dimension. J Dyn Differ Equ 1:245–267

Michalska H, Mayne DQ (1995) Moving horizon observers and observer-based control. IEEE Trans Autom Control 40(6):995–1006

Mortensen R (1968) Maximum-likelihood recursive nonlinear filtering. J Optim Theory Appl 2:386–394

Orner PA, Foster AM (1971) A design procedure for a class of distributed parameter control system. Trans ASME Ser G J Dyn Syst Meas Control 93:86–93

Pritchard AJ, Zabczyk J (1981) Stability and stabilizability of infinite dimensional systems. SIAM Rev 23:25–52

Rabier F, Jarvinen H, Klinker E, Mahfouf J-F, Simmons A (2000) The ECMWF operational implementation of four dimensional variational assimilation. Part I: experimental results with simplified physics. Q J R Meteorol Soc 126:1143–1170

Rebarber R (1999) Spectral assignability for distributed parameter systems with unbounded scalar control. SIAM J Control Optim 27:148–169

Russell DL (1968) Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems. J Math Anal Appl 62:182–225

Russell DL (1978) Controllability and stabilizability theory for linear partial differential equations: recent progress and open problems. SIAM Rev 20:639–739

Sakawa Y (1984) Feedback control of second order evolution equations with damping. SIAM J Control Optim 22:343–361

Sakawa Y, Matsushita T (1975) Feedback stabilization for a class of distributed systems and construction of a state estimator. IEEE Trans Autom Control AC-20:748–753

Schumacher JM (1983) A direct approach to compensator design for distributed parameter systems. SIAM J Control Optim 21:823–836

Smyshlyaev A, Krstic M (2008) Boundary control of PDEs: a course on backstepping designs. SIAM, Philadelphia

Smyshlyaev A, Krstic M (2009) Boundary control of an anti-stable wave equation with anti-damping on the uncontrolled boundary. Syst Control Lett 58:617–623

Sorenson HW (1985) Kalman filtering: theory and applications, IEEE, New York

Spurgeon SK (2008) Sliding mode observers: a survey. Int J Syst Sci 39(8):751–764

Sun SH (1981) On spectrum distribution of complete controllable systems. SIAM J Control Optim 19:730–743

Temam R (1997) Infinite-dimensional dynamical systems in mechanics and physics. Springer, New York

Triggiani R (1975) On the stabilization problem in Banach space. J Math Anal Appl 52:383–403

Tsinias J (1989) Observer design for nonlinear systems. Syst Control Lett 13:135

Xia X, Gao W (1989) Nonlinear observer design by observer error linearization. SIAM J Control Optim 27:199–216

Xiao M, Başar T (1999) Finite-dimensional compensators of H-infinity-optimal control for infinite-dimensional systems via Galerkin-type approximation. SIAM J Control Optim 37(5):1614–1647

Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR. Formulation and initial tests of the linear problem. Tellus 57A:546–559

Zeitz M (1987) The extended Luenberger observer for nonlinear systems. Syst Control Lett 9:149

Zheng G, Boutat D, Barbot J-P (2007) Single output-dependent observability normal form. SIAM J Control Optim 46(6):2242–2255

Chapter 2
Nudging Methods: A Critical Overview

S. Lakshmivarahan and John M. Lewis

Abstract A review of the various methods used to implement the "nudging" form of data assimilation has been presented with the intension of identifying both the pragmatic and theoretical aspects of the methodology. Its appeal rests on the intuitive belief that forecast corrections can be made on the basis of feedback control where forecast error from earlier times is incorporated into the dynamics. Further, the methodology is easy to implement. However, its early-period implementation with a nudging coefficient based on pure empiricism with slight consideration of the time scales of motion lacked a firm theoretical foundation. This empirical approach is reviewed but then placed in the context of advances that have attempted to optimally choose the nudging coefficient based on a functional that fits model to data as well as fitting the coefficient to an *a priori* estimate of the coefficient. Original research in this review makes it clear that these "optimal" methods have unintentionally neglected the inherent presence of serially correlated error in the nudged model. And in the absence of account for this error, the results are non-optimal. Finally, the theories of observer-based nudging and forward-backward nudging are presented as promising avenues of research for the nudging process of dynamic data assimilation.

S. Lakshmivarahan (✉)
School of Computer Science, University of Oklahoma, Norman, OK 73019, USA
e-mail: varahan@ou.edu

J.M. Lewis
National Severe Storms Laboratory, Norman, OK, USA

Desert Research Institute, Reno, NV, USA

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_2,

2.1 Introduction

To avoid confusion and repetition, we begin by establishing basic notation and definitions. Let $x \in R^n$ refer to the state vector of the forecast model, and $M : R^n \to R^n$ denote the one-step transition map. A discrete time nonlinear dynamic model is given by

$$x(k+1) = M(x(k)) \tag{2.1}$$

with $x(0)$ the initial condition. Given $x(0)$, the sequence of states $\{x(k)\}_{k \geq 0}$ is called the model forecast.

Let $z \in R^m$ and $h : R^n \to R^m$ where

$$z(k) = h(\bar{x}(k)) + V(k) \tag{2.2}$$

denote the observation at time k, where $\bar{x}(k)$ is the "true" unknown state of the system captured by the model in (2.1), $V(k)$ is the Gaussian white noise sequence $V(k) \sim N(0, \mathrm{R})$ where $\mathrm{R} \in R^{m \times m}$ is a known symmetric and positive definite covariance matrix of $V(k)$. It is assumed that the unknown true state evolves according to the dynamics

$$\bar{x}(k+1) = \bar{M}(\bar{x}(k)) \tag{2.3}$$

with $\bar{x}(0)$ as the initial condition. The differences

$$\tilde{M}(x) = M(x) - \bar{M}(x) \tag{2.4}$$

and

$$\tilde{x}(0) = x(0) - \bar{x}(0)$$

denote the model error and the error in the initial condition, respectively.

A modern version of the standard dynamic data assimilation problem (Lewis et al. (2006)) may be stated as follows: Given a set $\{z(k) : 1 \leq k \leq N\}$ of N observations, find the optimal initial condition $x^*(0)$ that minimizes the cost functional

$$J_1(x(0) = \frac{1}{2}\sum_{k=1}^{N} < e(k), \mathrm{R}^{-1}e(k) > \tag{2.5}$$

where

$$e(k) = z(k) - h(x(k)) \tag{2.6}$$

is the forecast error and $< a, b >= a^T b$ is the standard inner product of two vectors $a, b \in R^n$ where T denotes the transpose. The importance of this problem stems from the fact that the model forecast starting from $x^*(0)$ "best fits" the observation that in turn is a surrogate of the truth.

In the parlance of dynamic meteorology, the optimal forecast problem has a rich and cherished history. Wilhelm Bjerknes (1904), father of modern-day dynamic meteorology, was the first scientist to formulate forecasting as an initial value problem. And in the early 1920s, British meteorologist Lewis Fry ("L. F.") Richardson (1922) paved the foundation for modern numerical weather prediction (NWP) with his bold effort to use discrete mathematics to make a single-step advance of the atmospheric state (Lynch 2006). Although unsuccessful for reasons discussed in Lynch (2006), Richardson's work inspired a team of meteorologists and mathematicians at Princeton University's Institute for Advanced Study. And under the leadership of Jule Charney, this team made two successful 24-h NWP forecasts of the transient features of the large-scale flow (initialized on 30 January and 13 February 1949) using a filtered model that excluded the fast moving gravity-inertial waves while retaining the slower Rossby waves (Charney et al. 1950; Platzman 1979). The calculations were made on the ENIAC (Electronic Numerical Integrator And Computer), the very first generation of stored program digital computers, housed at the Aberdeen Proving Grounds, Maryland between 1947 and 1955.

With the advent of more-powerful digital computers along with advances in numerical analysis techniques, NWP and the associated numerical simulation of atmospheric flow have become dominant themes of research in meteorology. Indeed, Bjerknes's dream has been realized albeit tempered by the uncertainty of extended range forecasting in response to the chaotic nature of atmospheric flow (see review in Lewis 2005). The melding of observations and dynamics into the construction of an initial state of the system, the data assimilation (DA) phase of NWP, has been central to advances in NWP (See Lewis and Lakshmivarahan (2008) for a comprehensive historical review of meteorological data assimilation from the mid-1950s to the present day).

From roughly the early-1970s to the present day, variational calculus and optimization theory have assumed central roles in the solution to the dynamic data assimilation problem. The well-known 4D-Var (four-dimensional variational method), based on use of the adjoint model to determine the gradient of the cost function, has both esthetic appeal and pragmatic utility for assimilating data into a deterministic model (strong constraint where the model is assumed perfect). For a stochastic approach where the model is assumed imperfect, a "Kalman filter" type approach, referred to in meteorology as optimal or statistical interpolation or more recently referred to as 3D-Var, Lorenc (1986) has enjoyed wide appeal (Again, see Lewis and Lakshmivarahan (2008) for the historical development of these ideas). The books by Daley (1991), Kalnay (2003), Evensen (2007), and Lewis et al. (2006), offer pedagogical explanations and discussions of these various methods. We hasten to add that both the 4D-Var and Kalman-filter methods (Kalman 1960b) have enjoyed widespread development and use in the control theory literature.

Anthes (1974) and Hoke (1976) introduced a method of data assimilation that differed from the classic methods mentioned above. Fundamentally, this methodology has its roots in control theory where an empirical forecast error term is added to the dynamical constraint—essentially a feedback control Wiener (1948). More formally, the forecast error $e(k)$ in (2.6) is used as an artificial forcing to the model as follows:

$$x(k+1) = M\,(x(k)) + G^0(k)e(k) \tag{2.7}$$

where $G^0(k) \in R^{n\times m}$ is called the time varying nudging coefficient matrix. Since the correction term in (2.7) is proportional to $e(k) \in R^m$ (in the observation space), this form of nudging is called *observation nudging* where $G^0(k)$ is the associated nudging coefficient. Instead, let $x_a(k) \in R^n$ be the state vector on the computational grid obtained from $z(k)$ using any one of the DA schemes. Then,

$$x(k+1) = M(x(k)) + G^a(k)\,[x_a(k) - x(k)] \tag{2.8}$$

is called *analysis nudging* where $G^a(k) \in R^{n\times n}$ is a time varying analysis nudging coefficient. In either form, an appropriate measure of the forecast error is used to force the model state towards the observation. The nudging method has also been viewed as a case of Newtonian relaxation or "repeated insertion of data" (Macpherson (1991)).

The notion of using the error to drive a model towards a desired state is a basic principle underlying the design of feedback control systems. Refer to Bennett (1996), Bryson (1996), and Sussmann and Willems (1997) for historical overviews of these techniques.

The literature on nudging covers nearly four decades (since 1974) and can be broadly divided into parts or divisions as follows:

1. The nudging coefficient is empirically determined through examination of dynamical simulation over a broad range of coefficients. The coefficient G is positive and may be time varying, but its magnitude is controlled in part by the smallest time scale of the typical multi-scale phenomenon captured by the model.
2. The coefficient matrix G is optimally determined through minimization of a functional that combines the standard fit (equation 2.5) augmented by a term that fits the coefficient to an *a priori* estimate of that coefficient. The resulting constrained minimization is solved by the 4D-Var method mentioned earlier.
3. A class of methods that exploit the similarity between nudged dynamics (2.7–2.8) and feedback control in observer theory (Luenberger 1964).
4. A process labeled "back-and-forth nudging" that uses the same model in a forward and backward mode to obtain a good match between the forecast model and the observations (Auroux (2009)).

Nudging based dynamic data assimilation has been applied to a variety of problems including the following:

1. Initialization of a dynamic model as originally proposed by Anthes (1974) and Hoke (1976) where Hoke (1976) recommended an analysis-based nudging process [as found in (2.8)] as opposed to observation-based nudging [as found in (2.7)]. Application has been made to forecast of the Indian Monsoon [Krishnamurti et al. (1991), Ramamurthy and Carr (1987, 1988)].

2. Diagnostic studies of synoptic-scale and mesoscale processes in mid-latitude weather systems [Brill et al. (1991), Stauffer et al. (1985), Stauffer and Seaman (1990), Yamada and Bunker (1989), and Warner (1990)].
3. Observation System Simulation Experiments (OSSE's) using observations from wind profilers as found in the work of Kuo and Guo (1989).
4. Application of the nudging assimilation method to operational prediction has been made in both meteorology and oceanography. In meteorology, the following publications have explored nudging: Bell and Dickinson (1987) and Lorenc et al. (1991). In oceanography, Derber and Rosati (1989) and Derber et al. (1990) have used nudging.

Beyond this divisional breakdown of nudging processes in research and operations, the following studies are noteworthy: A non-operational application of nudging to analyses from FGGE (First GARP Global Experiment) as found in Stern et al. (1985), a sensitivity of assimilation and prediction to the nudging coefficient by Bao and Errico (1997), and a series of explorations into the "back-and-forth" nudging method by Auroux and collaborators [Auroux (2009), Auroux and Nodet (2010), and Auroux and Blum (2005) and (2008)].

We begin our review by providing a historical examination of the empirical methods used in nudging. This is followed by a study of the work that searched for optimal nudging coefficients including an account for serial correlation errors in the nudging process. We then examine the observer-based methods and explore the ideas behind the back-and-forth nudging process. We summarize and discuss the work on nudging in the final section of the paper.

2.2 Early Empirical Method

For completeness and to give a flavor of the ideas used in the early era, in this section we describe a method for determining the scalar nudging coefficient G. Following Brill et al. (1991) consider the analysis nudging scheme in continuous time. Let

$$\dot{x}(t) = F\left(x(t)\right) + Gf(t)\left[x_a(t) - x(t)\right] \tag{2.9}$$

where $F : R^n \to R^n$, $G \in R$ is an unknown positive scalar to be estimated, $f : [0, T] \to R$is a non-negative real valued function such that

$$0 \leq f(t) \leq 1 \;\; f(0) = 0 = f(T) \tag{2.10}$$

and $x_a(0)$ and $x_a(T)$ are the known analyses at times $t = 0$ and $t = T$ obtained from the available observations at these times. Brill et al. (1991) postulate that $x_a(t)$ in (2.9) varies linearly and is given by

$$x_a(t) = x_a(0) + \frac{t}{T}\left[x_a(T) - x_a(0)\right] \tag{2.11}$$

The dynamics is then integrated from the initial condition $x(0) = x_a(0)$.

Integrating (2.9), we get

$$x(T) - x(0) = \int\limits_0^T F\,(x(t))\,dt + G\int\limits_0^T f(t)\,[x_a(t) - x(t)]\,dt \tag{2.12}$$

Brill et al. (1991) further postulate that the first integral which is the contribution of the model accounts for the fraction $(1-\alpha)\,[x(T) - x(0)]$ of the total change in the solution $x(t)$ from time 0 to T as given by the left-hand side of (2.12), where $0 < \alpha < 1$. Consequently, the second integral accounts for the remainder of the change leading to the following relation:

$$a\,[x(T) - x(0)] = G\int_0^T f(t)\,[x_a(t) - x(t)]\,dt \tag{2.13}$$

To further simplify the evaluation of the integral on the right-hand side of (2.13), Brill et al. (1991) make one more assumption; namely, the nudged solution $x(t)$ varies linearly from $x(0) = x_a(0)$ to $x(T)$. That is,

$$x(t) = x(0) + \frac{t}{T}\,[x(T) - x(0)] \tag{2.14}$$

Now subtracting (2.14) from (2.11),

$$[x_a(t) - x(t)] = \frac{t}{T}\,[x_a(T) - x_a(0)] - \frac{t}{T}\,[x(T) - x(0)] \tag{2.15}$$

Substituting (2.15) in (2.13) and simplifying, we readily obtain

$$G = \frac{\alpha\beta}{(1-\beta)\frac{1}{T}\int_0^T tf(t)dt} \tag{2.16}$$

where

$$\beta = \frac{x(T) - x(0)}{x_a(T) - x_a(0)} = \frac{x(T) - x(0)}{x_a(T) - x(0)} \tag{2.17}$$

is a fraction of the change in the nudged forecast to that of the analysis. For the case when

$$f(t) = -6.75\left(\frac{t}{T}\right)^3 + 6.75\left(\frac{t}{T}\right)^2 \tag{2.18}$$

Brill et al. (1991) in their Appendix provide the values of G that range from $4.1\cdot 10^{-4}$ to $2.6\cdot 10^{-3}$. They also examine the contour plots of

$$\frac{G}{T}\int_0^T tf(t)dt = \frac{\alpha\beta}{1-\beta} \tag{2.19}$$

in the $\alpha - \beta$ plane.

In summary, it can be seen that this heuristic analysis rests firmly on two assumptions—namely, that the large fraction of the change in the solution is due to the model and that the evolution of the solution and the analysis from $t = 0$ to $t = T$ can be approximated linearly. A necessary condition for this latter assumption to hold is that the time horizon $[0, T]$ must be small. Brill et al. (1991) take $t = 3h$ in their analysis.

By discretizing (2.9) using an Euler scheme, we get

$$x(N) - x(0) = \left(\sum_{k=0}^{N-1} F\left(x(k)\right) \right) \Delta t + G \sum_{k=0}^{N-1} f(k) \left[x_a(k) - x(k) \right] \Delta t \qquad (2.20)$$

which is a direct analog of (2.12) in discrete form. By following the above arguments, we leave it to the reader to arrive at an expression for G similar to (2.16).

2.3 Estimating Optimal Nudging Coefficient: Problems and Challenges

There are two basic approaches to the problem of estimating the optimal value of the nudging matrix G. The first approach is due to Stauffer and Seaman (1990), Stauffer and Bao (1993) and Zou et al. (1992). Using the classic four-dimensional variational (4D-Var) data assimilation method, they independently found the optimal G. The second approach is due to Vidard et al. (2003) where a combination of Kalman filter and 4D-Var is used to estimate the optimal G. In this section we provide a summary of these two approaches. As will be seen, these approaches are incomplete in the sense that they do not account for the inherent serial correlation of forecast errors that constitute the basis for the estimation algorithm. A direct impact of excluding the underlying correlation introduces a bias into the problem that directly affects the value of the so-called optimal estimate.

For definiteness in the following development, we use the observation-based nudging scheme that easily extends to the analysis-based nudging scheme.

2.3.1 Estimation of G *Using the Variational Approach*

Let the observation and the nudged dynamics be given by (2.2) and (2.7), respectively. Let the forecast error $e(k) \in R^m$ be given by (2.6). Define a vector

$$e(1:N) = \left(e^T(1), e^T(2), \ldots, e^T(N)\right)^T \in R^{Nm} \qquad (2.21)$$

consisting of the forecast errors at times $1 \le k \le N$.

Define a cost function

$$J_2(G) = \frac{1}{2} < e(1:N), R^{-1}(N)e(1:N) > \tag{2.22}$$

which is an analog of the cost function $J_1(x(0))$ in (2.5), where

$$R(N) = I \otimes R \in R^{Nm \times Nm} \tag{2.23}$$

where $A \otimes B = [a_{ij} B]$ where a_{ij} is the ij$^{\text{th}}$ element of the matrix A, is called the Kroneker product of A and B. Clearly, $R(N)$ is a block diagonal matrix whose diagonal blocks are R and the off-diagonal blocks are zero matrices. Also, define

$$J_p(G) = \frac{\beta}{2} \left|\left| G - \hat{G} \right|\right|_F^2 \tag{2.24}$$

where $\hat{G}$ is a prior estimate of G, $\beta > 0$ is a penalty parameter (the larger its value the closer the estimate of G is to $\hat{G}$) and $\left|\left| A \right|\right|_F = \left[\sum_{i,j=1}^{n} a_{ij}^2 \right]^{\frac{1}{2}}$ is called the Frobenius matrix norm (which is an extension of the Euclidean norm for the matrix A).

Zou et al. (1992) and Stauffer and Seaman (1990) seek to minimize

$$Q_1(G) = J_2(G) + J_p(G) \tag{2.25}$$

using the nudged dynamics in (2.7) as a strong constraint.

This equality constrained minimization problem can be solved in one of two ways: using a Lagrangian formulation (Thacker and Long (1988)) or using the first-order variational formulation (Lewis et al. (2006)).

In either approach, the gradient $\nabla_G Q_1(G) \in R^{n \times m}$ is computed which is then used in a minimization algorithm to obtain a G that minimizes $Q_1(G)$.

There are two difficulties associated with the above formulation. First is the question related to the choice of the prior value $\hat{G}$ of the unknown nudging coefficient. The second and more serious problem is the inherent need to account for the temporally correlation of the forecast errors $e(1), e(2), \ldots, e(N)$. Exclusion of this correlation introduces a bias in the optimal estimate G^* of G (Lakshmivarahan and Lewis (2011)).

In the following we provide a pathway to quantify this inherent temporal correlation. To this end, first rewrite (2.7) as

$$x(k+1) = f(x(k), G) + Gz(k) \tag{2.26}$$

or as

$$x(k+1) = F(x_k, \bar{x}_k, G) + GV(k) \tag{2.27}$$

where

$$f(x, G) = M(x) - Gh(x) \tag{2.28}$$

and

$$F(x, \bar{x}, G) = f(x, G) + Gh(\bar{x}) \tag{2.29}$$

which is separable in x and $\bar{x}$.

Since $V(k)$ is a vector white noise Gaussian process, it readily follows from (2.27) that $\{x(k)\}_{k\geq 0}$ is a first-order nonlinear auto-regressive process of order 1 (Hamilton (1994)). Thus, given $M(x), M(\bar{x}), h(x), x(0)$ and $\bar{x}(0)$, $x(k)$ is a function of G and the complete history noise sequence $V(1), V(2), \ldots, V(k)$ for $k \geq 1$ Assuming $h(x) = x$, the error in (2.6) namely

$$e(k) = z(k) - x(k) \tag{2.30}$$

depends on G and the noise vector

$$V(1:k) = \left(V^T(1), V^T(2), \ldots, V^T(k)\right)^T \in R^{km} \tag{2.31}$$

Consequently, there exists a covariance matrix $V \in R^{Nm \times Nm}$ such that

$$V_{ij} = \text{cov}(e(j), e(j)) \in R^{m \times m} \tag{2.32}$$

for all $1 \leq i, j \leq N$.

Now define

$$J_3(G) = \frac{1}{2} < e(1:N), V^{-1}e(1:N) > \tag{2.33}$$

which is a modified version of $J_2(G)$ in (2.22). Accordingly, the correct formulation is as follows: Find G that minimizes

$$Q_2(J) = J_3(G) + J_p(G) \tag{2.34}$$

instead of $Q_1(G)$ in (2.25).

We hasten to add that while (2.34) is the correct formulation of the optimal nudging problem, it is very difficult to compute the elements of the covariance matrix V for the case when the state transition map M in (2.7) is nonlinear. However, when the dynamics is linear, we can give an explicit expression for the elements of V that captures the underlying correlation structure of the forecast errors.

Example 2.1. **Linear Dynamics and Observations**

Consider the special case when $M(x) = Mx, \overline{M}(x) = \overline{M}x, h(x) = Hx$ where $M \in R^{n \times n}$, $\overline{M} \in R^{n \times n}$, and $H \in R^{m \times n}$. Then the observation equation (2.6) becomes

$$z(k) = H\bar{x}(k) + V(k) \tag{2.35}$$

and the nudged dynamics (2.7) takes the form

$$x(k+1) = Mx(k) + G\,[z(k) - Hx(k)] \tag{2.36}$$

We can rewrite (2.36) as

$$\begin{aligned} x(k+1) &= Ax(k) + Gz(k) \\ A &= (M - GH) \end{aligned} \tag{2.37}$$

Iterating (2.37), it follows that the solution is given by

$$\begin{aligned} x(k) = A^k x(0) &+ \sum_{j=0}^{k-1} A^j GH\bar{M}^{k-1-j}\bar{x}(0) \\ &- \sum_{j=0}^{k-1} A^j GV(k-1-j) \end{aligned} \tag{2.38}$$

where we have used the fact that the true state $\bar{x}(k)$ dynamics is given by

$$\bar{x}(k+1) = \overline{M}\bar{x}(k) \tag{2.39}$$

with $\bar{x}(0)$ as the initial condition and

$$\bar{x}(k) = \overline{M}^k \bar{x}(0) \tag{2.40}$$

Hence substituting (2.35) and (2.39) in

$$e(k) = z(k) - x(k)$$

and simplifying we readily obtain a decomposition into deterministic and stochastic parts as follows:

$$e(k) = \left[F(k)\bar{x}(0) - A^k x(0)\right] + \eta(k) \tag{2.41}$$

where

$$F(k) = \left[H\overline{M}^k - \sum_{j=0}^{k-1} A^j GH\overline{M}^{k-1-j} \right] \tag{2.42}$$

and

$$\begin{aligned} \eta(k) &= V(k) - \sum_{j=0}^{k-1} A^j GV(k-1-j) \\ &= V(k) - \left[GV(k-1) + AGV(k-2) + A^2GV(k-3) + \ldots A^{k-2}GV(1)\right] \end{aligned} \tag{2.43}$$

From the properties of $V(k)$ it follows that

$$E\left[\eta(k)\right] = 0 \text{ and } E\left[e(k)\right] = \left[F(k)\bar{x}(0) - A^k x(0)\right] \tag{2.44}$$

The covariance matrix $V = [V_{ij}]$ is now given by

$$V_{ii} = E\left[\eta(i)\eta^T(i)\right] \tag{2.45}$$

and

$$V_{jj} = E\left[\eta(i)\eta^T(j)\right] \text{ for } i \neq j$$

As an example, consider the case when $N = 4$. Then from (2.43),

$$\begin{aligned}
\eta(1) &= V(1) \\
\eta(2) &= V(2) - GV(1) \\
\eta(3) &= V(3) - [GV(2) + AGV(1)] \\
\eta(4) &= V(4) - \left[GV(3) + AGV(2) + A^2GV(1)\right]
\end{aligned}$$

Hence,

$$\begin{aligned}
V_{11} &= E\left[\eta(1)\eta^T(1)\right] = E\left[V(1)V^T(1)\right] = \mathrm{R} \\
V_{22} &= E\left[\eta(2)\eta^T(2)\right] = \mathrm{R} + G\mathrm{R}G^{\mathrm{T}} \\
V_{33} &= E\left[\eta(3)\eta^T(3)\right] = \mathrm{R} + G\mathrm{R}G^{\mathrm{T}} + AG\mathrm{R}G^{\mathrm{T}}A^{\mathrm{T}} \\
V_{12} &= E\left[\eta(1)\eta^T(2)\right] = -\mathrm{R}G^{\mathrm{T}} = V_{21} \\
V_{13} &= E\left[\eta(1)\eta^T(3)\right] = -\mathrm{R}G^{\mathrm{T}}A^T = V_{31} \\
V_{14} &= E\left[\eta(1)\eta^T(4)\right] = -\mathrm{R}G^{\mathrm{T}}(A^2)^T = V_{41} \\
V_{23} &= E\left[\eta(2)\eta^T(3)\right] = -\mathrm{R}G^{\mathrm{T}} + GRG^{\mathrm{T}}A^{\mathrm{T}} = V_{32} \\
V_{24} &= E\left[\eta(2)\eta^T(4)\right] = -\mathrm{R}G^{\mathrm{T}}A^T + \mathrm{R}G^{\mathrm{T}}(A^2)^T = V_{42} \\
V_{34} &= E\left[\eta(3)\eta^T(4)\right] = -\mathrm{R}G^{\mathrm{T}} + G\mathrm{R}G^{\mathrm{T}}A + AG\mathrm{R}G^{\mathrm{T}}(A^2)^T = V_{43}
\end{aligned}$$

Thus, the elements of V are polynomial matrices in G, A, and R.

Hence the shape of $J_3(G)$ in (2.33) in general depends on the (unknown) model error, $x(0)$, $\bar{x}(0)$, R, random realizations of the observational errors, and G. Clearly the problem of determining the optimal G is much more involved than implied in the literature. To simplify matters, we generally assume that the model is perfect, that is, $M = \overline{M}$.

The deterministic part $F(k)$ in (2.42) of the error $e(k)$ in (2.41) takes a much simpler form when the model is perfect; that is, $M = \overline{M}$ and $H = I$ in which case $z(k) = x(k)$. Substituting $M = \overline{M}$ and $H = I$ in (2.34), we obtain

$$F(k) = M^k - \sum_{j=0}^{k-1} (M-G)^j GHM^{k-1-j} \tag{2.46}$$

It can be verified $F(1) = (M-G) = A$, $F(2) = (M-G)^2 = A^2$. It is a simple exercise to prove by induction that

$$F(k) = (M-G)^k = A^k \tag{2.47}$$

Substituting (2.47) in (2.41) and simplifying

$$e(k) = A^k\,(\bar{x}(0) - x(0)) + \eta(k) \tag{2.48}$$

Thus, in this case the deterministic part of the error has a simple from and is essentially controlled by the error in the initial condition.

We now illustrate the impact of model error, error in the initial condition and the variance of the observation noise on the optimal estimate of the nudging coefficient for a simple scalar, linear, discrete time model.

Example 2.2. **Numerical Experiment**

Consider a scalar linear nudged model given by

$$x(k+1) = mx(k) + g\,(z(k) - x(k)) \tag{2.49}$$

starting from the initial condition $x(0)$ and $g \in R$ is a nudging parameter. The observation

$$z(k) = \bar{x}(k) + V(k) \tag{2.50}$$

That is, $h(x) = x$, where $V(k)$ is a white Gaussian noise, namely $V(k) \sim N(0,\ \sigma^2)$, and $x(k)$ is the state of the dynamics given by

$$\bar{x}(k+1) = \overline{m} \cdot \bar{x}(k) \tag{2.51}$$

with $\bar{x}(0)$ as the initial condition.

Let $m = \overline{m} + \delta$ where δ denotes the model error and $(x(0) - \bar{x}(0))$ is the error in the initial condition. Let

$$e(k) = z(k) - x(k) \tag{2.52}$$

Consider the scalar analogs of (2.22) and (2.33) given by

$$J_2(g) = \frac{1}{2\sigma^2} e^T(1:4)e(1:4) \tag{2.53}$$

and

$$J_3(g) = \frac{1}{2} e^T(1:4)V^{-1}e(1:4) \tag{2.54}$$

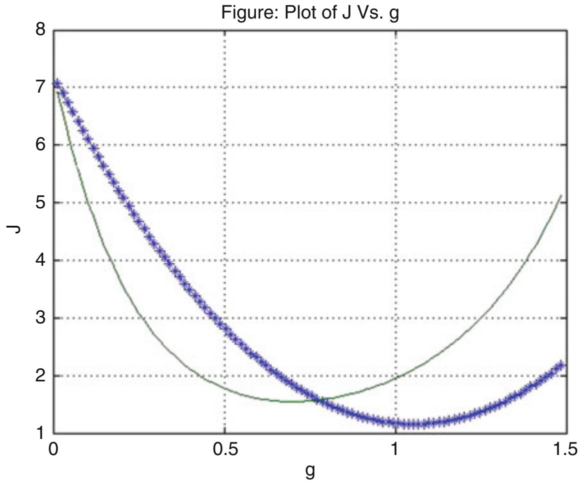


Fig. 2.1 Illustration of the cost functions from Example 2.2 with $\bar{m} = m = 1.1 (\delta = 0)$, $\bar{x}_0 = 1.1$, $h = 1$, $\sigma^2 = 0.01$, and $x_0 = 0.9$. $J_3(g)$ and $J_2(g)$ are represented by "xxx" and "——", respectively

where $e(1:4) = (e(1), e(2), e(3), e(4))^T \in R^4$ and the analog of the covariance matrix V in (2.45) is given by

$$V = \sigma^2 \begin{bmatrix} 1 & -g & -ag & -a^2g \\ -g & 1+g^2 & -g+ag^2 & -ag+a^2g^2 \\ -ag & -g+ag^2 & 1+g^2+a^2g^2 & -g+g^2a+g^2a^3 \\ -a^2g & -ag+a^2g^2 & -g+g^2a+g^2a^3 & 1+g^2+a^2g^2+a^4g^2 \end{bmatrix}$$

where $a = (m - g)$.

A comparison of the plots of $J_2(g)$ and $J_3(g)$ in (2.53) and (2.54) for the case when $\bar{m} = m = 1.1$ $(\delta = 0)$, $\bar{x}_0 = 1.1$, $h = 1$, $\sigma^2 = 0.01$, and $x_0 = 0.9$ is given in Fig. 2.1. It is easily seen that the minimum of $J_3(g)$ is to the right of the minimum of $J_2(g)$.

2.3.2 Estimation of G Using Kalman-Like Nudging Scheme

This approach is due to Vidard et al. (2003) which is a nice hybrid scheme that combines the Kalman filter like predictive part and the conventional nudging scheme to combine the innovation or the prediction error [Kalman (1960b)].

Let $x(0) = x^b(0) + \delta x(0)$ be the initial state for the nudged dynamics where $x^b(0)$ is the background/prior information about $x(0)$ and $\delta x(0)$ is the perturbation added to the background. Let B be the covariance of the background state $x^b(0)$. A two step nudging scheme is then given by

$$x^f(k) = M\,(x(k-1)) \tag{2.55}$$

and

$$x(k+1) = x^f(k) + G[z(k) - h(x^f(k))]$$

Define an innovation

$$d(k) = z(k) - h\left(x^f(k)\right) \in R^m \tag{2.56}$$

and define

$$d(1:N) = \left(d^T(1), d^T(2), \ldots, d^T(N)\right)^T \in R^{Nm} \tag{2.57}$$

Define

$$J_b\left(x(0)\right) = \frac{1}{2}\left(x(0) - x^b(0)\right)^T B^{-1}\left(x(0) - x^b(0)\right) \tag{2.58}$$

$$J_n(x(0), G) = \frac{1}{2} d^T(1:N) G^T (P^f)^{-1} G d(1:N) \tag{2.59}$$

Clearly, $J_b(x(0))$ measures the weighted squared distance between $x(0)$ and $x^b(0)$ and $J_n(x(0), G)$ is called the nudging term that measures the weighted square of the model error term in (2.55), and $P^f \in R^{Nm \times Rm}$ is the model error covariance matrix computed using the standard method used in the Kalman filter literature (Lewis et al. (2006)).

Vidard et al. (2003) then pose the estimation problem as one of minimizing

$$Q_3(x(0), G) = J_2(G) + J_b(x(0)) + J_n(x(0), G) \tag{2.60}$$

where $J_2(G)$ is defined in (2.22) and the nudged dynamics in (2.56) is used as a strong constraint. This minimization is again solved by invoking the standard adjoint method [Lewis et al. (2006)].

Following the arguments at the end of Sect. 2.3.1, it can be readily verified that the error vector e(1:N) which is a part of $J_2(G)$ in (2.22) is temporally correlated. Hence, the correct formulation $J_2(G)$ in (2.60) must be replaced by $J_3(G)$ in (2.33). Similarly, it can be verified that the innovation vector $d(1:N)$ is also temporally correlated and the $J_n(x(0), G)$ in (2.60) by a similar correct form of the functional that takes the serial correlation of $d(1:N)$ into account. Let $W \in R^{Nm \times Nm}$ be the serial correlation of $d(1:N)$. Then define

$$J_3(x(0), G) = \frac{1}{2} d^T(1:N) G^T W^{-1} G d(1:N) \tag{2.61}$$

Hence the correct formulation is to minimize

$$Q_3\left(x(0), G\right) = J_3(G) + J_b\left(x(0)\right) + J_3\left(x(0), G\right) \tag{2.62}$$

We leave the computation of the elements of W as an exercise to the reader.

2.4 Observability and Observer-Based Nudging

We start by reviewing some of the fundamental concepts related to observability that are key to the analysis of observer-based nudging. Loosely stated, observability relates to the goal of reconstruction of a past state, say $x(q)$ at time q, from a finite collection of N future observations $z(k)$ for $q \leq k \leq (N+q)$ of a system. The basic theory of controllability/reachability and its dual observability of linear deterministic dynamical systems was first introduced by Kalman (1960a). The notion of observer was introduced by Luenberger (1964, 1971). If the given dynamical system is observable, then the observer is a derived dynamical system that estimates the state of the original system. In this sense, observers are the deterministic counterpart of the well known Kalman filters which provide the "best" estimate of the state of a stochastic dynamical system. The notion of observability and the design of observers have been extended to nonlinear deterministic systems. Refer to the book by Isidori (1995) for a thorough treatment of this topic.

2.4.1 Conditions for Observability

Let $x(0) \in R^n$ be the initial state of a linear dynamical system

$$x(k+1) = Mx(k) \tag{2.63}$$

where $M \in R^{n \times n}$ is a nonsingular matrix. Iterating (2.63), it can be verified that

$$x(k+q) = M^k x(q) \tag{2.64}$$

for any integer $q \geq 0$ and $k \geq 0$. Let $z(k) \in R^m$ be the observation at time k given by

$$z(k) = Hx(k) + V(k) \tag{2.65}$$

where $H \in R^{m \times n}$ and $V(k) \sim N(0, R)$ is a white Gaussian noise with $R \in R^{m \times n}$, a known symmetric and positive definite matrix.

Assume that we are given a set $\{z(k) : q \leq k \leq N+q-1\}$ of N observations. Substituting (2.64) in (2.65), this set of N observations can be collectively represented by

$$\begin{bmatrix} z(q) \\ z(q+1) \\ z(q+2) \\ \cdot \\ \cdot \\ \cdot \\ z(q+N-1) \end{bmatrix} = \begin{bmatrix} H \\ HM \\ HM^2 \\ \cdot \\ \cdot \\ \cdot \\ HM^{N-1} \end{bmatrix} x(q) + \begin{bmatrix} V(q) \\ V(q+1) \\ V(q+2) \\ \cdot \\ \cdot \\ \cdot \\ V(q+N-1) \end{bmatrix} \tag{2.66}$$

To simplify the notation, $z(q : q + n - 1) \in R^{Nm}$ denotes the column vector of observations on the l. h. s. of (2.66) and $V(q : q + N - 1)$ denotes the column vector of observation noise in the second term on the r. h. s. of (2.66). Consequently (2.66) becomes

$$z(q : q + N - 1) = H(0 : N - 1)x(q) + V(q : q + N - 1) \tag{2.67}$$

mathematically, observability relates to solving the linear least squares problem (2.67) for $x(q)$.

From the standard linear least squares theory (Chap. 5, Lewis et al. (2006)), the best $x(q)$ is the one that minimizes

$$f\,(x(q)) = \frac{1}{2} < e(q : q + N - 1), (I \otimes \mathrm{R})^{-1} e(q : q + N - 1) > \tag{2.68}$$

where

$$e(q : q + N - 1) = z(q : q + N - 1) - H(0 : N - 1)x(q) \tag{2.69}$$

is the vector of residuals, $I \otimes \mathrm{R}$ is the Kronecker product of $I \in R^{N \times N}$, and $\mathrm{R} \in R^{m \times m}$.

It can be verified (Chap. 5, Lewis et al. (2006)) that the minimizer is given by

$$\begin{aligned} x_{ls}(q) = {} & \left[H^T(0 : N - 1)(I \otimes \mathrm{R})^{-1} H(0 : N - 1)\right]^{-1} \cdot \\ & \left[H^T(0 : N - 1)(I \otimes \mathrm{R})^{-1} z(q : q + N - 1)\right] \end{aligned} \tag{2.70}$$

Hence the solution exists and is unique if the observability matrix

$$\begin{aligned} O_N = {} & H^T(0 : N - 1)(I \otimes R)^{-1} H(0 : N - 1) \\ = {} & \sum_{k=0}^{N-1} (M^{k-1})^T (H^T R^{-1} H) M^{k-1} \end{aligned} \tag{2.71}$$

is nonsingular. A necessary and sufficient condition for O_N to be nonsingular is that the matrix (Bernstein (2009))

$$H(0 : N - 1) = \begin{bmatrix} H \\ HM \\ HM^2 \\ \cdot \\ \cdot \\ \cdot \\ HM^{N-1} \end{bmatrix} \in R^{Nm \times n} \tag{2.72}$$

must be of full rank, that is $Rank(H(0:N-1)) = n$. Consequently, if $H(0:N-1)$ satisfies this condition, the pair (M, H) is said to be observable. By Cayley-Hamilton theorem since M^n can be expressed as a linear combination of M^k for $0 \leq k \leq n-1$, it follows that $N = n$ observations would suffice. Hence, the pair (M, H) is observable if $H(0:n-1)$ is of rank n.

We now quote a mathematical fact that we need in the analysis of observer-based nudging considered in Sect. 2.4.2.

Fact 2.1: If the system (2.63) and (2.65) is such that the pair (M, H) is observable then there exists a matrix $G \in RH^{n\times m}$ such that

$$(M - GH) \text{ is a Hurwitz matrix} \tag{2.73}$$

That is, the eigenvalues λ_i, $1 \leq i \leq n$, of $(M - GH)$ are such that $|\lambda_i| < 1$ for all $1 \leq i \leq n$ where $|a|$ denotes the absolute value of the complex number a. That is, the eigenvalues of $(M - GH)$ lie within the unit circle in the complex plane. Refer to Chap. 12 in Bernstein (2009) for a proof of this fact.

Example 2.3. Let $n = 2$ and $m = 1$. Then $x(k) = (x_1(k), x_2(k))^T$, $z(k) \in R$. Let $H = [0, 1]$ and $M = \begin{bmatrix} 1 & 1 \\ 0 & a \end{bmatrix}$. Then $x(k+1) = Mx(k)$ in component form is given by

$$x_1(k+1) = x_1(k) + x_2(k)$$
$$x_2(k+1) = ax_2(k)$$

and

$$z(k) - x_2(k) + V(k)$$

Where $V(k) \sim N(0, \sigma^2)$. It can be verified that

$$H[0:1] = \begin{bmatrix} H \\ HM \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & a \end{bmatrix}$$

is of rank 1 and hence the pair (M, H) is not observable.

The import of the above example can be interpreted from another angle by using the standard data assimilation point of view. Let $z(1)$ and $z(2)$ be the two observations and let

$$e(k) = z(k) - Hx(k) = z(k) - HM^k x(0) \tag{2.74}$$

be the residuals for $k = 1$ and 2. Consider the sum of the squared residuals

$$f(x(0)) = \frac{1}{2\sigma^2}\left[e^2(1) + e^2(2)\right] \tag{2.75}$$

then it can be verified that

$$\nabla_{x(0)} f\ (x(0)) = -\frac{1}{\sigma^2}\left[e(1)M^T H^T + e(2)(M^2)^T H^T\right] \tag{2.76}$$

But since $M^T H^T = \begin{bmatrix} 0 \\ a \end{bmatrix}$ and $(M^2)^T H^T = \begin{bmatrix} 0 \\ a^2 \end{bmatrix}$, we get

$$\nabla_{x(0)} f = -\frac{1}{\sigma^2}\begin{bmatrix} 0 \\ ae(1) + a^2 e(2) \end{bmatrix} \tag{2.77}$$

this in turn implies that $f(x(0))$ is a constant with respect to the first component of $x(0)$. Hence, the initial condition $x(0)$ cannot be recovered from $z(1)$ and $z(2)$

2.4.2 Observer-Based Nudging: Linear Dynamics

Let

$$\bar{x}(k+1) = M\bar{x}(k) \tag{2.78}$$

where $\bar{x}(k)$ is the true linear dynamical state with true initial condition $\bar{x}(0)$ and

$$z(k) = H\bar{x}(k) + V(k) \tag{2.79}$$

the observations. It is assumed that the pair (M, H) is observable (See Sect. 2.4.1). Let the observer be given by (Luenberger 1964, 1971) the dynamics

$$x(k+1) = Mx(k) + G\left(z(k) - Hx(k)\right) \tag{2.80}$$

where $G \in R^{n\times m}$. The idea of the observer is that the observer state $x(k)$ is an estimate of the true state $\bar{x}(k)$. In the parlance of meteorology this observer is called the nudged dynamics (Anthes (1974)) and the matrix G is called the nudging coefficient.

To analyze the behavior of (2.80), using (2.79) and simplifying, we obtain

$$e(k) = x(k) - \bar{x}(k) \tag{2.81}$$

Subtracting (2.78) from (2.80), using (2.81) and simplifying, we obtain

$$e(k+1) = (M - GH)e(k) + GV(k) \tag{2.82}$$

Since it is given that the pair $(M.H)$ is observable, by the fact 2.1 in Sect. 2.4.1, there exists a matrix $G \in R^{n\times m}$ such that $(M - GH)$ is a Hurwitz matrix. Then setting $A = (M - GH)$, from (2.82) we obtain

$$e(k) = A^k e(0) + \sum_{j=0}^{k-1} A^j G V(k-1-j) \tag{2.83}$$

Since A is Hurwitz, it follows that the spectral norm $\left|\left|A\right|\right|_2$ of A is less than 1 and $\lim_{k\to\infty} A^k = 0$ and

$$\lim_{k\to\infty} \sum_{j=0}^{k-1} A^j G = \lim_{k\to\infty} (I + A + A^2 + \ldots + A^{k-1})G$$
$$= (I - A)^{-1} G \tag{2.84}$$

in analogy with the expansion of $(1-x)^{-1}$ when $|x| < 1$. Hence $E\,[e(k)] \equiv 0$ for all $k \geq 0$ and

$$Cov(e(k)) = E\left\{\left[\sum_{j=0}^{k-1} A^j GV(k-1-j)\right]\left[\sum_{j=0}^{k-1} A^j GV(k-1-j)\right]^T\right\} \tag{2.85}$$
$$= \sum_{j=0}^{k-1} A^j G\mathrm{R}G^T (A^j)^T$$

Hence,

$$\left|\left|\mathrm{cov}\,(e(k)\right|\right|_2 \leq \left|\left|G\mathrm{R}G^T\right|\right|_2 \sum_{j=0}^{\infty} \left|\left|A\right|\right|_2^{2j} \tag{2.86}$$
$$= \left|\left|G\mathrm{R}G^T\right|\right|_2 (1 - \left|\left|A\right|\right|_2^2)^{-1}$$

In the special case when the observations are noise free, the second term on the r. h. s. of (2.83) vanishes identically and in this case

$$\lim_{k\to\infty} e(k) = 0 \text{ or } \lim_{k\to\infty} x(k) = \bar{x}(k) \tag{2.87}$$

Clearly, the rate of convergence is controlled by the choice of G. If the norm $\left|\left|A\right|\right|_2$ is close to zero, the convergence is way too fast which should be avoided.

In the following, we illustrate the choice of G using a simple example.

Example 2.4. Let $n = 2$ and $m = 1$ with

$$M = \begin{bmatrix} 2 & -1 \\ -2 & 2 \end{bmatrix} \text{ and } H = [1, 0]$$

It can be verified that the eigenvalues of M are $\lambda_1 = (2 + \sqrt{2}) > 1$ and $\lambda_2 = (2 - \sqrt{2}) < 1$. Hence, the true system $\bar{x}(k+1) = M\bar{x}(k)$ is unstable with one growing mode and one decaying mode. Let $G = [g_1, g_2]^T \in R^{2\times 1}$ and consider

$$A = M - GH = \begin{bmatrix} (2 - g_1) & -1 \\ -(2 + g_2) & 2 \end{bmatrix}$$

the eigenvalues $\mu = (\mu_1, \mu_2)$ are given by the roots of

$$0 = p(\mu) = \det(A = \mu I) = \mu^2 - \mu(4 - g_1) + (2 - 2g_1 - g_2)$$

Setting $g_1 = 3$ and $g_2 = -\frac{17}{4}$, it follows that

$$0 - \mu^2 - \mu + \frac{1}{4} = \left(\mu\frac{1}{2}\right)^2$$

Hence, $\mu_1 = \mu_2 = \frac{1}{2}$ are the eigenvalues which in turn implies that $(M - GH)$ is a Hurwitz matrix.

2.4.3 Observer Based Nudging: Nonlinear Dynamics

There is a vast corpus of books and papers in control literature relating to the design of observers for nonlinear dynamical systems, simply known as nonlinear observers (Isidori (1995), Marquez (2003), Bonnabel et al. (2009), Auroux (2011)). While it is tempting to provide a comprehensive survey of results from this area, it turns out that nonlinear observer design theory is deeply rooted in some of the fundamental results from differential geometry. Even an elementary introduction to these beautiful results will take us too far from our stated goals. So, quite reluctantly, we content ourselves with a very simple approach based on the classical Lyapunov theory of stability.

Let the given nonlinear dynamical system be given by

$$\bar{x}(k+1) = M\bar{x}(k) + F\left(\bar{x}(k)\right) \tag{2.88}$$

with $\bar{x}(0)$ as the initial condition where the right hand side of (2.88) is the sum of the linear part Mx and the nonlinear part $F(x(k))$ where the map $F : R^n \to R^n$ is assumed to satisfy the (global) Lipschitz condition

$$\left\|F(x_1) - F(x_2)\right\|_2 \le L_F \left\|x_1 - x_2\right\|_2 \tag{2.89}$$

for all $x_1 x_2 \in R^n$ where $L_F > 0$ is called the Lipschitz constant.

Let

$$z(k) = h\left(\bar{x}(k)\right) \tag{2.90}$$

be the noise-free observations where the map (called the forward operator) $h : R^n \rightarrow R^m$ is also assumed to be Lipschitz with L_h as its Lipschitz constant.

Borrowing the ideas from the linear observer design in Sect. 2.4.2, we consider an observer of the form

$$x(k+1) = M\left(x(k)\right) + F\left(x(k)\right) + G\left[z(k) - h\left(x(k)\right)\right] \tag{2.91}$$

where $G \in R^{n \times m}$ is the unknown nudging matrix.

Subtracting (2.88) and (2.91) and using the definition of the error $e(k)$ in (2.81) and (2.90), we get

$$e(k+1) = Me(k) + F(x(k)) - F(\bar{x}(k)) - G\left[h(x(k)) - h(\bar{x}(k))\right] \tag{2.92}$$

Taking the norms of both sides, we get

$$\begin{aligned} ||e(k+1)|| \leq [||M|| \cdot ||e(k)|| + ||F\left(x(k)\right) - F\left(\bar{x}(k)\right)||] \\ + \ ||G|| \cdot ||h\left(x(k) - h\left(\bar{x}(k)\ \ ||\right]\right. \end{aligned} \tag{2.93}$$

where we have used the following facts:

$$\begin{aligned} ||Ax|| \ &\leq \ ||A|| \cdot ||x|| \\ ||a - b|| \ &\leq \ ||a|| + ||b|| \end{aligned}$$

Since F and h are Lipschitz, the above inequality becomes

$$||e(k+1)|| \ \leq \ (||M|| + L_F + ||G||L_h)\ ||e(k)|| \tag{2.94}$$

Clearly, $\left|\left|e(k)\right|\right| \rightarrow 0$ only when

$$(||M|| + L_F + ||G||L_h) < 1 \tag{2.95}$$

Since M, L_F, L_h are given, there is only a limited choice for G such that (2.95) holds.

We can get a better idea if the observation is linear, that is

$$z(k) = H\bar{x}(k) \tag{2.96}$$

In this case, (2.91) becomes

$$x(k+1) = (M - GH)x(k) + F(x(k)) \tag{2.97}$$

Subtracting (2.88) from (2.97), we obtain

$$e(k+1) = (M - GH)e(k) + F\,(x(k)) - F\,(\bar{x}(k)) \tag{2.98}$$

Using the Lipschitz property of $F(x)$ in (2.98), it becomes

$$e(k+1) \le (M - GH + L_F I_n)e(k) \tag{2.99}$$

where I_n is the identity matrix. Taking the norms of both sides, we obtain

$$||e(k+1)|| \;\le\; ||(M + L_F I_n) - GH|| \cdot ||e(k)|| \tag{2.100}$$

Thus, if $((M + L_F I_n), H)$ is observable, then there exists G such that $[(M + L_F I_n) - GH]$ is a Hurwitz matrix.

Clearly, if $F(x) \equiv 0$, then and we obtain the results of Sect. 2.4.2.

2.5 Back and Forth Nudging Scheme

Recently Auroux (2011) and his collaborators have introduced a nudging scheme wherein the same set of observations are inserted into the model that runs forward in time and then backward in time. Starting from an arbitrary initial condition, say $x_0^{(0)} = x_0$, let $x_N^{(0)}$ be the nudged model state at the final forecast time N. The nudged forecast is made using observations $\{z_0, z_1, \ldots, z_{N-1}\}$. Then the model is run backwards starting from the final state which is now denoted by $\tilde{x}_N^0 (= x_N^0)$. Let $\tilde{x}_0^{(0)}$ be the state at time $k = 0$ resulting from the backward run. Then a new forward run is initiated from the initial condition $x_0^{(1)} (= \tilde{x}_0^{(0)})$, the initial state computed by the backward run just completed. This cycle is repeated. It is shown by Auroux (2011) that the sequence of initial states for the forward run: $x_0^{(0)}, x_0^{(1)}, x_0^{(2)}, \ldots$ converges to the true initial state—that is, the initial state used to create the observations in the numerical experiment.

In the following we illustrate the power of this idea using a simple dynamics for both the cases of observations being noiseless and noisy.

Consider a linear advection equation

$$u_t + c u_x = 0 \tag{2.101}$$

where u_t and u_x are the first partial derivatives of $u = u(x, t)$ with respect to the time variable t and the space variable x where it is assumed that $x \in [-1, 1]$ and $t \ge 0$. It is also assumed that

$$u(x, 0) = \sin(\pi x) \;\; \text{(initial condition)} \tag{2.102}$$

and

$$u(-1,t) = u(1,t) = 0 \text{ (boundary condition)} \tag{2.103}$$

the parameter c is the constant phase velocity of the sinusoidal curve.

A useful way to characterize the solution of (2.101) is that the time derivative of u along the characteristics is zero, that is,

$$\frac{du}{dt} = 0 \left(\text{along the characteristic} : \frac{dt}{dx} = \frac{1}{c} \right) \tag{2.104}$$

We use this latter property to illustrate the back and forth nudging scheme and its properties.

The forward nudged dynamics in continuous time is given by

$$\frac{du}{dt} = g(z-u) \tag{2.105}$$

and where $g > 0$. The corresponding backward dynamics is given by (Auroux (2011))

$$\frac{dw}{dt} = -g(z-w) \tag{2.106}$$

Discrete form of (2.105) using Euler scheme is

$$u(k+1) = (1-g)u(k) + gz(k) \tag{2.107}$$

where $u(0)$ is the initial condition. Similarly, the discrete form of the backward dynamics is given by

$$w(k) = (1-\alpha)w(k+1) + \alpha z(k) \tag{2.108}$$

where $\alpha = \frac{g}{1+g} > 0$ and $w(N)$ is the starting condition for the backward integration.

We use the same set of observations

$$\{z(j) : 0 \leq j \leq N-1\} \tag{2.109}$$

in the nudging analysis, where it is tacitly assumed that $z(j)$ is model generated starting from a true initial state $u^T(0)$. Let

$$\frac{du^T(t)}{dt} = 0 \tag{2.110}$$

be the true model whose solution is given by

$$u^T(t) = u^T(0) \tag{2.111}$$

Then the observation $z(j)$ is given by

$$\begin{aligned} z(j) &= u^T(j\Delta t) + V(j) \\ &= u^T(0) + V(j) \end{aligned} \tag{2.112}$$

where $V(j) \sim N(0, \sigma^2)$ is the Gaussian noise affecting the observations and Δt is the time discretization used in the Euler scheme.

2.5.1 Analysis of Forward Nudging

Iterating (2.107), it can be verified that the forward solution of (2.107) at anytime is given by

$$u(k) = (1-g)^k u(0) + g \sum_{j=0}^{k-1} (1-g)^j z(k-1-j) \tag{2.113}$$

where $u(0)$ is the arbitrary initial condition used to start the forward run.

Substituting (2.112) into (2.113) we get

$$u(N) = DPF + SPF \tag{2.114}$$

where the deterministic part, DPF, is given by

$$\begin{aligned} DPF &= (1-g)^N u(0) + g u^T(0) \sum_{j=0}^{N-1} (1-g)^j \\ &= u^T(0) + (1-g)^N \left[u(0) - u^T(0)\right] \end{aligned} \tag{2.115}$$

where $[u(0) - u^T(0)]$ is the error in the initial condition. Similarly, the stochastic part is

$$SPF = g \sum_{j=0}^{N-1} (1-g)^j V(k-1-j) \tag{2.116}$$

whose mean is zero and the variance is given by

$$\frac{1}{\sigma^2} Var(SPF) = g^2 \left[\frac{1-(1-g)^{2N}}{1-(1-g)^2}\right] \tag{2.117}$$

Combining (2.115) and (2.117), it follows that $u(N)$ is a Gaussian random variable whose mean is DPF and variance is given by $Var(SPF)$.

2.5.2 Analysis of Backward Nudging

Iterating (2.108), it can be verified that the backward solution, at any time k, is given by

$$\mathrm{W(N-k)} = (1-\alpha)^{\mathrm{k}}\mathrm{W(N)} + \alpha \sum\nolimits_{\mathrm{j=1}}^{\mathrm{k}} (1-\alpha)^{\mathrm{k-j}}\mathrm{Z(N-j)} \tag{2.118}$$

where W(N) is the final condition from which the backward nudging starts.

Substituting (2.112) in (2.118) and simplifying we get

$$\mathrm{W(0)} = \mathrm{DPB} + \mathrm{SPB} \tag{2.119}$$

where the deterministic part, DPB is given by

$$\begin{aligned} \mathrm{DPB} &= (1-\alpha)^{\mathrm{N}}\mathrm{W(N)} + \alpha \sum\nolimits_{\mathrm{j=1}}^{\mathrm{k}} (1-\alpha)^{N-j}\mathrm{u}^{\mathrm{T}}(0) \\ &= \mathrm{u}^{\mathrm{T}}(0)(1-\alpha)^{\mathrm{N}}\left[\mathrm{W(N)} - \mathrm{u}^{\mathrm{T}}(0)\right] \end{aligned} \tag{2.120}$$

The stochastic part, SPB is given by

$$\mathrm{SPB} = \alpha \sum\nolimits_{\mathrm{j=1}}^{\mathrm{N}} (1-\alpha)^{\mathrm{N-j}}\mathrm{V(N-j)} \tag{2.121}$$

whose mean is zero and the variance is given

$$\begin{aligned} \frac{1}{\sigma^2}\mathrm{Var(SPB)} &= \alpha^2\left[\frac{1-(1-\alpha)^{2N})}{1-(1-\alpha)^2}\right] \\ &= \left[\frac{g^2}{(1+g)^2-1}\right]\left[1-\frac{1}{(1+g)^{2N}}\right] \end{aligned} \tag{2.122}$$

Thus, W(0) is a Gaussian random variable whose mean is given by DPB and variance is equal to Var(SPB).

2.5.3 Back and Forth Nudging Scheme

Against this background, we now close the loop between the forward and the backward steps to get the so called back and forth nudging scheme.

Let $\mathrm{u}^{(\mathrm{j})}(0)$ be the starting initial condition for the jth forward run of the model that leads to the sequence of rates given by $\{\mathrm{u}^{(\mathrm{j})}(0), \mathrm{u}^{(\mathrm{j})}(1), \mathrm{u}^{(\mathrm{j})}(2), \ldots\ldots.\mathrm{u}^{(\mathrm{j})}(\mathrm{N})\}$ obtained by running the forward model (2.107) where $\mathrm{u}^{(\mathrm{j})}(\mathrm{N})$ is the final state. In the jth backward run of the model, the starting final state $\mathrm{W}^{(\mathrm{j})}(\mathrm{N})$ is set to be equal to the final state $\mathrm{u}^{(\mathrm{j})}(\mathrm{N})$ of the jth forward run just completed.

Let $\{W^{(j)}(0),\ W^{(j)}(1),\ W^{(j)}(2),\ \ldots\ldots.W^{(j)}(N) = u^{(j)}(N)\}$ be the sequence of backward states obtained by running the backward model (2.108).

The new initial condition, $u^{(j+1)}(0)$ for the $(j+1)$th run of the forward model is set to be equal to the initial rate, $W^{(j)}(0)$ of the backward run just completed.

To start the overall iterative process at the 0th run of the forward model, the initial condition $u^{(0)}(0) = u(0)$, an arbitrary choice.

Our goal is to characterize the limiting behavior of the sequence $\{u(0) = u^{(0)}(0),\ u^{(1)}(0),\ u^{(2)}(0),\ \ldots\ldots.u^{(p)}(0)\ \ldots\ldots.\}$ of initial state of the forward run induced by the feed–back process between the forward and the backward runs described above.

We consider two cases.

CASE A: Observations are Noise – free

Under this assumption, $V(k) \equiv 0$ in (2.112). Consequently, the stochastic part SPF in (2.116) and SPB in (2.121) are identically zero.

We now derive a recurrence relation that relates the evolution of the required initial conditions $u^{(j)}(0)$. The final rate $u^{(j)}$ (N) starting from $u^{(j)}(0)$ is given by (2.115) as

$$u^{(j)}(N) = u^T(0) + (1-g)^N\left[u^{(j)}(0) - u^T(0)\right] \tag{2.123}$$

Similarly, referring to (2.120) the initial rate $W^{(j)}(0)$ of the jth backward run is given by

$$W^{(j)}(0) = u^T(0) + (1-\alpha)^N\left[W^{(j)}(N) - u^T(0)\right] \tag{2.124}$$

Since $W^{(j)}(N) = u^{(j)}(N)$, substituting (2.123) into (2.124) and simplifying we get,

$$\begin{aligned} u^{(j+1)}(0) &= W^{(j)}(0) \\ &= u^T(0) + (1-\alpha)^N(1-g)^N\left[u^{(j)}(0) - u^T(0)\right] \end{aligned} \tag{2.125}$$

That is,

$$u^{(j+1)}(0) - u^T(0) = (1-\alpha)^N(1-g)^N\left[u^{(j)}(0) - u^T(0)\right] \tag{2.126}$$

Thus, if $0 < g < 1$, then so is α and (2.126) becomes

$$|u^{(j+1)}(0) - u(0)| = \beta|u^{(j)} - u^T(0)| \tag{2.127}$$

when $\beta = (1-\alpha)^N(1-g)^N$ and $0 < \beta < 1$ for any fixed numbers $N(> 0)$ of observations.

Iterating (2.127), we obtain

$$|u^{(p)}(0) - u(0)| = \beta^p|u^{(0)} - u(0)| \tag{2.128}$$

That is, $u^{(p)}(0)$ converges to the true but unknown initial state exponentially, That is,

$$\lim_{p\to\infty} u^{(p)}(0) = u^{(T)}(0) \tag{2.129}$$

Referring to Chap. 10, Lewis et al. (2006) we can restate that $u^{(p)}(0)$ converges to $u^{T}(0)$ asymptotically at a linear rate.

CASE B: Noisy Observations

In this case, from (2.114, 2.115, 2.116 and 2.117) it follows that

$$u^{(j)}(N) = u^{(T)}(0) + (1-g)^{N}\left[u^{(j)}(0) - u^{T}(0)\right] + \eta^{(j)}(N) \tag{2.130}$$

where

$$\eta^{(j)}(N) \sim N\left(0, \mathrm{Var(SPF)}\right) \tag{2.131}$$

Similarly, from (2.118)–(2.112), we get

$$W^{(j)}(0) = u^{(T)}(0) + (1-\alpha)^{N}\left[W^{(j)}(N) - u^{T}(0)\right] + \varepsilon^{(j)}(0) \tag{2.132}$$

where

$$\varepsilon^{(j)}(0) \sim N\left(0, \mathrm{Var(SPB)}\right) \tag{2.133}$$

with $W^{(j)}(N) = U^{(j)}(N)$. Substituting (2.130) into (2.132) and using the feed—back law of back and forth nudging, we get,

$$\begin{aligned} u^{(j+1)}(0) &= W^{(j)}(0) \\ &= u^{(T)}(0) + (1-\alpha)^{N}(1-g)^{N}\left[u^{(j)}(0) - U^{T}(0)\right] + (1-\alpha)^{N}\eta^{(j)}(N) + \varepsilon^{(j)}(0) \end{aligned} \tag{2.134}$$

which on rewriting becomes

$$\left[u^{(j+1)}(0) - u^{T)}(0)\right] = \beta\left[u^{(j)}(0) - u^{(T)}(0)\right] + \psi^{(j)}(0) \tag{2.135}$$

where

$$\psi^{(j)}(0) = (1-\alpha)^{N}\eta^{(j)}(N) + \varepsilon^{(j)}(0) \tag{2.136}$$

Substituting for $\eta^{(j)}(N) = \mathrm{SPF}$ in (2.116) and $\varepsilon^{(j)}(0) = \mathrm{SPM}$ in (2.121) in (2.136), it can be verified that $\psi^{(j)}(0)$ is a mean—zero Gaussian random variable whose variance is given by

$$\frac{1}{\sigma^{2}}\mathrm{Var}\left[\psi^{(j)}(0)\right] = (1-\alpha)^{2N}\mathrm{Var(SPB)} + \mathrm{Var(SPF)} \tag{2.137}$$

Now, iterating (2.135), we get, for any integer $p > 0$,

$$\left[u^{(p)} - u^{(T)}(0)\right] = D + S \tag{2.138}$$

where

$$D = \beta^{p}\left[u^{(0)}(0) - u^{T}(0)\right] \tag{2.139}$$

which reads to zero as p grows since $0 < \beta < 1$, and S is given by

$$S = \sum\nolimits_{j=0}^{p-1} \beta^{p-1-j} \psi^{(j)}(0) \tag{2.140}$$

It can be verified that S is a mean zero Gaussian random variable whose variance is given by

$$\begin{aligned} \mathrm{Var}(S) &= \mathrm{Var}\left[\psi^{(j)}(0)\right] \sum\nolimits_{j=0}^{p-1} \beta^2(p-1-j) \\ &= \mathrm{Var}\left[\psi^{(j)}(0)\right] \frac{\left[1-\beta^{2p}\right]}{\left[1-\beta^2\right]} \end{aligned} \tag{2.141}$$

Thus, for a fixed number N of observations,

$$\mathrm{Var}(S) \rightarrow \frac{\mathrm{Var}\left[\psi^2(0)\right]}{\left[1-\beta^2\right]} \tag{2.142}$$

as the number, p of back and forth iterations increase.

2.6 Discussion and Conclusions

There is an ever-growing literature on the applications of nudging as a simple viable method for dynamic data assimilation. It is attractive to the geophysical science community because of its ease of implementation and its intuitive appeal—in essence, the use of the earlier known error in prediction to alter subsequent prediction appeals to common sense. Yet, in its earliest stage of development where empiricism was the theme, search for a suitable nudging coefficient exhibited great computational demand through numerous simulations and validation against the evolution of dynamical systems. And the final choice of the nudging coefficient was always subject to debate—linked to the question: isn't there a better coefficient? It naturally led to an effort to find a coefficient that exhibited optimality under a specific form of the cost functional that forced the coefficient toward an *a priori* estimate. And again, this brought up other questions concerning the "heavy handedness" by producing a cost function that was forced to remain close to the *a priori* estimate. Further, these methods have unintentionally omitted an important aspect of the nudging problem—nudging dynamics carries with it the presence of a serially correlated forecast error and this error must be accounted to find the optimal coefficient. It is computationally demanding to find the structure of this correlated error. For sure, its influence on the optimal nudging process is an important area of investigation that remains open. Our review also indicates that the well-established theory of observer design (as a practice in the contemporary control theory) deserves further attention from those involved in data assimilation for numerical prediction in the geophysical sciences. And the "back-and-forth" nudging offers promise for application to operational prediction, but where attention must be focused on the results for irreversible processes that are ubiquitous in the ocean-atmosphere system.

Acknowledgements Our thanks are due to an anonymous reviewer for his detailed comments that improved the quality and the readability of this chapter. We wish to thank Mr. Dung Phan for his timely help in typesetting this chapter. S. Lakshmivarahan's efforts are supported in part by NSF EPSCOR RII Track 2 Grant 105-155900 and by NSF Grant 105-15400.

References

Anthes RA (1974) Data assimilation and initialization of hurricane prediction model. J Atmos Sci 31:702–719

Auroux D (2009) The back and forth nudging algorithm applied to a shallow water model: comparison and hybridization with the 4D-VAR. Int J Num Meth In Fluids 61:911–929

Auroux D (2011) Back and forth nudging methods in geophysical data assimilation. In: Lecture notes, monsoon workshop on mathematical and statistical foundations of data assimilation, TIFR, Bangalore, 04–23 July 2011

Auroux D, Blum J (2005) Back and forth nudging algorithm for data assimilation problems. C R Acad Sci, Paris, Series I 340:873–878

Auroux D,Blum J (2008) A nudging based data assimilation method: the back and forth nudging (BFN) algorithm. Nonlinear Process Geophys 15:305–319

Auroux D, Nodet M (2010) The back and forth nudging algorithm for data assimilation problems: theoretical results on transport equations. Control Optimsation Calculus Var 18:1–25

Bao J-W, Errico RM (1997) An adjoint examination of a nudging method for data assimilation. Mon Weather Rev 125:1355–1373

Bell RS, Dickinson A (1987) The Meteorological Office operational numerical weather prediction system. Meteorological Office, Scientific paper, Vol 41. H.M.S.O, London

Bennett S (1996) A brief history of automatic control. IEEE Contr Syst Mag 16: 17–25

Bernstein D (2009) Matrix mathematics theory, facts and formulas, 2nd edn. Princeton University Press, Princeton, 1139 pp

Bjerknes V (1904) Das Problem der Wettervorhersage, betrachet vom Stanpunkt der Machanik und der Physik. Meteor Zeits 21:1–7

Bonnabel S, Martin P, Rouchon P (2009) Nonlinear symmetry preserving observers on lie groups. IEEE Trans Automat Contr 54:1709–1713

Brill KF, Uccellini LW, Manobianco J, Kocin PJ, Horman JH (1991) The use of successive dynamic initialization to simulate cyclogenesis during GALE IOP1. Meteorol Atmos Phys 45:15–40

Bryson AE (1996) Optimal control—1950 to 1985. IEEE Contr Syst 16:26–33

Charney J, Fjortoft R, von Neumann J (1950) Numerical Integration of the barotropic vorticity equation. Tellus 2:237–254

Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge, 457 pp

Derber J, Rosati A (1989) A global oceanic data assimilation system. J Phys Oceanogr 19:1333–1347

Derber J, Leetma A, Ji M, Shinn R (1990) Oceanic data assimilation at the National Meteorological Center.In: WMO international symposium on assimilation of observations in meteorology and oceanography, Clermont-Ferrand, France, 59–61

Evensen G (2007) Data assimilation: the ensemble Kalman filter.Springer, Berlin, 279 pp

Hamilton JD (1994) Time series analysis.Princeton University Press, Princeton, 799 pp

Hoke JE (1976) Initialization of models for numerical weather prediction by dynamic- initialization technique. Ph.D. thesis, The Pennsylvania State University, 202 pp

Hoke JE, Anthes RA (1976) The initialization of numerical models by a dynamic initialization technique. Mon Weather Rev 104:1551–1556

Isidori K (1995) Nonlinear control systems, 3rd edn.Springer, New York, 549

Kalman RE (1960a) On the general theory of control systems. In: Proceedings of the first international congress on automatic control, Butterworth's, London, 481–493

Kalman RE (1960b) A new approach to linear filtering and prediction problems. Transaction of the ASME, Series D. J Basic Eng 82:35–45

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, 341 pp

Krishnamurti TN, Jirshan X, Bedi HS, Ingles K, Dosterhof D (1991) Physical initialization of numerical weather prediction over tropics. Tellus 43A:53–81

Kuo Y-H, Guo Y-R (1989) Dynamic initialization using observations from a hypothetical network of profilers. Mon Weather Rev 117:1975–1998

Lakshmivarahan S, Lewis JM (2011) When is nudging optimal. Technical Report, School of Computer Science, University of Oklahoma, Norman, 73019

Lewis JM (2005) Roots of ensemble forecasting. Mon Weather Rev 133:1865–1885

Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, Cambridge, 654 pp

Lewis JM, Lakshmivarahan S (2008) Sasaki's pivotal contribution: calculus of variations applied to weather map analysis. Mon Weather Rev 136:3553–3567

Lorenc AC (1986) Analysis methods for numerical weather prediction. Q J Roy Meteorol Soc 112:1177–1194

Lorenc AC, Bell RS, Macpherson B (1991) The meteorological office analysis correction data assimilation scheme. Q J R Meteorol Soc 117:59–89

Luenberger DG (1964) Observing the state of a linear system. IEEE Trans Military Electron ME-8:74–80

Luenberger DG (1971) An introduction to observers. IEEE Trans Automat Contr AC-16:596–603

Lynch J (2006) The emergence of numerical weather Prediction (Richardson's dream). Cambridge University Press, Cambridge, 279 pp

Macpherson B (1991) Dynamic initialization by repeated insertion of data. Q J Roy Meteorol Soc 117:965–991

Marquez HJ (2003) Nonlinear control systems: analysis and design.Wiley, Hoboken, 352 pp

Platzman G (1979) The ENIAC computations of 1950—Gateway to numerical weather prediction. Bull Am Meteorol Soc 60:302–312

Ramamurthy MH, Carr FH (1987) Four-dimensional data assimilation in monsoon region. I: Experiments with wind data. Mon Weather Rev 115:1678–1706

Ramamurthy MH, Carr FH (1988) Four-dimensional data assimilation in the monsoon region Part II: Role of temperature and moisture data. Mon Weather Rev 116:1896–1913

Richardson LF (1922) Weather prediction by numerical prlocess. Cambridge University Press, London, 236 pp (Reprinted by Dover (1965, New York) with a new introduction by Sydney Chapman)

Stauffer DR, Warner TT, Seaman NL (1985) A Newtonian "nudging" approach to four-dimensional data assimilation: use of SES-AME-IV data in a mesoscale model. Preprints, seventh conference on numerical weather prediction, American Meteorological Society, Montreal, 77–82

Stauffer DR, Seaman NL (1990) Use of four-dimensional data assimilation in a limited-area mesoscale model, I: experiments with synoptic-scale data. Mon Weather Rev 118:1250–1277

Stauffer DR, Bao J-W (1993) Optimal determination of nudging coefficients using adjoint equations. Tellus 45A:358–369.

Stern WF, Ploshay JJ, Miyokoda K (1985) Continuous data assimilation at GFDL during FGGE. In: Proceedings of the ECMWF seminar/workshop: data assimilation system and observing system experiments with particular emphasis in FGGE, Shinfield Park, Reading, England, September

Sussmann HJ, Willems JC (1997) 300 years of optimal control from brachystochrone to the maximum principle. IEEE Contr Syst 17:32–44

Thacker WC, Long BR (1988) Fitting dynamics to data. J Geophys Res 93:1227–1240

Vidard PA, Dimet FXLE, Piacentini A (2003) Determination of optimal nudging coefficients. Tellus 55A:1–15

Warner TT (1990) Assimilation of data with mesoscale meteorological models. In: WMO international symposium of observations in meteorology and oceanography, Clermont-Ferrand, France, pp 165–170

Wiener N (1948) Cybernetics: control and communication in the animal and machine. Wiley, New York, 194 pp

Yamada T, Bunker S (1989) A numerical model study of nocturnal drainage flows with strong wind and temperature gradients. J Appl Meteorol 28:545–554

Zou X, Navon IM, Le-Dimet FX (1992) An optimal nudging data assimilation scheme using parameter estimation. Q J Roy Meteorol Soc 128:1163–1186

Chapter 3
Markov Chain Monte Carlo Methods: Theory and Applications

Derek J. Posselt

Abstract Markov chain Monte Carlo algorithms constitute flexible and powerful solutions to Bayesian inverse problems. They return a sample of the unapproximated posterior probability density, and make no assumptions as to linearity or the form of the prior or likelihood. MCMC algorithms are in principle easy to construct, however, they can prove difficult to implement in practice. This chapter describes the theory that underlies MCMC simulation, provides guidance for its practical implementation, and presents examples of applications of MCMC to satellite retrievals and model uncertainty characterization. Though the high dimensionality of Earth system datasets and the complexity of atmospheric, oceanic, and hydrologic models present significant challenges, continued advances in theory and practice are making application of MCMC algorithms increasingly feasible.

3.1 Introduction and History

Markov chain Monte Carlo (MCMC) algorithms were born out of investigations into numerical integration at Los Alamos national laboratory in the 1940s and 1950s. These were initially centered around the development of Monte Carlo (MC) methods, which comprise a class of algorithms designed to compute numerical solutions to integrals using random draws from a specified probability distribution. MC algorithms were developed as a way to numerically solve neutron diffusion problems during the development of the atomic bomb, but were generally limited to low-dimensional problems. Shortly after the advent of Monte Carlo based random simulation, Metropolis et al. (1953) developed an extension that allowed MC to be used to evaluate integrals over large dimensional spaces. The method was used to

D.J. Posselt (✉)
Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan,
2455 Hayward Street, Ann Arbor, MI 48103, USA
e-mail: dposselt@umich.edu

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_3,

simulate the movement of interacting particles in equilibrium with each other. In essence, the Metropolis algorithm is composed of a random sequence of particle moves, each of which depends only on the current position (thus forming a Markov chain). In each new configuration of particles, the energy of the system is computed and compared with the energy of the previous configuration. If the new state has lower energy than the old, it is accepted as a valid particle configuration. If not, it is accepted with a probability that depends on the difference in energy between the two states. The collection of states that results from a sufficiently long sequence of proposed configurations describes a system of particles in equilibrium with correct relationship between pressure, temperature, and volume. Hastings (1970) later generalized the Metropolis algorithm for use in integrating probability distributions, and this constituted the first use of Markov chain Monte Carlo algorithms for posterior probability simulation.

It is thought that the relative lack of computational power limited the general applicability of MCMC for statistical computation until the 1980s. Geman and Geman (1984) applied a variant of MCMC to the problem of image restoration and drew an analogy between image processing and computation of posterior probability densities. The use of Gibbs random fields in this study resulted in their algorithm being termed the “Gibbs sampler”. Inspired by the work of Geman and Geman (1984) and Gelfand and Smith (1990) generalized MCMC-based computation of posterior probability densities from of a collection of algorithms that included Geman and Geman (1984)’s Gibbs sampler, data augmentation methods (Tanner and Wong 1987), and importance sampling (Rubin 1987). It is generally acknowledged that the Gelfand and Smith (1990) paper led to widespread use of MCMC for Bayesian posterior sampling in the statistical community, and that Tierney (1994) was the first to thoroughly describe the necessary convergence properties of the underlying Markov chains. Building on the foundations laid by Metropolis et al. (1953), Hastings (1970), Gelfand and Smith (1990), Tierney (1994), and Green (1995) further generalized the MCMC algorithm to the exploration of probability spaces with variable dimensions. Since the mid-1990s, a host of variants of MCMC have been proposed, most aimed at increasing the efficiency with which MCMC samples the posterior probability space.

Though the use of MCMC in statistical computation is now common (Gelman et al. 2011), its adoption in the atmospheric and oceanic sciences has been slow. This is in part due to the exceedingly large dimensionality of most atmospheric and oceanic state estimation problems, as well as to the complexity and computational expense of geophysical process (forward) models. In the remainder of this chapter, we will briefly present the theory that underlies MCMC algorithms, describe the practical issues users encounter when implementing MCMC for a new application, highlight the successful use of MCMC in satellite retrievals and model parameter estimation, and finish with concluding remarks as to future applications of MCMC in the atmospheric sciences.

3.2 Theoretical Basis of MCMC in Bayesian Inference

MCMC algorithms are, in essence, a form of Bayesian inference and hence prior to a discussion of the details of MCMC itself, it is useful to briefly review the foundations of Bayesian theory. The general stochastic inverse problem is described in great detail by Tarantola (2005). According to inverse problem theory, there are three distinct sources of quantitative information about a system: measurements of the properties of the system, prior knowledge, and a model of the system. Each piece of information is associated with a space that contains all possible values. Placing the Earth system in this framework, all possible physical states (x) occupy one space and all possible outcomes of observations (y) made of any and all properties of the physical system occupy another space. The role of the model $y = F(x)$ is to map information from one space into another. If we allow for uncertainty in our knowledge of the state x, then any event (realization of a possible physical state) occupies a sub-region of the overall state space and can be associated with a probability distribution $P(x)$ over that region. Similarly, allowing for uncertainty in observations of the system of study gives rise to similar definition of an observation event that occupies a sub-region of the observation space. This sub-region is also associated with a probability distribution $P(y)$. Let us assume for the moment that the observations contain information about the portion of the state space we are interested in. In this case, the forward problem defines the process of mapping from state to observation space (producing an analogue of the observations using the model) and the inverse problem consists of determining information about the state space from information contained in the observations. Each exercise (the forward and inverse problems) consists of a conjunction of the two information spaces that can be represented in the joint probability distribution $P(x, y)$. Clearly, if the two information spaces are disjoint, then the forward and inverse calculations are meaningless.

It is common to make the problem more specific by assuming that one of two events (a realization of either a set of specific state properties x or set of observations y) has occurred. In this case, the problem becomes characterization of the observation space *given* a subset of the state space, or inference of the state space *given* a set of observations. These are formalized by introducing the definition of conditional probability, for which the probability of state x conditioned on a set of observations y is

$$P(x|y) = \frac{P(x, y)}{P(y)}, \tag{3.1}$$

and the corresponding probability of observations y conditioned on the state x is

$$P(y|x) = \frac{P(y, x)}{P(x)} = \frac{P(x, y)}{P(x)}. \tag{3.2}$$

Note that $P(x, y) = P(y, x)$ because in this case they simply reflect the intersection of two probability events. Bayes' theorem results from solving (3.2) for the joint probability $P(x, y)$ and inserting the result in (3.1). Given, as is typically the case, discrete state and observation vectors $\mathbf{x}$ and $\mathbf{y}$, Bayes' theorem can be written

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})} \tag{3.3}$$

In practice, we are always dealing with a system for which either a state (or set of states) has already been determined, or of which observations have already been taken, and it is common to begin with Bayes' theorem. The benefit of generalized inverse theory is its flexibility–it makes no assumption that any event has taken place and considers only the conjunction of information spaces. Standard Bayesian estimation theory (e.g., Jazwinski 1970; Bennett 1992; Evensen 2006) treats each quantity that contributes to a total quantitative knowledge of each component of the system as a stochastic quantity with an associated probability density function (PDF). The conjunction of separate PDFs associated with contributions from the model and the associated modeling errors, the observation errors, and the prior information result in a joint posterior PDF. The properties of the joint posterior PDF determine the characteristics and tractability of the inverse problem (e.g., whether or not the solution is unique). The explicit separation of information into components provided by the model, observation, and prior estimate facilitates determination of the unique contribution of each to the posterior PDF.

The inverse problem, according to Bayes' theorem, consists in computing each individual PDF on the right hand side of (3.3) and combining them to obtain the properties of the state of interest given prior knowledge and a set of observations related through the model. It is well known that computation of the posterior PDF is straightforward if the model is linear and all PDFs can be assumed to be Gaussian. Approximate solutions can be obtained in the case of a nonlinear model by either linearizing the model (e.g., extended Kalman filter (Gelb 1974), three dimensional variational data assimilation (3DVAR; Sasaki 1970; Lorenc 1986), four dimensional variational data assimilation (4DVAR; Courtier 1997)) or via Monte Carlo methods using a stochastically generated ensemble of states (e.g., the ensemble Kalman filter, Evensen 2006). In the case of models for which the dimension of the state space is large and the mapping between state and observation space is nonlinear, it is not computationally feasible to compute numerical solutions to (3.3). The fundamental result of the work of Metropolis et al. (1953) and Hastings (1970) is that it is not necessary to compute a solution to the posterior PDF if one can construct a Markov chain that has the same equilibrium distribution. The goal of MCMC is to sample the posterior PDF $P(\mathbf{x}|\mathbf{y})$ up to a normalizing constant (i.e., by computing only the numerator on the RHS of (3.3)) using a Markov chain that has a stationary transition probability $q(\mathbf{x_i}, \mathbf{x_{i+1}})$; the probability of moving to a set of states $\mathbf{x_{i+1}}$ from the current state $\mathbf{x_i}$. Robust samples of the posterior PDF are ensured in MCMC via the application of an *update* to the Markov chain that determines whether a proposed transition from the current state $\mathbf{x_i}$ to a proposed state $\hat{\mathbf{x}} \sim q(\mathbf{x}, \cdot)$ (consistent with the specified transition probability) is *accepted*. In the formulation originally introduced by Metropolis et al. (1953) and generalized by Hastings (1970), the update is done in the following way:

(i) Given a current state $\mathbf{x_i}$, propose a move to $\hat{\mathbf{x}}$ with probability

$$q(\mathbf{x_i}, \hat{\mathbf{x}}). \tag{3.4}$$

(ii) Now, compute forward observations $\hat{\mathbf{y}}$ by running the forward model on the proposed state $\hat{\mathbf{x}}$.

(iii) Compute the *Hastings ratio*

$$\rho\,(\mathbf{x_i}, \hat{\mathbf{x}}) = \frac{P(\hat{\mathbf{y}}|\hat{\mathbf{x}})P(\hat{\mathbf{x}})q(\hat{\mathbf{x}}, \mathbf{x_i})}{P(\mathbf{y_i}|\mathbf{x_i})P(\mathbf{x_i})q(\mathbf{x_i}, \hat{\mathbf{x}})}, \tag{3.5}$$

(iv) And accept the proposed move to $\hat{\mathbf{x}}$ with probability

$$Q\,(\mathbf{x_i}, \hat{\mathbf{x}}) = \min(1, \rho\,(\mathbf{x_i}, \hat{\mathbf{x}})). \tag{3.6}$$

Note that if the transition probability is symmetric and stationary, then the probability $q(\mathbf{x_i}, \hat{\mathbf{x}})$ is identical to $q(\hat{\mathbf{x}}, \mathbf{x_i})$ and the Hastings ratio simplifies to the Metropolis update

$$\rho\,(\mathbf{x_i}, \hat{\mathbf{x}}) = \frac{P(\hat{\mathbf{y}}|\hat{\mathbf{x}})P(\hat{\mathbf{x}})}{P(\mathbf{y_i}|\mathbf{x_i})P(\mathbf{x_i})} \tag{3.7}$$

Note the similarity between the numerator and denominator of (3.7) and the numerator in (3.3). It is precisely the accept/reject criterion that allows MCMC to sample from the un-normalized posterior $P(\mathbf{x}|\mathbf{y})$.

It is common to construct proposals of the form $\hat{\mathbf{x}} = \mathbf{x_i} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_{\mathbf{x}})$. Here, $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the variance of the proposal distribution–specification of this is one of the subtleties involved in constructing an MCMC algorithm. The choice of transition probability $q(\mathbf{x}, \cdot)$ is flexible and the likelihood $P(\mathbf{y}|\mathbf{x})$ may assume any form, consistent with the characteristics of the system of study. The requirement for any MCMC algorithm is that the Markov chain with (Metropolis or Hastings) updates converges (in the limit of infinite number of proposal steps) to sampling the stationary invariant posterior distribution (ergodicity). Ergodicity is guaranteed if the MCMC algorithm is properly constructed. In practice, convergence is not assured if the form and properties of the posterior distribution are unknown (black box MCMC). The construction of an MCMC algorithm involves a number of decisions as to the form and characteristics of the transition probability distribution, as well as the specifics of the Metropolis-Hastings update. We discuss such practical issues in the next section.

3.3 Practical Issues

The only fundamental requirement for proper MCMC simulation is that the user construct a chain that is Markov with transition probabilities and updates that ensure ergodicity to the invariant posterior (target) distribution. The practical usability of an

MCMC algorithm, however, is a function of how rapidly and thoroughly it samples the posterior space. As such, while constructing a Metropolis-Hastings algorithm is simple, ensuring its efficiency is not. In this section, we outline several practical issues encountered in adapting an MCMC algorithm to a new problem and highlight a number of best practices and potential pitfalls along the way.

3.3.1 Choice and Tuning of Proposal Distribution

It is clear from the formulation of the Metropolis-Hastings update (3.5) that the proposal distribution $q(\mathbf{x}, \cdot)$ plays an important role in the MCMC algorithm. Proposals that result, on average, in large deviations from the current position will in general lead to smaller Hastings ratio (3.5) and lower probability of acceptance, and vice versa. A desirable property of any MCMC algorithm is that it mix thoroughly and rapidly, not being confined to a small region of the state space. Because there is often little knowledge of the shape of the posterior distribution, it is necessary in practice to adjust the width (e.g., (co)variance) of the proposal distribution to strike a balance between sampling rapidly enough to mix thoroughly (large moves through the state space; large proposal width) and sampling finely enough to resolve details of the probability distribution (small moves through the state space; small proposal width). Because of this, much of the subtlety involved in constructing a MCMC algorithm centers around (1) choice of a suitable proposal distribution, and (2) tuning the distribution width. As mentioned above, choice of a symmetric proposal distribution leads to $q(\mathbf{x_i}, \hat{\mathbf{x}}) = q(\hat{\mathbf{x}}, \mathbf{x_i})$, which simplifies the Metropolis-Hastings update. While there are variants of MCMC that use non-symmetric proposal distributions (e.g., Langevin-Hastings MCMC; Roberts and Rosenthal 1998), in all of the discussion that follows we will assume the use of a symmetric proposal distribution. A common choice of proposal is Uniform, centered on the current estimate. In this case, the tunable parameter is simply the width of this Uniform distribution. The advantage of this is its simplicity, and indeed this was the choice originally made by Metropolis et al. (1953). It is now common to use a zero mean multivariate Normal as the proposal distribution. This has the advantage of consistently producing moves of about one standard deviation, but with finite probability of much larger or smaller moves as well, allowing the chain to more easily move between regions of the space containing localized probability maxima.

Once a suitable proposal distribution has been chosen the question naturally arises as to how to successfully tune it to thoroughly and efficiently sample the posterior distribution. In essence, the question is what makes one Markov chain "better" than another? Desirable properties are: rapid exploration of the space, fast convergence to the target distribution, and production of a thorough and accurate sample; no regions of the state space containing probability mass are missed. It is clear from (3.5) that if the width of the proposal distribution (3.4) is small, virtually all proposed moves will be accepted, but the movements will be very small and

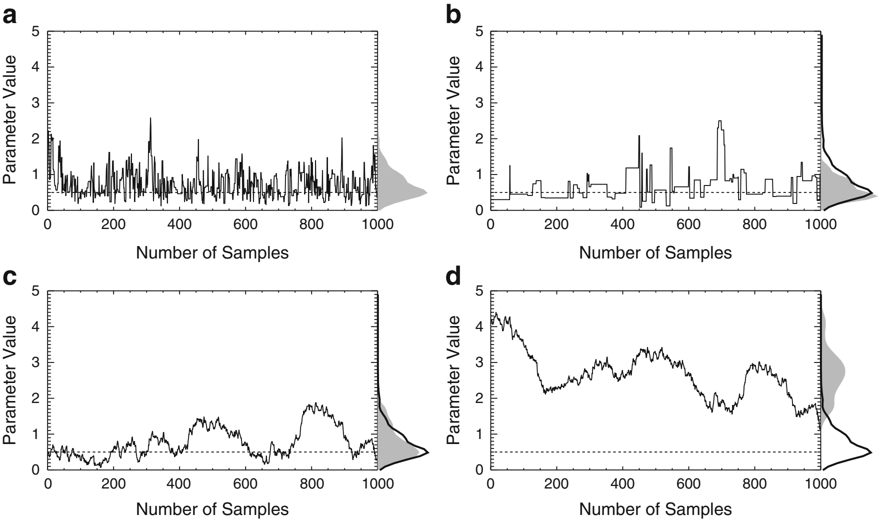


Fig. 3.1 Timeseries plots of parameter values in MCMC chains with (**a**) well tuned proposal, (**b**) proposal variance that is too large, (**c**) proposal variance that is too small, and (**d**) proposal variance that is too small and a chain that is started far from the mode of the target distribution. The *dashed horizontal line* corresponds to the true parameter maximum likelihood value (= 0.5), and the marginal distribution of each parameter is plotted in *gray* along the ordinate axis of each plot on the *right hand side*. For reference, each marginal distribution is overlaid with a *black line* depicting the distribution obtained by sampling with the well-tuned proposal (**a**)

the chain will mix very slowly. Conversely, if the width of the proposal distribution is large, most proposed moves will be rejected (due to small Hastings ratio) and the chain may not move at all. An example is shown in Fig. 3.1, in which a single parameter is estimated using three different proposal widths. The first (Fig. 3.1a) is tuned to an optimal 25 % acceptance rate (see below for a discussion on the theory of optimal tuning), while the proposal width in the second and third are set to an order of magnitude larger (Fig. 3.1b) and an order of magnitude smaller (Fig. 3.1c), respectively. It can be seen that each explores the same portion of the parameter space, and that their marginal PDFs (plotted on the right hand ordinate axis) are all very similar. Even so, the case with large proposal variance gets stuck for many successive iterations at the same parameter value, while the chain with small proposal variance moves very slowly through the parameter space. A small proposal width can also lead to problems if the chain is started at a point outside of a region with sufficient probability density (Fig. 3.1d); in this case, slow mixing may cause the chain not to encounter a region with significant probability density for many iterations and in the worst case may produce an erroneous sample of the target distribution.

The trade-off between thorough and rapid sampling gives rises to the so-called *Goldilocks principle*; it is desirable to find a proposal scale that is not too large and not too small, but just right. What is “just right”? In general, the ratio of accepted moves to proposed moves should neither be close to 0 or close to 1. If the target distribution is multivariate Gaussian, then the optimal acceptance rate is 0.23 (Gelman et al. 2004) and a value of approximately 20 % has been shown to work well for non-Gaussian posterior PDFs as well (Geyer and Thompson 1995). Roberts et al. (1997) determined that, in the limit of large dimensional state space and for posterior distribution in which each variable is i.i.d., the optimal acceptance rate is precisely 23.4 %. Roberts and Rosenthal (2001) showed that the optimal multivariate Normal proposal covariance Σ_p should be proportional to the covariance Σ of the target distribution; $\Sigma_p = k\Sigma$. Given a multivariate Normal target density with multivariate covariance Σ and dimension d, the optimal Normal proposal covariance is

$$\Sigma_p = \left[\frac{(2.38)^2}{d}\right]\Sigma. \tag{3.8}$$

Of course, the difficulty is that (1) Σ is typically not known a priori and (2) there is no guarantee the target distribution is multivariate Normal. Though care must be taken not to blindly tune to the "optimal" acceptance rate, the theory laid out in Roberts et al. (1997) and Roberts and Rosenthal (2001) serves as a useful starting point.

In practice, the following procedure has proven to work well for most problems:

(i) Run a pilot MCMC chain that generates an ensemble of realizations of $P(\mathbf{y}|\mathbf{x})$ and compute an approximate Σ, assuming the posterior is multivariate Normal.
(ii) Construct an initial Σ_p from (3.8) above.
(iii) Monitor the acceptance rate in the early stages of the algorithm and ensure it stabilizes between 10 and 60 %.

If the target distribution is far from multivariate Gaussian, then the result will still be slow mixing, but the chain will be more efficient than if left un-tuned.

The question remains: in the absence of knowledge of the true covariance of the target distribution, how can one optimally choose the proposal covariance? Adaptive algorithms are the most commonly used solution to this problem and are nearly uniformly used during a period that is referred to by most authors as "burn-in", though the term burn-in is rather confusing as it has been applied both to the practice of proposal tuning and rejection of the initial portion of the chain (see Sect. 3.3.2 below). Increasingly, adaptive algorithms are used over the length of the chain (e.g., Haario et al. 2001, 2006; Roberts and Rosenthal 2007, 2009; Vrugt et al. 2009; Vrugt and Ter Braak 2011), but a full discussion of this topic is beyond the scope of this paper. Adaptive proposal tuning is typically done as follows. The user first selects a proposal covariance (typically diagonal) under the assumption that the individual parameter proposal variances are proportional to the realistic range of values of each. A starting point for the chain is selected and Metropolis-Hastings sampling commences. After a set of n iterations, the *sample* covariance Σ_n is computed and the proposal covariance matrix is updated using these values. The new proposal covariance is then held fixed for the next m iterations, after which the most recent set of n values is used to produce an updated proposal covariance.

According to Roberts and Rosenthal (2001), the optimal proposal covariance update is

$$\Sigma_p = \left[\frac{(2.38)^2}{d}\right]\Sigma_n \tag{3.9}$$

where, as above, d is the dimension of the system of interest. Note that the subsample size n (e.g., the number of iterations used to update the covariance) and the frequency of update m are not necessarily equal, and reasonable choices (assuming a Gaussian target distribution) are documented in Haario et al. (1999). Once the acceptance rate converges to something between 10 and 60 %, then the proposal covariance is held fixed and the algorithm is allowed to run freely. It is standard practice to discard the iterations used in adaptively tuning the proposal covariance. A note of caution is merited here: for highly complex parameter covariances and large dimensions a very large number of samples may be required to robustly estimate the posterior covariance matrix. The author has found that, in cases involving a complicated posterior structure and infrequent covariance update, the wrong covariance may be specified, leading to inefficient sampling. In practice, a safe choice is to update the proposal variances only (neglecting any information on the parameter covariances), though this may result in a slightly less efficient algorithm. It should also be noted that the choice of starting point may also influence the effectiveness of parameter tuning. For example, imagine the chain starts in a very low-probability portion of the space within which the gradient in probability mass is also small. In this case, each proposed point will have very similar likelihood, the Hastings ratio will always be close to 1, and nearly all moves will be accepted. It follows that, because of the large acceptance fraction, the proposal variance will become tuned too large. Just as importantly, the sample covariances will be unrepresentative of those in the true posterior PDF.

3.3.2 The Initial Sample

There are two issues that must be considered when initiating the Markov chain at the core of the MCMC algorithm: (1) the characteristics of the posterior sample should not be sensitive to the values of the state variables at the start of the chain, and (2) it is desirable to start the chain in a region that contains relatively large probability mass. This is not only because these are the regions the sampling algorithm is designed to characterize, but also because it is not desirable to include samples in the chain that are associated with very low probability. This problem can be illustrated by considering a tutorial example in which two parameters are estimated from two observations. A random collection of 20,000 points drawn from a posterior sample generated with a high-probability start point and well-tuned proposal is shown in Fig. 3.2a. Three experiments are conducted. In the first, the start point is located in a high probability region; in the second, the start point is ten standard deviations outside the mean; and in the third, the start point is ten standard deviations

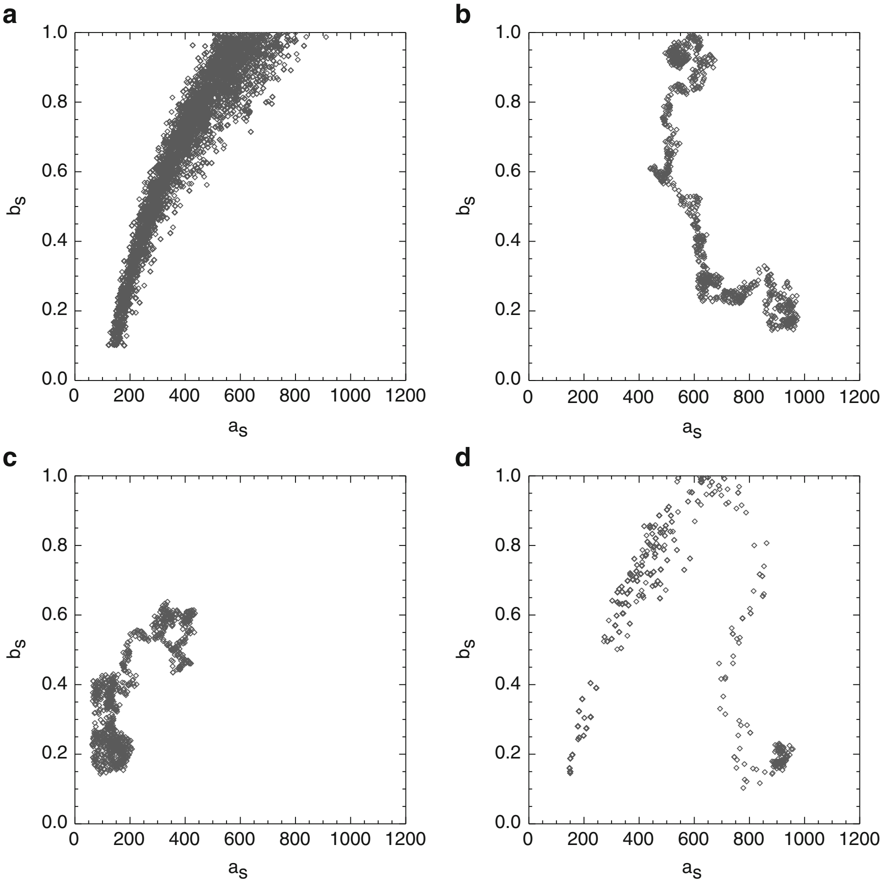


Fig. 3.2 Scatter plots of (**a**) 20,000 samples from an MCMC chain with a well tuned proposal, (**b**)–(**d**) 1,000 samples from three test MCMC chains. In (**b**), the proposal variance is too small, and the chain is started far from the posterior mode. In (**c**), the proposal variance is too small, but the chain is started near the posterior mode. In (**d**), the chain is started at a point far from the posterior mode, and with proposal variance that is initially too small. However, the proposal variance is allowed to vary according to the characteristics of the sample during the first 1,000 iterations

outside of the mean, but the proposal variance is adaptively tuned according to (3.9). The transition probability is a zero-mean multivariate Normal distribution with standard deviation equal to 0.5 % of the commonly observed range of parameter values. The proposal width is intentionally set too small to illustrate the potential problems encountered when a poor start point is chosen. Because the proposal width is small, nearly all proposed transitions are accepted (the Hastings ratio (3.5) is everywhere close to 1). It thus takes some time for the chain with initially poor start position (Fig. 3.2b) to enter a region with reasonable probability density.

As a consequence, the initial set of points in the chain is not representative of the posterior PDF. In contrast, the chain initiated within a region of relatively high probability (Fig. 3.2c) immediately begins sampling regions with relatively large probability, however, the slow rate of mixing leads to inefficient sampling. When the proposal variance is allowed to adapt (Fig. 3.2d), the chain rapidly converges to efficient posterior sampling, though it is clear that the first portion of the chain is still not representative. This example illustrates the fact that, even with a poorly chosen start point, effective tuning of the proposal distribution can produce a robust sample.

To eliminate dependence of the posterior PDF on the start point, and to ensure posterior samples are representative of the target distribution, it is common practice to discard a portion of the beginning of the Markov chain. However, there is disagreement as to how much of the chain should in general be thrown out. Some authors suggest running a chain until convergence has been determined, then discarding the first *half* of the chain (Gelman et al. 2004). Others note that, if a suitable starting point is chosen, then there is no need to discard any samples at all (Haario et al. 1999; Geyer 2011). In practice, it is difficult to know a priori whether the chosen starting point lies in a region of sufficient probability density. Diagnostic examination of the properties of the chain after a suitable number of iterations typically reveals how many samples should be discarded. In general, the Markov chain can be said to have "forgotten" the initial position once the lagged autocorrelation between a given point in the chain and the starting point is sufficiently close to zero. The resulting number of samples constitutes the minimum number that should be discarded. Comparison of the likelihood ($P(\mathbf{y}|\mathbf{x})$) of the posterior mean with likelihood values near the beginning of the chain often reveals which values near the start of the chain are associated with very low probability and should be removed. Note that we are drawing a distinction between the practice of discarding a number of initial samples and the practice of so-called burn-in, which is often conflated with tuning of the proposal distribution.

3.3.3 Single Versus Multiple Chains

Multi-core computing has become common in most research environments, and it is now standard practice to run multiple Markov chains in parallel (Gelman et al. 2004). The motivation behind doing this is primarily computational efficiency; once each chain has converged to sampling the target distribution, samples from all chains can be combined together and the sample size greatly increased in the process. In theory, this practice can be effective, provided each chain is constructed as carefully as would be done with a single chain. There are, however, a number of potential pitfalls. The first is the temptation to replace a single long chain with multiple short chains, with the goal of obtaining the same sample size in a shorter period of time. This has been used to great utility in computationally demanding problems, and for applications that require rapid solutions (e.g., Delle Monache et al. 2008). However, the possibility arises that one or more of the chains may not have converged to sampling the target distribution, and the resulting sample will not be representative of the true posterior PDF. This is related to the problem of "pseudo-convergence", which is discussed in greater detail below (and it should also be noted that multiple chains can be effectively used to diagnose convergence). Another common practice is to start multiple chains in as widely dispersed locations in the parameter space as possible, creating an "over-dispersed" initial sample (Gelman et al. 2004). As discussed above, it is not desirable to start a chain in a region of the space that contains low probability values as it may then take more time for the algorithm to find a region with relatively large probability density, and in the process the proposal may be badly mis-tuned. In most computing environments, and for most applications, it is not possible to disperse the initial points in such a way as they span the full range of possible parameter combinations. As such, some (potentially vast) regions of the space will not contain a Markov chain start point and may be left unexplored.

3.3.4 Pseudo-convergence

Because the true structure of the posterior PDF is not typically known in advance, the possibility exists that a chain that has appeared to have converged may in fact have simply spent a very long time sampling a localized probability structure. After running for a suitably long interval, the chain may make a sudden transition to a new and previously unexplored high probability region of the space. This situation is most common for posterior PDFs that exhibit multiple highly localized modes and is referred to as pseudo-convergence; the tendency for the chain in this case to appear to have converged to sampling the invariant target distribution when in reality it has not. In practice, there is no way to guarantee the chain has truly converged; the best way to safeguard against pseudo-convergence is to run very long chains. That said, it is possible that the use of a heavy-tailed proposal distribution may help to avoid this problem by making large proposed moves with greater frequency than the centered multivariate Normal. It may also be possible to make clever use of proposals from multiple chains to increase the likelihood of jumping between widely dispersed modes (e.g., Vrugt et al. 2009; Vrugt and Ter Braak 2011).

3.3.5 Diagnosing Convergence

As mentioned above, in cases for which the shape of the true posterior PDF is unknown, it is impossible to know with absolute certainty that the Markov chain has converged to sampling the invariant target distribution. Even so, there are several diagnostic tools that can be brought to bear in assessing whether the chain has at the very least pseudo-converged. In most practical applications, this may be the best one can hope for.

3.3.5.1 Time Series Plots

This is one of the simplest, yet also most effective convergence diagnostics as it leverages the powerful pattern recognition capabilities built into the human brain. It is performed by simply plotting a long time series of one or more parameters from the multidimensional parameter set. A chain can be said to have converged when it varies rapidly about a stable central value, not exhibiting a trend in the mean or changes in spread. The drawback to this technique is that it is qualitative rather than quantitative. An illustration of the sort of time series plots that are used in this analysis can be see in Fig. 3.1.

3.3.5.2 Running or Batch Moments

In addition to time series plots of the parameter values, convergence can also be diagnosed from time series of the *moments* of the sample computed in batches as the chain runs. This leverages the fact that the chain should eventually converge to sampling the stable (invariant) target distribution and as such should produce stable values of the posterior moments. Alternatively, comparison of moments for randomly selected sets of sub-samples (batches) can be done and the result should be similar to that of the running moments.

3.3.5.3 Multi-chain Convergence Diagnostics: The R-Statistic

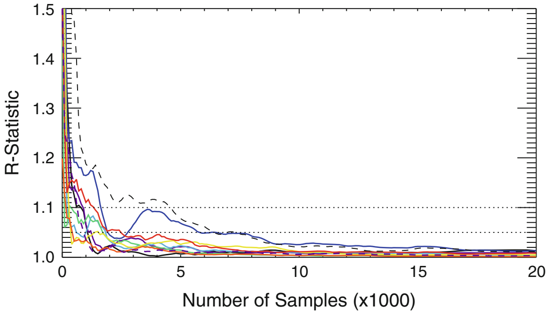
Another method of assessing convergence leverages the information contained in the differences between chains of a multi-chain MCMC simulation. This method is based on a comparison of the variance (or other moments) *within* each chain to the variance *between* chains for each estimated parameter, and is done in the following manner (Gelman et al. 2004). Consider m chains, each of length n samples. First, the within-chain variance is computed for each parameter x as

$$W = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2\right], \tag{3.10}$$

where $\overline{x_j}$ is the mean of each parameter x within each chain

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,. \tag{3.11}$$

Fig. 3.3 Plots of the R-statistic for successively greater numbers of samples (see text for derivation and explanation) for each of ten estimated parameters



The between-chain variances are computed as

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\overline{x_j} - \overline{x} \right)^2, \tag{3.12}$$

where $\overline{x}$ is the mean of the given parameter across all chains

$$\overline{x} = \frac{1}{m} \sum_{j=1}^{m} \overline{x_j}. \tag{3.13}$$

An unbiased estimate of the marginal posterior variance of each parameter x conditioned on the set of observations $\mathbf{y}$ can be obtained from a weighted combination of B and W as

$$\widehat{var}^{+}(x|\mathbf{y}) = \frac{n-1}{n} W + \frac{1}{n} B. \tag{3.14}$$

This quantity tends to overestimate the true marginal posterior variance, but converges to the true variance as $n \to \infty$. Proper chain mixing is assessed by comparing the variance estimate to the within-chain variance, and computing the *R-statistic*, $\hat{R}$, an estimate of the factor by which the dispersion in the current sample would be reduced if each chain were allowed an infinite length

$$\hat{R} = \sqrt{\frac{\widehat{var}^{+}(x|\mathbf{y})}{W}}. \tag{3.15}$$

It can be readily seen from (3.14) that this estimate will converge to 1 in the limit as $n \to \infty$. According to Gelman et al. (2004), there is no specific value of the r-statistic for which chains can be said to have sufficiently mixed, though a value of $\hat{R}$ less than 1.1 for each parameter is generally deemed acceptable. An illustration of convergence of within and between-chain variance for increasing numbers of samples for an 8-chain MCMC experiment is depicted in Fig. 3.3. It can be seen that by approximately 5,000 iterations, the chains can be assumed to have mixed sufficiently as their $\hat{R}$ value drops below 1.1, and between 5,000 and 10,000 iterations $\hat{R}$ drops below 1.05 and levels off.

3.3.6 Working with the Posterior Sample

After tuning the proposal distribution, assessing convergence, and ensuring insensitivity to the Markov chain starting point, the end result of MCMC integration is a sample of the target posterior PDF. As with any discrete sample, moments and quantiles can be computed, and the maximum likelihood estimate (or maximum a posteriori estimate; mode) can be obtained. In addition, while chain monitoring and convergence diagnostics lend confidence that the algorithm has sampled a stationary posterior distribution, it is also useful to present an estimate of the error (termed the Monte Carlo error) in the posterior sample associated with the discrete nature of the data. While it is impossible to precisely determine the error if the exact posterior distribution is unknown, it is possible to approximate it by examining the variance of the asymptotic distribution of, for example, the mean of the distribution for increasing numbers of samples in the chain. For a clear and comprehensive discussion of Monte Carlo error, the reader is referred to Flegal et al. (2008). Summary statistics may obscure some of the relevant features of the posterior sample (i.e., multi-modality), and it is common practice to analyze the data from multiple different perspectives. Plots of one-dimensional histograms of parameters **x** provide an initial indication of the center of mass and dispersion in the posterior distribution, however, integration over the remaining $d-1$ dimensions can mask inter-parameter relationships and multiple modes. For this reason, it is common practice to examine two-dimensional marginals for every pair of parameters. These are typically presented either as scatter plots or contour plots of the posterior PDF. To obtain a more robust estimate of the posterior mode and structure from the discrete sample, most studies apply a kernel density estimate (KDE) to the posterior data (Wand and Jones 1995; Tamminen and Kyrola 2001). KDE consists of multiplying every data point by a kernel function (e.g., Gaussian) with width determined from the sample (Jones et al. 1996). The result is a smoothed representation of the posterior sample that does not suffer from potential errors introduced in the specification of histogram bin widths and locations.

Because MCMC affords near infinite flexibility in the specification of the PDFs in (3.3), the posterior sample can be used to examine the error introduced by approximations made in the implementation of simpler and/or more computationally efficient posterior estimates (e.g., optimal estimation-type satellite retrievals; Rodgers 2000). In applications that involve uncertainty quantification, the posterior PDF represents the variability in a set of model output variables **y** associated with changes in a set of input parameters **x**. The posterior PDF can thus be used to examine the sensitivity of model output to changes in parameters, as well as the relationships between parameters. The degree of sensitivity in a parameter or set of parameters is directly related to the reduction in the dispersion of the prior PDF. For parameters that exert large influence over the model state, a small change in parameter values will produce a relatively large change in model output. As such, the posterior PDF will narrow relative to the prior. The degree of sensitivity can be formalized via computation of the Shannon Information content (Shannon and Weaver 1949; Rodgers 2000; Cooper et al. 2006), which is computed as the reduction in the entropy of the estimate due to the addition of information from observations. In discrete form, the entropy of state P is defined as:

$$S(P) = \sum_{i=1}^{K} p_i \log_2(p_i), \tag{3.16}$$

where p_i is the discretized PDF and K is the number of discrete bins the PDF is divided into. Shannon information content is then defined as the difference between the entropy of the a priori state and the entropy of the retrieved state

$$H = S(\mathbf{x_a}) - S(\hat{\mathbf{x}}) = \left[-\sum_{i=1}^{N} p_i(\mathbf{x_a}) \log_2(p_i(\mathbf{x_a})) \right] - \left[-\sum_{i=1}^{N} p_i(\hat{\mathbf{x}}) \log_2(pi(\hat{\mathbf{x}})) \right], \tag{3.17}$$

which can be interpreted as the extent to which the number of allowable states is reduced by the addition of information from the measurements. Because MCMC algorithms return a sample of the full PDF, H can be computed directly from the above relationship. Select examples of application of MCMC to satellite retrievals and model uncertainty evaluation are presented below.

3.4 Select Applications of MCMC in the Atmospheric Sciences

3.4.1 Retrieval of Atmospheric State Variables from Satellite Measurements

The surface-based Earth observing network is generally limited to land and to regions close to major population centers and consequently does not sufficiently sample the global atmosphere. For this reason, studies of the Earth's weather and climate have increasingly depended on observations from satellites, which have the potential to provide a more complete temporal and spatial observational record. In contrast to many in-situ measurements, satellite-based observations are *indirect* measurements of the quantities of interest. Extracting geophysical quantities from satellite radiances requires the solution to an inverse problem: inference of the state of the atmosphere from observations of radiative properties via some functional relationship (e.g., a forward radiative transfer model, regression, etc.; Rodgers 2000; Miller et al. 2000 and references therein). As such, the geophysical quantities obtained from satellite are said to be *retrieved* from what the satellite actually measures. The magnitude and characteristics of the uncertainty in a retrieval depend on the characteristics of uncertainty in the observations, their sensitivity to the parameters of interest, forward model accuracy, and the quality of available prior

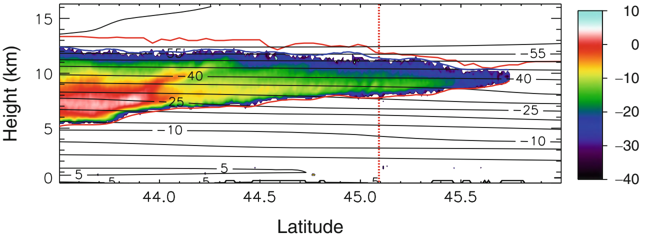


Fig. 3.4 Cross section of CloudSat radar reflectivities (*color shaded*) and CALIPSO retrieved cloud top height (*red line*). The *vertical dotted line* depicts the location of the single pixel used in the retrieval analysis (Adapted from Posselt et al. (2008a), Fig. 1b)

information about the retrieved state (Rodgers 2000; Cooper et al. 2003, 2007; L'Ecuyer et al. 2006). Though the relationship between satellite measurements and geophysical quantities may be complex and nonlinear, computational constraints often restrict retrievals to simplified frameworks. MCMC can be effectively used to robustly diagnose the uncertainty in a satellite retrieval, to improve the implementation of more efficient algorithms, and in some cases, to perform the retrieval itself (Tamminen and Kyrola 2001; Tamminen 2004).

In this section, we briefly highlight the use of an MCMC algorithm for diagnosing uncertainties in an ice cloud property retrieval. Vertically integrated ice mass (ice water path; IWP) and ice particle effective radius are obtained from relative differences in absorption of infrared radiation by clouds at two channels in the infrared window (wavelengths between 8 and 14 μm, Inoue 1985 and Prabhakara et al. 1988). Split-window retrievals provide global estimates of climate-relevant characteristics of widespread thin cirrus clouds under both daytime and nighttime conditions. Retrieved cloud properties from the split-window technique are known to be sensitive to uncertainties in cloud top height (Miller et al. 2000; Cooper et al. 2003), cloud geometric thickness (Hong et al. 2007), and ice crystal shape (Cooper et al. 2003; Baum et al. 2005). Posselt et al. (2008a) examined these sources of uncertainty by applying an MCMC algorithm to a cloud scene observed by the Moderate Resolution Imaging Spectroradiometer (MODIS). A portion of the results corresponding to analysis of the retrieval solution for a single pixel (Fig. 3.4), is presented below. Additional details on the case, as well as results from the entire scene, can be found in Posselt et al. (2008a,b).

The cloud of interest was associated with the warm frontal portion of an extratropical cyclone off the United States East Coast at 1730 UTC 22 November 2006 (Fig. 3.4). In-cloud temperatures were uniformly below −25 °C and the cloud was approximately 4 km thick. Cloud top and base were obtained from CloudSat and CALIPSO radar and lidar profiles, respectively. The forward radiative transfer model consisted of a combination of OPTRAN for gaseous transmission (Kleespies et al. 2004), and the Successive Order of Interaction (SOI) model for cloud particle scattering and absorption (Heidinger et al. 2006; O'Dell et al. 2006). Cloud properties were retrieved from MODIS brightness temperatures at 11.0 and 13.3 μm wavelengths.

The MCMC algorithm used to perform the retrieval was a straightforward implementation of Metropolis-Hastings sampling with uncorrelated zero mean Gaussian proposal distribution for each of five unknown retrieved parameters: the cloud top and base height, effective radius, ice water path, and ice crystal shape. Proposed values were generated for all parameters simultaneously and proposal variance was tuned during the initial portion of the Markov chain to converge to an acceptance rate of approximately 25 %. A Gaussian PDF was assumed for the satellite brightness temperature uncertainty, and the error standard deviation was assumed to be 1.5 and 1.0 K for the 11 μm brightness temperature and 11–13.3 μm brightness temperature difference. Cloud top and base height were obtained from CloudSat reflectivity profiles and assumed to have a Gaussian uncertainty with standard deviation of 2 km. Note that we also tested a log-Normal error distribution for cloud top height uncertainty. These results are presented in Posselt et al. (2008a) and will not be discussed here.

The prior distribution for all five retrieved parameters was assumed to be bounded Uniform with bounds set to [0,100] μm for the effective radius, [0,200] g m^{-2} for IWP, and [0,15] km for cloud top base and height. Ice crystal shape was varied by allowing the proposal to sample from all real numbers in the range [0.5,4.5] then rounding to the nearest integer value. The ice crystal shape was corresponding to this integer value was then used in the forward radiative transfer model (reordering the crystal shape index was found to have no influence on the outcome of the retrieval). Cloud base was constrained to lie below cloud top by treating any proposal with cloud base > cloud top height as an automatic rejection.

Posterior PDFs of IWP and effective radius retrieved for the pixel of interest are shown in Fig. 3.5. In Fig. 3.5a, b, the ice crystal shape is assumed to be solid columns, and the cloud top and base height are fixed (Fig. 3.5a) and allowed to vary (Fig. 3.5b). It can be seen that the characteristics of the posterior solution do not change significantly when cloud top and base are allowed to vary–the maximum a posteriori estimate (mode) is unchanged and the functional relationship between IWP and effective radius is consistent. The primary effect of variability in cloud top and base location is to increase the variance in the solution. When ice crystal shape is allowed to vary, the characteristics of the posterior PDF change markedly. Solid columns and droxtals (Fig. 3.5a, c) produce posterior PDFs with similar characteristics, but with slightly larger retrieved IWP and Re for droxtals. The PDF corresponding to bullet rosettes (Fig. 3.5d) is bimodal with a primary mode that is much more compact and circular in shape than for the other crystal shapes and with a solution that returns far smaller IWP and effective radius. In contrast, when aggregates are the assumed crystal shape (Fig. 3.5e), the posterior mode is elongated and centered at much larger values of IWP and effective radius. When the algorithm is allowed to adaptively choose a crystal shape (Fig. 3.5f), the result is bimodal with the primary mode associated with aggregates and the secondary mode a combination of bullet rosettes, droxtals, and columns.

The results demonstrate the utility of MCMC for examining the properties of a retrieval with unknown uncertainty characteristics. In the above case, while it is generally acknowledged that ice crystal shape is an important contributor to ice

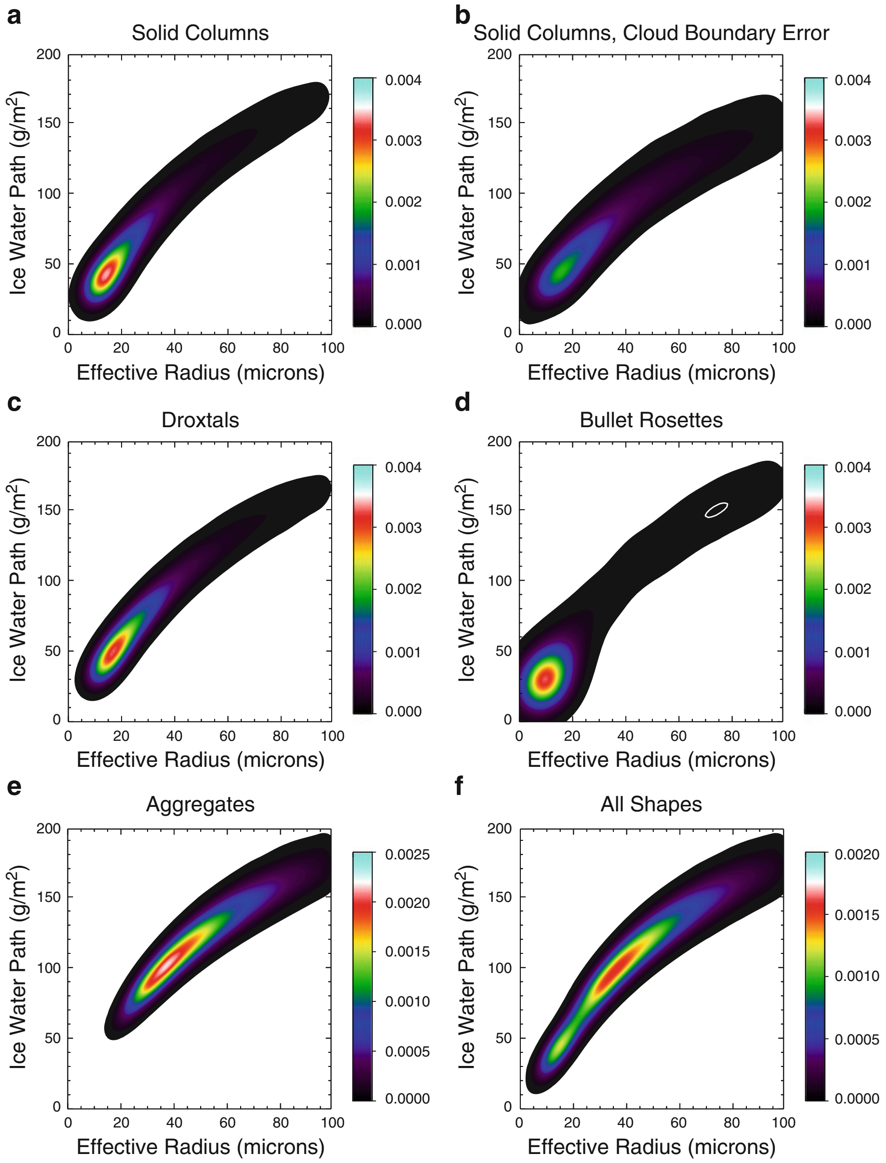


Fig. 3.5 Joint PDFs of IWP and effective radius in a single pixel for the case in which cloud top and base height are fixed (**a**, **c**, **d**, **e**, and **f**) and varied (**b**). Plots correspond to cases in which (**a**) only solid columns are allowed, (**b**) only solid columns are allowed, and cloud top and base height is allowed to vary by ±1 km, (**c**) only droxtals are allowed, (**d**) only bullet rosettes are allowed, (**e**) only aggregates are allowed, and (**f**) all crystal shapes are considered. Note that a white line has been added to (**d**) to highlight the secondary mode in the joint PDF of bullet rosettes (Adapted from Figs. 4b and 6 in Posselt et al. (2008a))

cloud retrieval error, the characteristics of the error (e.g., changes in the shape of the posterior distribution with changes in crystal type and the presence of discrete multiple modes when all types are allowed) would have been difficult to determine in advance. It is important to recognize that an MCMC-based retrieval method provides the unapproximated joint posterior PDF of all retrieved quantities for *every pixel in the scene* and thus can yield a much more robust estimate of the scene dependent characteristics of the error. This topic is explored in more detail in Posselt et al. (2008a).

3.4.2 Model Parameter Estimation and Uncertainty Analysis

Errors and/or uncertainty in model physics parameterizations are increasingly recognized to be an important source of forecast error in weather and climate prediction (Murphy et al. 2004; Palmer et al. 2005; Stainforth et al. 2005; Berner et al. 2011; Jarvinen et al. 2010, 2012; Laine et al. 2012). Specifically, empirically specified parameters associated with simplifying assumptions about the form of the particle size distribution of ice and liquid condensate have an important effect on the details of cloud and precipitation development and feed back on the radiative fluxes, heating rates, and thermodynamic environment (Tao et al. 1995; Grabowski et al. 1999; Wu et al. 1999; Petch and Gray 2001; Gilmore et al. 2004; van den Heever and Cotton 2004). It is reasonable to expect certain sets of parameters to produce model trajectories that are consistent with observations. However, due to nonlinearity in the parameter-state relationship and errors in observations, there may not exist one optimal set of parameter values. The issue of how to quantitatively represent parameterization uncertainties presents a significant challenge, and has implications for the efficacy of ensemble weather and climate forecasting, data assimilation, and model physics development.

In this section, we demonstrate how MCMC can be used to understand the functional relationship between model physics parameters and model output variables. The outcome is an estimate of the sensitivity of the simulation output to the model formulation, as well as information on how to properly account for parameter uncertainty in a data assimilation system. The parameters of interest define the particle shape, density, and size distribution in a bulk cloud microphysical parameterization (Lin et al. 1983; Rutledge and Hobbs 1983, 1984; Tao et al. 2003; Lang et al. 2007), and are listed in Table 3.1. To evaluate parameter uncertainty in isolation from the complications introduced by feedback to the flow field and thermodynamic state, the physical parameterization is driven with specified vertical motion and water vapor tendencies that vary sinusoidally with height, and change in magnitude with time. Particles are allowed to settle according to their mass weighted fall speed and interact fully with long and shortwave radiation. The thermodynamic environment, water vapor forcing, and vertical motion are set consistent with a vertical column passing through a tropical deep convective squall line. As such, the model demonstrates two distinct regimes: convective and

Table 3.1 Cloud microphysical parameters used in the MCMC-based parameter sensitivity experiments, along with truth values for the simulated observation experiment and parameter ranges. Note that all values are reported in CGS units to be consistent with what is used in the model formulation and inverse method

Parameter description	Abbreviation	Units	Truth	Min	Max
Snow fall speed coefficient	a_s	cm^{1-b_s}	200.0	50.0	1,000.0
Snow fall speed exponent	b_s	None	0.3	0.1	1.0
Graupel fall speed coefficient	a_g	cm^{1-b_g}	400.0	50.0	1,200.0
Graupel fall speed exponent	b_g	None	0.4	0.1	0.9
Slope intercept of the rain particle size distribution	N_{0r}	cm^{-4}	0.5	0.0	5.0
Slope intercept of the snow particle size distribution	N_{0s}	cm^{-4}	0.5	0.0	5.0
Slope intercept of the graupel particle size distribution	N_{0g}	cm^{-4}	0.5	0.0	5.0
Snow particle density	ρ_s	$\mathrm{g \cdot cm^{-3}}$	0.2	0.1	1.0
Graupel particle density	ρ_g	$\mathrm{g \cdot cm^{-3}}$	0.4	0.1	1.0
Threshold cloud mass mixing ratio for autoconversion to rain	q_{c0}	$\mathrm{g \cdot kg^{-1}}$	1.0	0.1	3.0

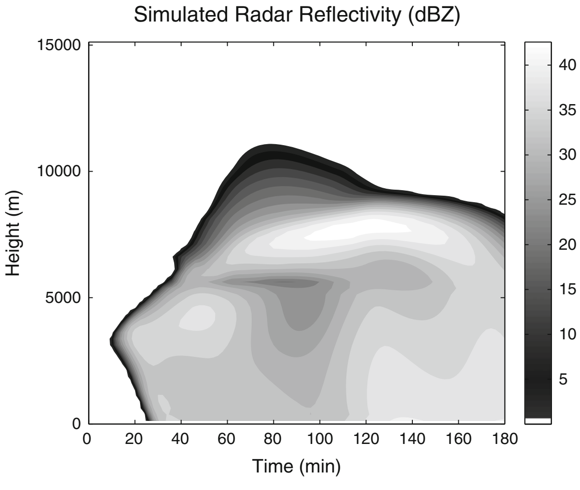


Fig. 3.6 Simulated 10-cm wavelength radar reflectivity (dBZ) for the 1D emulated squall line (Adapted from van Lier-Walqui et al. (2012), Fig. 1)

stratiform (Fig. 3.6–simulated radar reflectivity from van Lier-Walqui et al. 2012). Note that though there is no bright-band simulator in the radar forward model, the effects of melting snow and graupel are accounted for in the model. For additional details on the model configuration, the reader is referred to Posselt and Vukicevic (2010), Posselt and Bishop (2012), and van Lier-Walqui et al. (2012).

A MCMC algorithm very similar in form to that implemented for the aforementioned ice cloud property retrieval is used to examine how changes in each of ten cloud microphysical parameters affect output precipitation, liquid and ice water path, and radiative fluxes (Table 3.2) for the idealized deep convective squall line.

Table 3.2 Observations used in the MCMC-based parameter sensitivity experiments, along with their units and error estimates

Observation	Units	Error
Precipitation rate	$\mathrm{mm \cdot h^{-1}}$	$2.0\,\mathrm{mm \cdot h^{-1}}$
Liquid water path	mm	$0.5\,\mathrm{mm}$
Ice water path	mm	$1.0\,\mathrm{mm}$
TOA LW radiative flux	$\mathrm{W \cdot m^{-2}}$	$10\,\mathrm{W \cdot m^{-2}}$
TOA SW radiative flux	$\mathrm{W \cdot m^{-2}}$	$20\,\mathrm{W \cdot m^{-2}}$

As in the retrieval problem, all ten parameters are perturbed simultaneously using a Gaussian proposal distribution centered on the current parameter value and with variance that is adaptively tuned early in each Markov chain so that the acceptance rate is approximately 25 %. Parameter prior ranges are obtained from observations of cloud particle properties (Locatelli and Hobbs 1974; Mitchell 1996; Tokay and Short 1996; Heymsfield et al. 2002; Roy et al. 2005). A set of specified parameters is used to produce a true state, from which observations are drawn at 30, 60, 90, 120, 150, and 180 min of simulated time. Observations consist of precipitation rate, liquid and ice water path, and outgoing longwave and shortwave radiative fluxes, and measurement uncertainty is set equal to values consistent with error estimates on Tropical Rainfall Measuring Mission (TRMM) retrievals. Each Markov chain in the MCMC parameter estimation experiment was run for 4×10^6 iterations. Further details of the parameter values, observations, and simulation output can be found in Posselt and Vukicevic (2010).

Two dimensional marginal PDFs for select sets of parameters are shown in Fig. 3.7. The parameter sets depicted in this plot are chosen because they exhibited multi-mode posterior PDFs and a non-trivial influence on each of the output variables of interest. Each row in Fig. 3.7 corresponds to assimilation of observations with different characteristics and each demonstrates the utility of MCMC for assessing observation impact as well as the vagaries of the assimilation algorithm. Note that obtaining the results depicted in each row also required a new run of the MCMC algorithm. It can be seen from Fig. 3.7a–d that observations of precipitation rate and liquid and ice water path are not sufficient to constrain the parameter values; the mode of the joint PDF does not line up with the true parameter values. When radiative flux observations are added to the likelihood function (Fig. 3.7e–h), the number of possible solutions is reduced and the mode in the PDF lies at approximately the true value. However, there are clearly multiple modes in the solution space. In an attempt to reduce the solution to a single set of most likely parameter values, the algorithm is re-run under the assumption that more accurate observations of outgoing long and shortwave radiation are available. This serves only to exacerbate the multimodality (Fig. 3.7i–l). In essence, in the presence of a multimode solution, more accurate observations cannot serve to eliminate one of the modes, they only serve to make the modes more distinct. This may actually make the data assimilation problem more difficult, and Posselt et al. (2008a) found that this was also true of cloud property retrievals. Note that including more observation *times* may help to constrain the problem via reduction in the number of possible solutions (Vukicevic and Posselt 2008).

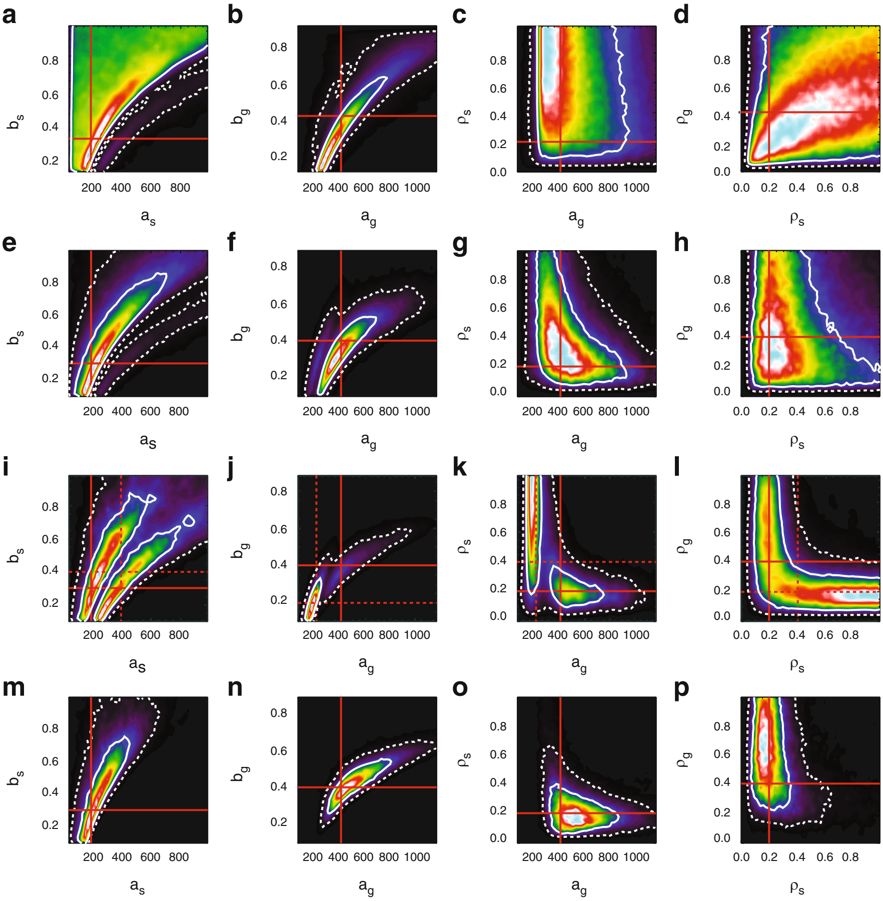


Fig. 3.7 Plots of select joint PDFs for observations of precipitation rate, liquid water path, and ice water path (PLI; *row 1*), PLI plus observations of outgoing long and shortwave radiation (PLIOO; *row 2*), PLIOO with reduced error on the radiative fluxes (*row 3*), and PLIOO with reduced error on the radiative fluxes and the constraint that the graupel fall speed coefficient exceed the snowfall speed coefficient (ag > as). *White solid* and *dashed curves* correspond to the 68 and 95 % probability contours, respectively; *red solid lines* indicate the position of the parameter truth value; and *red dashed lines* indicate the truth value for the situation in which snow and graupel parameters are interchanged (Adapted from Posselt and Vukicevic (2010), Fig. 12)

Close inspection of the posterior PDFs in Fig. 3.7i–l indicates that nonuniqueness in the relationship between snow and graupel may be the underlying cause of multiple modes in the solution. If snow and graupel are essentially interchangeable in the code, then identical solutions will be obtained if snow parameters are set to graupel truth values and vice versa. This possibility was eliminated by applying the constraint that graupel fall faster than snow; note that this is easily tested in the MCMC framework by adding another criterion to the Metropolis update that rejects

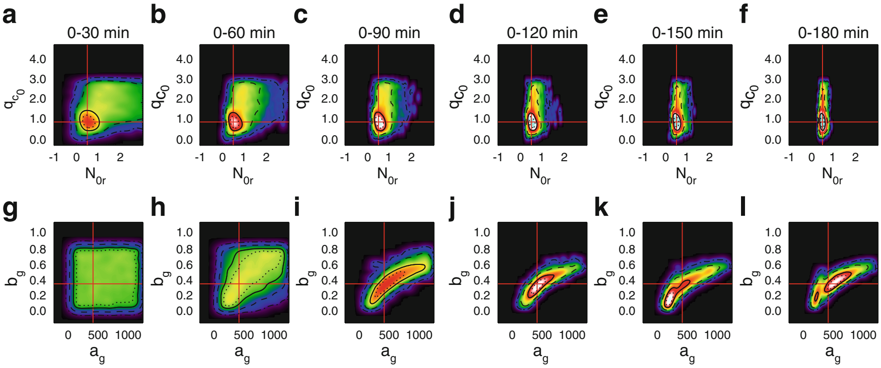


Fig. 3.8 Posterior two-dimensional marginal parameter PDFs for increasing numbers of observations assimilated in an MCMC parameter estimation experiment. Each row corresponds to a different pair of parameters, while each column represents a different set of assimilated observation times. (**a**)–(**f**) are 2D marginal PDFs of the slope intercept of the rain particle size distribution (N_{0r}) and threshold cloud mixing ratio for autoconversion to rain (q_{c0}). (**g**)–(**l**) are 2D marginal PDFs of the coefficient and exponent in the graupel diameter – fall velocity relationship. True parameter values are indicated in the *red cross-hairs*. The *thin dash-dot*, *dashed*, *solid*, and *dotted black contours* enclose the 99.7, 95, 68.3, and 38.3 % probability contours, respectively

any parameter values for which snow falls faster than graupel. The result is shown in Fig. 3.7m–p: the solution reduces to a single mode centered on the true parameter values for the snow and graupel fall speed parameters. Graupel density (Fig. 3.7p) continues to exhibit a mode that stretches from its true value to the maximum allowable value, indicating a loss of sensitivity to changes in this parameter above a certain value. The 2D marginals contained in Fig. 3.7 demonstrate the utility of MCMC for exploring the relationship (e.g., covariance) between parameters, the effect on parameter estimates of adding additional and/or more accurate observations, and the need for physical constraints in the data assimilation procedure.

Now, the MCMC algorithm is a static inverse method; control variables are not updated sequentially but are instead assumed fixed over the length of the simulation. We may address the question of when non-uniqueness in the posterior parameter PDF arises by running several MCMC experiments, and changing the number of observation times included in each one. The results of such an experiment, in which we included observations from 30, 30–60, 30–90, 30–120, 30–150, and 30–180 min in the MCMC algorithm, are presented in Fig. 3.8. Posterior PDFs for the cloud-rain autoconversion threshold and the slope intercept of the rain particle size distribution (Fig. 3.8a–f), and the graupel fall speed parameters (Fig. 3.8g–l) are shown. Several conclusions can be drawn from this figure. First, changes in the warm rain parameters affect the solution most during the convective phase of the simulation (0–90 min) with influence that wanes as the system is forced to make the transition to stratiform. The opposite is true of the graupel fall speeds, which strongly influence the solution at stratiform times, but have limited effect early in the simulation. In addition, it is clear that multiple modes do not arise in the solution until the final two observation times are used–the posterior PDF that results from assimilation of observations at 30, 60, 90, and 120 min has a single mode. The fact that multimodality arises suddenly in the solution upon incorporation of a single additional observation time is interesting, and has implications for the performance of ensemble data assimilation methods. Note that the sample of the joint posterior distribution produced by MCMC, in addition to its utility for parameter optimization and uncertainty quantification, can also be used as a benchmark for examining the characteristics of approximate solutions to inverse problems. A comparison of the posterior PDF produced by MCMC with that obtained using deterministic and stochastic versions of an Ensemble Transform Kalman Filter (ETKF, Bishop et al. 2001) is presented in Posselt and Bishop (2012).

3.5 Concluding Remarks

MCMC is now widely used for Bayesian inference in the statistical research community, and is gaining popularity in the physical and social sciences. The large dimensionality of Earth system datasets and complexity of process models present a significant computational and algorithmic challenge. This is particularly true in the case of global Earth system models and high resolution process models, which may require weeks of compute time for a single integration. For these models, running tens of thousands of integrations in a Markov chain is simply not feasible. Nevertheless, the development of efficient sampling algorithms, judicious application of simplified models, and widespread availability of multicore computing make application of MCMC to problems in atmospheric, oceanic, and hydrologic sciences increasingly feasible. As demonstrated above, significant progress is already possible in the areas of model uncertainty quantification and satellite retrievals. In particular, use of MCMC to evaluate simpler (and computationally more efficient) satellite retrieval algorithms and data assimilation schemes appears to be particularly promising. In addition, the PDF returned by MCMC can be effectively used to assess the information content in new and future observing systems, and to examine which types of observations might be used to constrain uncertain parameters in numerical models. It is likely that the next few years will see continued development of adaptive and hybrid algorithms, as well as innovative uses of MCMC for exploration of processes in the physical system. Experiments with sequential MCMC will allow more sophisticated evaluation of ensemble filters, as well as advance the development of nonlinear, non-Gaussian data assimilation techniques.

On a final note, though it is still common for scientists and statisticians to write their own MCMC software, open source codes are becoming more widely available. These include (among others) the R package "mcmc" (http://www.stat.umn.edu/geyer/mcmc/), Bayesian inference Using Gibbs Sampling (BUGS, Lunn et al. 2009, http://www.openbugs.info), the Delayed Reject Adaptive Metropolis code of Haario et al. (2006, http://www.helsinki.fi/~mjlaine/dram/), and the Differential Evolution

Adaptive Metropolis algorithm (DREAM, Vrugt and Ter Braak 2011, http://jasper.eng.uci.edu/software.html). (The author is not affiliated with or funded by any of these research efforts, and makes no claims as to the utility or effectiveness of the software.)

Acknowledgements The author gratefully acknowledges the support of the NASA Modeling, Analysis, and Prediction program under grants NNX09AJ43G and NNX09AJ46G, the Office of Naval Research Broad Agency Announcement program under grant N00173-10-1-G035, and the National Science Foundation Physical and Dynamic Meteorology Program under grant AGS 1005454. Tomislava Vukicevic (NOAA-AOML), Craig Bishop (NRL-Monterey), Marcus van Lier-Walqui (U. Miami-RSMAS), Tristan L'Ecuyer (U. Wisconsin), Steve Cooper (U. Utah), and Graeme Stephens (NASA-JPL) all contributed to this research. Luca Della Monache (NCAR), Dan Hodyss (NRL-Monterey), Peter Norris (NASA Goddard Space Flight Center), and an anonymous reviewer all contributed valuable feedback on this manuscript.

References

Baum BA, Heymsfield AJ, Yang P, Bedka ST (2005) Bulk scattering properties for the remote sensing of ice clouds: part I: microphysical data and models. J Appl Meteorol 44:1885–1895

Bennett A (1992) Inverse methods in physical oceanography. Cambridge University Press, Cambridge/New York, p 346

Berner J, Ha S-Y, 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

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

Cooper SJ, L'Ecuyer TS, Stephens GL (2003) The impact of explicit cloud boundary information on ice cloud microphysical property retrievals from infrared radiances. J Geophys Res 108. doi:10.1029/2002JD002611

Cooper S, L'Ecuyer T, Gabriel P, Baran KAJ, Stephens G (2006) Objective assessment of the information content of visible and infrared radiance measurements for cloud microphysical property retrievals over the global oceans. Part 2: ice clouds. J Appl Meteorol 45:42–62

Cooper SJ, L'Ecuyer TS, Gabriel P, Baran AJ, Stephens GL (2007) Performance assessment of a five-channel estimation-based ice cloud retrieval scheme for use over the global oceans. J Geophys Res 112:D04207. doi:10.1029/2006JD007122

Courtier P (1997) Dual formulation of four-dimensional variational assimilation. Q J R Meteorol Soc 123:2449–2461

Delle Monache L et al (2008) Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale. J Appl Meteorol climatol 47: 2600–2613

Evensen G (2006) Data assimilation: the ensemble Kalman filter. Springer, New York, p 279

Flegal JM, Haran M, Jones GL (2008) Markov chain Monte Carlo: can we trust the third significant figure? Stat Sci 23:250–260

Gelb A (1974) Applied optimal estimation. The Analytic Science Corporation/MIT, Cambridge, p 374

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC, New York

Gelman A, Brooks S, Jones G, Meng XL (2011) Handbook of Markov chain Monte Carlo: methods and applications. Chapman and Hall/CRC, New York

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741

Geyer CJ (2011) Introduction to Markov chain Monte Carlo. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo. Chapman and Hall/CRC, Boca Raton

Geyer CJ, Thompson EA (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. J Am Stat Assoc 90:909–920

Gilmore MS, Straka JM, Rasmussen EN (2004) Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. Mon Weather Rev 132:2610–2627

Grabowski WW, Wu X, Moncrieff MW (1999) Cloud-resolving modeling of tropical cloud systems during phase III of GATE. Part III: effects of cloud microphysics. J Atmos Sci 56:2384–2402

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

Haario H, Saksman E, Tamminen J (1999) Adaptive proposal distribution for random walk Metropolis algorithm. Comput Stat 14:375–395

Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis algorithm. Bernoulli 7:223–242

Haario H, Laine M, Mira A, Saksman E (2006) DRAM: efficient adaptive MCMC. Stat Comput 16:339–354

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Heidinger A, O'Dell CW, Greenwald T, Bennartz R (2006) The successive-order-of-interaction radiative transfer model. Part I: model development. J Appl Meteorol Climatol 45:1388–1402

Heymsfield AJ, Bansemer A, Field PR, Durden SL, Smith JL, Dye JE, Hall W, Grainger CA (2002) Observations and parameterizations of particle size distributions in deep tropical cirrus and stratiform precipitating clouds: results from in situ observations in TRMM field campaigns. J Atmos Sci 59:3457–3491

Hong G, Yang P, Huang H-L, Baum BA, Hu Y, Platnick S (2007) The sensitivity of ice cloud optical and microphysical passive satellite retrievals to cloud geometrical thickness. IEEE Trans Geosci Remote Sens 45:1315–1323

Inoue T (1985) On the temperature and effective emissivity determination of semi-transparent cirrus clouds by bispectral measurements in the 10 mm window region. J Meteorol Soc Jpn 63:88–99

Jarvinen H, Raisanen P, Laine M, Tamminen J, Ilin A, Oja E, Solonen A, Haario H (2010) Estimation of ECHAM5 climate model closure parameters with adaptive MCMC. Atmos Chem Phys 10:9993–10002

Jarvinen H, Laine M, Solonen A, Haario H (2012) Ensemble prediction and parameter estimation system: the concept. Q J R Meteorol Soc 138:281–288

Jazwinski H (1970) Stochastic processes and filtering theory. Mathematics in science and Engineering, vol 64. Academic, New York, pp 376

Jones MC, Marron JS, Sheather SJ (1996) A brief survey of bandwidth selection for density estimation. J Am Stat Assoc 91:401–407

Kleespies J, van Delst P, McMillin LM, Derber J (2004) Atmospheric transmittance of an absorbing Gas. 6. OPTRAN status report and introduction to the NESDIS/NCEP community radiative transfer model. Appl Opt 43:3103–3109

Laine M, Solonen A, Haario H, Jarvinen H (2012) Ensemble prediction and parameter estimation system: the method. Q J R Meteorol Soc 138:289–297

Lang S, Tao W-K, Cifelli R, Olson W, Halverson J, Rutledge S, Simpson J (2007) Improving simulations of convective systems from TRMM LBA: easterly and westerly regimes. J Atmos Sci 64:1141–1164

L'Ecuyer TS, Gabriel PM, Leesman K, Cooper SJ, Stephens GL (2006) Objective assessment of the information content of visible and infrared radiance measurements for cloud microphysical property retrievals over the global oceans. Part I: liquid clouds. J Appl Meteorol 6:20–41

Lin Y-L, Farley RD, Orville HD (1983) Bulk parameterization of the snow field in a cloud model. J Clim Appl Meteorol 22:1065–1092

Locatelli JD, Hobbs PV (1974) Fall speeds and masses of solid precipitation particles. J Geophys Res 79:2185–2197

Lorenc AC (1986) Analysis methods for numerical weather prediction. Q J R Meteorol Soc 112:1177–1194

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions (with discussion). Stat Med 28:3049–3082

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Miller SD, Stephens GL, Drummond CK, Heidinger AK, Partain PT (2000) A multisensor diagnostic satellite cloud property retrieval scheme. J Geophys Res 105:19955–19971

Mitchell DL (1996) Use of mass- and area-dimensional power laws for determining precipitation particle terminal velocities. J Atmos Sci 53:1710–1723

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772.

O'Dell CW, Heidinger AK, Greenwald T, Bauer P, Bennartz R (2006) The successive-order-of-interaction radiative transfer model. Part II: model performance and applications. J Appl Meteorol Climatol 45:1403–1413

Palmer TN, Shutts GJ, Hagedorn R, Doblas-Reyes FJ, Jung T, Leutbecher M (2005) Representing model uncertainty in weather and climate prediction. Annu Rev Earth Planet Sci 33:163–193

Petch JC, Gray MEB (2001) Sensitivity studies using a cloud-resolving model simulation of the tropical west Pacific. Q J R Meteorol Soc 127:2287–2306

Posselt DJ, Bishop CH (2012) Nonlinear parameter estimation: comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. Mon Weather Rev 140:1957–1974

Posselt DJ, Vukicevic T (2010) Robust characterization of model physics uncertainty for simulations of deep moist convection. Mon Weather Rev 138:1513–1535

Posselt DJ, L'Ecuyer TS, Stephens GL (2008a) Exploring the error characteristics of thin ice cloud property retrievals using a Markov chain Monte Carlo algorithm. J Geophys Res 113:D24206. doi:10.1029/2008JD010832

Posselt DJ, Stephens GL, Miller M (2008b) CloudSat: adding a new dimension to a classical view of extratropical cyclones. Bull Am Meteorol Soc 89:599–609

Prabhakara C, Fraser RS, Dalu G, Wu MLC, Curran RJ (1988) Thin cirrus clouds: seasonal distribution over oceans deduced from Nimbus-4 IRIS. J Appl Meteorol 27:379- 399 (1988)

Roberts GO, Rosenthal JS (1998) Optimal scaling of discrete approximations to Langevin diffusions. J R Stat Soc Ser B (Stat Methodol) 60:255–268

Roberts GO, Rosenthal JS (2001) Optimal scaling for various Metropolis-Hastings algorithms. Stat Sci 16:351–367

Roberts GO, Rosenthal JS (2007) Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J Appl Probab 44:458–475

Roberts GO, Rosenthal JS (2009) Examples of adaptive MCMC. J Comput Graph Stat 18:349–367

Roberts GO, Gelman A, Gilks WR (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann Appl Probab 7:110–120

Rodgers CD (2000) Inverse methods for atmospheric sounding, theory and practice. World Scientific, Singapore

Roy SS, Datta RK, Bhatia RC, Sharma AK (2005) Drop size distributions of tropical rain over south India. Geofizika 22:105–130

Rubin DB (1987) The calculation of posterior distributions by data augmentation: comment: a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. J Am Stat Assoc 82:543–546

Rutledge SA, Hobbs PV (1983) The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. VIII: a model for the "seeder-feeder" process in warm-frontal rainbands. J Atmos Sci 40:1185–1206

Rutledge SA, Hobbs PV (1984) The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. Part XII: a diagnostic modeling study of precipitation development in narrow cold frontal rainbands. J Atmos Sci 41:2949–2972

Sasaki Y (1970) Some basic formalisms in numerical variational analysis. Mon Weather Rev 98:875–883

Shannon C, Weaver W (1949) The mathematical theory of communication, University of Illinois Press, Champaign, p 144

Stainforth DA et al (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433:403–406

Tamminen J (2004) Validation of nonlinear inverse algorithms with Markov chain Monte Carlo method. J Geophys Res 109:D19303. doi:10.1029/2004JD004927

Tamminen J, Kyrola E (2001) Bayesian solution for nonlinear and non-Gaussian inverse problems by Markov chain Monte Carlo method. J Geophys Res 106:14377–14390

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–540

Tao W-K, Scala JR, Ferrier BS, Simpson J (1995) The effect of melting processes on the development of a tropical and a midlatitude squall line. J Atmos Sci 52:1934–1948

Tao W-K, Simpson J, Baker D, Braun S, Chou M-D, Ferrier B, Johnson D, Khain A, Lang S, Lynn B, Shie C-L, Starr D, Sui C-H, Wang Y, Wetzel P (2003) Microphysics, radiation, and surface processes in the Goddard Cumulus Ensemble (GCE) model. A special issue on non-hydrostatic mesoscale modeling. Meteorol Atmos Phys 82:97–137

Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, Philadelphia

Tierney L (1994) Markov chains for exploring posterior distributions. Ann Stat 22:1701–1762

Tokay A, Short DA (1996) Evidence from tropical raindrop spectra of the origin of rain from stratiform versus convective clouds. J Appl Meteorol 35:355–371

van den Heever SC, Cotton WR (2004) The impact of hail size on simulated supercell storms. J Atmos Sci 61:1596–1609

van Lier-Walqui M, Vukicevic T, Posselt DJ (2012) Quantification of cloud microphysical parameterization uncertainty using radar reflectivity. Mon Wea Rev 140:3442–3466

Vrugt JA, Ter Braak CJF (2011) DREAM(D): an adaptive Markov chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. Hydrol Earth Syst Sci 15:3701–3713

Vrugt J, Ter Braak C, Diks C, Robinson B, Hyman J, Hygdon D (2009) Accelerating Markov chain Monte Carlo simulation using self-adaptative differential evolution with randomized subspace sampling. Intl J Nonlinear Sci Numer Simul 10:1–12

Vukicevic T, Posselt DJ (2008) Analysis of the impact of model nonlinearities in inverse problem solving. J Atmos Sci 65:2803–2823

Wand MP, Jones MC (1995) Kernel smoothing. Chapman and Hall/CRC, London

Wu X, Hall WD, Grabowski WW, Moncrieff MW, Collins WD, Kiehl JT (1999) Long-term behavior of cloud systems in TOGA COARE and their interactions with radiative and surface processes. Part II: effects of ice microphysics on cloud-radiation interaction. J Atmos Sci 56:3177–3195

Chapter 4
Observation Influence Diagnostic of a Data Assimilation System

Carla Cardinali

Abstract The influence matrix is used in ordinary least-squares applications for monitoring statistical multiple-regression analyses. Concepts related to the influence matrix provide diagnostics on the influence of individual data on the analysis, the analysis change that would occur by leaving one observation out, and the effective information content (degrees of freedom for signal) in any sub-set of the analysed data. In this paper, the corresponding concepts are derived in the context of linear statistical data assimilation in Numerical Weather Prediction. An approximate method to compute the diagonal elements of the influence matrix (the self-sensitivities) has been developed for a large-dimension variational data assimilation system (the 4D-Var system of the European Centre for Medium-Range Weather Forecasts). Results show that, in the ECMWF operational system, 18 % of the global influence is due to the assimilated observations, and the complementary 82 % is the influence of the prior (background) information, a short-range forecast containing information from earlier assimilated observations. About 20 % of the observational information is currently provided by surface-based observing systems, and 80 % by satellite systems.

A toy-model is developed to illustrate how the observation influence depends on the data assimilation covariance matrices. In particular, the role of high-correlated observation error and high-correlated background error with respect to uncorrelated ones is presented. Low-influence data points usually occur in data-rich areas, while high-influence data points are in data-sparse areas or in dynamically active regions. Background error correlations also play an important role: high correlation diminishes the observation influence and amplifies the importance of the surrounding real and pseudo observations (prior information in observation space). To increase the observation influence in presence of high correlated background

C. Cardinali (✉)
Data Assimilation Section, European Centre for Medium-Range Weather Forecast,
Shinfield Park, Reading, Berks, RG2 9AX, UK
e-mail: c.cardinali@ecmwf.int

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_4,

error is necessary to introduce the observation error correlation but also observation and background error variances must be of similar size. Incorrect specifications of background and observation error covariance matrices can be identified, interpreted and better understood by the use of influence matrix diagnostics for the variety of observation types and observed variables used in the data assimilation system.

4.1 Introduction

Over the years, data assimilation schemes have evolved into very complicated systems, such as the four-dimensional variational system (4D-Var) (Rabier et al. 2000) at the European Centre for Medium-Range Weather Forecasts (ECMWF). The scheme handles a large variety of both space and surface-based meteorological observations. It combines the observations with prior (or background) information of the atmospheric state and uses a comprehensive (linearized) forecast model to ensure that the observations are given a dynamically realistic, as well as statistically likely response in the analysis.

Effective monitoring of such a complex system, with the order of 10^9 degrees of freedom and more than 10^7 observations per 12-h assimilation cycle, is a necessity. The monitoring cannot be restricted to just a few indicators, but a complex set of measures is needed to indicate how different variables and regions influence the data assimilation (DA) scheme. Measures of the observational influence are useful for understanding the DA scheme itself: How large is the influence of the latest data on the analysis and how much influence is due to the background? How much would the analysis change if one single influential observation were removed? How much information is extracted from the available data? It is the aim of this work to provide such analytical tools.

We turn to the diagnostic methods that have been developed for monitoring statistical multiple regression analyses. In fact, 4D-Var is a special case of the Generalized Least Square (GLS) problem (Talagrand 1997) for weighted regression, thoroughly investigated in the statistical literature.

The structure of many regression data sets makes effective diagnosis and fitting a delicate matter. In robust (resistant) regression, one specific issue is to provide protection against distortion by anomalous data. In fact, a single unusual observation can heavily distort the results of ordinary (non-robust) LS regression (Hoaglin et al. 1982). Unusual or influential data points are not necessarily bad data points: they may contain some of the most useful sample information. For practical data analysis, it helps to judge such effects quantitatively. A convenient diagnostic measures the effect of a (small) change in the observation y_i on the corresponding predicted (estimated) value $\widehat{y}_i$. In LS regression this involves a straightforward calculation: any change in y_i has a proportional impact on $\widehat{y}_i$. The desired information is available in the diagonal of the *hat matrix* (Velleman and Welsch 1981), which gives the estimated values $\widehat{y}_i$ as a linear combination of the observed values y_i. The term *hat matrix* was introduced by J.W. Tukey (Tukey 1972) because the matrix maps the observation vector $\boldsymbol{y}$ into $\hat{y}$, but it is also referred to as the *influence matrix* since its elements indicate the data influence on the regression fit of the data. The matrix elements have also been referred to as the *leverage* of the data points: in case of high *leverage* a unit y-value will highly disturb the fit (Hoaglin and Welsch 1978). Concepts related to the influence matrix also provide diagnostics on the change that would occur by leaving one data point out, and the effective information content (degrees of freedom for signal) in the data.

These influence matrix diagnostics are explained in Sect. 4.2 for ordinary least-squares regression. In Sect. 4.3 the corresponding concepts for linear statistical DA schemes is derived. It will be shown that observational influence and background influence complement each other. Thus, for any observation y_i either very large or very small influence could be the sign of inadequacy in the assimilation scheme, and may require further investigation. A practical approximate method that enables calculation of the diagonal elements of the influence matrix for large-dimension variational schemes (such as ECMWF's operational 4D-Var system) is described in Cardinali et al. (2004) and therefore not shown here. In Sect. 4.4 results and selected examples related to data influence diagnostics are presented, including an investigation into the effective information content in several of the main types of observational data. Conclusions are drawn in Sect. 4.5.

4.2 Classical Statistical Definition of Influence Matrix and Self-Sensitivity

The ordinary linear regression model can be written:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\mathbf{y}$ is an $m \times 1$ vector for the response variable (predictand); $\mathbf{X}$ is an $m \times q$ matrix of q predictors; $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters to be estimated (the regression coefficients) and $\boldsymbol{\varepsilon}$ is an $m \times 1$ vector of errors (or fluctuations) with expectation $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and covariance $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_\mathrm{m}$ (that is, uncorrelated observation errors). In fitting the model (4.1) by LS, the number of observations m has to be greater than the number of parameters q in order to have a well-posed problem, and $\mathbf{X}$ is assumed to have full rank q.

The LS method provides the solution of the regression equation as $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The fitted (or estimated) response vector $\mathbf{y}$ is thus:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \tag{4.2}$$

where

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{4.3}$$

is the $m \times m$ *influence matrix* (or hat matrix). It is easily seen that

$$\mathbf{S} = \frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}} \tag{4.4}$$

and that

$$
\begin{aligned}
S_{ij} &= \frac{\partial \hat{y}_i}{\partial y_j} \\
S_{ii} &= \frac{\partial \hat{y}_i}{\partial y_i}
\end{aligned}
\tag{4.5}
$$

for the off-diagonal ($i \neq j$) and the diagonal ($i = j$) elements, respectively. Thus, S_{ij} is the rate of change of $\hat{y}_i$ with respect to y_j variations. The diagonal element S_{ii}, instead, measures the rate of change of the regression estimate $\hat{y}_i$ with respect to variations in the corresponding observation y_i. For this reason the *self-sensitivity* (or self-influence, or leverage) of the ith data point is the ith diagonal element S_{ii}, while an off-diagonal element is a *cross-sensitivity* diagnostic between two data points.

Hoaglin and Welsch (1978) discuss some properties of the influence matrix. The diagonal elements satisfy

$$
0 \leq S_{ii} \leq 1 \, \ldots\ldots\ldots \, i = 1, 2, \ldots, m \tag{4.6}
$$

as $\mathbf{S}$ is a symmetric and idempotent projection matrix ($\mathbf{S} = \mathbf{S}^2$). The covariance of the error in the estimate $\widehat{\mathbf{y}}$, and the covariance of the residual $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ are related to $\mathbf{S}$ by

$$
\begin{aligned}
\mathrm{var}(\hat{\mathbf{y}}) &= \sigma^2 \mathbf{S} \\
\mathrm{var}(\mathbf{r}) &= \sigma^2 (I_m - \mathbf{S})
\end{aligned}
\tag{4.7}
$$

The trace of the influence matrix is

$$
tr(\mathbf{S}) = \sum_{i=1}^{m} S_{ii} = q = rank(\mathbf{S}) \tag{4.8}
$$

(in fact $\mathbf{S}$ has m eigenvalues equals to 1 and $m - q$ zeros). Thus, the trace is equal to the number of parameters. The trace can be interpreted as the amount of information extracted from the observations or *degrees of freedom for signal* (Wahba et al. 1995). The complementary trace, $tr(\mathbf{I} - \mathbf{S}) = m - tr(\mathbf{S})$, on the other hand, is the *degree of freedom for noise*, or simply the degree of freedom (*df*) of the error variance, widely used for model checking (F test).

A zero self-sensitivity $S_{ii} = 0$ indicates that the ith observation has had no influence at all in the fit, while $S_{ii} = 1$ indicates that an entire degree of freedom (effectively one parameter) has been devoted to fitting just that data point. The average self-sensitivity value is q/m and an individual element S_{ii} is considered 'large' if its value is greater than three times the average (Velleman and Welsch 1981). By a symmetrical argument a self-sensitivity value that is less than one-third of the average is considered 'small'.

Furthermore, the change in the estimate that occurs when the ith observation is deleted is

$$\hat{y}_i - \hat{y}_i^{(-i)} = \frac{S_{ii}}{(1 - S_{ii})} r_i \tag{4.9}$$

where $\hat{y}_i^{(-i)}$ is the LS estimate of y_i obtained by leaving-out the ith observation of the vector $\mathbf{y}$ and the ith row of the matrix $\mathbf{X}$. The method is useful to assess the quality of the analysis by using the discarded observation, but impractical for large systems. The formula shows that the impact of deleting $(y_i, \mathbf{x}_i)$ on $\hat{y}_i$ can be computed by knowing only the residual r_i and the diagonal element S_{ii} – the nearer the self-sensitivity S_{ii} is to one, the more impact on the estimate$\hat{y}_i$. A related result concerns the so-called cross-validation (CV) score: that is, the LS objective function obtained when each data point is in turn deleted (Wahba 1990, Theorem 4.2.1):

$$\sum_{i=1}^{m} (y_i - \hat{y}_i^{(-i)})^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{(1 - S_{ii})^2} \tag{4.10}$$

This theorem shows that the CV score can be computed by relying on the all-data estimate $\hat{\mathbf{y}}$ and the self-sensitivities, without actually performing m separate LS regressions on the leaving-out-one samples. Moreover, (4.9) shows how to compute self-sensitivities by the leaving out one experiment.

The definitions of influence matrix (4.4) and self-sensitivity (4.5) are rather general and can be applied also to non-LS and nonparametric statistics. In spline regression, for example, the interpretation remains essentially the same as in ordinary linear regression and most of the results, like the CV-theorem above, still apply. In this context, Craven and Wahba (1979) proposed the generalized-CV score, replacing in (4.10) S_{ii} by the mean $\mathrm{tr}(\mathbf{S})/q$. For further applications of influence diagnostics beyond usual LS regression (and further references) see Ye (1998) and Shen et al. (2002). The notions related to the influence matrix that it has introduced here will in the following section be derived in the context of a statistical analysis scheme used for data assimilation in numerical weather prediction (NWP).

4.3 Observational Influence and Self-Sensitivity for a DA Scheme

4.3.1 Linear Statistical Estimation in Numerical Weather Prediction

Data assimilation systems for NWP provide estimates of the atmospheric state $\mathbf{x}$ by combining meteorological observations $\mathbf{y}$ with prior (or background) information $\mathbf{x}_\mathrm{b}$. A simple Bayesian Normal model provides the solution as the posterior expectation for $\mathbf{x}$, given $\mathbf{y}$ and $\mathbf{x}_\mathrm{b}$. The same solution can be achieved from a classical *frequentist* approach, based on a statistical linear analysis scheme providing the Best Linear Unbiased Estimate (Talagrand 1997) of **x**, given **y** and $\mathbf{x}_{\mathrm{b}}$. The optimal GLS solution to the analysis problem (see Lorenc 1986) can be written

$$\mathbf{x}_a = \mathbf{K}\mathbf{y} + (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{x}_b \tag{4.11}$$

The vector $\mathbf{x}_{\mathrm{a}}$ is the 'analysis'. The gain matrix $\mathbf{K}(n \times m)$ takes into account the respective accuracies of the background vector $\mathbf{x}_{\mathrm{b}}$ and the observation vector **y** as defined by the $n \times n$ covariance matrix **B** and the $m \times m$ covariance matrix **R**, with

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} \tag{4.12}$$

Here, **H** is a $m \times n$ matrix interpolating the background fields to the observation locations, and transforming the model variables to observed quantities (e.g. radiative transfer calculations transforming the models temperature, humidity and ozone into brightness temperatures as observed by several satellite instruments). In the 4D-Var context introduced below, **H** is defined to include also the propagation in time of the atmospheric state vector to the observation times using a forecast model.

Substituting (4.12) into (4.11) and projecting the analysis estimate onto the observation space, the estimate becomes

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{x}_a = \mathbf{H}\mathbf{K}\mathbf{y} + (\mathbf{I}_m - \mathbf{H}\mathbf{K})\mathbf{H}\mathbf{x}_b \tag{4.13}$$

It can be seen that the analysis state in observation space ($\mathbf{H}\mathbf{x}_{\mathrm{a}}$) is defined as a sum of the background (in observation space, $\mathbf{H}\mathbf{x}_{\mathrm{b}}$) and the observations **y**, weighted by the $m \times m$ square matrices $\mathbf{I} - \mathbf{H}\mathbf{K}$ and $\mathbf{H}\mathbf{K}$, respectively.

Equation (4.13) is the analogue of (4.1), except for the last term on the right hand side. In this case, for each unknown component of $\mathbf{H}x$, there are two data values: a real and a 'pseudo' observation. The additional term in (4.13) includes these pseudo-observations, representing prior knowledge provided by the observation-space background $\mathbf{H}\mathbf{x}_{\mathrm{b}}$. From (4.13) and (4.4), the analysis sensitivity with respect to the observations is obtained

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \mathbf{K}^T\mathbf{H}^T \tag{4.14}$$

Similarly, the analysis sensitivity with respect to the background (in observation space) is given by

$$\frac{\partial \hat{\mathbf{y}}}{\partial (\mathbf{H}\mathbf{x}_{\mathbf{b}})} = \mathbf{I} - \mathbf{K}^T\mathbf{H}^T \quad = \mathbf{I}_m - \mathbf{S} \tag{4.15}$$

Let's focus here on the expressions (4.14) and (4.15). The influence matrix for the weighted regression DA scheme is actually more complex (see Appendix), but it obscures the dichotomy of the sensitivities between data and model in observation space.

The (projected) background influence is complementary to the observation influence. For example, if the self-sensitivity with respect to the ith observation is $\mathbf{S}_{ii}$, the sensitivity with respect the background projected at the same variable, location and time will be simply $1 - S_{ii}$. It also follows that the complementary trace, $\mathrm{tr}(\mathbf{I} - \mathbf{S}) = m - \mathrm{tr}(\mathbf{S})$, is not the df for noise but for background, instead. That is the weight given to prior information, to be compared to the observational weight $\mathrm{tr}(\mathbf{S})$. These are the main differences with respect to standard LS regression. Note that the different observations can have different units, so that the units of the cross-sensitivities are the corresponding unit ratios. Self-sensitivities, however, are pure numbers (no units) as in standard regression. Finally, as long as **R** is diagonal, (4.6) is assured (see Sect. 4.3.2), but for more general non-diagonal **R**-matrices it is easy to find counter-examples to that property.

Inserting (4.12) into (4.14), we obtain

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T \tag{4.16}$$

As $(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}$ is equal to the analysis error covariance matrix **A**, we can also write $\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}\mathbf{A}\mathbf{H}^{\mathrm{T}}$.

4.3.2 R Diagonal

In this section it is shown that as long as **R** is diagonal (4.6) is satisfied. Equation (4.16) can be written as

$$\begin{aligned}\mathbf{S} &= \mathbf{R}^{-1}\mathbf{H}[\mathbf{B} - \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}]\mathbf{H}^T \\ &= \mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^T - \mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^T\end{aligned} \tag{4.17}$$

Let's introduce the matrix $\mathbf{V} = \mathbf{H}\mathbf{B}\mathbf{H}^T$, (4.17) becomes

$$\begin{aligned}\mathbf{S} &= \mathbf{R}^{-1}\mathbf{V} - \mathbf{R}^{-1}\mathbf{V}(\mathbf{V} + \mathbf{R})^{-1}\mathbf{V} \\ &= \mathbf{R}^{-1}\mathbf{V}[\mathbf{I} - (\mathbf{V} + \mathbf{R})^{-1}\mathbf{V}] \\ &= \mathbf{R}^{-1}\mathbf{V}[(\mathbf{V} + \mathbf{R})^{-1}(\mathbf{V} + \mathbf{R}) - (\mathbf{V} + \mathbf{R})^{-1}\mathbf{V}] \\ &= \mathbf{R}^{-1}\mathbf{V}(\mathbf{V} + \mathbf{R})^{-1}\mathbf{R} \\ &= \mathbf{R}^{-1}[(\mathbf{V} + \mathbf{R})(\mathbf{V} + \mathbf{R})^{-1} - \mathbf{R}(\mathbf{V} + \mathbf{R})^{-1}]\mathbf{R} \\ &= \mathbf{R}^{-1}[\mathbf{I} - \mathbf{R}(\mathbf{V} + \mathbf{R})^{-1}]\mathbf{R} \\ &= \mathbf{I} - (\mathbf{V} + \mathbf{R})^{-1}\mathbf{R} \\ &= (\mathbf{V} + \mathbf{R})^{-1}\mathbf{V}\end{aligned} \tag{4.18}$$

Since $\mathbf{V}$ and $\mathbf{R}$ are positive definite covariance matrices, the matrix $(\mathbf{V}+\mathbf{R})$ is positive definite as well. In fact by definition for a non-zero vectors $\mathbf{z}$ with real entries the quantity $\mathbf{z}^{\mathrm{T}}(\mathbf{V}+\mathbf{R})\mathbf{z} = \mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{R}\mathbf{z} > \mathbf{0}$.

Let's consider the following theorem: If $\mathbf{D}$ is positive definite matrix then $\mathbf{D}^{-1}$ is positive definite and defining

$\mathbf{D}^{-1} = \{\delta_{\mathrm{ij}}\}$, $\mathrm{D} = \{\mathrm{d_{ij}}\}$ we have: $\delta_{\mathrm{ii}} \geq 1/\mathrm{d_{ii}}$ where the equality holds if and only if $\mathrm{d_{i1}} = \cdots = \mathrm{d_{ii-1}} = \mathrm{d_{ii+1}} = \cdots = \mathrm{d_{in}} = 0$.

The diagonal elements of $\mathbf{D}^{-1} = (\mathbf{V}\ +\mathbf{R})^{-1} = \{\delta_{\mathrm{ij}}\}$ are then larger than the diagonal elements of $(\mathbf{V}+\mathbf{R})$. Moreover, if $\mathbf{V} = \{\mathrm{v_{ij}}\}$ and $\mathbf{R} = \mathrm{diag}(\mathrm{r_i})$ we obtain

$$\delta_{ii} \geq \frac{1}{v_{ii} + r_i} \tag{4.19}$$

And since the i-diagonal element of $(\mathbf{V}+\mathbf{R})^{-1}\mathbf{R}$ is $(\delta_{i1}, \ldots, \delta_{in}) \begin{pmatrix} 0 \\ \vdots \\ r_i \\ \vdots \\ 0 \end{pmatrix} = \delta_{ii} r_i$

$$\delta_{ii} r_i \geq \frac{r_i}{v_{ii} + r_i} \tag{4.20}$$

From (4.18) considering that the product of two positive definite matrix is still a positive definite matrix

$$0 < S_{ii} = 1 - \delta_{ii} r_i \leq 1 - \frac{r_i}{v_{ii} + r_i} = \frac{v_{ii}}{v_{ii} + r_i} < 1 \tag{4.21}$$

(4.21) proves that the diagonal elements of the influence matrix for the weighted regression DA scheme are bound between (0,1).

4.3.3 Toy Model

Let's assume a simplified model with two observations, each coincident with a point of the background – that is $\mathbf{H} = \mathbf{I}_2$. Assume the error of the background at the two locations have correlation α, that is $\mathrm{B} = \sigma_b^2 \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$, with variance σ_b^2, and that similarly $\mathrm{R} = \sigma_o^2 \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$ with variance σ_o^2 and correlation β. For this simple case S is obtained from (4.14)

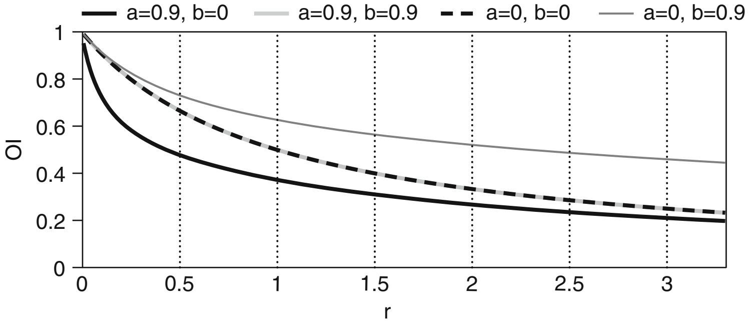


Fig. 4.1 Self-Sensitivities or Observation Influence (*OI*) as a function of the ratio between the observation error variance and the background error variance. Four different cases are shown: highly correlated **R** and uncorrelated **B** (*thick black line*). Highly correlated **R** and highly correlated **B** (*thick grey line*). Uncorrelated **R** and highly correlated **B** (*thin grey line*). Uncorrelated **R** and uncorrelated **B** (*dashed black line*)

$$S_{11} = S_{22} = \frac{\sigma_b^2\sigma_o^2(1-\alpha\beta) + \sigma_b^4(1-\alpha^2)}{\sigma_b^4(1-\alpha^2) + \sigma_o^4(1-\beta^2) + 2\sigma_b^2\sigma_o^2(1-\alpha\beta)} \tag{4.22}$$

$$S_{12} = S_{21} = \frac{\sigma_b^2\sigma_o^2(\alpha-\beta)}{\sigma_b^4(1-\alpha^2) + \sigma_o^4(1-\beta^2) + 2\sigma_b^2\sigma_o^2(1-\alpha\beta)} \tag{4.23}$$

For α ≠ ±1 and β ≠ ±1(**R** and **B** are full rank matrices). Let's define $r = \sigma_o^2/\sigma_b^2$, (4.22) and (4.23) reduce to

$$S_{11} = S_{22} = \frac{r(1-\alpha\beta) + 1 - \alpha^2}{r^2(1-\beta^2) + 1 - \alpha^2 + 2r(1-\alpha\beta)} \tag{4.24}$$

$$S_{12} = S_{21} = \frac{r(\alpha-\beta)}{r^2(1-\beta^2) + 1 - \alpha^2 + 2r(1-\alpha\beta)} \tag{4.25}$$

Figure 4.1 shows the diagonal elements of the influence matrix as a function of r, $\mathrm{S}_{\mathrm{ii}} = \mathrm{S}_{\mathrm{ii}}(r)$ (4.24). From now on, S_{ii} is also indicated as Observation Influence (*OI*). In general, the observation influence decreases with the increase of r. For highly correlated ($\alpha = 0.9$, $\beta = 0.9$) **R** and **B** and diagonal ($\alpha = 0$, $\beta = 0$) **R** and **B**, the observation influence as a function of r is the same (solid grey line and dash thick line, respectively). Maximum observation influence is achieved when **B** is diagonal ($\alpha = 0$) and **R** is highly correlated ($\beta = 0.9$) (thin black line). The observation influence will constantly decrease from the 'maximum curve' with the decrease of the correlation degree in **R** (**B** still diagonal). And the minimum observation influence curve is achieved when **R** is diagonal ($\beta = 0$) and **B** is highly correlated ($\alpha = 0.9$) (thick solid line). It is worth to notice that if the observation error variance is larger than the background error variance ($\sigma_o^2 > \sigma_b^2$) introducing the observation error correlation will slightly increase the observation influence and for $\sigma_o^2 \gg \sigma_b^2$ the observations will not be more influent in the analysis despite R is not diagonal.

(i) ***R** diagonal and **B** non-diagonal* ($\alpha \neq 0,\ \beta = 0$). Equations (4.24) and (4.25) reduce respectively to

$$S_{11} = S_{22} = \frac{r + 1 - \alpha^2}{r^2 + 1 - \alpha^2 + 2r} \tag{4.26}$$

$$S_{12} = S_{21} = \frac{r\alpha}{r^2 + 1 - \alpha^2 + 2r} \tag{4.27}$$

It can be seen that if the observations are very close compared to the scale-length of the background error correlation, i.e. $\alpha \sim 1$ (data dense area), then

$$S_{11} = S_{22} = S_{12} = S_{21} \simeq \frac{1}{r + 2} \tag{4.28}$$

Furthermore, if $\sigma_b = \sigma_o$, that is $r = 1$, we have three pieces of information with equal accuracy and $S_{11} = S_{22} = 1/3$. The background sensitivity at both locations is $1 - S_{11} = 1 - S_{22} = 2/3$. If the observation is much more accurate than the background ($\sigma_b >> \sigma_o$), that is $r \sim 0$, then both observations have influence $S_{11} = S_{22} = 1/2$, and the background sensitivities are $1 - S_{11} = 1 - S_{22} = 1/2$.

Let's now turn to the dependence on the background-error correlation α, for the case $\sigma_b = \sigma_o(r = 1)$. It is

$$S_{11} = S_{22} = \frac{2 - \alpha^2}{4 - \alpha^2} \tag{4.29}$$

$$S_{12} = S_{21} = \frac{\alpha}{4 - \alpha^2} \tag{4.30}$$

If the locations are far apart, such that $\alpha \sim 0$, then $S_{11} = S_{22} = 1/2$, the background sensitivity is also $1/2$ and $S_{12} = S_{21} = 0$. It can be concluded that where observations are sparse, S_{ii} and the background-sensitivity are determined by their relative accuracies (r) and the off-diagonal terms are small (indicating that surrounding observations have small influence). Conversely, where observations are dense, S_{ii} tends to be small, the background-sensitivities tend to be large and the off-diagonal terms are also large.

It is also convenient to summarize the case $\sigma_b = \sigma_o(r = 1)$ by showing the projected analysis at location 1

$$\hat{y}_1 = \frac{1}{4 - \alpha^2}\left[(2 - \alpha^2)y_1 + 2x_1 - \alpha(x_2 - y_2)\right] \tag{4.31}$$

The estimate $\hat{y}_1$ depends on y_1, x_1 and an additional term due to the second observation. It is noticed that, with a diagonal **R**, the observational contribution is generally devalued with respect to the background because a group of correlated background values count more than the single observation $[\alpha \rightarrow \pm 1, (2-\alpha^2) \rightarrow 1]$. From the expression above we also see that the contribution from the second observation is increasing with the correlation's absolute value, implying a larger contribution due to the background x_2 and observation y_2 nearby observation y_1.

4.4 Results

The diagonal elements of the influence matrix have been computed for the operational 4D-Var assimilation system at T159 spectral truncation 91 model levels for October 2011. For the calculation details see Cardinali et al. (2004). The observation departures ($\mathbf{y} - \mathbf{Hx}_b$) were calculated by comparing the observations with a 12-h forecast integration at T511 resolution. The assimilated observations for each main observation type are given in Table 4.1. A large proportion (∼98 %) of the used data is provided by satellite systems.

4.4.1 Trace diagnostic: Observation Influence and DFS

The global average Observation Influence (*OI*) is defined as

$$OI = \frac{tr(\mathbf{S})}{m} \tag{4.32}$$

where m is the total number of observations. For October 2011 $OI = 0.18$. Consequently, the average background global influence to the analysis at observation points is equal to 0.82 (see 4.15). It is clear that in the ECMWF system the global observation influence is quite low.

In Fig. 4.2 the *OI* for the all different observation types is plotted. In general, *OI* of conventional observations (SYNOP, DRIBU, PROFILER, PILOT, DROP, TEMP, Aircraft) is larger than the satellite one. The largest *OI* is provided by DRIBU surface pressure observations because they are located over the oceans that are in general very poor observed (less than continental areas). Moreover, DRIBU and SYNOP observations are very high quality measurements and the observation error variances is quite small, likely smaller than the background error variance (see 'toy model' in Sect. 4.3.3). Similarly, the *OI* ∼0.4–0.5 of the remaining conventional data is due to their quite small observation error variance. In Sect. 4.3.3 it has been proved that if R is diagonal the *OI* is bounded between (0,1) but from Fig. 4.2, we can see that DRIBU *OI* is higher than 1. This is due to the approximation of the numerical solution and, in particular, the use in the influence matrix calculation of

Table 4.1 Observation type assimilated on October 2011. The total number of data in one assimilation cycle is on average m ∼ 25,000,000

Data name	Data kind	Information
OZONE (O_3)	Backscattered solar UV radiation, retrievals	Ozone, stratosphere
GOES-Radiance	US geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
MTSAT-Rad	Japanese geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
MET-rad	EUMETSAT geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
AMSU-B	Microwave sounder radiances	Moisture, troposphere
MHS	Microwave sounder radiances	Moisture, troposphere
MERIS	Differential reflected solar radiation, retrievals	Total column water vapour
GPS-RO	GPS radio occultation bending angles	Temperature, surface pressure
IASI	Infrared sounder radiances	Temperature, moisture, ozone
AIRS	Infrared sounder radiances	Temperature, moisture, ozone
AMSU-A	Microwave sounder radiances	Temperature
HIRS	Infrared sounder radiances	Temperature, moisture, ozone
ASCAT	Microwave scatterometer backscatter coefficients	Surface wind
MODIS-AMV	US polar atmospheric motion vectors, retrievals	Wind, troposphere
Meteosat-AMV	EUMETSAT geostationary atmospheric motion vectors, retrievals	Wind, troposphere
MTSAT-AMV	Japanese geostationary atmospheric motion vectors, retrievals	Wind, troposphere
GOES-AMV	US geostationary atmospheric motion vectors, retrievals	Wind, troposphere
PROFILER	American, European and Japanese Wind profiles	Wind, troposphere
PILOT	Radiosondes from land stations	Wind, troposphere
DROP	Dropsondes from aircrafts	Wind, temperature, moisture, pressure, troposphere
TEMP	Radiosondes from land and ships	Wind, temperature, moisture, pressure, troposphere
Aircraft	Aircraft measurements	Wind, temperature, troposphere
DRIBU	Drifting buoys	Surface pressure, temperature, moisture, wind
SYNOP	Surface observations at land stations and on ships	Surface pressure, temperature, moisture, wind

an estimate of the analysis covariance matrix A (see Cardinali et al. 2004 for details). On the contrary, the *OI* influence of satellite data is quite small. The largest influence is provided by GPS-RO observations (∼0.4) which again are accurate data (Healy and Thépaut 2006), followed by AMSU-A measurements (∼0.3). All the other observations have an influence of about 0.2. Recently, changes in the assimilation

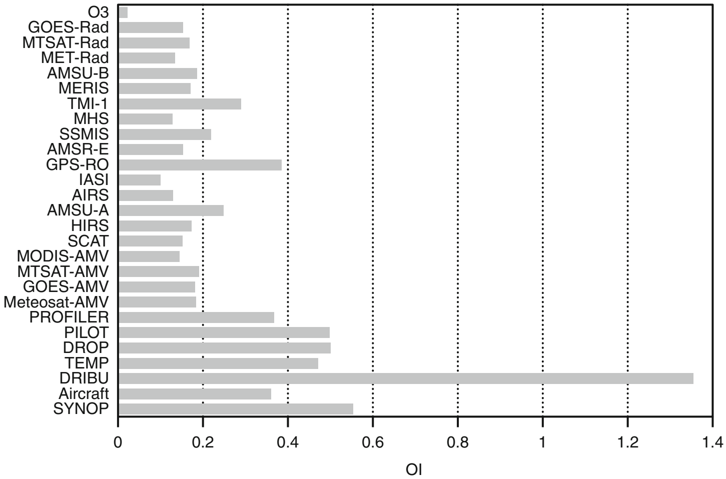


Fig. 4.2 Observation Influence (*OI*) of all assimilated observations in the ECMWF 4DVar system in October 2011. Observation types are described in Table 4.1

of 'All-Sky' observations (TMI and SSMIS) have increased their influence in the analysis (Cardinali and Prates 2011; Geer and Bauer 2011).

In Sect. 4.2 it has been shown that tr(**S**) can be interpreted as a measure of the amount of information extracted from the observations. In fact, in non-parametric statistics, tr(**S**) measures the 'equivalent number of parameters' or *degrees of freedom for signal (DFS)*. Having obtained values of all the diagonal elements of **S** (using 4.16) we can now obtain reliable estimates of the information content in any subset of the observational data. However, it must be noted that this theoretical measure of information content does not necessarily translate on value of forecast impact. Figure 4.3 shows the information content for all main observation types. It can be seen that AMSU-A radiances are the most informative data type, providing 23 % of the total observational information, IASI follows with 17 % and AIRS with 16 %. The information content of Aircraft (10 %) is the largest among conventional observations, followed by TEMP and SYNOP (∼4 %). Noticeable is the 7 % of GPS-RO (4th in the satellite *DFS* ranking) that well combines with the 0.4 value for the average observation influence. In general, the importance of the observations as defined by e.g. the *DFS* well correlates with the recent data impact studies by Radnoti et al. (2010).

Similar information content of different observation types may be due to different reasons. For example, DRIBU and OZONE information content is similarly small but whilst OZONE observations have a very small average influence (Fig. 4.2) and dense data coverage, DRIBU observations have large mean influence but much lower data counts (Fig. 4.2). Anyhow, the OZONE data are important for the ozone assimilation in spite of their low information content per analysis cycle. In fact, OZONE is generally a long-lived species, which allows observational information to be advected by the model over periods of several days.

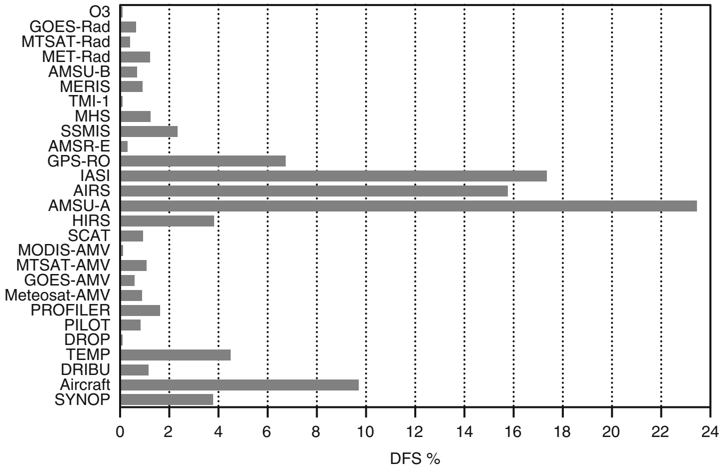


Fig. 4.3 Degree of Freedom for Signal (*DFS*) of all observations assimilated in the ECMWF 4DVar system in October 2011. Observation types are described in Table 4.1

The difference between *OI* and *DFS* comes from the number of observation assimilated. Therefore, despite the generally low observation influence of satellite measurements, they show quite large *DFS* because of the large number assimilated. A large discrepancy between *OI* and *DFS* points on those observation types where a revision of the assigned covariance matrices **R** and **B** will be beneficial: more information extracted from e.g. satellite measurements.

Another index of interest is the partial Observation Influence (OI_m) for any selected subset of data

$$OI_m = \frac{\sum\limits_{i \in I} S_{ii}}{m_I} \tag{4.33}$$

where m_I is the number of data in subset I. The subset I can represent a specific observation type, a specific vertical or horizontal domain or a particular meteorological variable. In Fig. 4.4 the *OI* of Aircraft data (I) is plotted as a function of pressure layers and for all observed parameters: temperature (t), zonal (u) and meridional (v) component of the wind. The largest *OI* is provided by temperature observations (∼0.4) similar distributed on the different pressure layers. Wind observations have larger influence (0.4) on the top of the atmosphere (above 400 hPa) than on the bottom one (0.2) due to the fact that there are very few wind observations on the troposphere and lower stratosphere mainly over the oceans. At those levels, temperature information is also provided by different satellite platforms (in terms of brightness temperature or radiance). In Fig. 4.5 the Aircraft *DFS* with respect to different pressure levels and observed parameters is shown. The largest *DFS* in the lower troposphere (below 700 hPa) for temperature measurements (∼10 % with respect to the total Aircraft *DFS*) with respect to wind ones is due to

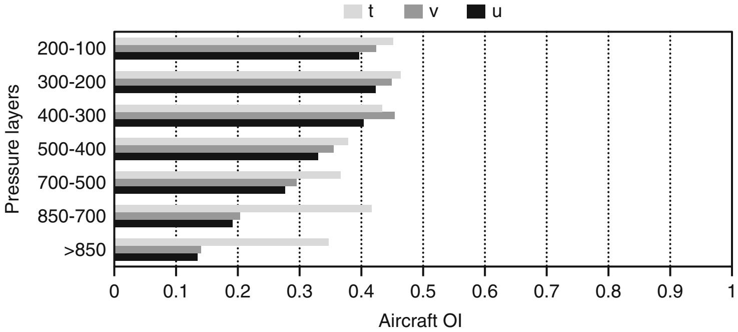


Fig. 4.4 Observation Influence (*OI*) for Aircraft observations and for October 2011 grouped by pressure layer and observed parameter. Parameters are temperature (t) *light grey bar*; meridional wind (v) *dark grey bar* and zonal wind (u) *black bar*

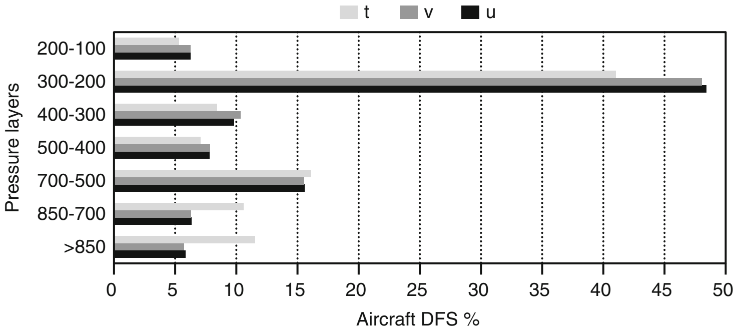


Fig. 4.5 Degree of Freedom for Signal (*DFS*) in percentage for aircraft observations and for October 2011 grouped by pressure layer and observed parameter. Parameters are temperature (t) *light grey bar*; meridional wind (v) *dark grey bar* and zonal wind (u) *black bar*. The percentage is relative to the total Aircraft DFS

the largest temperature influence. For all the other levels, the *DFS* is quite similar to the *OI* distribution with the exception of the layer from 200 to 300 hPa where the increase to ∼50 % is due to the increase of number of observations assimilated. Figures 4.6 and 4.7 shows the AMSU-A *OI* and *DFS*, respectively, for all the channels assimilated. A large part of the AMSU-A information is with respect to stratospheric temperature and the largest *OI* at that atmospheric layer is from channel 9 to 10 (∼0.4) (Fig. 4.6). Channel 5 (∼700 hPa) shows a very large ∼0.8 *OI*, the largest influence among all the channels. The reason of this large *OI* is unclear, and investigation is in due course to understand the cause. The channels observation influence distribution is similar to the *DFS* distribution (Fig. 4.7): channel 9 and 10 count for 18 % of the AMSU-A *DFS* and channel 5 for 24 %.

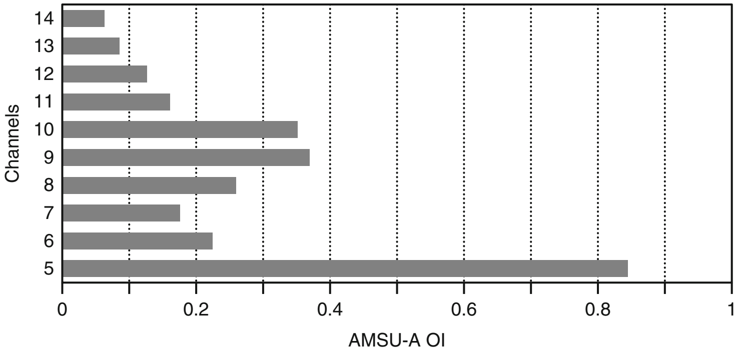


Fig. 4.6 Observation Influence (*OI*) for AMSU-A observations and for October 2011 grouped by channels

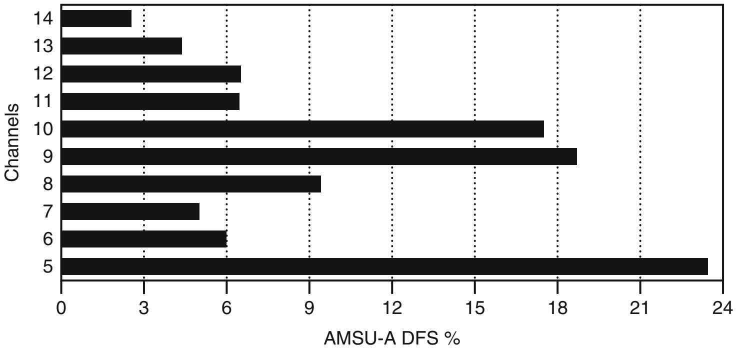


Fig. 4.7 Degree of Freedom for Signal in percentage (*DFS*) for AMSU-A observations and for October 2011 grouped by channels. The percentage is relative to the total *AMSU-A DFS*

4.4.2 Geographical Map of OI

The geographical map of observation influence for SYNOP and DRIBU surface pressure observations is shown in Fig. 4.8. Each box indicates the observation influence per observation location averaged among all the October 2011 measurements. Data points with influence greater than one are due to the approximation of the computed diagonal elements of influence matrix (see Cardinali et al. 2004). Low-influence data points have large background influence (see 4.14 and 4.15), which is the case in data-rich areas such as North America and Europe (observation influence ∼0.2) (see also Sect. 4.3.3). In data-sparse areas individual observations have larger influence: in the Polar regions, where there are only few isolated observations, the *OI* is very high (theoretically ∼1) and the background has very small influence on the analysis.

In dynamically active areas (Fig. 4.8: e.g. North Atlantic and North Pacific), several fairly isolated observations have large influence on the analysis. This is

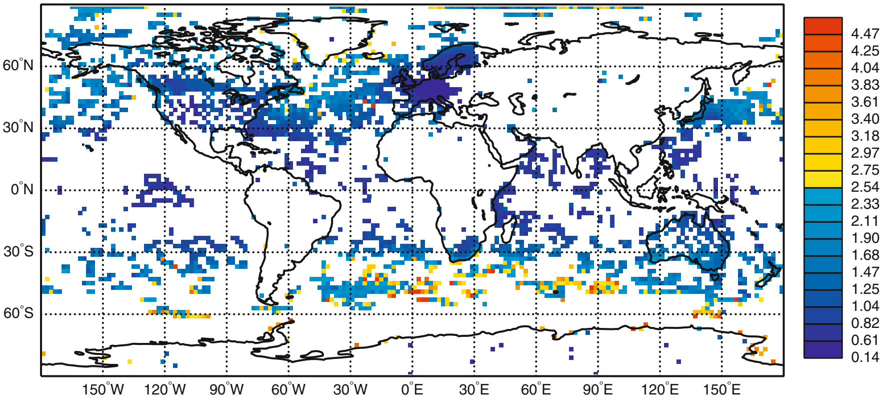


Fig. 4.8 Observation Influence (*OI*) of SYNOP and DRIBU surface pressure observations for October 2011. High influential points are close to 1 and low influential points are close to 0

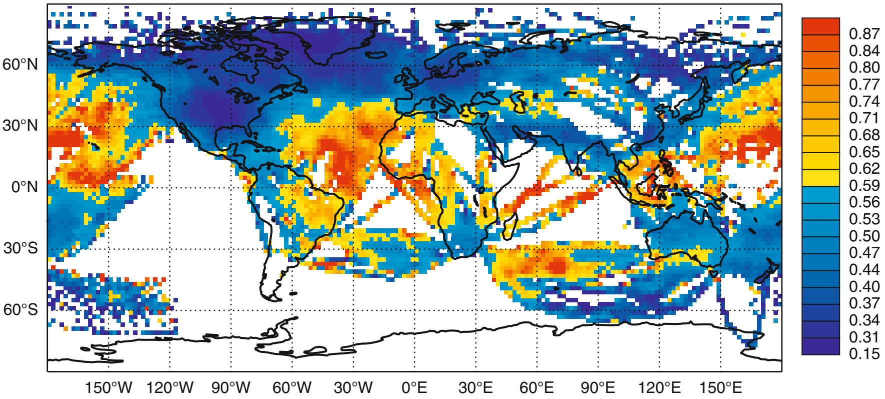


Fig. 4.9 Observation Influence (*OI*) of Aircraft zonal wind component above 400 hPa for October 2011. High influential points are close to 1 and low influential points are close to 0

also due to the evolution of the background-error covariance matrix as propagated by the forecast model in 4D-Var (Thépaut et al. 1993, 1996). As a result, the data assimilation scheme can fit these observations more closely.

Similar features can be seen in Fig. 4.9, which shows the influence of u-component wind observations for Aircraft data above 400 hPa. Isolated flight tracks over Atlantic and Pacific oceans show larger influences than measurements over data-dense areas over America and Europe. The flight tracks over North Atlantic and North Pacific are also in dynamically active areas where the background error variances are implicitly inflated by the evolution of the background-error covariance matrix in the 4D-Var window. Figure 4.10 shows the geographical distribution of AMSU-A channel 8 observation influence. The largest influence is noticed in the extra-tropics and polar areas ($\sim$0.4) whilst in the

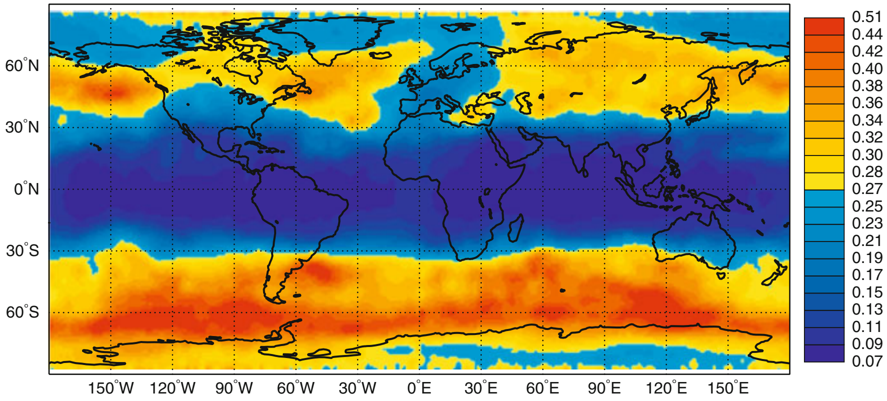


Fig. 4.10 Observation Influence (*OI*) of AMSU-A channel 8 for October 2011. High influential points are close to 1 and low influential points are close to 0

tropics the maximum *OI* is ∼0.12. Since channel 8 observation error variances are geographically constant the main difference in the observed *OI* pattern is likely due to the **B** covariance matrix. It looks that the background error correlation are higher in the tropics than in the extra-tropics.

4.5 Conclusions

The influence matrix is a well-known concept in multi-variate linear regression, where it is used to identify influential data and to predict the impact on the estimates of removing individual data from the regression. In this paper the influence matrix in the context of linear statistical analysis schemes has been derived, as used for data assimilation of meteorological observations in numerical weather prediction (Lorenc 1986). In particular an approximate method to compute the diagonal elements of the influence matrix (the self-sensitivities or observation influence) in ECMWF's operational data assimilation system (4D-Var) has been derived and implemented. The approach necessarily approximates the solution due to the large dimension of the estimation problem at hand: the number of estimated parameters is of the order 10^9, and the number of observational data is around $25{\cdot}10^6$.

The self-sensitivity provides a quantitative measure of the observation influence in the analysis. In robust regression, it is expected that the data have similar self-sensitivity (sometimes called leverage) – that is, they exert similar influence in estimating the regression line. Disproportionate data influence on the regression estimate can have different reasons: First, there is the inevitable occurrence of incorrect data. Second, influential data points may be legitimately occurring extreme observations. However, even if such data often contain valuable information, it is constructive to determine to which extent the estimate depends on these data.

Moreover, diagnostics may reveal other patterns e.g. that the estimates are based primarily on a specific sub-set of the data rather than on the majority of the data.

In the context of 4D-Var there are many components that together determine the influence given to any one particular observation. First there is the specified observation error covariance **R**, which is usually well known and obtained simply from tabulated values. Second, there is the background error covariance **B**, which is specified in terms of transformed variables that are most suitable to describe a large proportion of the actual background error covariance. The implied covariance in terms of the observable quantities is not immediately available for inspection, but it determines the analysis weight given to the data. Third, the dynamics and the physics of the forecast model propagate the covariance in time, and modify it according to local error growth in the prediction. The influence is further modulated by data density. Examples for surface pressure and aircraft wind observations have been shown indicating that low influence data points occur in data-rich areas while high influence data points are in data-sparse regions or in dynamically active areas. Background error correlations also play an important role. In fact, very high correlations drastically lessen the observation influence (it is halved in the idealized example presented in Sect. 4.3.3) in favour of background influence and amplify the influence of the surrounding observations. The observation influence pattern of AMSU-A channel 8 suggests some affectation of the correlation expresses by the **B** covariance matrix.

The global observation influence per assimilation cycle has been found to be 18 %, and consequently the background influence is 82 %. Thus, on average the observation influence is low compared to the influence of the background (the prior). However, it must be taken into account that the background contains observation information from the previous analysis cycles. The theoretical information content (the degrees of freedom for signal) for each of the main observation types was also calculated. It was found that AMSU-A radiance data provide the most information to the analysis, followed by IASI, AIRS, Aircraft, GPS-RO and TEMP. In total, about 20 % of the observational information is currently provided by surface-based observing systems, and 80 % by satellite systems. It must be stressed that this ranking is not an indication of relative importance of the observing systems for forecast accuracy. Nevertheless, recent studies on the 24-h observation impact on the forecast with the adjoint methodology have shown similar data ranking (Langland and Baker 2004; Zhu and Gelaro 2008; Cardinali 2009; Cardinali and Prates 2011).

If the influence matrix were computed without approximation then all the self-sensitivities would have been bounded in the interval zero to one. With the approximate method used, out-of-bound self-sensitivities occur if the Hessian representation based on an eigen-vector expansion is truncated, especially when few eigen-vectors are used. However, it has been shown that this problem affects only a small percentage of the self-sensitivities computed, and in particular those that are closer to one.

Self-sensitivities provide an objective diagnostic on the performance of the assimilation system. They could be used in observation quality control to protect against distortion by anomalous data; this aspect has been explored by

Junjie et al. (2009) in the context of Ensemble Kamlan Filter where the **B** is well known and the solution for the diagonal element of the Influence matrix is therefore very accurate. Junjie et al. have shown that the leaving-out-one observation, not practical for large system dimension, can be replaced by the Self-sensitivities (4.9) that provide a similar diagnostic without performing separate least square regressions. Self-sensitivities also provide indication on model and observation error specification and tuning. Incorrect specifications can be identified, interpreted and better understood through observation influence diagnostics, partitioned e.g. by observation types, variable, levels, and regions.

In the near future more satellite data will be used and likely be thinned. Thinning has to be performed either to reduce the observation error spatial correlation (Bormann et al. 2003) or to reduce the computational cost of the assimilation. The observation influence provides an objective way of selecting observations dependent on their local influence on the analysis estimate to be used in conjunction with forecast impact assessments. Recently, Bauer et al. (2011) have shown that satellite measurements in sensitive areas as defined by singular vectors methodology have larger impact in the forecast than measurements in different regions and also larger or similar impact than the full amount of data. In this case, a dynamical thinning can be thought that selects, at every assimilation cycle, the most influent measurements partition of a particular remote sensing instrument, from information based on the previous cycle (see also Rabier et al. 2002). Clearly, it can be assumed that components of the observing network remain constant and the background error variances remain almost unchanged for close assimilation cycles.

Acknowledgements The author thanks Olivier Talagrand and Sergio Pezzulli for the fruitful discussions on the subject. Many thanks are given to Mohamed Dahoui and Anne Fouilloux for their precious technical support.

Appendix

Influence Matrix Calculation in Weighted Regression Data Assimilation Scheme

Under the *frequentist* approach, the regression equations for observation

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}_o$$

and for background

$$\mathbf{x}_b = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_b$$

are assumed to have uncorrelated error vectors $\boldsymbol{\varepsilon}_\mathbf{o}$ and $\boldsymbol{\varepsilon}_\mathrm{b}$, zero vector means and variance matrices **R** and **B**, respectively. The $\boldsymbol{\theta}$ parameter is the unknown system state (**x**) of dimension n. These regression equations are summarized as a weighted regression

$$\mathbf{z} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\mathbf{z} = \left[\mathbf{y}^T \mathbf{x}_b^T\right]^T$ is $(m+n) \times 1$; $\mathbf{X} = \left[\mathbf{H}^T \mathbf{I}_n\right]^T$ is $(m+n) \times n$ and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_b]^T$ is $(m+n) \times 1$ with zero mean and variances matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{B} \end{pmatrix}$$

The generalized LS solution for $\boldsymbol{\theta}$ is BLUE and is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{z} \tag{4.34}$$

see Talagrand (1997). After some algebra this equation equals (4.11). Thus

$$\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\theta}} = \left[\mathbf{H}^T \mathbf{x}_a^T \mathbf{x}_a^T\right]^T = \mathbf{X}(\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{z}$$

and by (4.5) the influence matrix becomes

$$\mathbf{S}_{\mathrm{zz}} = \frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}} = \frac{\partial \hat{\boldsymbol{\theta}}}{\partial \mathbf{z}} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yb} \\ \mathbf{S}_{by} & \mathbf{S}_{bb} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1}\mathbf{H}\mathbf{A}\mathbf{H}^T & \mathbf{R}^{-1}\mathbf{H}\mathbf{A} \\ \mathbf{B}^{-1}\mathbf{A}\mathbf{H}^T & \mathbf{B}^{-1}\mathbf{A} \end{pmatrix}$$

where $\mathbf{S}_{yy} = \frac{\partial \mathbf{H}\mathbf{x}_a}{\partial \mathbf{y}}$; $\mathbf{S}_{yb} = \frac{\partial \mathbf{x}_a}{\partial \mathbf{y}}$; $\mathbf{S}_{by} = \frac{\partial \mathbf{H}\mathbf{x}_a}{\partial \mathbf{x}_b}$; $\mathbf{S}_{bb} = \frac{\partial \mathbf{x}_a}{\partial \mathbf{x}_b}$. Note that $\mathbf{S}_{\mathrm{yy}} = \mathbf{S}$ as defined in (4.4).

Generalized LS regression is different from ordinary LS because the influence matrix is not symmetric anymore. For idempotence, using (4.33) it easy to show that $\mathbf{S}_{\mathrm{zz}}\mathbf{S}_{\mathrm{zz}} = \mathbf{S}_{\mathrm{zz}}$. Finally,

$$\mathbf{S}_{bb} = \mathbf{B}^{-1}\mathbf{A} = \mathbf{I}_n - \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}\mathbf{A}$$

hence,

$$tr(\mathbf{S}_{bb}) = n - tr(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}\mathbf{A}) = n - tr(\mathbf{S}_{yy})$$

it follows that

$$tr(\mathbf{S}_{zz}) = tr(\mathbf{S}_{yy}) + tr(\mathbf{S}_{bb}) = n$$

The trace of the influence matrix is still equal to the parameter's dimension.

References

Bauer P, Buizza R, Cardinali C, Thépaut J-l (2011) Impact of singular vector based satellite data thinning on NWP. Q J R Meteorol Soc 137:286–302

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 Wea Rev 131:706–718

Cardinali C (2009) Monitoring the forecast impact on the short-range forecast. Q J R Meteorol Soc 135:239–250

Cardinali C, Prates F (2011) Performance measurement with advanced diagnostic tools of all-sky microwave imager radiances in 4D-Var.Q J R Meteorol Soc 137(Issue 661, Part B):2038–2046

Cardinali C, Pezzulli S, Andersson E (2004) Influence matrix diagnostics of a data assimilation system. Q J R Meteorol Soc. 130:2767–2786

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403

Geer AJ, Bauer P (2011) Observation errors in all-sky data assimilation. Q J R Meteorol Soc 137(Issue 661, Part B):2024–2037

Healy SB, Thépaut J-N (2006) Assimilation experiments with CHAMP GPS radio occultation measurements. Q J R Meteorol Soc 132:605–623

Hoaglin DC, Welsch RE (1978) The hat matrix in regression and ANOVA. Am Stat 32:17–22, and Corrigenda 32:146

Hoaglin DC, Mosteller F, Tukey JW (1982) Understanding robust and exploratory data analysis. Wiley Series in Probability and Statistics. Wiley, New York

Junjie L, Kalnay E, Miyoshi T, Cardinali C (2009) Analysis sensitivity calculation within an ensemble Kalman filter. Q J R Meteorol Soc 135:1842–1851

Langland R, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201

Lorenc A (1986) Analysis methods for numerical weather prediction. Q J R Meteorol Soc 112:1177–1194

Rabier F, Järvinen H, Klinker E, Mahfouf JF, Simmons A (2000) The ECMWF operational implementation of four-dimensional variational assimilation. Part I: experimental results with simplified physics. Q J R Meteorol Soc 126:1143–1170

Rabier F, Fourrié N, Chafaï D, Prunet P (2002) Channel selection methods for infrared atmospheric sounding interferometer radiances. Q J R Meteorol Soc 128:1011–1027

Radnoti G, Bauer P, McNally A, Cardinali C, Healy S, de Rosnay P (2010) ECMWF study on the impact of future developments of the space-based observing system on Numerical Weather Prediction. ECMWF Tec. Memo 638

Shen X, Huang H, Cressie N (2002) Nonparametric hypothesis testing for a spatial signal. J Am Stat Ass 97:1122–1140

Talagrand O (1997) Assimilation of observations, an introduction. J Meteorol Soc Japan 75(1B):191–209

Thépaut JN, Hoffman RN Courtier P (1993) Interactions of dynamics and observations in a four-dimensional variational assimilation. Mon Weather Rev 121:3393–3414

Thépaut JN, Courtier P, Belaud G Lemaître G (1996) Dynamical structure functions in four-dimensional variational assimilation: a case study. Q J R Meteorol Soc 122:535–561

Tukey JW (1972) Data analysis, computational and mathematics. Q Appl Math 30:51–65

Velleman PF, Welsch RE (1981) Efficient computing of regression diagnostics. Am Stat 35:234–242

Wahba G (1990) Spline models for observational data. SIAM, CBMS-NSF. Regional conference series in applied mathematics, vol 59. Society for Industrial and Applied Mathematics, Philadelphia, p 165

Wahba G, Johnson DR, Gao F, Gong J (1995) Adaptive tuning of numerical weather prediction models: randomized GCV in three- and four-dimensional data assimilation. Mon Weather Rev 123:3358–3369

Ye J (1998) On measuring and correcting the effect of data mining and model selection. J Am Stat Ass 93:120–131

Zhu Y, Gelaro R (2008) Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. Mon Weather Rev 136:335–351

Chapter 5
A Question of Adequacy of Observations in Variational Data Assimilation

John M. Lewis and S. Lakshmivarahan

Abstract The adequacy of observations to locate the minimum of the standard cost function for variational data assimilation under strong constraint has been investigated. A simplified yet meaningful Lagrangian air/sea interaction model that captures key aspects of air mass modification over the Gulf of Mexico in wintertime is the dynamical tool used to examine this question of adequacy. Two mathematically different yet equivalent variational schemes are used in numerical experiments with a fixed number of observations along a prior known trajectory over the Gulf. Research clearly indicates that sensitivity of model output to elements of control (initial condition, boundary condition, and physical parameter) is key to placement of observations in order to minimize the cost function and determine optimal corrections to control.

5.1 Introduction

From the early days of numerical weather map analysis as an aid to numerical weather prediction (NWP) (Wiin-Nielsen 1991), the adequacy of observations to produce an analysis faithful to the weather and consistent with the dynamical model has been an ever-present concern. In the earliest numerical map analysis by Bergthórsson and Döös (1955) that spanned the North Pacific Ocean and the bounding continental areas (northern Europe and eastern Canada), a climatological

J.M. Lewis (✉)
National Severe Storms Laboratory, Norman, OK, USA

Desert Research Institute, Reno, NV, USA
e-mail: jlewis@dri.edu

S. Lakshmivarahan
School of Computer Science, University of Oklahoma, Norman, OK 73019, USA
e-mail: varahan@ou.edu

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_5,

background field for the 500 mb geopotential field was required to produce a meaningful upper-air analysis over the ocean. The observations alone were insufficient to yield a product useful for NWP.

The Bergthórsson-Döös analysis was non-optimal, yet it provided guidance for one of the first optimal data assimilation methods in meteorology—the optimal or statistical interpolation method [generally referred to as OI method (See Lewis et al. 2006)]. The OI method optimally fit the analysis to both background (forecast from an earlier time) and observations in accord with relative accuracy of these inputs. It fundamentally depends on the statistical structure of errors associated with the background forecast. Problems can develop with OI when the background error covariance matrix is "almost singular" which occurs when the observations are clustered together and lack any sense of distributional uniformity—related to the inadequacy of observations. This situation is made all the worse when the observational errors are small [See the carefully crafted quotation by J. Purser on the subject in Lewis and Lakshmivarahan (2008)].

When the four-dimensional data assimilation method using adjoint equations (4D-Var with Adjoint) arrived on the operational scene in the late 1980s [LeDimet and Talagrand (1986), Lewis and Derber (1985), and Thacker and Long (1988)], the minimization of the cost function through "steepest descent" was the philosophy to find the optimal control vector of the model (initial conditions, boundary conditions, and physical/empirical parameters)—the control that minimized the squared departure between forecast and observations under the "strong" dynamical constraint (exact satisfaction of the dynamical law). The method has esthetic appeal with its foundation in calculus of variations and it also possesses a utilitarian component through its efficiency in calculating the gradient of the cost function. However, in the presence of the complex dynamics of atmospheric flow, it is not unusual to encounter "flatness" in the geometric structure of the cost function in the space of control and this poses problems for steepest descent-type algorithms. As explored by Thacker (1989), the insufficiency issue benefits from an examination of the Hessian matrix, the matrix of second derivatives of the cost function with respect to control. The eigenvalues and associated eigenvectors of the Hessian about the terminal iterated state reveal the structure of the cost function. The eigenvectors point along the principal axes of the ellipse of constant cost and the eigenvalues determine the lengths of the semiaxes. As the eigenvalue approaches zero, the semiaxis approaches infinity and the Hessian approaches singularity. The cost function surface should curve upward steeply in directions associated with well-determined elements of control. When the surface is flat along some directions, this is a sign of ill-conditioning of the optimization problem—another way of saying the observations are inadequate for finding the optimal state. For the high-dimension nonlinear dynamics of weather prediction, it is most challenging to determine the characteristics of the Hessian in spite of efforts to simplify the problem and make it tractable (Lewis et al. 2006).

In those cases of ill-conditioning, and where acquisition of more observations is difficult or impossible, the use of prior knowledge such as climatology or a forecast from an earlier time is the most reasonable avenue of pursuit to rid the problem of ill-conditioning—an augmentation to the cost function including terms that fit the model to the prior in accord with the representativeness and accuracy of the prior. But as will be explored in this paper, if the number of observations is fixed but free to be moved (in space and/or time), there is a strategy that can generally remove the ill-conditioned nature of the problem.

We numerically test this idea of moveable observations in the context of a simplified Lagrangian air/sea interaction model with relevance to air mass modification over the Gulf of Mexico in wintertime. The strategy revolves around knowledge of the sensitivities of the model variables to elements of control—generally obtained by integration of equations similar to the dynamical equations of the model. The examination of sufficiency/insufficiency of observations in variational data assimilation will be explored through use of two schemes: Forward Sensitivity Method (FSM) and the variational method with $-\nabla J$ (negative gradient of the cost function) serving to determine the search direction. For simplicity, we refer to this latter method as the ∇J-method. In an earlier paper, Lakshmivarahan and Lewis (2010) have proved the equivalence of these two schemes. Since little is known about FSM, the mechanics of this scheme will be developed after rudiments of the air/sea interaction model are presented. Numerical experiments follow and the paper ends with conclusions and a discussion on applicability of these ideas to the more challenging dynamics of NWP.

5.2 Model Dynamics: Air/Sea Interaction

5.2.1 Background Physical Processes

We consider a persistent operational weather prediction problem that has plagued modelers and weather forecasters at the National Center for Environmental Prediction (NCEP)/Environmental Modeling Center (EMC) for several decades [from the late 1980s to the present day; reviewed in Lewis (2007)]. The problem occurs in association with air mass modification over the Gulf of Mexico in the cool season. These events occur 4–5 times per month from November to March. The phenomenon is labeled "return flow" since air that enters the Gulf exhibits anticyclonic turning and returns to the coast as the cold high pressure system moves eastward. A schematic diagram of the process is depicted in Fig. 5.1. In the top portion of this figure, the cold front is shown entering the Gulf with the attendant low/high pressure couplet to the north. As the front moves through the Gulf along with the eastward movement of the cyclone/anticyclone couplet, low-level southerly winds at LCH (Lake Charles, LA) and BRO (Brownsville, TX) shift to northerly and easterly, respectively. From this wind structure it is clear that the air entering the Gulf near LCH will move southward and then westward along an over-water trajectory. The persistent problem faced by forecasters is bias in the numerical prediction of low-level temperature and water vapor as the air is modified over

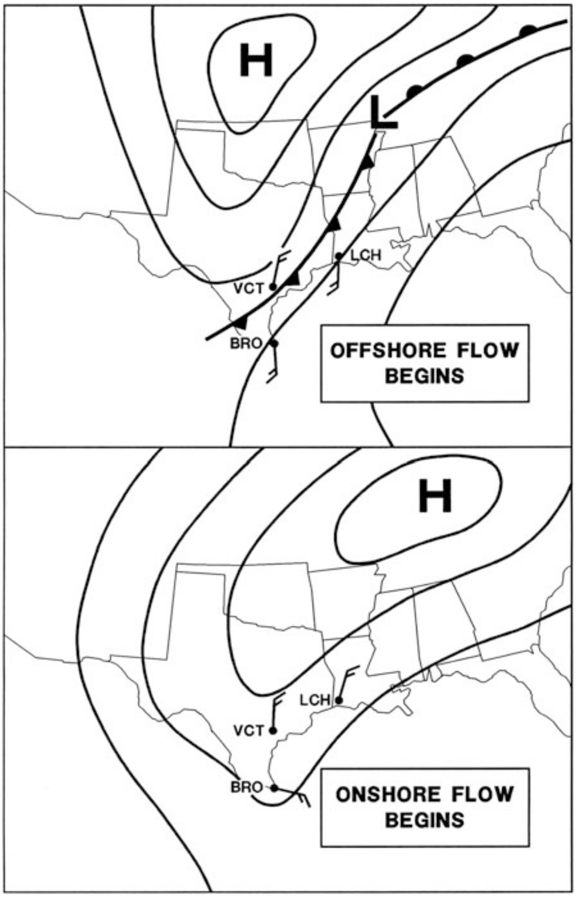


Fig. 5.1 Schematic diagram indicating the sequence of synoptic events associated with a typical return flow event over the Gulf of Mexico. Wind speeds are in knots (1 $\mathrm{ms}^{-1} \approx 2$ knots)

the ocean and returns to land (a bias that was cold/dry in the late-1980s through the early-to-mid 1990s but warm/moist since that time). The consequence of a poor forecast in these return-flow events is serious since slight changes in the moisture and heat content of the returning air leads to significantly different weather regimes—a range of weather that varies from mist and low stratus to shallow convection (without precipitation) to deep convection with thunderstorms.

From experience with return flow events during project GUFMEX (Lewis et al. 1989), we pattern our study after a typical return-flow event in the northwestern Gulf. The typical trajectory associated with a shallow intrusion of cold air into the Gulf is shown in Fig. 5.2. The bathymetry of the Gulf underpins this trajectory in Fig. 5.2 and indicates that the low-level airflow takes place over shelf water. When we assume surface winds of 15 ms^{-1} along the over-water path of $\sim$ 1,000 km, the time of transit over water is 18 h.

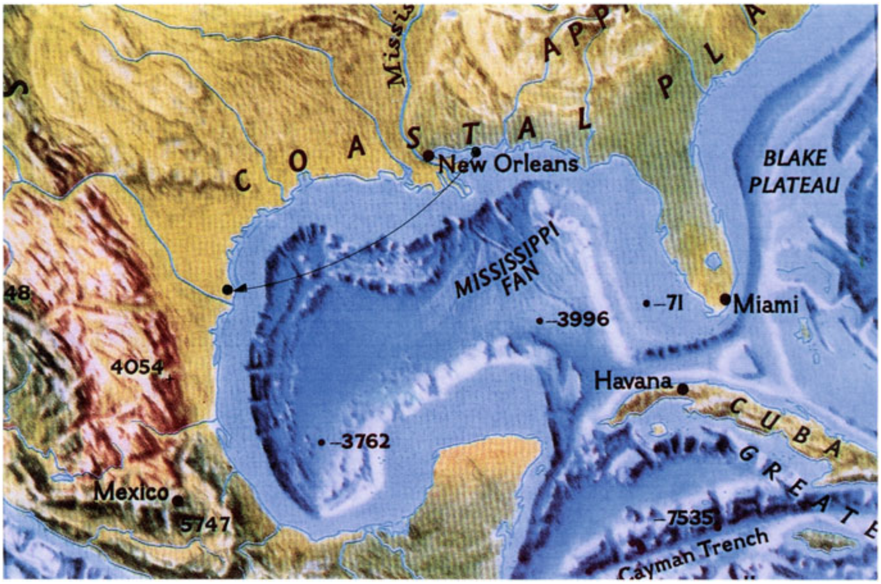


Fig. 5.2 The trajectory of air over the northwestern Gulf of Mexico superimposed over bathymetry of the Gulf. Ocean depths are recorded in meters

5.2.2 Governing Equation

Since candidates for the source of bias errors in return flow are uncertainties in initial conditions, sea surface temperature (SST: boundary condition), and turbulent transfer of heat and moisture from the sea to the air (turbulence parameterization), we consider a simplified yet physically meaningful air/sea interaction model that includes these three elements of control. We assume prior knowledge of the air trajectory over water (as shown in Fig. 5.2).

Our governing equation represents the Lagrangian forecast of air temperature along the known trajectory where elements of control are the initial temperature of the surface air just east of New Orleans, a sea surface temperature (SST) that is assumed constant along the trajectory (boundary condition), and a turbulent transfer coefficient that controls the turbulent heat exchange at the air/sea boundary.

The continuous form of the constraint is

$$\frac{dx}{dt} = C_T(\theta - x) \tag{5.1}$$

where x is air temperature, t is time, C_T is the turbulent transfer coefficient, and θ is sea surface temperature (SST). In Euler's form of finite differencing, the Lagrangian forecast of temperature at time step k (1 h time steps) is a weighted average of temperature at the previous time step and the SST that takes the form

$$\begin{aligned} x(k) &= x(k-1)(1 - C_T \Delta t) + C_T \Delta t \theta \\ &= x(k-1) \cdot (1-c) + c\theta \end{aligned} \tag{5.2}$$

where $c = C_T \Delta t$ is a nondimensional exchange coefficient (=0.25) based on typical values of parameters involved in the air/sea interaction [See Liu et al. (1992)]. A closed form solution to (5.2) is:

$$x(k) = (1-c)^k (x(0) - \theta) + \theta \tag{5.3}$$

5.2.3 Sensitivities

The solution (5.3) is nonlinear in the elements of control. The associated sensitivities, again nonlinear in control, are given by:

$$\begin{aligned} \frac{\partial x(k)}{\partial x(0)} &= (1-c)^k \\ \frac{\partial x(k)}{\partial \theta} &= [1 - (1-c)^k] \\ \frac{\partial x(k)}{\partial \nu} &= -\frac{k}{10}(1-c)^{k-1}[x(0) - \theta] \end{aligned} \tag{5.4}$$

where $\nu = 10c$. This change of control-element variable is a form of preconditioning that leads to faster convergence of the optimization process described below [See Gill et al. (1981) for a discussion of preconditioning]. The true control vector is taken to be $Y = [x(0), \theta, \nu] = [1°C, 11°C, 2.5]$ while the incorrect control is $Y' = [x'(0), \theta', \nu'] = [2°C, 10°C, 3.0]$. Thus, the difference between true and erroneous control is given by $Y - Y' = [-1°C, +1°C, -0.5]$. Entries in Table 5.1 exhibit the time evolution of the three sensitivities for correct and incorrect control (to be used in the numerical experiments). Under the assumption that true elements of control are unknown in practice, incorrect sensitivities are used to initiate the variational data assimilation process.

5.2.4 Cost Function for Data Assimilation

We assume that M observations of air temperature are made at a subset of the points along the trajectory displayed in Fig. 5.1. Indices associated with these observation points are represented by the sequence $\{\kappa_i : \kappa_1, \kappa_2, \ldots, \kappa_M\}$. With $z(\kappa_i)$ representing the observation at point κ_i, the cost function takes the form

Table 5.1 Sensitivity functions for the correct control (left-hand columns) where correct control is given by $x(0) = 1.0$, $\theta = 11.0$, and $\nu = 2.5$, and for the incorrect control (right-hand columns) where the incorrect control is given by $x(0) = 2.0$, $\theta = 10.0$, and $\nu = 3.0$. The bold numbers indicate errors greater than 40 %

	Correct Sensitivity			Incorrect Sensitivity		
k	$\frac{\partial x(k)}{\partial x_0}$	$\frac{\partial x(k)}{\partial \theta}$	$\frac{\partial x(k)}{\partial \nu}$	$\frac{\partial x(k)}{\partial x_0}$	$\frac{\partial x(k)}{\partial \theta}$	$\frac{\partial x(k)}{\partial \nu}$
0	1.000	0.000	0.000	1.000	0.000	0.000
1	0.750	0.250	1.000	0.700	0.300	0.800
2	0.562	0.438	1.500	0.490	0.510	1.120
3	0.422	0.578	1.688	0.343	0.657	1.176
4	0.316	0.684	1.688	0.240	0.760	1.098
5	0.237	0.763	1.582	0.168	0.832	0.960
6	0.178	0.822	1.424	0.118	0.882	**0.807**
7	0.134	0.866	1.246	0.082	0.918	**0.659**
8	0.100	0.900	1.068	**0.058**	0.942	**0.527**
9	0.075	0.925	0.901	**0.040**	0.960	**0.415**
10	0.056	0.944	0.751	**0.028**	0.972	**0.323**
11	0.042	0.958	0.619	**0.020**	0.980	**0.249**
12	0.032	0.968	0.507	**0.014**	0.986	**0.190**
13	0.024	0.976	0.412	**0.010**	0.990	**0.144**
14	0.018	0.982	0.333	**0.007**	0.993	**0.109**
15	0.013	0.987	0.267	**0.005**	0.995	**0.081**
16	0.010	0.990	0.214	**0.003**	0.997	**0.061**
17	0.008	0.992	0.170	**0.002**	0.998	**0.045**
18	0.006	0.994	0.135	**0.001**	0.998	**0.033**

$$J(x(0), \theta, \nu) = \frac{1}{2}\sum_{i=1}^{M} [z(\kappa_i) - x(\kappa_i)]^2 \tag{5.5}$$

5.3 Forward Sensitivity Method (FSM) Applied to Air/Sea Interaction Model

The strong-constraint forecast of air temperature at observation point κ_i is given by $x_f(\kappa_i)$. It is generally different than the observation due to incorrect control and error in the observation. The basic idea behind FSM is that corrections to control can annihilate the difference between forecast and observation. To first order in the Taylor series expansion about the forecasted state, the new estimate of air temperature at κ_i is given by

$$\begin{aligned} x(\kappa_i) = \{&x_f(\kappa_i) \\ &+[\tfrac{\partial x(k)}{\partial x(0)}]_{k=\kappa_i} \cdot \Delta x(0) + [\tfrac{\partial x(k)}{\partial \theta}]_{k=\kappa_i} \cdot \Delta\theta + [\tfrac{\partial x(k)}{\partial \nu}]_{k=\kappa_i} \cdot \Delta\nu\} \end{aligned} \tag{5.6}$$

where the sensitivities and forecast are assumed known and incremental changes to control are the unknowns. Optimal changes to control are those that minimize the cost function. In terms of (5.6), the cost function (5.5) is rewritten as

$$J(x(0),\theta,\nu)=\sum_{i=1}^{M}\{e(\kappa_i) \tag{5.7}$$

$$-[\frac{\partial x(k)}{\partial x(0)}]_{k=\kappa_i}\cdot\Delta x(0)-[\frac{\partial x(k)}{\partial\theta}]_{k=\kappa_i}\cdot\Delta\theta-[\frac{\partial x(k)}{\partial\nu}]_{k=\kappa_i}\cdot\Delta\nu\}$$

where $e(\kappa_i)=z(\kappa_i)-x_f(\kappa_i)$.

Let us define a sensitivity matrix $S\in R^{M\times 3}$:

$$S=\begin{bmatrix} [\frac{\partial x(k)}{\partial x(0)}]_{k=\kappa_1} & [\frac{\partial x(k)}{\partial\theta}]_{k=\kappa_1} & [\frac{\partial x(k)}{\partial\nu}]_{k=\kappa_1} \\ [\frac{\partial x(k)}{\partial x(0)}]_{k=\kappa_2} & [\frac{\partial x(k)}{\partial\theta}]_{k=\kappa_2} & [\frac{\partial x(k)}{\partial\nu}]_{k=\kappa_2} \\ \cdots & \cdots & \cdots \\ [\frac{\partial x(k)}{\partial x(0)}]_{k=\kappa_M} & [\frac{\partial x(k)}{\partial\theta}]_{k=\kappa_M} & [\frac{\partial x(k)}{\partial\nu}]_{k=\kappa_M} \end{bmatrix} \tag{5.8}$$

and an error vector $E\in R^{M\times 1}$:

$$E=\begin{bmatrix} z(\kappa_1)-x_f(\kappa_1) \\ z(\kappa_2)-x_f(\kappa_2) \\ \cdots \\ z(\kappa_M)-x_f(\kappa_M) \end{bmatrix} \tag{5.9}$$

and an incremental control vector $\Delta\varepsilon\in R^{3\times 1}$:

$$\Delta\varepsilon=\begin{bmatrix} \Delta x(0) \\ \Delta\theta \\ \Delta\nu \end{bmatrix} \tag{5.10}$$

Then

$$J=\frac{1}{2}(S\Delta\varepsilon-E)^T(S\Delta\varepsilon-E) \tag{5.11}$$

where superscript T indicates transpose. The necessary condition for minimization of J is vanishing of the derivative of J with respect to the increment of control $\Delta\varepsilon$. Satisfaction of this condition gives

$$\Delta\varepsilon=(S^TS)^{-1}S^TE \tag{5.12}$$

[See Lewis et al. (2006) for details].

Following correction to control based on this first-order Taylor expansion, a revised forecast is made and associated errors calculated. Iteratively the corrections to control are made until satisfaction of some empirical criteria — a criteria such as the norm of the incremental correction vector is smaller than a value commensurate with the expected error norm of control. However, in the numerical experiments to follow, corrections to control are found in a single step.

5.4 Numerical Experiments

5.4.1 Prelude

The problem we investigate assumes availability of surface air temperature observations along the known trajectory and an estimate of average SST along the trajectory. Observations from the moored buoys operated by the National Data Buoy Center (NDBC) and the U. S. Coast Guard serve as our guide in establishing a realistic distribution of surface observations over the Gulf of Mexico for the numerical experiments. For reasons related to economy of operation and maintenance of instruments, most buoys over the Gulf of Mexico are located in the shelf waters—roughly 50–100 km of the shoreline (Hamilton 1986). Given the position of NDBC buoys in the vicinity of the trajectory shown in Fig. 5.1, it is reasonable to assume that there are four instrumented buoys neighboring the trajectory (M = 4).

Our premise is that differential placement of observations along the trajectory is key to understanding the condition for sufficiency/insufficiency of observations—the condition that makes it possible/impossible to minimize the cost function.

The mechanics for correction to control by FSM have been discussed in Sect. 5.3. The standard procedure for finding ∇J in 4D-Var is backward integration of the model's adjoint. Under the simplified constraint of air/sea interaction, however, ∇J is found by straightforward differentiation of the cost function using knowledge of the sensitivities found in Table 5.1. Further, the conjugate gradient algorithm is used to determine search direction and step size (Lewis et al. 2006).

5.4.2 Forecast Errors

In our experiments, we assume observations are true—derived from the strong constraint (5.2) with true control (Y). The forecast error stems from incorrect control (Y'). The error vector E is displayed in Table 5.2 and the systematic nature of the error is obvious—an under-forecast the order of $-(0.01 - 1°C)$ up to t = 5 h and an over-forecast the order of $+(0.1 - 1°C)$ from t = 6 h until the end of the forecast at t = 18 h.

Table 5.2 Forecast error $E(k)$, $z(k)$: $observation$, and $x_f(k)$: $forecast$

k	$z(k)$	$x_f(k)$	$E(k) : z(k) - x_f(k)$
0	1.000	2.000	−1.000
1	3.500	4.400	−0.900
2	5.375	6.080	−0.705
3	6.781	7.256	−0.475
4	7.836	8.079	−0.243
5	8.627	8.655	−0.028
6	9.220	9.059	+0.161
7	9.665	9.341	+0.324
8	9.999	9.539	+0.460
9	10.249	9.677	+0.572
10	10.437	9.774	+0.663
11	10.587	9.842	+0.736
12	10.683	9.889	+0.794
13	10.762	9.922	+0.840
14	10.822	9.946	+0.876
15	10.866	9.962	+0.904
16	10.900	9.973	+0.927
17	10.925	9.981	+0.944
18	10.944	9.987	+0.957

5.4.3 Experiment 1: Insufficiency of Observations

From the sensitivities displayed in Table 5.1, air temperature at the last few hours of forecast is insensitive to the initial air temperature. This insensitivity is even made apparent when the forecast error is due only to incorrect initial air temperature, i.e., 2°C instead of 1°C. In this case, the error at t $=$ 18 is –0.005°C which is a very small fraction of the forecast error when all three elements are incorrect $(=+0.957)$. The sensitivity of air temperature to turbulent transfer coefficient at the last few times is also relatively small compared to sensitivity with respect to SST. From the FSM, it is clear that correction to a given element of control is tied to the fraction of forecast error due to incorrectness of the given element, i.e., the error relative to errors in the other elements. In view of this fact, when the four observations are located at the last four times (t $=$ 15, 16, 17, and 18 h), it is unlikely that meaningful corrections to $x(0)$ and ν can be found. Accordingly, we assume observations at these last four time steps as the input to Experiment 1.

The components of the gradient of the cost function at the end of the first iteration are:

$$\frac{\partial J}{\partial x(0)} = -0.011, \frac{\partial J}{\partial \theta} = -3.721, \text{ and } \frac{\partial J}{\partial \nu} = -0.203 \tag{5.13}$$

where the philosophy of steepest descent obviously indicates that the control state is being moved toward a higher SST as expected. But most apparent is the extreme flatness of the surface along the direction of the initial condition element. The

Table 5.3 Numerical experiment results

Erroneous control	Correct control
$x'(0), \theta', \nu'$	$x(0), \theta, \nu$
2.0, 10.0, 3.0	1.0, 11.0, 2.5

Experiment 1: Observations at $k = 15, 16, 17$, and 18

	∇J-method	FSM
Adjusted Control		
$(x(0), \theta, \nu)$	(2.00, 10.93, 3.1)	(1.23, 10.99, 2.0)
J (initial): J (final)	$1.74 : 8.3 \cdot 10^{-4}$	$1.74 : 6.1 \cdot 10^{-2}$
Eigenvalues of H	$-0.010,\ 0.42,\ 3.99$	$-0.002,\ 0.56,\ 4.99$

Experiment 2: Observations at $k = 1, 2, 17, 18$

	∇J-method	FSM
Adjusted control		
$(x(0), \theta, \nu)$	(1.39, 11.13, 2.2)	(1.05, 10.98, 2.3)
J (initial): J (final)	$1.56 : 2.0 \cdot 10^{-2}$	$1.56 : 6.0 \cdot 10^{-2}$
Eigenvalues of H	0.06, 1.82, 5.06	0.06, 1.94, 5.12

J (initial): Initial value of the cost function
J (final): Final value of the cost function
H: Hessian matrix at the point of adjusted control

surface is also relatively flat along the direction of the turbulent transfer coefficient. Not only is the surface flat in these directions, the negative gradient with respect to the initial condition and turbulent transfer coefficient will push the correction toward higher values of these elements whereas ideal corrections are toward lower values. Nevertheless, it is the end result of the search for the minimum that is critical to examine. Four iterations of the conjugate gradient method are executed before the empirical criterion for termination is satisfied (a change in the cost function less than 10^{-4} from one iteration to the next).

A summary of results for both ∇J-method and FSM are shown in the top portion of Table 5.3. Although the value of the cost function has significantly decreased via the ∇J-method, the corrections to initial condition and turbulent transfer are minimal. The reduction is obviously due to the SST correction alone. The search moves steadily toward the correct value of SST because of a significant negative gradient in that direction, but moves nary a bit in the other directions because of flatness of the surface in those directions. In short, the value of the cost function can be reduced significantly due to correction of SST alone—poor estimates of initial condition and turbulent exchange coefficient have little influence on the fit. What is the consequence of the poor fit to these elements? The consequence that can be determined in this idealized experiment is a poor forecast of temperature at the early times—based on observations that are known although unused in the functional.

Eigenvalues of the Hessian fundamentally determine the adequacy or inadequacy of observations in this case. The eigenvalue set for the ∇J-method is of mixed sign indicating a saddle point at the location of the final iteration. Positive eigenvalues are indicative of upward turning of the cost function whereas a negative eigenvalue is indicative of downward turning of the surface. In this circumstance of mixed signs in the Hessian's eigenvalue set, we conclude that the observations are insufficient to locate the minimum of J.

One iteration of the FSM for this case yields an adjusted control shown beside the results for the ∇J-method (again in the top portion of Table 5.3). The cost function is reduced but not as much as in the case of the ∇J-method. Further iterations with FSM would improve the fit. Whereas flatness in the cost function surface is the problematical aspect encountered by the ∇J-method, the corresponding problem for FSM is displayed as an absence of differences in the sets of sensitivities at the last four times (evident through examination of the last four rows of the sensitivity functions in Table 5.1). This lack of difference in the corresponding elements in these rows indicates near singularity of the $S^T S$ matrix. The near singularity is measured by the largeness of the condition number of the $S^T S$ matrix ($=10^9$)—the ratio of the largest to the smallest eigenvector. Inversion of $S^T S$ matrix is essential to finding corrections to control [See (5.12)]. As was the case for the ∇J-method, the FSM exhibits an eigenvalue set of mixed sign indicating insufficiency of the observations to locate the minimum of the cost function.

5.4.4 Experiment 2: Sufficiency of Observations

In this experiment, we replace the observations at t = 15 and 16 with observations at t = 1 and 2. We now have observations where the forecast is sensitive to initial conditions and turbulent exchange coefficient (times 1 and 2) as well as sensitive to SST (at all four times). As opposed to results from Experiment 1, the structure of the cost function in the vicinity of Y' is not flat in any direction. Further, the magnitudes of the various elements of $-\nabla J$ are comparable and they have signs consistent with ideal corrections to control (positive for SST and negative for initial condition and turbulent transfer coefficient). The first-iteration components of ∇J are

$$\frac{\partial J}{\partial x(0)} = +0.971, \frac{\partial J}{\partial \theta} = -1.267, \text{ and } \frac{\partial J}{\partial \nu} = +1.436 \tag{5.14}$$

Results from the optimization processes are found in the lower portion of Table 5.3. Adjusted controls for both FSM and ∇J-method are reasonably good and the eigenvalue set consists of positive eigenvalues for both forms of assimilation. Thus, the terminal points of the optimization process for both schemes are extremely close to the minimum of the cost function and the observation set is sufficient to find the minimum of J. Another indication of sufficiency is the small value of the condition number for the $S^T S$ matrix. In this case the condition number is 60.

5.5 Discussion and Conclusions

With the natural appeal that comes from application of variational data assimilation to weather analysis and prediction problems, the experiments conducted warn that an iterative process leading to a reduction in the cost functional is not necessarily heading for a minimum of the functional. In the low-order systems similar to the one we have investigated, where calculation of the eigenvalues of the Hessian about the terminal point of iteration is not overly challenging, the presence or absence of a minimum at the terminal point can be definitively determined.

In the more realistic systems associated with NWP, a meaningful structure of the true Hessian is difficult to determine despite a variety of innovative methods that have been developed to capture this structure.

In the absence of knowing the Hessian, this research has presented another view of the problem that holds promise for improving the chances of finding optimal correction to control through reliance on the forward sensitivity method (FSM)—knowledge of forward evolving sensitivity of model variable to elements of control in the context of a variational data assimilation scheme. Generally, this requires straightforward yet computationally demanding integration of the equations of sensitivity [described in detail in Lakshmivarahan and Lewis (2010)]. The FSM identifies those locations in space and time where observations are most likely to have little impact on the assimilation process—observations that generally lead to an ill-posed variational adjustment problem. Although the methodology does not give a recipe for an ideal or optimal placement of observations, valued locations are identified through their sensitivity to the various elements of control.

The methodology developed in this paper has application to the more-realistic NWP models used in operations. For example, in a post-mortem examination of the biased forecasts associated with return flow over the Gulf of Mexico, the FSM can identify those observational locations in space and time where the model variables exhibit sensitivity to elements of control. Calculation of error at these points, assuming availability of instrumented buoys near these locations, along with knowledge of sensitivity, provide the means to make corrections to control. Under the assumption that the model is faithful to the event (i.e., inclusion of reasonable representations of major physical processes), those elements that require the largest relative corrections are candidates for producing the bias.

Acknowledgements The authors thank Carlisle Thacker for a copy of his notes on variational data assimilation associated with his 1989 invited lectures at the Forschungscentrum in Hamburg, Germany. His discussion of adequacy/inadequacy of observations in variational analysis was especially lucid. We also thank research meteorologists Geoff Manikin and Zavisa Janjic of NCEP/EMC (National Center for Environmental Prediction/Environmental Modeling Center) for valued discussions related to the possible sources of bias in forecasts of return flow over the Gulf of Mexico.

We thank Joan O'Bannon, former graphics specialist at National Severe Storms Laboratory, for drafting Fig. 5.1, and we thank the National Geographic Society, particularly Eric Lindstrom, the Society's senior map editor and map library director, for permission to reproduce the bathymetric chart of the Gulf of Mexico that has been used as backdrop for the trajectory in Fig. 5.2. S. Lakshmivarahan's efforts are supported in part by two grants: NSF EPSCOR Track 2 grant 105-155900 and NSF grant 105-15400.

References

Bergthórsson P, Döös B (1955) Numerical weather map analysis. Tellus 7:329–340

Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic, London/New York, 401 pp

Hamilton G (1986) National Data Buoy Center programs. Bull Amer Meteorol Soc 67:411–415

Lakshmivarahan S, Lewis JM (2010) Forward sensitivity approach to dynamic data assimilation. Adv Meteorol. doi:10.1155/2010/375615

LeDimet F-X, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations. Tellus 38A:97–110

Lewis J (2007) Use of a mixed-layer model to investigate problems in operational prediction of return flow. Mon Weather Rev 135:2610–2628

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 (2008) Sasaki's pivotal contribution: calculus of variations applied to weather map analysis. Mon Weather Rev 136:3553–3567. doi: 10.1175/2008MWR2400.1

Lewis JM, Hayden C, Merrill R, Schneider J (1989) GUFMEX: a study of return flow in the Gulf of Mexico. Bull Am Meteorol Soc 70:24–29

Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, Cambridge, 654 pp

Liu Q, Lewis JM, Schneider JM (1992) A study of cold-air modification over the Gulf of Mexico using *in situ* data and mixed-layer modeling. J Appl Meteorol 31:909–924

Thacker WC (1989) The role of the Hessian matrix in fitting models to measurements. J Geophys Res 94:6177–6196

Thacker WC, Long RB (1988) Fitting dynamics to data. J Geophys Res 93:1227–1240

Wiin-Nielsen A (1991) The birth of numerical weather prediction. Tellus 43AB:36–52

Chapter 6
Quantifying Observation Impact for a Limited Area Atmospheric Forecast Model

Clark Amerault, Keith Sashegyi, Patricia Pauley, and James Doyle

Abstract Adjoint models calculate the first order sensitivity of a scalar output parameter to an input vector. Adjoint numerical weather prediction models have been used for a variety of sensitivity and data assimilation studies to provide a gradient for a measure of error with respect to the model's analysis variables. Recent work has shown that the adjoint of the data assimilation system can map the gradient information in analysis space onto individual observations to provide a quantitative estimate of an observation's influence on short-term forecast error. This chapter will review the framework of an adjoint observation impact system and some reported applications. Aspects of the framework particular to limited area atmospheric models will be the main focus of this chapter and results from a specific system will be presented. Issues discussed include: the effect of horizontal grid spacing on observation impact, the influence of lateral boundaries on forecast error, the relative importance of observations for different physical locations, and appropriate error metrics for limited area forecast models.

6.1 Adjoint Sensitivities

This chapter investigates the application of a limited area adjoint observation impact system. The adjoint operators of a numerical weather prediction (NWP) model and data assimilation (DA) system are combined to quantify the influence an observation has on short-term forecast error. This section reviews the previously developed framework of the system and its components, beginning with the adjoint NWP model. A description of the components of the limited area modeling system utilized for this work is given in Sect. 6.2, and observation impacts for the system are presented in Sect. 6.3. Future considerations are discussed in Sect. 6.4.

C. Amerault (✉) · K. Sashegyi · P. Pauley · J. Doyle
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. II)*, DOI 10.1007/978-3-642-35088-7_6,

6.1.1 Tangent Linear and Adjoint of the Forecast Model

The tangent linear and adjoint of an NWP model were introduced by Dimet and Talagrand (1986) for data assimilation experiments. An overview of tangent linear and adjoint models in meteorology and references for seminal works can be found in paper by Errico (1997). A tangent linear model $\mathbf{M}$ calculates the first order response of the output to a perturbation of the input of a nonlinear model M. If M is quasi-linear, the perturbation forecast by $\mathbf{M}$ will be similar to the difference in output between perturbed and non-perturbed runs of M. The tangent linear model is constructed by differentiating the nonlinear model with respect to the model's forecast state vector $\mathbf{x}$.

$$\mathbf{M} = \frac{\partial M(\mathbf{x})}{\partial \mathbf{x}}. \tag{6.1}$$

Adjoint models provide the gradient of some scalar function J of the forecast state vector within a numerical weather prediction NWP model with respect to the initial state vector (analysis). The forecast state vector depends on the initial conditions in the following way,

$$J(\mathbf{x}) = J(M(\mathbf{x_a})), \tag{6.2}$$

where $\mathbf{x_a}$ is initial state. The gradient of J with respect to the initial model state is

$$\frac{\partial J}{\partial \mathbf{x_a}} = \mathbf{M}^T \frac{\partial J}{\partial \mathbf{x}}, \tag{6.3}$$

where $\mathbf{M}^T$ is the adjoint model. The adjoint model maps the gradient of J with respect to the forecast state to the initial time, resulting in the sensitivity of J with respect to the analysis. The input to the adjoint model, $\frac{\partial J}{\partial \mathbf{x}}$ is calculated by differentiating J with respect to the forecast state. Since the adjoint model is derived from the tangent linear model, it's ability to calculate meaningful gradients is dependent on the quasi-linearity of the forecast model. Therefore, adjoint model fields in the atmosphere are most accurate over short time scales and for dry processes.

6.1.2 Observation Sensitivity

More recently, the adjoint of a DA system was formulated to compute gradients with respect to observations (Baker and Daley 2000). In DA, to obtain the analysis $\mathbf{x_a}$, the linear three dimensional analysis equations is written as

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x_b}). \tag{6.4}$$

The vector of observations $\mathbf{y}$ is length N, while the model space analysis $\mathbf{x_a}$ and background vectors $\mathbf{x_b}$ are both length L. The linear forward operator $\mathbf{H}$

transfers model space values to observation locations and is the Jacobian matrix corresponding to the nonlinear forward operator $H(\mathbf{x_b})$ linearized around $\mathbf{x_b}$. The Kalman gain $\mathbf{K}$ can be thought of as a weighting matrix and when expanded is written as,

$$\mathbf{K} = \mathbf{P_b}\mathbf{H}^T(\mathbf{H}\mathbf{P_b}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{6.5}$$

where $\mathbf{P_b}$ is the background error covariance matrix and $\mathbf{R}$ is the covariance matrix of the observations. A more detailed derivation of these equations and the assumptions utilized in the formulation is given in Daley (1991).

The analysis equation (6.4) can be rewritten as

$$\mathbf{x_a} = \mathbf{x_b} - \mathbf{K}\mathbf{H}\mathbf{x_b} + \mathbf{K}\mathbf{y} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{x_b} + \mathbf{K}\mathbf{y}, \tag{6.6}$$

with the $N \times N$ identity matrix $\mathbf{I}$. Differentiating Eq. 6.6 with respect to $\mathbf{y}$ gives an expression for the sensitivity of the analysis field with respect to the observations,

$$\frac{\partial \mathbf{x_a}}{\partial \mathbf{y}} = \mathbf{K^T}. \tag{6.7}$$

Likewise, the sensitivity with respect to the background field is obtained by differentiating Eq. 6.6 with respect to $\mathbf{x_b}$,

$$\frac{\partial \mathbf{x_a}}{\partial \mathbf{x_b}} = (\mathbf{I} - \mathbf{K}\mathbf{H})^{\mathbf{T}} = \mathbf{I} - \mathbf{H^T}\mathbf{K^T}. \tag{6.8}$$

The focus of this chapter is gradients with respect to observations, so $\frac{\partial \mathbf{x_a}}{\partial \mathbf{x_b}}$ will not be mentioned further.

An expression for the sensitivity of the scalar function in Eq. 6.2 with respect to the observations is obtained by applying the chain rule for derivatives to Eqs. 6.3 and 6.7 to obtain,

$$\frac{\partial J}{\partial \mathbf{y}} = \frac{\partial J}{\partial \mathbf{x_a}}\frac{\partial \mathbf{x_a}}{\partial \mathbf{y}} = \mathbf{K^T}\frac{\partial J}{\partial \mathbf{x_a}}. \tag{6.9}$$

Therefore, the sensitivity of a scalar function of a model's output with respect to the observations that were used to create the model's analysis field is obtained by applying the transpose of the Kalman gain to the result of the backward in time adjoint NWP integration.

Since $\mathbf{P_b}$ and $\mathbf{R}$ are symmetric matrices, $(\mathbf{H}\mathbf{P_b}\mathbf{H}^T + \mathbf{R})$ is also symmetric and Eq. 6.9 can be expanded to give,

$$\frac{\partial J}{\partial \mathbf{y}} = (\mathbf{H}\mathbf{P_b}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P_b}\frac{\partial J}{\partial \mathbf{x_a}}. \tag{6.10}$$

An examination of Eq. 6.10 reveals that most of the matrix operations performed in the original analysis procedure (Eq. 6.4) also appear in the gradient calculation, only in a different order. Therefore, the adjoint of the DA system is obtained by reordering the routines of the analysis scheme and is much easier to formulate than the adjoint of the NWP model.

6.1.3 Adjoint Observation Impact

A robust procedure for quantitatively evaluating the impact of an observation's assimilation on short-range forecast error utilizing the adjoint observation sensitivity framework was developed by Langland and Baker (2004). This procedure is responsible for most of the results presented in this chapter so the details found in Langland and Baker (2004) are summarized here for clarity.

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,

$$e_f = \big\langle (\mathbf{x}_f - \mathbf{x}_t), \mathbf{C}(\mathbf{x}_f - \mathbf{x}_t) \big\rangle , \tag{6.11}$$

and,

$$e_g = \big\langle (\mathbf{x}_g - \mathbf{x}_t), \mathbf{C}(\mathbf{x}_g - \mathbf{x}_t) \big\rangle . \tag{6.12}$$

The coefficients in $\mathbf{C}$ weight the model fields so that the error is measured in an energy norm. The forecast for g begins at an earlier time than f, and a short-term field (usually 6 or 12 h) 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. For global atmospheric NWP models, the value of e_f is generally less than e_g due to the assimilation of observations $\mathbf{y}$ to update $\mathbf{x_b}$. If no observations are assimilated to produce $\mathbf{x_a}$ for the f forecast, then $\mathbf{x_a}$ will be the same field as $\mathbf{x_b}$ and e_f will equal e_g. For limited area models, the lateral boundaries are also updated during the analysis procedure, which can lead to a change in e_f even if no observations are assimilated (Sect. 6.3.1).

To quantify the value of observations in reducing forecast error, an equation for the difference in e_f and e_g is defined,

$$\Delta e_f^g = e_f - e_g . \tag{6.13}$$

Using the adjoint NWP 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 f and g forecast trajectories,

$$J_f = \frac{1}{2} e_f , \tag{6.14}$$

$$J_g = \frac{1}{2} e_g , \tag{6.15}$$

$$\frac{\partial J_f}{\partial \mathbf{x}_f} = \mathbf{C}(\mathbf{x}_f - \mathbf{x}_t), \tag{6.16}$$

$$\frac{\partial J_g}{\partial \mathbf{x}_g} = \mathbf{C}(\mathbf{x}_g - \mathbf{x}_t). \tag{6.17}$$

Eqs. 6.11–6.12 and 6.16–6.17 can be used to rewrite Eq. 6.13 as,

$$\Delta e_f^g = \left\langle (\mathbf{x}_f - \mathbf{x}_g), \frac{\partial J_f}{\partial \mathbf{x}_f} + \frac{\partial J_g}{\partial \mathbf{x}_g} \right\rangle. \tag{6.18}$$

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 can be written as,

$$\delta e_f^g = \left\langle (\mathbf{x_a} - \mathbf{x_b}), \frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right\rangle. \tag{6.19}$$

Equation 6.19 is not an exact match to Δe_f^g because the adjoint model does not capture all of the processes of the nonlinear model used to calculate the error. To determine the impacts in observation space, the analysis increment is replaced in Eq. 6.19 in the following manner,

$$\delta e_f^g = \left\langle \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x_b}), \frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right\rangle. \tag{6.20}$$

Using the properties of an adjoint operator in an inner product, the following expression in observation space,

$$\delta e_f^g = \left\langle (\mathbf{y} - \mathbf{H}\mathbf{x_b}), \mathbf{K}^{\mathrm{T}} \left(\frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right) \right\rangle, \tag{6.21}$$

is obtained. The observation impacts are a product of the innovation vector components and the vector obtained from the adjoint DA process. The inner product in Eq. 6.21 gives a total estimate for all observations, but the inner product can be partitioned into any particular subset of interest. For example, the impact of a particular instrument, observation type (temperature, winds, etc.), or even a single measurement at a specific location can be determined using the method outlined above.

In practice, to calculate the observation impacts for a cycling NWP system the following steps are involved:

- Save appropriate forecast trajectories during the nonlinear NWP model run.
- When a verifying analysis $\mathbf{x}_t$ becomes available, calculate cost functions (Eqs. 6.14 and 6.15) and forcings for the adjoint NWP runs (Eqs. 6.16 and 6.17).
- Perform two adjoint NWP integrations along trajectories f and g back to the analysis time of f.
- Add the two resulting model space vectors together to create the input vector for the adjoint DA operator.
- Run the adjoint DA scheme to obtain the gradient in observation space, the inner product of this vector with the innovation will provide observation impacts.

The adjoint observation impact framework has also been derived in terms of a Taylor series expansion and shown to be a third-order metric by Errico (2007). The nonlinear nature of the metric means that cross terms with other observations may exist, which may make subsets of impacts difficult to interpret. However, the cross term effect was found to be small in a global system for the major observation networks (Gelaro et al. 2007). Although the cross terms will not be considered in this chapter, they may be important for smaller subsets of observations in a limited area model.

Another consideration not considered in this chapter is redundancy of information. Removing an observation from the DA process may result in a previously unimportant observation becoming important. Finally, the observation impact framework in this chapter applies to sequential DA, like three-dimensional variational systems. Approximations are needed to apply the framework to four-dimensional variational systems (Tremolet 2008).

The adjoint observation impact framework (Langland and Baker 2004) is a powerful tool for monitoring data assimilation performance and observation quality. Some of the subsequent applications (mainly to global NWP systems) of this framework will be discussed below (Sect. 6.1.4) and its application to a limited area model will be presented in Sect. 6.3.

6.1.4 Applications

In the seminal work by Langland and Baker (2004), the framework was applied to the Naval Research Laboratory's (NRL) global atmospheric modeling system. The NWP model was the Navy's Operational Global Atmospheric Predicition System (NOGAPS) and the accompanying DA component was the NRL Variational Data Assimilation System (NAVDAS). In the Northern Hemisphere, the largest error reductions were due to the assimilation of rawindsondes, satellite wind data, and aircraft observations, while in the Southern Hemisphere, satellite retrieved temperature profiles were important along with rawindsondes and satellite wind data. The framework has been implemented at a number of operational NWP center as a diagnostic monitoring system for the DA process (Langland 2005; Gelaro and Zhu 2009; Cardinali 2009). The system will indicate if a particular observation type or physical area is repeatedly increasing forecast error. These problem observations or areas can then be further investigated to find the cause of forecast degradation.

A comparison of observation impact systems for three distinct global modeling systems showed that observations have similar impact no matter the system on a global scale (Gelaro et al. 2010). Satellite sounding radiances provided the largest total impact in each of the forecast systems. Satellite winds, radiosondes, aircraft observations were also important contributors and are important components of the global atmospheric observing network. Slightly more than half of all the measurements (between 50 and 55 %) for each observation type are actually beneficial to forecast error reduction, meaning that a large number (between 45 and 50 %) are degrading the forecast.

The framework has also been accepted in the oceanographic community (Moore et al. 2011). More recent studies (Liu and Kalnay 2008) have investigated obtaining impact values with an ensemble of models (to replace the need for the adjoint operators) and obtaining sensitivity to other parameters of the DA system for performance tuning purposes (Daescu and Todling 2010).

6.2 COAMPS and NAVDAS

Results presented in this chapter were obtained from an adjoint observation impact system developed for NRL's limited area model. The system includes the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS®)[1] atmospheric model and its accompanying DA component, NAVDAS. Brief descriptions of these components as well as their accompanying adjoint operators are provided below.

6.2.1 COAMPS

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

$$\sigma_z = \frac{z_{\mathrm{t}}(z - z_{\mathrm{s}})}{z_{\mathrm{t}} - z_{\mathrm{s}}}, \tag{6.22}$$

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 NOGAPS. A detailed description of COAMPS is given in Hodur (1997).

6.2.2 NAVDAS

NAVDAS is currently used by the U.S. Navy for its operational regional DA system at the Fleet Numerical Meteorology and Oceanography Center (FNMOC).

[1] COAMPS® is a registered trademark of NRL.

It had also been used for the global DA system from October 2003 until October 2009, when it was extended to a four-dimensional variational system. NAVDAS uses an incremental update cycle to create the initial conditions for generating new model forecasts every 6 or 12 h. In the regional application, NAVDAS uses conventional and aircraft observations, geostationary satellite winds, Special Senor Microwave/Imager (SSM/I) winds speeds, satellite total precipitable water retrievals, scatterometer and passive microwave derived surface marine winds and satellite temperature retrievals. In addition, synthetic observations are used to define the wind and thermal structure of tropical cyclones. Satellite sounding radiances are currently under testing to replace the satellite temperature retrievals in near future. NAVDAS includes a geostrophic balance constraint and uniform analysis length scale.

The preprocessing and quality control software for the different observation types is built into NAVDAS. Innovations are first computed by interpolating the 6 or 12 h background forecast to the observation locations and then subtracting the result from the observations. A complex quality control (Collins and Gandin 1990; Gandin et al. 1993) is used for checking the rawinsonde observations. For aircraft data, the quality control includes sophisticated flight track checking and characteristic error detection. Vertical profiles of temperature and wind from the surface to 400 mb are created from the aircraft accents during takeoff for more efficient handling by the analysis algorithm. Satellite feature tracked winds from the geostationary satellites and polar-orbiting satellites, surface marine winds from scatterometer and WindSat instruments, and wind speed from SSM/I are all checked and averaged to create lower density "superobs" for each instrument type (Pauley 2003). The feature track satellite winds, scatterometer and WindSat winds are averaged in 1° by 1° boxes, while wind speed from SSM/I, total-precipitable water estimates from several instruments, and temperature retrievals are averaged in 2° by 2° boxes. The total-precipitable water superobs are used to generate vertical retrievals of pseudo relative humidity (ratio of observed mixing ratio to the saturation mixing ratio of the background).

NAVDAS uses Eq. 6.4 to update the background field (although the observation operator is not always linear as in Eq. 6.4). A preconditioned conjugate gradient algorithm is used to invert $(\mathbf{H}\mathbf{P_b}\mathbf{H}^T + \mathbf{R})$ followed by a post-multiplication step to obtain the analysis corrections $(\mathbf{x_a} - \mathbf{x_b})$ at COAMPS grid points. For increased efficiency, vertical profiles of observations from any single instrument are transformed into coefficients of the eigenvectors of the background error correlation. The number of vertical modes used in the analysis can depend on the type of instrument used for the observation; ten modes are used for the low vertical resolution of satellite temperature retrievals and a reduced set of modes can be used for satellite humidity retrievals. Predefined functions are used to define the horizontal and vertical variation of the background error covariance (Daley and Barker 2001). The horizontal and vertical length scales of the background error correlations may also vary with any combination of height, horizontal location and grid resolution. The covariances for the wind are based on those for the streamfunction and velocity potential (Daley 1991). For mesoscale analysis, a horizontal length scale of 385 km is currently used for the autocovariance of the streamfunction. The geopotential/wind background error covariances are then related by the geostrophic assumption. A latitude dependent coupling parameter (smaller in tropics compared to poles) is used to control the degree of geostrophic coupling between the streamfunction of the wind and the geopotential.

6.2.3 COAMPS Adjoint

In addition to the dynamical core, the COAMPS tangent linear and adjoint models include all of the respective components of the nonlinear model's physical parameterizations, with the exception of the radiation parameterization and the cumulus scheme (a less complex option is available in the tangent linear and adjoint models). More information on the adjoint COAMPS atmospheric model can be found in Amerault et al. (2008). The convective and moist parameterization adjoint operators are not robust for the length of the adjoint integrations (12 h) due to discontinuities and nonlinearities inherent in the corresponding nonlinear schemes. Therefore, current results are obtained with a "dry" adjoint model (although the nonlinear forecasts are run with all available physics options). The observation impacts calculated with a dry adjoint model will not capture all of the information in the nonlinear model's error. In practice, the estimate of Δe_f^g in analysis or observation space is roughly 80–90 % of the actual value (similar to global systems). Furthermore, COAMPS can be configured with multiple nests where the horizontal grid spacing is less than 10 km. The adjoint model can also be run with nests, but this option is computationally expensive due to the extra trajectories and not currently configured for the observation impact system. Therefore, all nonlinear forecasts and adjoint integrations are performed on a single domain for this work.

6.2.4 NAVDAS Adjoint

The adjoint NAVDAS adjoint operator is performed by reordering the operations of NAVDAS (Sect. 6.1.2). The adjoint operator does not involve any of the code responsible for quality control of the observations or creating the innovation vector. Therefore, the adjoint of NAVDAS for COAMPS was able to incorporate many of the components of the global system (Langland and Baker 2004). However, NAVDAS and COAMPS differ in their state variables and vertical coordinate. Additional code was needed to transfer the gradient field produced by the COAMPS adjoint model into NAVDAS analysis space.

For example, NAVDAS creates analyses of pseudo relative humidity which are converted and output as fields of dewpoint depression to be read by COAMPS. One more conversion to mixing ratio takes place before the COAMPS integration begins. In the adjoint observation impact system there is corresponding code to transfer the gradients through the mixing ratio – dewpoint depression – pseudo relative humidity sequence. The NAVDAS adjoint code is checked by comparing calculations of δe_f^g in observation and model space and with values of Δe_f^g.

6.3 Observation Impacts for COAMPS/NAVDAS

As discussed in Sect. 6.1.4, the adjoint observation impact framework has been primarily implemented for global modeling systems. This section will highlight some of the unique aspects of the framework for a limited area model such as COAMPS.

The flow of error information is shown in Fig. 6.1 for a 12 h COAMPS forecast with 60 km grid spacing valid at 1200 UTC 05 May 2010. The forecast error was computed on the lowest 20 model levels (out of 30) from the surface to the upper troposphere in a dry energy norm. All of the experiments in this chapter (except in Sect. 6.3.4) will use a dry energy norm in the lowest 20 model levels. The sum of the components of Δe_f^g in the vertical at each model grid point in the horizontal is presented in Fig. 6.1a. The blue shaded areas indicate where the error in the forecast from the analysis field is less than the background forecast valid at the same time. The COAMPS adjoint model integrates this information backward in time so that the components of δe_f^g in analysis space can be shown (Fig. 6.1b). Similar plots in observation space are shown for radiosondes and aircraft data (Fig. 6.1c, d) with the aid of the NAVDAS adjoint.

A small majority of observations contribute to the overall reduction in forecast error, which is why there are many observation locations with red shading in Fig. 6.1c, d. The percentage of beneficial observations in COAMPS/NAVDAS is similar to values for other global systems (Gelaro et al. 2010). The area of the forecast error calculation is smaller than the model domain, but the COAMPS adjoint model spreads error information to grid points outside the forecast error area. The smaller error area was chosen because of the influence of the lateral boundaries which will be discussed in the next section.

6.3.1 Lateral Boundaries

In the original framework for an adjoint observation impact system, Langland and Baker (2004) noted that the reduction in forecast error after the analysis procedure was due entirely to the assimilation of new observational information. This is true for a global NWP model, but for a limited area model like COAMPS, the lateral boundary conditions are also updated with the new analysis. Fortunately the framework allows for a quantitative estimation of the lateral boundary effects.

To illustrate, three impact experiments were conducted that only differed in the area over which the forecast error reduction was calculated. The dark boxes in

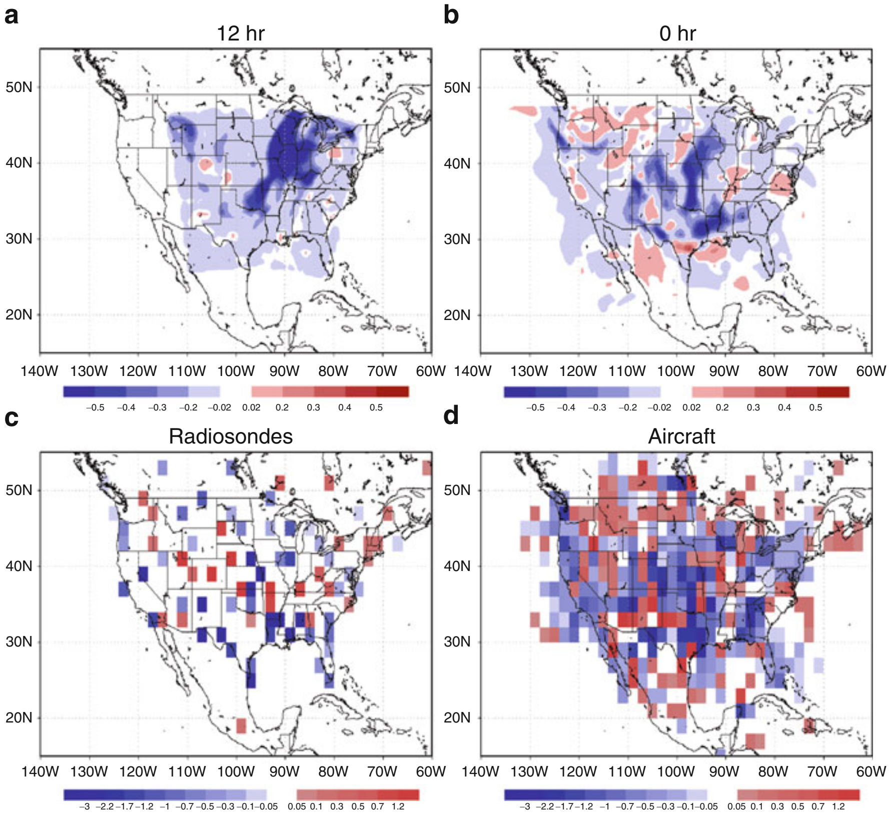


Fig. 6.1 The vertically integrated components of (**a**) Δe_f^g (**b**) and the model space estimate δe_f^g at the analysis time. Also, the subset of observation space estimate δe_f^g mapped to (**c**) radiosonde, and (**d**) aircraft observation locations. All calculations are for a 12 h COAMPS forecast valid 1200 UTC 05 May 2010. Values are in units of J kg^{-1}

Fig. 6.2 indicate the box edges for the error calculation in Cases 1–3. For Case 1, the error was calculated at every model grid point, while in Case 2, the outermost seven grid points along each edge were removed. In Case 3, the box was placed over the eastern United States to allow information to flow upstream in the COAMPS adjoint model and still stay on the model's grid. Impacts were computed for a 12 h COAMPS forecast valid 0000 UTC Sep 25 2011 with 90 km horizontal grid spacing.

The ratio of δe_f^g in model space to Δe_f^g is shown at each of the 12 h of the COAMPS adjoint integration by the black bars in Fig. 6.3 for each of the three cases. At 12 h, the value δe_f^g and Δe_f^g are exactly equal so their ratio is 1.0. However, the adjoint model is linear and does not account for all of the contributions in Δe_f^g. Therefore, the ratio drops as the COAMPS adjoint model is integrated backward in time. The ratio is just above 0.5 by the end of adjoint model run in Case 1. Removing some points near the boundary improves the situation in Case 2, and the

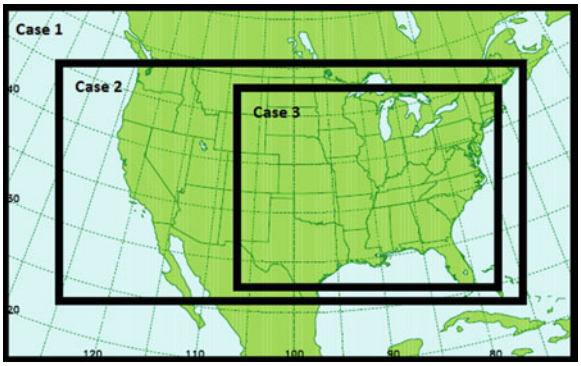


Fig. 6.2 Area over which forecast error is computed for Cases 1–3

ratio is even closer to 1.0 in Case 3. These results indicate that the lateral boundaries may effect the ratio calculation, and that the relatively large reduction in Case 1 may not be entirely due to the inability of the adjoint model to account for parts of the nonlinear error.

As the COAMPS adjoint model runs backward in time, some information is transferred from the adjoint model variables to the adjoint lateral boundary condition variables and is not included in the calculation of δe_f^g. Therefore, the calculation of δe_f^g was updated to include the effects of the lateral boundaries,

$$\delta e_f^g = \left\langle (\mathbf{x_a} - \mathbf{x_a}), \frac{\partial J_f}{\partial \mathbf{x_a}} + \frac{\partial J_g}{\partial \mathbf{x_b}} \right\rangle + \left\langle (\mathbf{l_a} - \mathbf{l_b}), \frac{\partial J_f}{\partial \mathbf{l_a}} + \frac{\partial J_g}{\partial \mathbf{l_b}} \right\rangle. \tag{6.23}$$

The updated expression includes a second inner product similar in form to the first term except that the model space vector $\mathbf{x}$ has been replaced by a vector of lateral boundary conditions $\mathbf{l}$. As long as the proper lateral boundary fields and gradients with respect to these fields are stored, the calculation of this second term is trivial.

The value of the second inner product in Eq. 6.23 for each of the three cases are indicated by the red bars in Fig. 6.3. With the addition of the lateral boundary contribution to δe_f^g, the ratios are all above 0.8 (similar to the values observed in global systems Langland and Baker 2004; Gelaro et al. 2010). As would be expected, the red bars are largest for Case 1, and smallest for Case 3.

The ability to quantify the effect of lateral boundary conditions on forecast error is a nice byproduct of this framework. However, only the information in the first term of Eq. 6.23 is passed to observation space, so the remainder of the experiments will be similar to Case 3 to maximize the impact of observations and minimize the lateral boundary effects.

6.3.2 Location

Observation impacts for a week of forecasts (Dec 24–31 2010) for three different locations (continental United States, eastern Pacific Ocean and western United

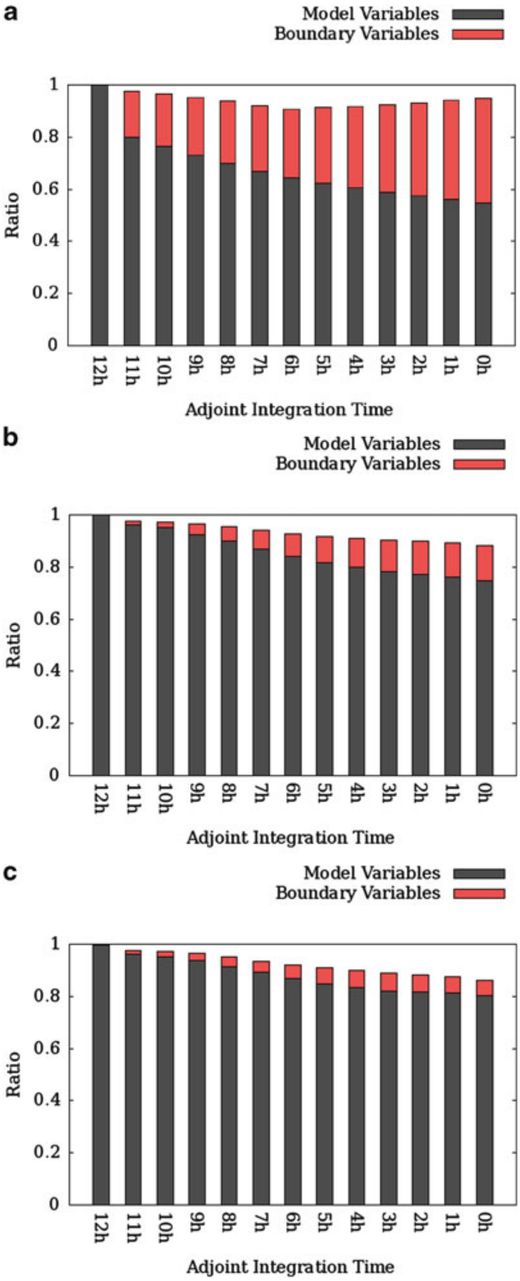


Fig. 6.3 Ratio of δe_f^g to Δe_f^g for every hour of the adjoint COAMPS integration for (**a**) Case 1, (**b**) Case 2, and (**c**) Case 3. The verification time was 0000 UTC Sep 25 2011. The contribution to δe_f^g from the model variables is indicated by the *black portion* of the bars, while the lateral boundary component is colored *red*

States, and southwest Asia) are shown in Fig. 6.4. The model's horizontal grid spacing was 60 km in each location. A brief description of the observation categories listed in Fig. 6.4 are provided in Table 6.1.

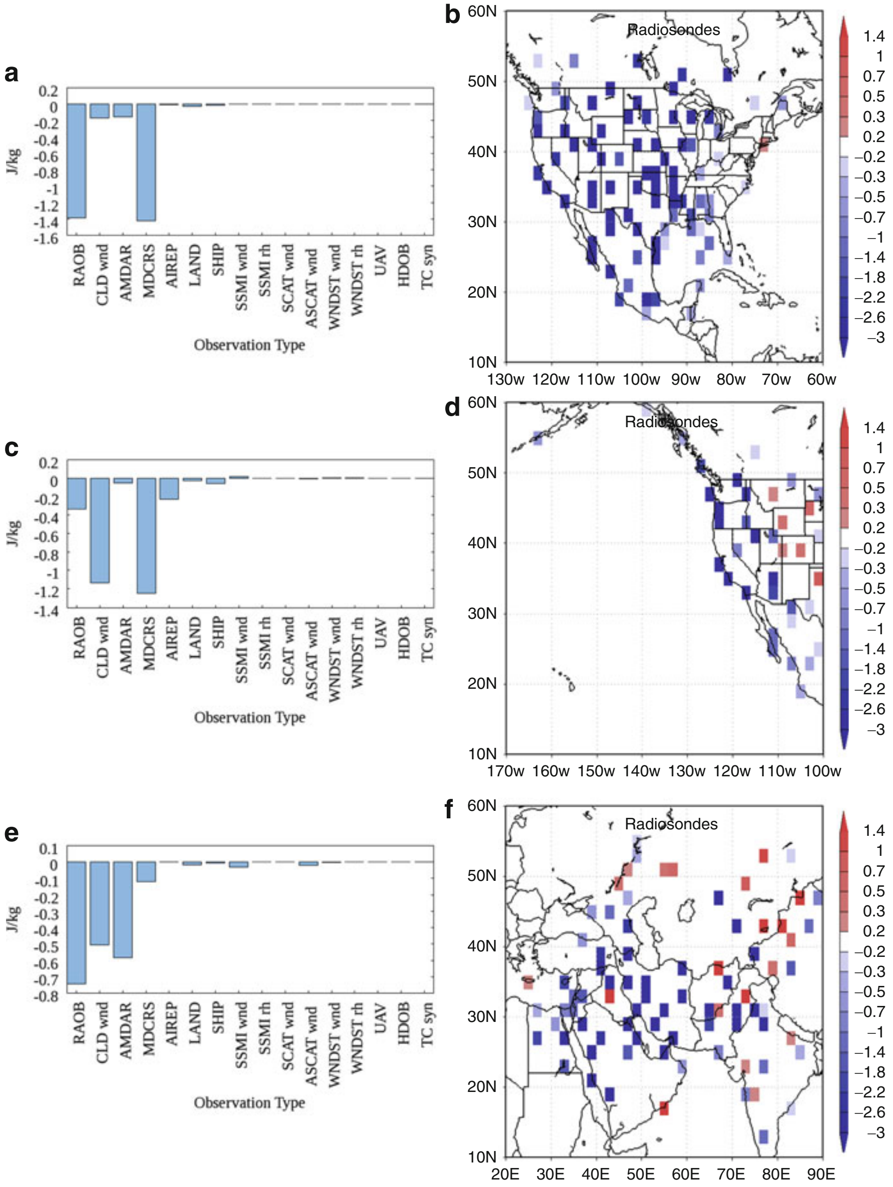


Fig. 6.4 Averaged observation impacts by type for each forecast cycle (*left*) and contribution at radiosonde locations (*right*, multiplied by 1,000) for domains covering (**a**) and (**b**) the continental United States, (**c**) and (**d**) the eastern Pacific Ocean and western United States, and (**e**) and (**f**) southwest Asia. The impacts were calculated for 12 h forecasts valid over the period 0000 UTC 24–0000 UTC 31 Dec 2010. Values are in units of J kg^{-1}

Table 6.1 Description of observation types

RAOB	Radiosonde
CLD wnd	Satellite feature track wind
AMDAR	Aircraft instrument
MDCRS	Aircraft instrument
AIREP	Aircraft instrument
LAND	Surface observation over land
SHIP	Surface observation over water
SSMI wnd	Surface wind from SSM/I
SSMI rh	Moisture retrieval from SSM/I
SCAT wnd	Surface wind vector from SCAT
ASCAT wnd	Surface wind vector from ASCAT
WNDST wnd	Surface wind from WNDST
WNDST rh	Moisture retrieval from WNDST
UAV	Observations from unmanned aircraft
HDOB	High density hurricane hunter aircraft
TC Syn	Synthetic TC data

For each forecast cycle, radiosondes (RAOB) and aircraft data (AMDAR, MDCRS, AIREP) are large contributors to forecast error reduction over the United States. Also note that the shading at the United States radiosonde locations is almost entirely blue, meaning that over time, each of these locations is reducing the forecast error. When the model's domain moves west to include a portion of the Pacific Ocean, the impact of the radiosondes decrease, while the feature track winds (CLD wnd) increase. In southwest Asia, error reduction is due to a combination of radiosondes, aircraft data, and feature track winds.

The importance of radiosondes, aircraft data, and feature track winds is also seen in global systems (Gelaro et al. 2010) along with satellite sounding radiances. NAVDAS for COAMPS does not currently assimilate radiances, but this system will prove to be a useful tool when they are added to the stream of assimilated data.

6.3.3 Model Resolution

The COAMPS adjoint observation impact system was run with two different grid configurations of differing horizontal grid spacing (90 and 30 km) for 3 days (Sep 22–25 2011). On a per observation basis, impacts are larger for the grid configuration with more points and smaller horizontal spacing (Fig. 6.5). This is partly due to larger impacts as seen in Fig. 6.6 by the darker blue shading off the southern California coast for the 30 km grid spacing. However, the larger per observation values can also be partly explained by the greater number of observations that are assimilated for the 90 km grid. NAVDAS creates a grid that is extended beyond the model domain by 15 points in every direction (1,350 km for the 90 km grid and 450 km for the 30 km grid). So even though the area that

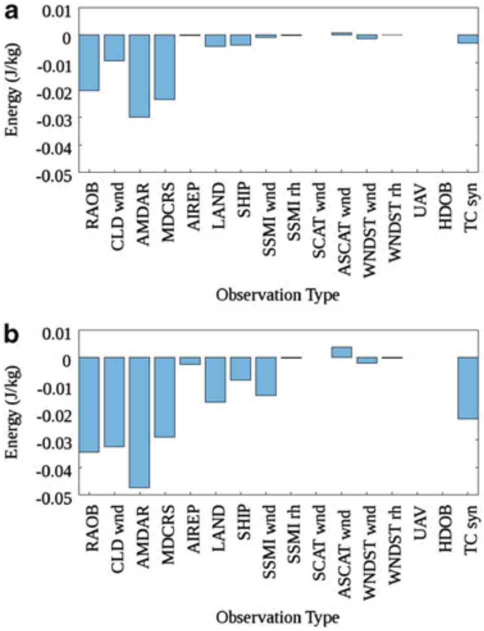


Fig. 6.5 Impacts by observation category on a per observation basis (multiplied by 1,000) for 3 days of 12 h COAMPS forecasts Sep 22–25 2011. The *horizontal grid* spacing was (**a**) 90 km and (**b**) 30 km. Values are in units of J kg^{-1}

the model domain covers is equal for both cases, the analysis area is larger for the 90 km grid. Therefore, more observations are being assimilated on the 90 km grid which reduce the per observation impact values. This can be seen by comparing the amount of shading south of 15° N in Fig. 6.6a, b.

Many more experiments need to be conducted before any definitive statements can be made about the effect of horizontal grid spacing on observation impact calculations. One other consideration is the manner in which the error is calculated. The truth is an analysis field on the same grid as the model forecasts, so it is not consistent between the two grid configurations (30 km forecasts are compared with 30 km analyses and 90 km forecasts are compared with 90 km analyses). Future experiments would benefit from comparisons with a consistent truth.

6.3.4 *Other Metrics*

As noted earlier, the error for all other experiments in this chapter was calculated in a dry energy norm for the lowest 20 model levels. This is a metric that is suitable for a global observation monitoring system; however, mesoscale applications may require different error considerations. This section demonstrates that a change in the error metric can also change the relative importance of observations in reducing that metric. Impacts for the same COAMPS 12 h forecast as in Sect. 6.3.1 were computed

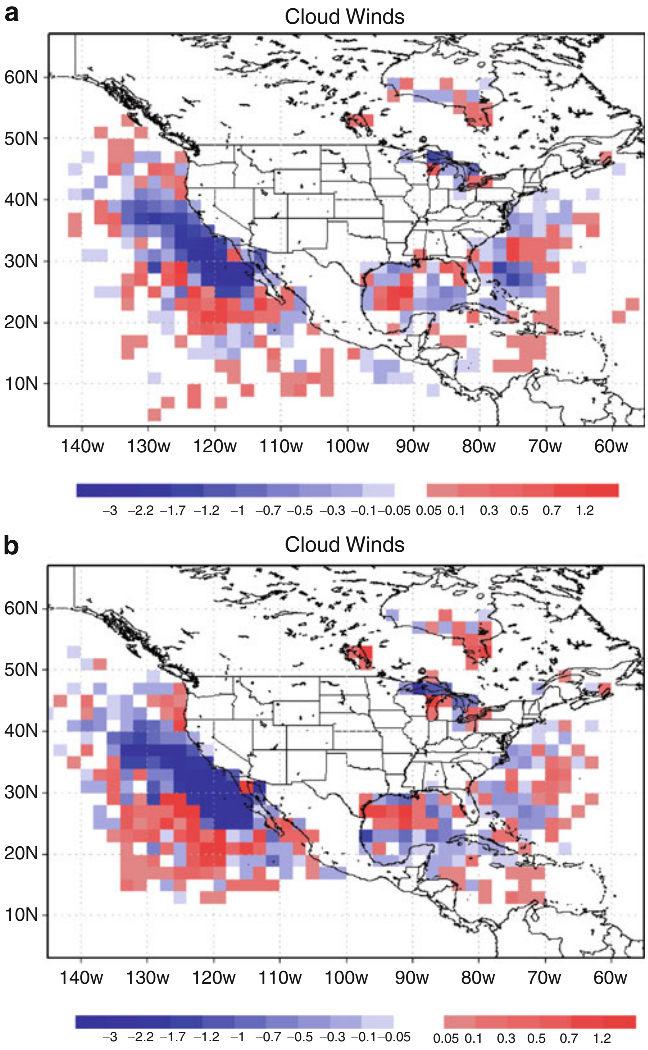
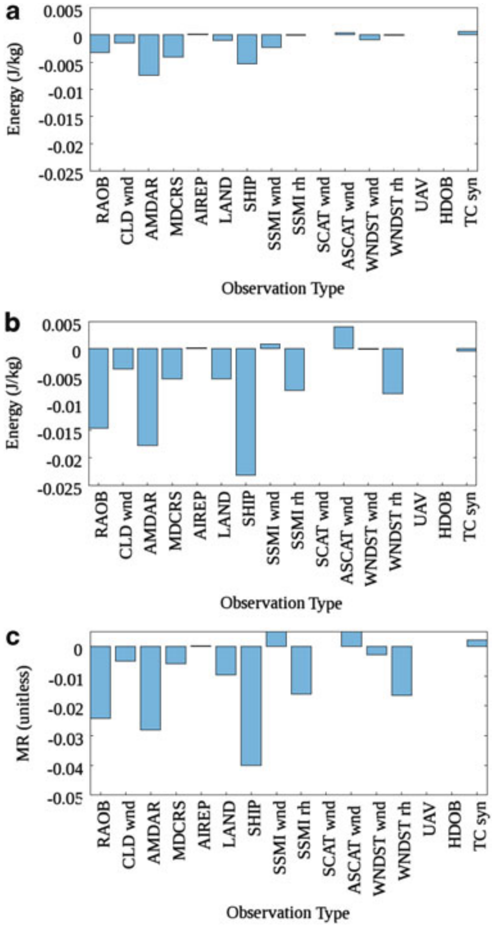


Fig. 6.6 Impacts of feature track satellite winds at observation locations (multiplied by 1,000) for the same forecasts in Fig. 6.5. The *horizontal grid* spacing was (**a**) 90 km and (**b**) 30 km. Values are in units of J kg^{-1}

for three different metrics. First, the error was calculated in a dry energy norm, but only for the the lowest ten model layers (roughly the extent of the planetary boundary layer). Then a moisture term was added to the energy norm calculated over the same area. Finally, the error was calculated in terms of modified refractive index (only depends on the temperature and humidity variables). The results for these different metrics are shown in Figs. 6.7 and 6.8. Although the units are different,

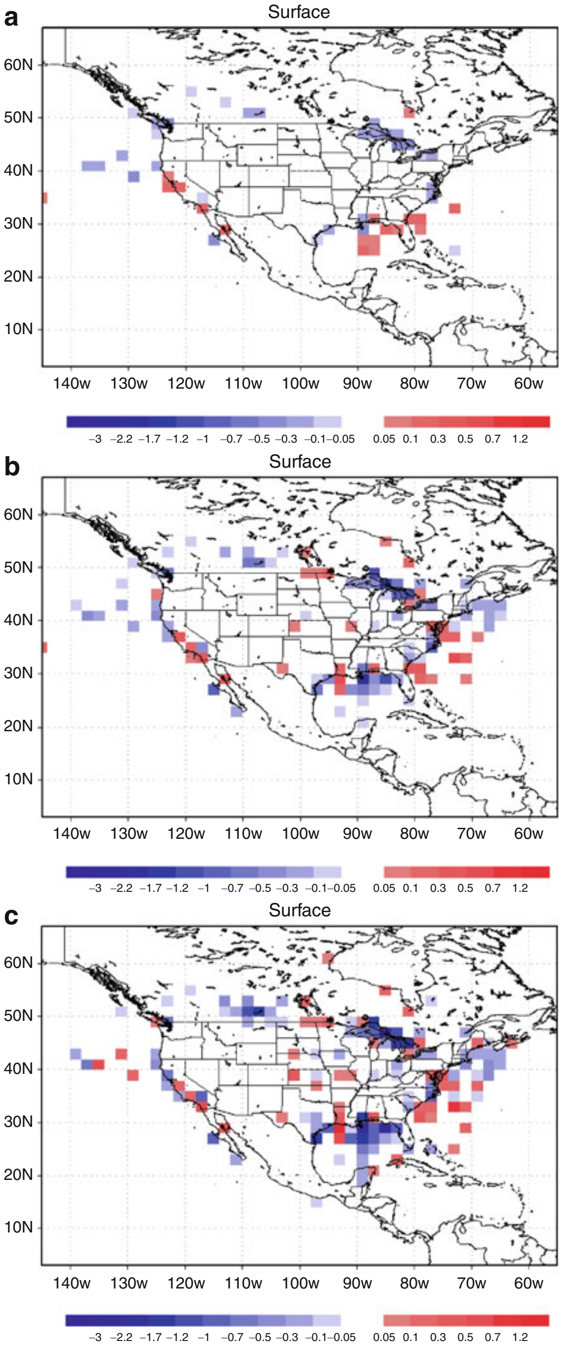
Fig. 6.7 Impacts by observation category on a per observation basis a 12 h COAMPS forecasts valid 0000 UTC Sep 25 2011. The error was calculated in a (**a**) dry energy norm, (**b**) moist energy norm, and (**c**) modified refractivity space. Values are in units of J kg^{-1} multiplied by 1,000 in (**a**) and (**b**) and unitless in (**c**)



the patterns of the moist energy norm and the modified refractive index are quite similar, even though the moist energy error depends on the winds, along with temperature and moisture. Not surprisingly, adding moisture to the error calculation increases the importance of the satellite relative humidity retrievals (SSMI rh and WNDST rh). Oceanic surface observations (SHIP) are also important for the moist metrics and in this case are due to the positive impacts in the Gulf of Mexico.

As in Sect. 6.3.3, more experimentation is needed before definitive statements can be made about these different error metrics. Still, it is easy to see that changing the metric and its location can have a dramatic effect on an observation's relative impact.

Fig. 6.8 Impacts of surface measurements at observation locations for the same forecast in Fig. 6.7. The error was calculated in a (**a**) dry energy norm, (**b**) moist energy norm, and (**c**) modified refractivity space. Values are in units of J kg^{-1} multiplied by 1,000 in (**a**) and (**b**) and unitless in (**c**)



6.4 Summary and Considerations

An adjoint observation impact system for a limited area model is a useful tool. The system built for COAMPS is similar to other global systems in that it indicates that radiosondes, aircraft data, and feature track satellite winds are all important in reducing forecast error measured in a dry energy norm throughout the depth of the troposphere. As new observation types such as satellite radiances are added to NAVDAS, the impact system can be utilized to ensure the data is being properly assimilated. A nice additional feature of this system is its ability to quantify the impact of lateral boundary conditions as well as observational data.

The relative importance of observations can vary with location and error metric. Choosing a suitable metric for a user's particular interest is key to properly evaluating an observation's importance. The COAMPS observation impact system contains much of the functionality inherent in both COAMPS and NAVDAS (relocatability, varying grid configurations) as well as the ability to easily change the volume over which the error is calculated and the metric. However, to investigate smaller scale atmospheric features ($<30\,\text{km}$) the nesting capability in the adjoint model will need to be added to the system. Also, an option for the truth other than the model's analysis field would be helpful for making comparisons between different experiments. For example, the error could be calculated with respect to radiosonde observations, but this will require more system development. As it stands, the COAMPS adjoint observations impact system is a valuable asset and is already providing important information on the performance of the entire atmospheric modeling system.

Acknowledgements This work was supported by the US Office of Naval Research's program element 0601153N. Computational resources of the Department of Defense High Performance Computing Modernization Program were vital to this work.

References

Amerault C, Zou X, Doyle J (2008) Tests of an adjoint mesoscale model with explicit moist physics on the cloud scale. Mon Weather Rev 136:2120–2132. doi: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

Collins W, Gandin L (1990) Comprehensive hydrostatic quality-control at the National Meteorological Center. Mon Weather Rev 118:2752–2767

Daescu D, Todling R (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. Q J R Meteorol Soc 136:2000–2012

Daley R (1991) Atmospheric data assimilation. Cambridge University Press, Cambridge, UK

Daley R, Barker E (2001) The NAVDAS source book 2001. Technical report NRL/PU/7530-01-441, Naval Research Laboratory

Dimet FL, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theroretical aspects. Tellus 38A:97–110

Errico R (1997) What is an adjoint model? Bull Am Meteorol Soc 78:2577–2591

Errico R (2007) Interpretations of an adjoint-derived observational impact measure. Tellus 59A:273–276

Gandin L, Morone L, Collins W (1993) Two years of operational comprehensive hydrostatic quality-control at the National Meteorological Center. Weather Forcast 8:57–72

Gelaro R, Zhu Y (2009) Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. Tellus 61A:179–193

Gelaro R, Zhu Y, Errico R (2007) Examination of various-order adjoint-based approximations of observation impact. Meteorol Z 16:685–692

Gelaro R, Langland R, Pellerin S, Todling R (2010) The THORPEX observation impact intercomparison experiment. Mon Weather Rev 138:4009–4025. doi:10.1175/2010MWR3393.1

Hodur R (1997) The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). Mon Weather Rev 125:1414–1430. doi:10.1175/1520-0493(1997)125<1414:TNRLSC>2.0.CO;2

Langland R (2005) Observation impact during the North Atlantic TReC-2003. Mon Weather Rev 133:2297–2309

Langland R, Baker N (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation system. Tellus 56A:189–201

Liu J, Kalnay E (2008) Estimating observation impact with adjoint model in an ensemble Kalman filter. Q J R Meteorol Soc 134:1327–1335

Moore A, Arango H, Broquet G, Edwards C, Veneziani M, Powell B, Foley D, Doyle J, Costa D, Robinson P (2011) The regional ocean modeling system (ROMS) 4-dimensional variational data assimilation systems Part iii – Observation impact and observation sensitivity in the California Current System. Prog Oceanogr 91:74–94

Pauley P (2003) Superobbing satellite winds for NAVDAS. Technical report NRL/MR/7530-03-8670, Naval Research Laboratory

Tremolet Y (2008) Computation of observation sensitivity and observation impact in incremental variational data assimilation. Tellus 60A:964–978. doi:10.1111/j.1600–0870.2008.00349

Chapter 7
Skewness of the Prior Through Position Errors and Its Impact on Data Assimilation

Daniel Hodyss and Alex Reinecke

Abstract Uncertainty in the position of a feature is a ubiquitous influence on data assimilation (DA) in geophysical applications. This chapter explores the properties of distributions arising from the uncertainty of the location of a flow feature. It is shown that distributions arising from phase uncertainty have surprisingly complex, non-Gaussian characteristics. These non-Gaussian characteristics are explored from an ensemble DA perspective in which the skewness (third-moment) is shown to be a significant contributor to the state-estimates obtained through Bayesian state estimation. Idealized examples, as well as an example in a real tropical cyclone using a state-of-the-art numerical weather prediction model, will be shown.

7.1 Introduction

Data assimilation (DA) is the combining of information from a model forecast and an observation to obtain an estimate of the state of a physical system that is generally better than either individually. One way DA is accomplished is through ensemble (Monte-Carlo) methods. The basic idea in this perspective is to perform regression of the state variables in need of updating against the observations of the state. This form of DA is rapidly becoming the technique of choice for the estimation of the state of a geophysical system. This popularity is largely due to the significant ease of implementation afforded by the use of Ensemble-based Kalman Filter (EnKF) DA systems. The EnKF is a state-estimation technique that makes use of the ensemble to estimate the first and second moments of the prior distribution, which are then used to estimate the posterior mean. This reliance upon just the first and second moments of the prior distribution allows for significant computational advantages

D. Hodyss (✉) · A. Reinecke
Naval Research Laboratory, Monterey, CA 93943, USA
e-mail: daniel.hodyss@nrlmry.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_7,

over more complex methods. Even with its reliance upon only the first two moments the application of the EnKF in the meteorological community has been met with considerable success in a wide-range of applications (e.g., Houtekamer et al. 2005; Szunyogh et al. 2008; Meng and Zhang 2008; Torn and Hakim 2008; Whitaker et al. 2008; Anderson et al. 2009).

There are, however, several unresolved issues with the application of the EnKF to the highly nonlinear dynamics inherent to meteorological flows at high resolution. Situations in which the EnKF is known to have some difficulty, and where nonlinearity may be significant, include the assimilation of: vortex position (Lawson and Hansen 2005; Chen and Snyder 2007), radar observations (Dowell et al. 2011), parameter estimation (Hacker et al. 2011), and observations over a long assimilation window (Khare et al. 2008). We speculate that one reason the assimilation in these situations is sometimes difficult is the fact that the relationship between the prior estimates of the observed variables and the state-vector is nonlinear. This nonlinearity may come about from nonlinearity in the model operator, which is described through the dynamics of the physical system, or from nonlinearity in the observation operator used to observe the system. In either case, this nonlinear relationship leads to skewed (non-zero third moment) posterior distributions that, as has been discussed by Hodyss (2011) and reviewed below, results in suboptimal behavior from the EnKF.

Previous work towards state estimation techniques for nonlinear modeling systems has been discussed by Kushner (1967); Jazwinski (1998; Chap. 9), Anderson and Anderson (1999), and Julier and Uhlmann (1997, 2004). However, in the geophysical sciences work towards explicitly accounting for the non-Gaussian aspects of the prior within an ensemble DA framework has emphasized particle filtering [See van Leeuwen (2009) for a comprehensive review]. In particle filtering one estimates the probability that a particular realization is the true state by comparing against observations. Given these probabilities (sometimes referred to as weights) one can do such things as make a state estimate based on the mean of the pdf or randomly sample from the pdf to generate an ensemble consistent with observational and prior uncertainties. Another way to get non-normal distributions is through nonlinearity in the observation operators or the non-normality of the observation likelihood. Two examples of ensemble filters that are aimed at the situation where the observation operator is nonlinear or the observation likelihood is non-Gaussian (but the prior distribution is normal) are the work of Zupanski (2005) and Fletcher and Zupanski (2006).

The focus here however will be on neither model nor observation operator induced nonlinearity; here we will focus on the nonlinearity, or more specifically the non-Gaussianity, that arises from phase errors. This "nonlinearity" that results from the uncertainty in the location of the feature will be shown to arise from a nonlinear relationship between the uncertainty in the location of the feature and the state variables describing the physical system. We show below that prior distributions whose uncertainty arises from errors in the location of a feature leads to significant skewness and hence significantly non-Gaussian distributions. The distributions that arise from phase error uncertainty will be shown to have surprisingly complex multivariate structures. Previous work (e.g., Lawson and Hansen 2005; Chen and Snyder 2007) examining the relationship between phase error uncertainty and data assimilation have noted the role of the non-Gaussian shape of the distribution and some of the potential avenues for failure of the EnKF. In Lawson and Hansen (2005) a two-step procedure was suggested in which the DA is first performed to account for the position errors and then after shifting the ensemble consistent with the updated positions (which attempts to reduce the non-Gaussianity by reducing the variance in position-space) assimilate a second set of observations to account for the structural update. In Chen and Snyder (2007) it was suggested that assimilating observations of vortex shape and intensity as well as position helps reduce the errors made by the EnKF. This chapter intends to first provide a detailed examination of the specific structure of distributions that arise from phase uncertainty and then show how incorporating information about the third moments of the prior impacts the DA. The focus here will therefore be to understand and then attempt to use the non-Gaussian information in the prior distribution rather than find ways to avoid or eliminate it.

The organization of this paper is as follows: In Sect. 7.2 we will describe DA through a Bayesian perspective and illustrate both linear and nonlinear regression. In Sect. 7.3 we show how an error in the location of a feature in the fluid leads to a non-Gaussian distribution. In Sect. 7.4 we will apply Kalman and higher-order DA methods to the assimilation of observations in which the prior uncertainty is described by errors in location. Section 7.5 closes the manuscript with a summary of the most important results, conclusions, and points out avenues of future research presently being investigated.

7.2 Understanding Data Assimilation Through Bayes' Rule

7.2.1 Bayes' Rule: The Posterior Distribution

We imagine the true state, $\mathbf{x}$, to be an N-vector and that it is drawn from a distribution whose pdf we label $\rho(\mathbf{x})$. In addition, we will collect the sum total of all previous information about this true state in a previous estimate we label $\mathbf{x}_f$. At the present time we have available a p-vector of observations $\mathbf{y}$ such that we may use Bayes' rule to obtain a density that describes the combined knowledge of the likely distribution of states:

$$\rho\left(\mathbf{x}|\,\mathbf{y},\mathbf{x}_f\right) = \frac{\rho\left(\mathbf{y}|\,\mathbf{x}\right)\rho\left(\mathbf{x}|\,\mathbf{x}_f\right)}{\int\limits_{-\infty}^{\infty}\rho\left(\mathbf{y}|\,\mathbf{x}\right)\rho\left(\mathbf{x}|\,\mathbf{x}_f\right)d\mathbf{x}}. \tag{7.1}$$

The density $\rho(\mathbf{y}|\mathbf{x})$ describes the conditional distribution of observations given a particular value of the truth (observation likelihood), while $\rho(\mathbf{x}|\mathbf{x}_f)$ describes the conditional distribution of truth given a previous estimate of the true state; this last density will hereafter be referred to as the "prior." The density $\rho(\mathbf{x}|\mathbf{y},\mathbf{x}_f)$ describes the conditional distribution of the truth given a particular observation and previous estimate; this density will hereafter be referred to as the "posterior."

7.2.2 Data Assimilation as a Problem in Nonlinear Regression

The presentation below relates the estimation of the posterior mean to the well-known methods of nonlinear regression and is presented as a review of Hodyss (2011). A standard estimation technique for the true state given the posterior density is to find its mean, i.e.

$$\bar{\mathbf{x}}\left(\mathbf{y},\mathbf{x}_f\right) = \int_{-\infty}^{\infty} \mathbf{x}\rho\left(\mathbf{x}|\,\mathbf{y},\mathbf{x}_f\right) d\mathbf{x}. \tag{7.2}$$

This estimate of the true state has the property that it is unbiased and that it minimizes the posterior error variance (Jazwinski 1998).In ensemble-based estimation techniques it is often assumed that our previous estimate of the truth is the mean of the prior distribution, $\bar{\mathbf{x}}_f$. Note that a random draw from the prior distribution behaves as $\mathbf{x} = \bar{\mathbf{x}}_f + \boldsymbol{\varepsilon}_f$, where $\boldsymbol{\varepsilon}_f$ is a random variable with zero mean. Adding and subtracting the prior mean allows (7.2) to be written as

$$\bar{\mathbf{x}}\left(\mathbf{y},\bar{\mathbf{x}}_f\right) = \bar{\mathbf{x}}_f + \int_{-\infty}^{\infty} \left(\mathbf{x}-\bar{\mathbf{x}}_f\right)\rho\left(\mathbf{x}|\,\mathbf{y},\bar{\mathbf{x}}_f\right) d\mathbf{x}. \tag{7.3}$$

Equation (7.3) shows that the "correction" to the mean of the prior that produces the mean of the posterior is the expected error given the distribution of errors conditioned on today's observation and prior mean. Without loss of generality we may consider the right-hand side of (7.3) as simply an unknown function of today's observation and prior mean. By taking this perspective (7.3) may be written concisely as

$$\bar{\mathbf{x}} - \bar{\mathbf{x}}_f = \mathbf{f}\left(\mathbf{y},\bar{\mathbf{x}}_f\right). \tag{7.4}$$

The vector-function $\mathbf{f}$ is assumed smooth and is the object of central interest. One way to understand the structure of $\mathbf{f}$ is through an expansion in terms of the observation about the prior estimate of that observation (Jazwinski 1998, pp. 340–346), i.e.

$$\bar{\mathbf{x}} - \bar{\mathbf{x}}_f = \mathbf{f}_0 + \mathbf{M}_1\mathbf{v} + \mathbf{M}_2\mathbf{v}^2 + \ldots \tag{7.5}$$

where the innovation, $\mathbf{v} = \mathbf{y} - \mathbf{H}\bar{\mathbf{x}}_f$, the matrix $\mathbf{H}$ is the observation operator that takes the N-dimensional state vector, $\bar{\mathbf{x}}_f$, into the p-dimensional observation space, and the vector $\mathbf{f}_0 = \mathbf{f}(\mathbf{v} = \mathbf{0})$ and the matrix coefficients of the expansion, $\mathbf{M}_i$, are to be determined below.

The unusual vector notation $\mathbf{v}^2$ represents a vector of length p^2 such that,

$$\mathbf{v}^2 = \mathbf{v} \otimes \mathbf{v} = \left[v_1\mathbf{v}^T \; v_2\mathbf{v}^T \; \cdots \; v_p\mathbf{v}^T \right]^T , \tag{7.6}$$

where the symbol "⊗" refers to the Kronecker product and v_i is the ith element of the innovation vector. A similar representation applies to the p^3-vector $\mathbf{v}^3$ and so on.

The entire polynomial expansion in (7.5) may formally be represented as

$$\bar{\mathbf{x}} - \bar{\mathbf{x}}_f = \mathbf{f}_0 + \mathbf{G}\hat{\mathbf{v}} \tag{7.7}$$

where

$$\mathbf{G} = [\mathbf{M}_1 \; \mathbf{M}_2 \ldots \mathbf{M}_\infty] , \tag{7.8}$$

$$\hat{\mathbf{v}} = \left[\mathbf{v}^T \; \mathbf{v}^{2T} \; \cdots \; \mathbf{v}^{\infty T} \right]^T . \tag{7.9}$$

Equation (7.7) describes how to calculate the mean of the posterior using the "linear" update involving the gain matrix $\mathbf{G}$ operating on the predictor vector $\hat{\mathbf{v}}$, where $\hat{\mathbf{v}}$ is comprised of an infinite number of predictors formed from today's innovation.

To determine the vector $\mathbf{f}_0$ we note that a random draw from the posterior distribution is $\mathbf{x} = \bar{\mathbf{x}} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a random variable with zero mean. Therefore, an equation for the "error" in estimating the true state as the posterior mean may be obtained by subtracting $\mathbf{x}$ from both sides of (7.7):

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_f - (\mathbf{f}_0 + \mathbf{G}\hat{\mathbf{v}}) . \tag{7.10}$$

Because the expected value of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_f$ must vanish this implies that

$$\mathbf{f}_0 = -\mathbf{G} \langle \hat{\mathbf{v}} \rangle , \tag{7.11}$$

where

$$\langle \hat{\mathbf{v}} \rangle = \left[\langle \mathbf{v} \rangle^T \; \left\langle \mathbf{v}^2 \right\rangle^T \; \cdots \right]^T , \tag{7.12}$$

and the notation $\langle\rangle$ represents the expected value of a random variable. Note that if we assume that $\langle \mathbf{v} \rangle = \mathbf{0}$ this implies that we have assumed that the observation and the mean of the prior are accurate in so far as the distribution of truth about them is unbiased.

Given (7.11) we may now re-write (7.10) as

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_f - \mathbf{G}\hat{\mathbf{v}}', \tag{7.13}$$

where $\hat{\mathbf{v}}' = \hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle$, which now clearly has the property that the "errors" have zero mean.

The standard technique to derive the minimum error variance estimate is to find the gain matrix in (7.13) that minimizes the trace of the *expected* posterior error covariance matrix, i.e.

$$\bar{\mathbf{P}} = \int_{-\infty}^{\infty} \mathbf{P}(\mathbf{v})\rho(\mathbf{v})d\,\mathbf{v}, \tag{7.14}$$

where $\mathbf{P}(\mathbf{v})$ is the posterior error covariance matrix,

$$\mathbf{P}\left(\mathbf{v}\right) = \int_{-\infty}^{\infty} \left(\mathbf{x} - \bar{\mathbf{x}}\right)\left(\mathbf{x} - \bar{\mathbf{x}}\right)^T \rho\left(\mathbf{x}\middle|\,\mathbf{v}\right)d\,\mathbf{x} \tag{7.15}$$

and $\rho(\mathbf{v})$ is the pdf that describes the distribution of innovations. As we will discuss in detail below the *expected* posterior error covariance matrix (7.14) is generally different from the actual posterior error covariance matrix (7.15) and these differences will be shown to be significant for distributions with significant third moments.

Upon making use of (7.13, 7.14, and 7.15) the *expected* posterior error covariance matrix may be calculated:

$$\bar{\mathbf{P}} = \mathbf{P}_f - \mathbf{G}\left\langle \hat{\mathbf{v}}'\boldsymbol{\varepsilon}_f^T \right\rangle - \left\langle \boldsymbol{\varepsilon}_f \hat{\mathbf{v}}'^T \right\rangle \mathbf{G}^T + \mathbf{G}\left\langle \hat{\mathbf{v}}'\hat{v}'^T \right\rangle \mathbf{G}^T. \tag{7.16}$$

The matrices in (7.16) are defined in Appendix 1.

By minimizing the trace of the *expected* posterior error covariance matrix (7.16) with respect to the matrix, $\mathbf{G}$, one may determine an expression for $\mathbf{G}$ that minimizes the expected posterior error variance, i.e.

$$\mathbf{G} = \left\langle \boldsymbol{\varepsilon}_f \hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle \left\langle \hat{\mathbf{v}}'\hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle^{-1} \tag{7.17}$$

We may subsequently use (7.17) in (7.16) to obtain the expected posterior error covariance matrix:

$$\bar{\mathbf{P}} = \mathbf{P}_f - \left\langle \boldsymbol{\varepsilon}_f \hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle \left\langle \hat{\mathbf{v}}'\hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle^{-1} \left\langle \hat{\mathbf{v}}'\boldsymbol{\varepsilon}_f^T \right\rangle. \tag{7.18}$$

By truncating the expansion in (7.5) we may obtain approximations of various levels that coincide with polynomial regression. The first two levels of approximation are explored next.

7.2.2.1 Linear Regression

By truncating the polynomial expansion in (7.5) at the linear term and minimizing the trace of the expected error covariance matrix we find

$$\mathbf{f}_0 = \mathbf{0}, \tag{7.19}$$

$$\mathbf{M}_1 = \mathbf{P}_f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{7.20}$$

where it should be recognized that we have derived the Kalman (1960) and Kalman-Bucy (1961) filter, which in geophysics is commonly implemented using ensemble methods and referred to as the EnKF:

$$\bar{\mathbf{x}} \approx \bar{\mathbf{x}}_a = \bar{\mathbf{x}}_f + \mathbf{K}\mathbf{v}, \tag{7.21}$$

$$\bar{\mathbf{P}} \approx \bar{\mathbf{P}}_a = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}_f, \tag{7.22}$$

where

$$\mathbf{K} = \mathbf{P}_f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R}\right)^{-1}. \tag{7.23}$$

Two issues should be noted. First, the linear filter is incapable of estimating the constant term $\mathbf{f}_0$ of the expansion and therefore the estimate of the mean of the posterior for vanishingly small innovations is simply that of the mean of the prior. As we will show, this is a good assumption for symmetric prior distributions but not for strongly skewed prior distributions. Second, while (7.21) is correctly an approximation to the true posterior mean the updated posterior error covariance matrix in (7.22) is an approximation to (7.14) and not an approximation to the actual posterior error covariance matrix in (7.15). This second issue will be shown below to be the reason why EnKF ensemble generation algorithms are unable to produce the appropriate posterior distribution in the presence of phase errors.

7.2.2.2 Quadratic Nonlinear Regression

By truncating the polynomial at the quadratic term and minimizing the trace of the expected error covariance matrix we find

$$\mathbf{f}_0 = -\,(\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{T}_f\mathbf{H}_2^T\mathbf{\Pi}^{-1}\left\langle \mathbf{v}^2\right\rangle, \tag{7.24}$$

$$\mathbf{M}_1 = \mathbf{K} - (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{T}_f\mathbf{H}_2^T\mathbf{\Pi}^{-1}\mathbf{H}_2\mathbf{T}_f^T\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{7.25}$$

$$\mathbf{M}_2 = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{T}_f\mathbf{H}_2^T\mathbf{\Pi}^{-1}. \tag{7.26}$$

In (7.24, 7.25, and 7.26) we can see that the inclusion of the quadratic term has not just added a new term, $\mathbf{M}_2$, but also corrected the previous terms. Hence, one may think of this procedure of adding additional terms and finding a gain matrix that minimizes the expected error covariance as developing an “expansion” for each of the terms in the exact gain in (7.17). The linear filter of the previous section consists of the lowest order terms of the expansion of the coefficients in (7.5). It is important to recognize however that for vanishingly small innovations and a prior distribution with a significant third moment that the leading term of the complete expansion

(7.5) is $\mathbf{f}_0$. Hence, in the limit of vanishingly small innovation the linear filter (7.21) fails to include the largest term of the complete expansion (7.5), while the quadratic filter makes an *estimate* of $\mathbf{f}_0$.

We may write the terms in (7.24, 7.25, and 7.26) as in (7.5) to derive the "quadratic" ensemble filter:

$$\bar{\mathbf{x}} \approx \bar{\mathbf{x}}_a = \bar{\mathbf{x}}_f + \mathbf{K}\mathbf{v} + (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{T}_f\mathbf{H}_2^T\boldsymbol{\Pi}^{-1}\left[\mathbf{v}^{2\prime} - \mathbf{H}_2\mathbf{T}_f^T\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R}\right)^{-1}\mathbf{v}\right], \tag{7.27}$$

$$\bar{\mathbf{P}} \approx \bar{\mathbf{P}}_a = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}_f - (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{T}_f\mathbf{H}_2^T\boldsymbol{\Pi}^{-1}\mathbf{H}_2\mathbf{T}_f^T\left(\mathbf{I} - \mathbf{H}^T\mathbf{K}^T\right), \tag{7.28}$$

Where $\mathbf{v}^{2\prime} = \mathbf{v}^2 - \langle\mathbf{v}^2\rangle$. Note that equations (7.27 and 7.28) are simply the linear filter of section (3.a) with a correction term that is proportional to the third moment of the prior distribution. Hence, if the prior distribution is perfectly symmetric the quadratic filter reduces to the linear filter. In this case it would prove more effective to truncate the polynomial expansion at the cubic term in order to retain a correction to the linear filter. In addition, note that the expected error variance (7.28) is that of the linear filter (7.22) with a new term that for all non-zero third moments reduces the trace of the analysis error covariance to less than that of the linear filter.

Another way to view the quadratic ensemble filter is through (7.7) as a kind of "Kalman filter" in an extended state-space:

$$\bar{\mathbf{x}} \approx \bar{\mathbf{x}}_a = \bar{\mathbf{x}}_f + \mathbf{Z}\mathbf{w}, \tag{7.29}$$

$$\hat{\bar{\mathbf{P}}}_a = \left(\mathbf{I} - \hat{\mathbf{P}}_f\hat{\mathbf{H}}^T\left[\hat{\mathbf{H}}\hat{\mathbf{P}}_f\hat{\mathbf{H}}^T + \hat{\mathbf{R}}\right]^{-1}\hat{\mathbf{H}}\right)\hat{\mathbf{P}}_f, \tag{7.30}$$

where $\mathbf{Z}$ is the first N rows of $\hat{\mathbf{Z}}$ (Please see Eq. 7.45 of Appendix 1) and the weights of the prior ensemble may be calculated from

$$\mathbf{w} = \hat{\mathbf{Z}}^T\hat{\mathbf{H}}^T\left[\hat{\mathbf{H}}\hat{\mathbf{P}}_f\hat{\mathbf{H}}^T + \hat{\mathbf{R}}\right]^{-1}\hat{\mathbf{v}}'. \tag{7.31}$$

Again, please see Appendix 1 for notation. The equations (7.29 and 7.30) are in fact identical to that of (7.27 and 7.28) but because its form is identical to that of a Kalman filter this representation allows a practical algorithm to be developed that may be easily incorporated into an already constructed EnKF (Hodyss 2012). The formulation as (7.29 and 7.30) makes clear that the quadratic ensemble filter can be viewed as "linear" regression in an extended state-space in which the implied nonlinearities have been linearized by extending the state-space to include them. Please see Appendix A in Hodyss (2011) and also Hodyss (2012) for more discussion on this extended state-space.

Writing the quadratic ensemble filter as in (7.27 and 7.28) is interesting because it illustrates the potential errors the EnKF (linear regression) may make when applied to situations with skewed prior distributions. The potential errors that an

EnKF may make is simply noted by asking when the new terms in (7.27 and 7.28) dominate over those of the EnKF; there are two important situations when this happens. The first situation in which skewness is a significant issue for an EnKF is when the innovation is very large when compared with its variance. The EnKF makes a correction to the prior mean that goes linearly with the innovation. However, depending on the direction of the skewness of the posterior (sign of the third moment) and the sign of the innovation this correction will either be too large or too small. This is due to the fact that the true posterior mean is a curved (nonlinear) function of the innovation whenever the posterior has significant skewness (Hodyss 2011). Therefore, the EnKF's linear (in the innovation) estimate of the curved posterior mean will always contain significant error whenever the innovation and the skewness of the posterior are large.

The second situation illustrated by (7.27 and 7.28) in which skewness is a significant issue for an EnKF occurs when the innovation is very small when compared with its variance. The fact that the EnKF makes significant errors when the innovation is small is rather surprising and is a result of the fact that when the innovation is small and the prior third moments are large the $\mathbf{f}_0$ term dominates the expansion. Note that when the innovation is small the estimate of the posterior mean from the EnKF (See 7.21) is just that of the prior mean. However, when there is significant skewness in the prior distribution the prior mean is not a good estimate of the posterior mean. This can be seen in (7.7) through the importance of the $\mathbf{f}_0$ term for small innovation. The point to be made here is that whenever there is significant skewness, and the innovation is very small, the posterior mean and the prior mean will differ. However, in this situation the EnKF will not make a correction to the prior mean and in terms of its estimate of the posterior mean it will behave as if there were no observation to assimilate. More about these issues will be discussed below in Sect. 7.4

7.3 Distributions Arising from Phase Errors

Imagine a localized disturbance to a fluid, such as a tropical cyclone (TC). Suppose that this disturbance has a pressure field that appears as in Fig. 7.1a. Further suppose that the center (point of lowest pressure) of this disturbance is situated at a grid point ($x = x_0 = 0$) of our model and we are, for now, interested in the moments of the prior distribution at this grid point. The uncertainty in the prior is assumed to arise from a normally distributed random variable $\varphi \sim N(0, \sigma^2)$ denoting the location of the disturbance. No amplitude (structural) uncertainty will be considered.

The uncertainty in phase has been assumed to be normally distributed in order to show that non-Gaussianity in a state variable such as pressure can arise even from normally distributed phase errors. There is however no reason to expect that phase errors in complex fluid dynamical systems would be normally distributed. This point is explored further in Appendix 2. In Appendix 2 we show through the method of characteristics that in general sheared flows the variable translation speed

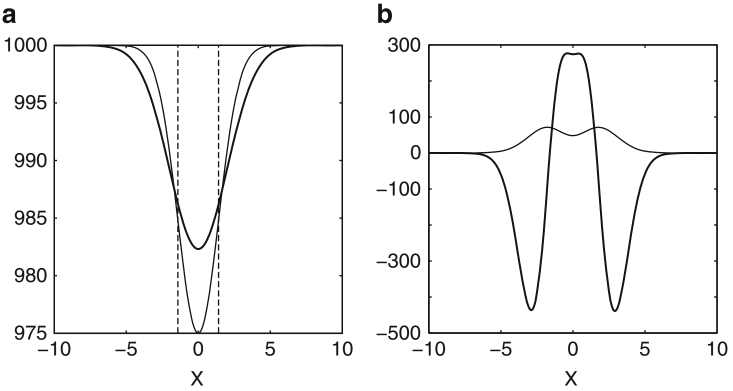


Fig. 7.1 Gaussian phase error model. In (**a**) is shown a single member (*thin solid*) of the distribution centered at the origin as well as the mean of the distribution (*thick solid*). The *vertical dashed lines* denote the location of the inflection points. In (**b**) is shown the variance (*thin solid*) and the third moment (*thick solid*)

of a disturbance may lead to non-Gaussian phase uncertainty. The strength of the shearing of the flow is shown to be the determining factor as to whether the resulting phase distribution will be Gaussian or non-Gaussian.

Returning to our phase error model of the pressure field of a TC we may write a Taylor-series approximation of the pressure field at x_0, of the form:

$$p\left(x_0;\varphi\right) = p\left(x_0\right) + \beta\left(x_0 - \varphi\right)^2 + \ldots, \tag{7.32}$$

where

$$\beta = \frac{1}{2}\left.\frac{d^2 p}{dx^2}\right|_{x=x_0} > 0. \tag{7.33}$$

Note that the term at leading-order that depends on the phase of the disturbance is quadratic because the term proportional to dp/dx vanishes at the center of the TC. Hence, the distribution of p at the mean (center) location of the disturbance is therefore approximately chi-square[1], which can be seen in the values of its scalar moments:

$$\langle p\rangle = p_0 + \beta\sigma^2, \tag{7.34}$$

$$\left\langle (p - \langle p\rangle)^2\right\rangle = 2\beta^2\sigma^4, \tag{7.35}$$

$$\left\langle (p - \langle p\rangle)^3\right\rangle = 8\beta^3\sigma^6. \tag{7.36}$$

[1] A chi-square distribution with one-degree of freedom is constructed by squaring each random draw from a Gaussian distribution.

Note that we have assumed that the variance of the phase uncertainty is small compared with the characteristic length scale of the disturbance such that the truncation of the Taylor-series to the quadratic term is sensible. Physically, the reason the distribution of p is approximately chi-square at this location arises because the values of pressure at this location can never be below p_0 but can vary as high as the far-field values of pressure allow. This type of hard lower limit characterizing the distribution of pressures is a characteristic of distributions like that of the chi-square distribution and always leads to significant third moments.

Similarly, we may perform a Taylor-expansion around the inflection point ($x = x_1$), which lies along the edge of the storm (See Fig. 7.1):

$$p\left(x_1;\varphi\right) = p\left(x_1\right) + \alpha\left(x_1 - \varphi\right) + \ldots, \tag{7.37}$$

where

$$\alpha = \left.\frac{dp}{dx}\right|_{x=x_1}. \tag{7.38}$$

Note that at the inflection point of the TC the second derivative vanishes while the first derivative (α) is large. Hence, the distribution in the vicinity of the inflection point is nearly Gaussian and therefore has vanishingly small third moment, which can be seen in the values of its moments:

$$\langle p\rangle = p_1 + \alpha x_1, \tag{7.39}$$

$$\left\langle (p - \langle p\rangle)^2 \right\rangle = \alpha^2\sigma^2, \tag{7.40}$$

$$\left\langle (p - \langle p\rangle)^3 \right\rangle = 0. \tag{7.41}$$

Therefore, we have so far seen that at the mean position of the phase distribution the structure of the pressure distribution is strongly non-Gaussian, but as we make our way away from the center of the phase distribution we find that the pressure distribution appears to become approximately Gaussian.

This variation in the structure of the scalar distributions at each grid point along the disturbance can be summarized by eliminating the Taylor-expansion and simply randomly sampling from a known distribution. We define a function that is vaguely similar to the pressure field of a tropical cyclone,

$$p\left(x;\varphi\right) = 1000 - 25\exp\left[-\frac{\left(x-\varphi\right)^2}{4}\right], \tag{7.42}$$

and proceed to randomly sample this function by drawing values of $\varphi \sim N(0,2)$. This function represents a Gaussian-shaped depression that is positioned at random locations and is in fact the function plotted in Fig. 7.1a. By calculating the scalar moments of this distribution at each grid point we find a pattern as in Fig. 7.1b.

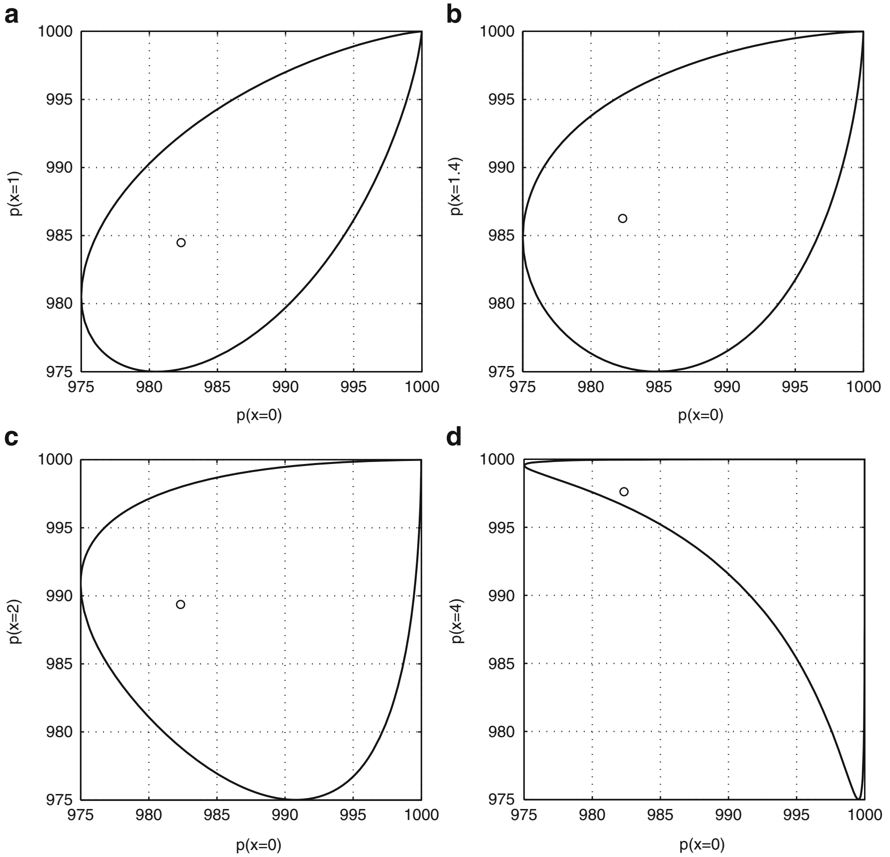


Fig. 7.2 The structure of phase error distribution as loops. In (**a**) is the distribution of pressure between x = 0 and x = 1, (**b**) is for x = 0 and x = 1.4, (**c**) is for x = 0 and x = 2, and (**d**) is for x = 0 and x = 4. The *open circle* in each panel denotes the location of the mean

The pattern found in Fig. 7.1b is consistent with the analysis presented above in so far as the third moment is positive at the center of the distribution and decreases to zero near the inflection point. Figure 7.1b reveals however that the third moment actually turns negative outside of the inflection points leading to a tri-pole structure. This tri-pole structure in the third moments depends on the strength of the phase error variance of φ. When the phase error variance is large, rendering the assumptions about truncating the Taylor-series invalid, the tri-pole pattern in the third moments is replaced with a relatively wide monopole negative region (not shown).

To gain understanding of the multivariate structure we plot in Fig. 7.2 the distribution of pressure values at the center of the distribution ($x = x_0$) against the values of pressure at various locations for the same Gaussian phase error

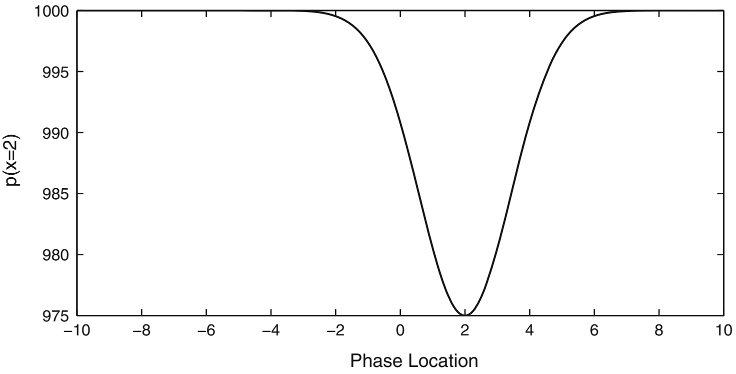


Fig. 7.3 The distribution of pressure at x = 2 as a function of the location of the feature

model (7.42). Note that the structure of the distribution in this plane focuses all the members along a single loop structure. This loop arises because for a value of pressure at, say $x = x_0$, there are always two different phase values, φ, that are consistent, which leads to two different possible pressure values at $x = x_1$. We emphasize that the loop structure is not restricted to just the relationship between the center and the inflection point, but also occurs between all other locations as well; the loop structure shown in Fig. 7.2 between other locations differs only in the shape of the loops. The existence of these loops has important implications to the accuracy of the DA. Because the relationship between any two locations is not only nonlinear but also multi-valued the DA must be able to choose which side of the loop is the correct side. Subsequently, a single observation of pressure cannot discern the location of the feature because the distribution is always multi-modal. Moreover, as shown in Fig. 7.2 the loop structure of the prior distribution assures that the prior mean will not be a state representative of any particular member. The ensemble mean value in this plane is denoted in Fig. 7.2 and can be seen to be well away from the loop. This loop structure makes clear the complexity of these distributions and implies significant high-order multivariate moments.

In tropical cyclone data assimilation the location (or phase, φ) of the minimum central pressure is an often used observation. Therefore, it is of interest to examine the prior distribution of phase locations (φ) plotted against the values of pressure at some location we might update with that observation of location. In this respect this tells us the relationship between the state and the prior estimates of the observed variables. Figure 7.3 shows what this looks like for phase locations (φ) plotted against the values of pressure at $x = 2$ for our Gaussian phase error model (7.42). Because the disturbance structure is simply a function of the phase the distribution here is not a loop like Fig. 7.2 but simply the actual structure of the Gaussian phase error model (7.42). Nevertheless, this functional dependence is clearly nonlinear and as we show next will lead to significant difficulties with Kalman filter-based DA.

7.4 DA in the Presence of Phase Errors

7.4.1 Idealized Cases

In this section we will perform DA on the distribution (7.42) studied in Sect. 7.3. In particular, we set the distribution shown in Fig. 7.3 to be our prior. This implies that we will simply observe the location of the feature and attempt to update the state. As a baseline to compare linear regression (EnKF) against quadratic nonlinear regression (quadratic ensemble filtering) we will employ a particle filter [See van Leeuwen 2009) for a review]. Because we will be using an ensemble of size 20,000 in this section the posterior mean obtained from the particle filter will be for all intents and purposes identical to that of the un-approximated application of Bayes' rule. Because we have one observation (the location of the feature) and 20,000 ensemble members in this particle filter these experiments will not be contaminated by the detrimental issues of limited ensemble size discussed in Snyder et al. (2008). In any event, the goal in these experiments will be to understand when linear and/or quadratic nonlinear regression can and cannot get close to the result of the un-approximated application of Bayes' rule.

7.4.1.1 Estimating the Posterior Mean

To begin we will examine the situation of observing the location of the feature with very low observation error variance, $R = 0.01$. When the observation error is very low an EnKF based DA system will make substantial corrections for large innovations. However, because the relationship between the observation and the state space is nonlinear (See Fig. 7.3) the EnKF correction will not be accurate for all values of the innovation. We proceed to illustrate this in detail next.

In Fig. 7.4a, c, e is shown the results of the experiments with low observation errors. In each figure is shown the true state that is observed. In Fig. 7.4a the true state is the Gaussian phase error model (7.42) for a $\varphi = 0$ Because the distribution of phase errors is $\varphi \sim N(0, 2)$, the prior mean is identically zero and because the observation in this case is zero the innovation is also identically zero. For this innovation the particle filter's estimate of the posterior mean is for all intents and purposes identical to the true state and hence indistinguishable from the true state in this figure. What this means is that a high quality observation of the location of the feature should identify that feature precisely because there is no uncertainty in the structure of the feature. Recall that for each value of location (the observation in this case) the distribution in Fig. 7.3 is a nonlinear but *single-valued* function. This implies that if one's DA system can handle the nonlinearity, then one observation can identify the feature quite closely. However, in Fig. 7.4a one can see the obvious result of the zero innovation being that the EnKF estimate of the posterior mean is actually identical to the prior mean, i.e. *no correction has been made*. In contrast, the Quadratic Ensemble Filter makes a correction even when the innovation is zero,

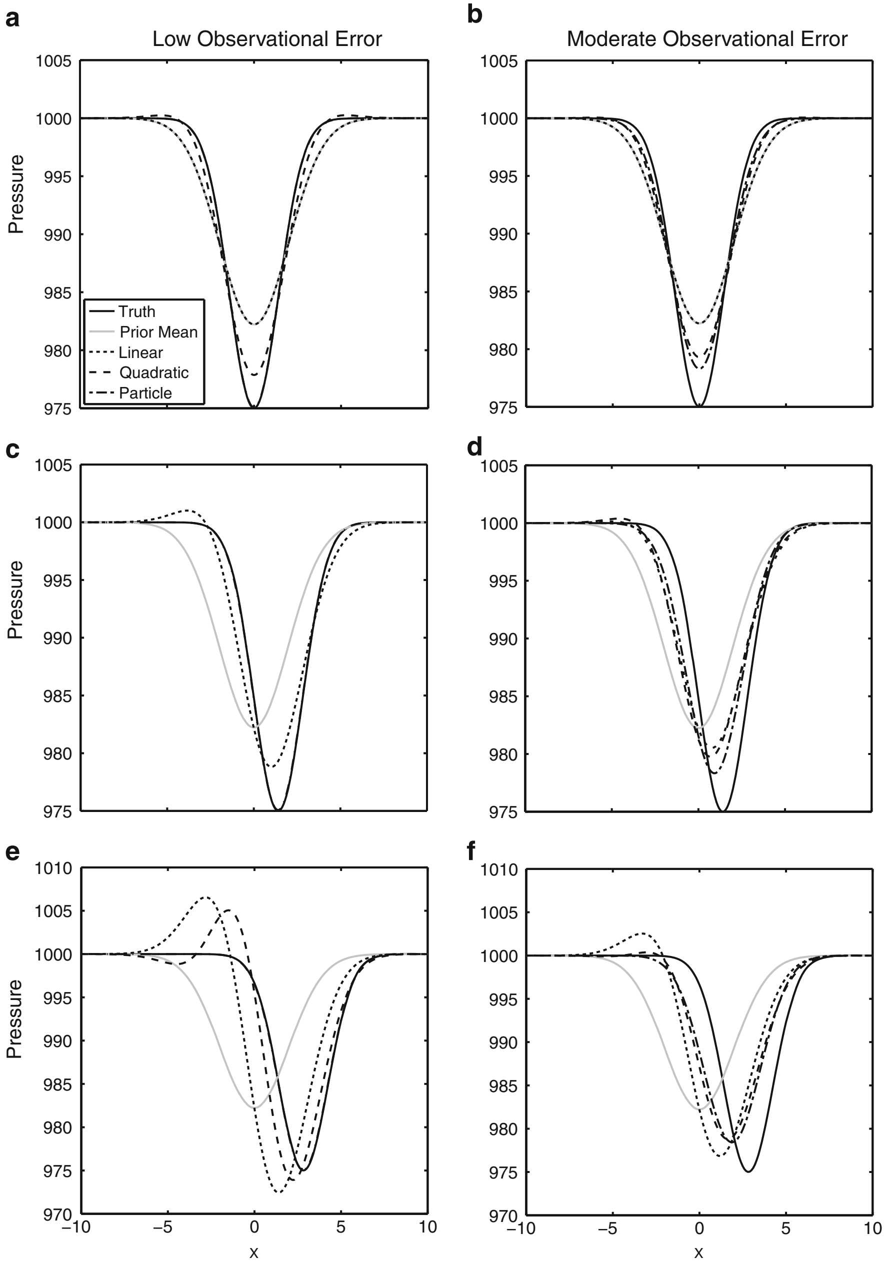


Fig. 7.4 Estimates of the posterior mean from phase error distributions. The *left column* shows results for low observation errors and the *right column* shows results for moderate observation errors. The *first row* shows results for the true state located at the origin; the *second row* is for the true state at one standard deviation; the *third row* is for the true state at two standard deviations

as was already discussed in detail at the end of Sect. 7.2. Note that this result is rather remarkable in so far as the phase location of the feature is located precisely at its expected value and one might hope that this would be the location that an EnKF DA system would perform well.

In fact, an EnKF DA system performs better (e.g., similar to quadratic nonlinear regression), not when the innovation is very small, but actually when the innovation is equal to its expected value (Hodyss 2011). This result can be seen in Fig. 7.4c. In Fig. 7.4c the true state is the Gaussian phase error model (7.42) for a $\varphi = \sigma$, i.e. the phase location observation is taken as situated at one standard deviation and therefore the amplitude of the innovation is essentially its expected value. Again, the particle filter's estimate of the posterior mean is indistinguishable from the true state in this case. However, the EnKF result is now basically identical to the Quadratic Ensemble Filter result. Mathematically, we can understand that this occurs when the quantity in square brackets in (7.27) vanishes, and this happens for innovations that have the amplitude of their expected value ($\langle \mathbf{v}^2 \rangle \approx \sigma^2$) because the third moment in observation space (location observations) is zero for Gaussian phase errors.

In Fig. 7.4e we show an example for which the true state is the Gaussian phase error model (7.42) for a $\varphi = 2\sigma$, i.e. the phase location observation is taken as situated at two standard deviations and therefore the amplitude of the innovation is significantly larger than its expected value. Again, the particle filter's estimate of the posterior mean is indistinguishable from the true state in this figure. However, in this case the EnKF's estimate of the posterior mean and that of the Quadratic Ensemble Filter is now quite different. When the observation errors are low but the innovation is larger than its expected value a common way in which the EnKF is in error is to produce an estimate of the posterior mean that has too much amplitude. This can be seen in Fig. 7.4e by the large values in the estimate of the posterior mean above 1,000 mb and below 975 mb. Note that the true posterior mean does not go above 1,000 mb or below 975 mb. In addition, the phase location of the estimate from the EnKF typically does not shift far enough towards the true location. The Quadratic Ensemble Filter produces an estimate that has significantly reduced issues with the amplitude of the estimate of the posterior mean and the phase location and this is obviously due to its ability to account for the nonlinearity seen in Fig. 7.3.

Next, we will consider the situation where the observation error variance is substantially larger: $R = 1$. In this case the true posterior mean from the particle filter will not be particularly close to the true state as the observation is not accurate enough to distinguish the location of the feature exactly. This can be seen in Fig. 7.4b. In this case the innovation is again zero because the observation is taken to be located at the mean phase location. Again, the estimate of the posterior mean from the EnKF is just that of the prior mean. In contrast, the Quadratic Ensemble Filter now produces an estimate of the posterior mean that is very close to that of the true posterior mean from the particle filter. In Fig. 7.4d is shown the case where the location of the observation is at one standard deviation. As one can see from Fig. 7.4d the state estimate from the EnKF is now again very close to that of the Quadratic Ensemble Filter. Therefore, for this size of innovation both linear and quadratic nonlinear regression give very similar estimates of the posterior mean and they are quite close to the true posterior mean. In Fig. 7.4f is shown the case where the location of the observation is at two standard deviations. Here, again we see that the EnKF estimate has the wrong phase location (with respect to the true posterior mean) and still has substantially too large values (greater than 1,000 mb) and also too small of values that are less than that of the true posterior mean minimum value of approximately 978 mb. In contrast, the Quadratic Ensemble Filter matches the location and amplitude of the true posterior mean quite closely.

7.4.1.2 Ensemble Update

Estimating the posterior mean is only one-half of the calculations necessary to update the ensemble posterior distribution. The other half of the calculation is to produce a set of ensemble perturbations that appropriately perturb the estimated posterior mean. Traditional ensemble generation schemes, like perturbed observations (Houtekamer and Mitchell 1998; Burgers et al. 1998; Evensen 1994) and square root filters (Anderson 2001; Bishop et al. 2001; Tippet et al. 2003), are all based on determining the actual posterior error variance (7.15) from its expected value (7.14). As discussed by Hodyss (2011), this is an accurate technique for normal distributions because the posterior error variance is not a function of the innovation (or observation) for normal distributions. Furthermore, Hodyss (2011) showed that the posterior error variance (7.15) is a function of the innovation (observation) whenever the posterior distribution is skewed, i.e.

$$\frac{dP}{dv} = \frac{T}{R}, \tag{7.43}$$

where P is the posterior error variance, T is the posterior third moment, and R is the observation error covariance and v is the innovation. [For clarity (7.43) was written for a scalar system; multivariate generalizations may be found in Appendix 2 of Hodyss (2011).] We show next that this issue is particularly significant for phase error distributions and location observations because the location of maximum posterior error variance must shift with the location of the observation.

In Fig. 7.5 is shown the diagonal of the posterior error covariance matrix from (7.22) for the EnKF and (7.28) for the Quadratic Ensemble Filter. Recall that these equations are the object that is constraining perturbed observation and square root update algorithms. It is important to keep in mind that (7.22 and 7.28) are not attempting to approximate the true posterior error variance (7.15), but in fact are an approximation to the *expected* posterior error variance (7.14). This fact can be seen in Fig. 7.5. The panels in Fig. 7.5 correspond to the same experiments as Fig. 7.4. The major result to be noted from Fig. 7.5 is that in each column the posterior error variance from an EnKF (7.22) and quadratic ensemble filter (7.28) algorithm does not move with the location of the true state. Indeed, equations (7.22 and 7.28) do not include information from the actual observation (innovation) of location. Note in contrast we may use the particle filter to calculate the true posterior error variance as a function of the innovation (7.15). These curves are plotted in Fig. 7.5 and show that the true posterior error variance as a function of the innovation does in fact shift with

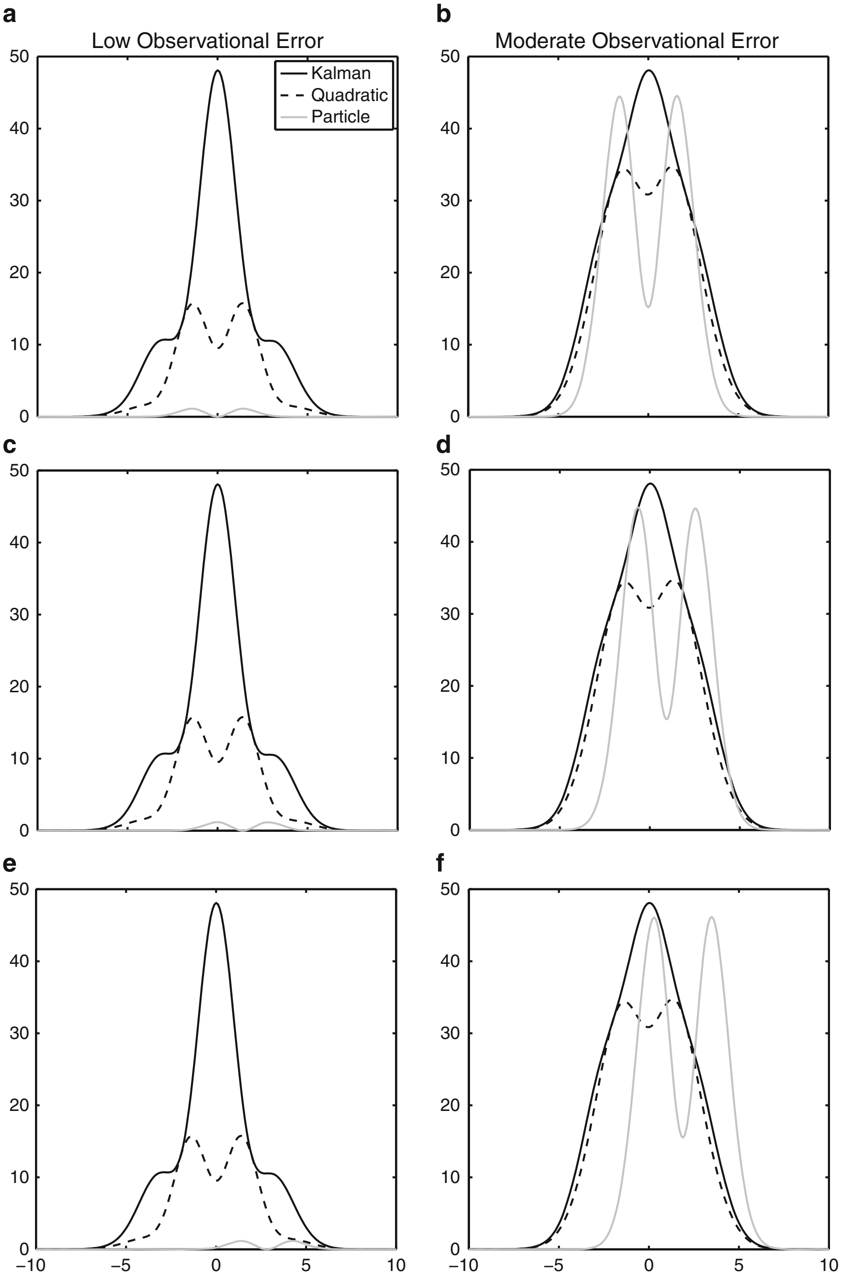


Fig. 7.5 Same as Fig. 7.4 except now each panel shows the estimates of the expected posterior error variance from the Kalman and Quadratic techniques as well as the posterior error variance as a function of the innovation from the particle filter technique. Note that the posterior error variance from the particle filter moves with the location observation while that from the other techniques does not

the location of the observation. We note in passing that we have also compared the structure of the true posterior covariance matrices (7.15) against those covariance matrices determined from (7.22 and 7.28). The result is that the covariance's are also similarly in error when compared to the structure of the true posterior covariance matrix (not shown). Because a covariance between two variables may be positive or negative, the error in (7.22 and 7.28) when compared to (7.15) may be wrong in both the magnitude as well as the sign of that covariance. Hence, to produce an ensemble consistent with the "errors of the day" requires the consideration of the posterior error covariance matrix as a function of the innovation (7.15). Unfortunately, traditional ensemble generation schemes make use of the *expected* posterior error variance (7.14), which does not properly account for information from the innovation.

7.4.2 Hurricane Katia (2011)

In this section we will illustrate that the idealized results found in the previous section can be found in a real ensemble DA experiment with a tropical cyclone. We utilize a prior distribution from an 80-member ensemble for Hurricane Katia (2011) generated with the Coupled Ocean/Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS®-TC; Doyle et al. 2012) ensemble data assimilation system. The COAMPS-TC system is a limited area model designed specifically for the simulation and prediction of tropical cyclones. It is comprised of a suite of packages and parameterizations that represent physical processes unique to the tropical environment. COAMPS-TC uses three horizontally nested domains with the horizontal resolution decreasing from 45 km on the outer basin-scale domain to 5-km on the inner vortex-scale domain. All calculations below are done on the 5-km inner vortex-scale domain. In order for COAMPS-TC to remain computationally efficient, the inner two domains are designed to track the location of the storm. The system utilizes the Data Assimilation Research Testbed (DART; Anderson et al. (2009) developed at the National Center for Atmospheric Research to assimilate observations with a square-root version of the EnKF as well as adaptive prior inflation.

The results presented in this section are based upon a prior distribution for Hurricane Katia valid on 12 UTC, 2 September 2011. The ensemble was initialized 00 UTC, 30 August by interpolating the Global Forecasting System ensemble (Hamill et al. 2011) to the three COAMPS-TC nested domains. The COAMPS-TC ensemble was then cycled by using DART to assimilate observations of radiosondes, cloud-track winds, surface observations, and aircraft data every 6-h until 12 UTC, 2 September. The fact that Katia developed away from land in the central Atlantic Ocean makes it an ideal case to test the various filter algorithms because any skewness will be a direct result of phase and intensity variability and not interactions with land.

This ensemble will be used in precisely the same way as in the previous section to compare the levels of approximation resulting from DA with linear and quadratic

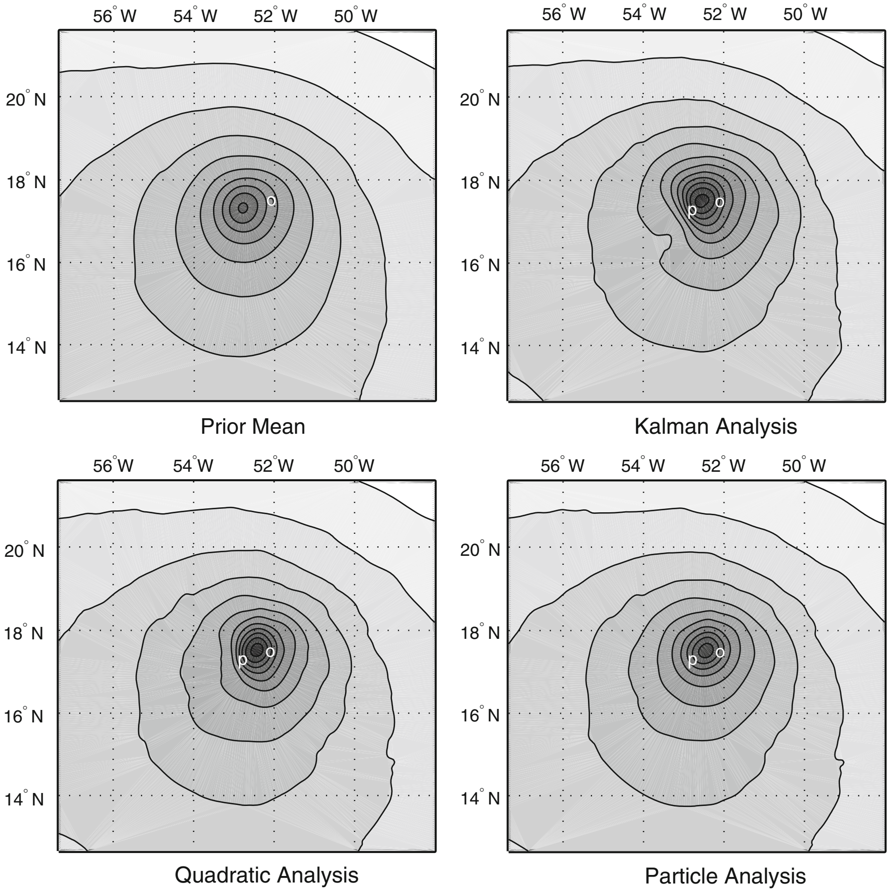


Fig. 7.6 Estimates of the posterior mean for Hurricane Katia (2011). The sea-level pressure is plotted in each panel with a contour from 990 to 1,016 mb at intervals of 2 mb. The "*o*" denotes the location of the observation and the "*p*" denotes the location of the center of the sea-level pressure in the prior mean

nonlinear regression against that resulting from the application of Bayes' rule as seen through particle filtering. In Fig. 7.6a is shown the prior mean sea level pressure that will be used in the DA experiments. In Fig. 7.6d is the true posterior mean from the particle filter after assimilation of just the position observations, which consists of two observations; one of which is the location in longitude and the other is the location in latitude. One can see that the result of the position observations was to shift the mean to the Northeast. Note however that the posterior mean was not shifted all the way to the observation location. The observation location is at about one standard deviation from the prior mean in latitude and two standard deviations from the prior mean in longitude. Because this location observation is at a location that is greater than one standard deviation in longitude and the observation error variance is

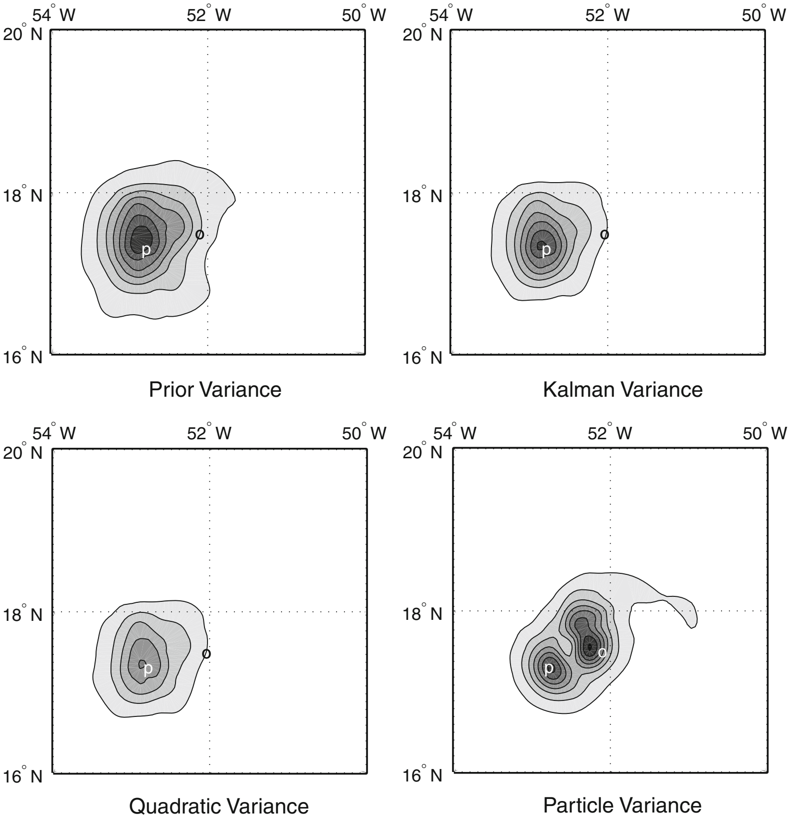


Fig. 7.7 Estimates of the posterior error variance for Hurricane Katia (2011). In each panel the contour interval is 0–80 mb^2 at intervals of 10, except for the particle filter which is 0–16 mb^2 at intervals of 2

of moderate value this example is most consistent with Fig. 7.4f from the idealized examples. Recall that in Fig. 7.4f the location of the posterior mean was also not centered on the position observation location, but rather was located approximately halfway between the observation location and the prior mean location. Similarly, note that the EnKF estimate in Fig. 7.5b is shifted less towards the prior mean location than when compared with that of the particle filter, and includes a small "high" to the Southwest of the prior mean location. This small high can be seen in the ridging to the Southwest of the prior mean location and also in the very flat appearance of the contours on the Southwest side of the TC. Again, recall that these features of the Kalman estimate were also seen in Fig. 7.4f. In contrast, the Quadratic Ensemble Filter estimate of the posterior mean is very close (in both position and structure) to that of the true posterior mean, which is again very similar to that of Fig. 7.4f.

In Fig. 7.7a is shown the prior error variance. The prior error variance has a single maximum in the vicinity of the prior mean location. In contrast, the true posterior error variance as a function of the position observation, which is of course the estimate of the posterior error variance obtained from the particle filter, has two maxima. One maximum is at the location of the prior mean and the other is near to the location of the observation. The structure of the true posterior error variance, in particular its multi-modal character, is again very similar to that found in the idealized experiment (Fig. 7.5f), which gives confidence to our estimate of the true posterior error variance using the limited ensemble size of 80 members. In Fig. 7.7b, c are the EnKF and quadratic ensemble filter estimates of the expected posterior error variance. Note that while the quadratic posterior error variance is less than that estimated by the EnKF neither has the correct structure of the posterior error variance; both have a single maximum in the vicinity of the prior mean location much like that of the idealized experiment in Fig. 7.5f. The fact that the ensemble mean from the quadratic ensemble filter is much better than that from the EnKF but that the ensemble generation is not significantly better underscores the fact that contemporary ensemble generation techniques do not properly account for the latest set of observations.

7.5 Summary and Conclusions

This chapter has explored in detail the issues surrounding the impact of phase errors on the ability of traditional ensemble-based Kalman filtering (EnKF) algorithms to accurately reproduce the posterior mean and perturbations to that mean that sample the posterior distribution. We began by illustrating the relationships between EnKF algorithms and linear and nonlinear regression. Here, we saw that quadratic nonlinear regression is simply the EnKF with a correction term that provides some accounting for the prior third moment. The prior third moment turns out to be of some significance as prior distributions whose uncertainty arises from uncertainty in the location of the feature have large third moments. This third moment (or skewness) of the prior distribution was shown to lead to difficulties for EnKF algorithms.

We have shown that an important issue with the estimation of the posterior mean from an EnKF algorithm is the size of the innovation. In situations with non-zero prior third moments, such as phase uncertainty, the posterior distribution is almost always a curved (nonlinear) function of the innovation (Hodyss 2011). This implies that for large innovations the EnKF will always produce significant error because it is a linear function of the innovation. More surprising however is the fact that the EnKF will also make a significant error in its estimate of the posterior mean whenever the innovation is very small. This is because whenever the posterior third moment is large and the innovation is small the posterior mean curves away from the prior mean. This leads to significant error in the estimate of the posterior mean because for small innovation the EnKF estimate of the posterior mean is always the prior mean. In fact, the size of the innovation for which an EnKF algorithm is most accurate is when the magnitude of the innovation is approximately equal to its expected value.

We have shown that a significant issue with the generation of an ensemble that correctly samples the posterior distribution is in the calculation of the posterior error variance. One way to view ensemble generation is simply as a method to estimate the posterior error covariance matrix. Because normal distributions have posterior covariance matrices that are independent of the innovation (observation) this has led to a significant number of ensemble generation algorithms based on the *expected* posterior error covariance matrix. This unfortunately is an assumption and it fails most strongly whenever the posterior distribution has non-zero third moments (skewness). The generation of an ensemble in feature-based systems with observations of location leads to this issue becoming extremely important. We have seen in both idealized cases and in a real tropical cyclone that the correct posterior error covariance matrix knows the location of the observation whereas the *expected* posterior error covariance matrix does not. This leads to two problems: (1) the uncertainty is not appropriately centered in the correct location and (2) the structure of the variances and co-variances, even if they were correctly located through a shift in their position, are simply not correct. The first problem leads to the uncertainty being in the wrong location and subsequently not accurately predicting the probability of particular events. The second problem implies that the ensemble generation cannot produce structures that are self-consistent with the structures within the prior ensemble. This means that the TCs that are produced by the ensemble will not have the correct relationships between variables because the co-variances are simply incorrect. It is important to realize that this issue was not corrected by the use of quadratic nonlinear regression. This is because while the polynomial expansion in the innovation suggested in Sect. 7.2 provides a convergent estimate of the posterior mean with higher-order approximations, it does not in fact provide a convergent estimate of the posterior error covariance matrix as a function of the innovation when combined with the ensemble generation algorithms referred to as perturbed observations (Houtekamer and Mitchell 1998; Burgers et al. 1998; Evensen 1994) or square root filters (Anderson 2001; Bishop et al. 2001; Tippet et al. 2003). The reason for this is because these algorithms for generating ensemble perturbations are based on an estimate of the posterior error covariance matrix that is incorrect in situations with significant third moments. Both of these ensemble generation algorithms are based on a covariance matrix that is obtained by estimating the weighted averaged over all possible posterior covariance matrices rather than the one associated with the latest innovation.

Nevertheless, from a practical point of view, the fact that the estimate of the posterior mean from the quadratic ensemble filter is more accurate than that from an EnKF and subsequently that the posterior error variance is smaller may in fact translate into a better performing data assimilation system in the presence of phase uncertainty. Research in this direction with tropical cyclones and other atmospheric phenomena is ongoing.

Acknowledgements We gratefully acknowledge support from the Office of Naval Research through PE-0601153N.

Appendix 1: Matrix and Notation Definitions

The prior error covariance matrix is extended to include the higher moments:

$$\hat{\mathbf{P}}_f = \frac{\hat{\mathbf{Z}}\hat{\mathbf{Z}}^T}{K-1} = \begin{bmatrix} \mathbf{P}_f & \mathbf{T}_f \\ \mathbf{T}_f^T & \mathbf{F}_f - \mathbf{p}_f \mathbf{p}_f^T \end{bmatrix}, \tag{7.44}$$

whose square root form may be approximated with an ensemble as

$$\hat{\mathbf{Z}} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 & \cdots & \boldsymbol{\varepsilon}_K \\ \boldsymbol{\varepsilon}_1 \otimes \boldsymbol{\varepsilon}_1 - \mathbf{p}_f & \boldsymbol{\varepsilon}_2 \otimes \boldsymbol{\varepsilon}_2 - \mathbf{p}_f & \cdots & \boldsymbol{\varepsilon}_K \otimes \boldsymbol{\varepsilon}_K - \mathbf{p}_f \end{bmatrix} \tag{7.45}$$

and the vectorized covariance matrix, $\mathbf{p}_f = vec(\mathbf{P}_f)$, is an N^2-vector constructed from the concatenation of the N columns of $\mathbf{P}_f$ and whose organization follows that of the Kronecker product "⊗", and K is the ensemble size.

The extended observation operator takes the following form:

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{H} \\ \mathbf{H} \otimes \mathbf{H} \end{bmatrix}. \tag{7.46}$$

Note that for nonlinear observation operators one would not use a linearized form of the operator. Instead, the correct procedure is to operate the nonlinear observation operator on each member of the ensemble and then perform linear or nonlinear regression on this new distribution of predicted prior observations against the state variables needing update (Houtekamer and Mitchell 2001).

The covariance matrices in the extended state-space takes the form:

$$\left\langle \boldsymbol{\varepsilon}_f \hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle = \begin{bmatrix} \mathbf{P}_f \mathbf{H}^T & \mathbf{T}_f \mathbf{H}_2^T & \cdots \end{bmatrix}, \tag{7.47}$$

$$\left\langle \hat{\mathbf{v}}' \hat{\mathbf{v}}'^{\mathbf{T}} \right\rangle = \left\langle \hat{\mathbf{v}} \hat{\mathbf{v}}^T \right\rangle - \left\langle \hat{\mathbf{v}} \right\rangle \left\langle \hat{\mathbf{v}}^T \right\rangle, \tag{7.48}$$

$$\left\langle \hat{\mathbf{v}} \hat{\mathbf{v}}^T \right\rangle = \begin{bmatrix} \mathbf{H}\mathbf{P}_f \mathbf{H}^T + \mathbf{R} & \mathbf{H}\mathbf{T}_f \mathbf{H}_2^T & \cdots \\ \mathbf{H}_2 \mathbf{T}_f^T \mathbf{H}^T & \mathbf{H}_2 \mathbf{F}_f \mathbf{H}_2^T + \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{R}_4 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{7.49}$$

$$\left\langle \hat{\mathbf{v}} \right\rangle \left\langle \hat{\mathbf{v}}^T \right\rangle = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \left\langle \mathbf{v}^2 \right\rangle \left\langle \mathbf{v}^{2T} \right\rangle & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{7.50}$$

and $\mathbf{0}$ is the $p \times p$ zero matrix. The matrices in (7.17) are defined as:

$$\mathbf{P}_f = \left\langle \boldsymbol{\varepsilon}_f \boldsymbol{\varepsilon}_f^T \right\rangle, \tag{7.51a}$$

$$\mathbf{T}_f = \left\langle \boldsymbol{\varepsilon}_f \boldsymbol{\varepsilon}_f^{2T} \right\rangle, \tag{7.51b}$$

$$\mathbf{F}_f = \left\langle \boldsymbol{\varepsilon}_f^2 \boldsymbol{\varepsilon}_f^{2T} \right\rangle, \tag{7.51c}$$

$$\mathbf{R} = \left\langle \boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_o^T \right\rangle, \tag{7.52a}$$

$$\mathbf{R}_4 = \left\langle \boldsymbol{\varepsilon}_o^2 \boldsymbol{\varepsilon}_o^{2T} \right\rangle, \tag{7.52b}$$

$$\mathbf{A} = \left\langle \boldsymbol{\varepsilon}_o^2 \boldsymbol{\varepsilon}_f^{2T} \right\rangle \mathbf{H}_2^T + \mathbf{H}_2 \left\langle \boldsymbol{\varepsilon}_f^2 \boldsymbol{\varepsilon}_o^{2T} \right\rangle, \tag{7.53}$$

$$\mathbf{B} = \mathbf{R} \otimes \mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{H}\mathbf{P}_f\mathbf{H}^T \otimes \mathbf{R}, \tag{7.54}$$

$$\mathbf{C} = \left\langle \left(\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_f^T \mathbf{H}^T \right) \otimes \left(\mathbf{H} \boldsymbol{\varepsilon}_f \boldsymbol{\varepsilon}_o^T \right) \right\rangle + \left\langle \left(\mathbf{H} \boldsymbol{\varepsilon}_f \boldsymbol{\varepsilon}_o^T \right) \otimes \left(\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_f^T \mathbf{H}^T \right) \right\rangle, \tag{7.55}$$

and $\mathbf{H}_2 = \mathbf{H} \otimes \mathbf{H}$ is the matrix operator that takes an N^2 vector into the p^2 predictor space and copious use of the identity, $(\mathbf{H}\boldsymbol{\varepsilon}_f) \otimes (\mathbf{H}\boldsymbol{\varepsilon}_f) = (\mathbf{H} \otimes \mathbf{H})(\boldsymbol{\varepsilon}_f \otimes \boldsymbol{\varepsilon}_f)$, has been made.

In (7.51a), $\mathbf{P}_f$ is a square matrix listing the necessary second moments of the prior distribution. In (7.51b), $\mathbf{T}_f$ is a rectangular matrix listing the necessary third moments of the prior distribution. In (7.51c), $\mathbf{F}_f$ is a square matrix listing the necessary fourth moments of the prior distribution. In (7.52a), $\mathbf{R}_4$ is a square matrix listing the necessary fourth moments of the observation likelihood. Note that even when the observation error covariance matrix, $\mathbf{R}$, is diagonal $\mathbf{R}_4$ is not diagonal. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are sparse, square matrices that represent various combinations of observation error covariances and forecast error covariances.

The covariance matrix of squared innovations is

$$\boldsymbol{\Pi} = \mathbf{H}_2\mathbf{F}_f\mathbf{H}_2^T + \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{R}_4 - \mathbf{H}_2\mathbf{T}_f^T\mathbf{H}^T \left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R} \right)^{-1} \mathbf{H}\mathbf{T}_f\mathbf{H}_2^T - \left\langle \mathbf{v}^2 \right\rangle \left\langle \mathbf{v}^{2T} \right\rangle. \tag{7.56}$$

Appendix 2: Non-Gaussian Phase Uncertainty from Variable Shear Flows

To isolate the effects of phase uncertainty we focus on the one-dimensional advection equation:

$$\frac{\partial p}{\partial t} + c\,(x)\,\frac{\partial p}{\partial x} = 0, \tag{7.57}$$

where $p = p(x,t)$ is the wavefield being advected at the speed $c(x)$. We attach to (7.57) an unbounded domain in x as well as a localized initial condition such as the functions (7.42), which was our prototype for the surface pressure field of members of a prior distribution of tropical cyclones. By the method of characteristics we know that the solutions to (7.57) move away from their initial location according to,

$$\frac{dX}{dt} = c(X), \tag{7.58}$$

where $X = X(t)$ is the location of, say, the minimum central pressure of the function (7.42). We may immediately note that even though equation (7.57) is linear in the amplitude of the disturbance, equation (7.58) may be non-linear if the function $c(x)$ is non-linear. Hence, the motion of the location of the minimum central pressure may evolve non-linearly and therefore even if the initial distribution of phase uncertainty is Gaussian it still may evolve into a non-Gaussian phase distribution owing to the non-linearity in (7.58). Two examples follow:

(1) Linear Shear

In the case where the shear is a linear increasing function of distance from the origin, i.e.

$$c(x) = c_0 \frac{x}{L}, \tag{7.59}$$

where c_0 is a characteristic phase speed and L is characteristic length scale. By inserting (7.59) into (7.58) and solving finds

$$X(t) = X_0 \exp\left(\frac{c_0 t}{L}\right), \tag{7.60}$$

where X_0 is the initial location of the minimum central pressure. In Sect. 7.3 the parameter X_0 was denoted as φ and was normally distributed. Notice that if the location of the minimum central pressure is normally distributed with mean $\bar{x}$ and variance σ^2, then at a later time, t, the phase distribution will be normal with

$$X \sim N\left(\bar{x} \exp\left(\frac{c_0 t}{L}\right), \sigma^2 \exp\left(\frac{2c_0 t}{L}\right)\right). \tag{7.61}$$

Hence, linear shear produces disturbances that move away from their initial location exponentially with time. Nevertheless, disturbances in linear shear preserve the Gaussian character of their initial phase uncertainty.

(2) Quadratic Shear

In the case where the shear is a quadratic function of distance from the origin, i.e.

$$c(x) = c_0 \left(\frac{x}{L}\right)^2, \tag{7.62}$$

the solution to (7.58) is

$$X(t) = \frac{L}{\frac{L}{X_0} - \frac{c_0 t}{L}}. \tag{7.63}$$

In this case, the phase locations of the disturbances move away from their initial locations faster than exponential with time. This can be seen by the fact that disturbances in linear shear flow approach infinity, i.e. $X \rightarrow \infty$, as $t \rightarrow \infty$, but disturbances in quadratic shear flow reach infinity in finite-time $[t_\infty = L^2/(c_0 X_0)]$. Note that in the limit of small time, i.e., $t \ll L^2/(c_0 X_0)$ then equation (7.63) behaves approximately as

$$X(t) \approx X_0 + \frac{c_0 t}{L^2} X_0^2, \tag{7.64}$$

which is obviously non-Gaussian even when X_0 is normally distributed and becomes increasingly non-Gaussian as time goes on.

Hence, the structure of the shear flow the disturbances are being advected within will determine whether the resulting phase distribution will be Gaussian or non-Gaussian at some time later. Moreover, even though (7.57) is an equation linear in the amplitude of the disturbance, its characteristics maybe be non-linear, which could lead to non-Gaussian distributions. This implies that it is not sufficient to simply note that the physical system (7.57) is linear in amplitude in order to assess whether a Kalman filter will be optimal or not. The correct condition is that both the model and its characteristics must be linear and the initial distribution one draws from must also be Gaussian. Given the severity of these conditions it would appear that it is unlikely that the phase distributions of actual flow features found in nature would always maintain a normally distributed character. Rather, it would seem more likely that the atmosphere would evolve through time and flow configurations in which a Gaussian initial phase distribution would be altered to have non-Gaussian characteristics.

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

Anderson J, Hoar T, Raeder K, Liu H, Collins N, Torn R, Avellano A (2009) The data assimilation research testbed: a community facility. Bull Amer Meteor Soc 90:1283–1296

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

Chen Y, Snyder C (2007) Assimilating vortex position with an ensemble Kalman filter. Mon Weather Rev 135:1828–1845

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 Weather Rev 139:272–294

Doyle JD, Jin Y, Hodur R, Chen S, Jin H, Moskaitis J, Reinecke A, Black P, Cummings J, Hendricks E, Holt T, Liou C, Peng M, Reynolds C, Sashegyi K, Schmidt J, Wang S (2012) Real time tropical cyclone prediction using COAMPS-TC. In: Chun-Chieh Wu, Jianping Gan (eds) Advances in geosciences, vol 28. World Scientific, Singapore, pp 15–28

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99:143–162

Fletcher SJ, Zupanski M (2006) A data assimilation method for log-normally distributed observational errors. Q J R Meteorol Soc 132:2505–2519

Hacker JP, Snyder C, Ha S-Y, Pocernich M (2011) Linear and non-linear response to parameter variations in a mesoscale model. Tellus A 63:429–444

Hamill TM, Whitaker JS, Fiorino M, Benjamin SG (2011) Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. Mon Weather Rev 139:668–688

Hodyss D (2011) Ensemble state estimation for nonlinear systems using polynomial expansions in the innovation. Mon Weather Rev 139:3571–3588

Hodyss D (2012) Accounting for skewness in ensemble data assimilation. Mon Weather Rev doi:10.1175/MWR-D-11-00198.1 140, 2346–2358

Houtekamer PL, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. Mon Weather Rev 126:796–811

Houtekamer PL, Mitchell HL (2001) A sequential ensemble Kalman filter for atmospheric data assimilation. Mon Weather Rev 129:123–137

Houtekamer PL, Mitchell HL, Pellerin G, Buehner M, Charron M, Spacek L, Hansen B (2005) Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. Mon Weather Rev 133:604–620

Jazwinski AH (1998) Stochastic processes and filtering theory. Academic, New York, 376pp

Julier SJ, Uhlmann JK (1997) New extension of the Kalman filter to nonlinear systems. Proc of SPIE 3068:182–193

Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc of the IEEE 92:401–422

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

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. Trans of the ASME Series D J of Basic Eng 83:95–107

Khare SP, Anderson JL, Hoar TJ, Nychka D (2008) An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. Tellus A 60:97–112

Kushner HJ (1967) Approximations to optimal nonlinear filters. IEEE Trans Auto Control AC-12:546–556

Lawson GW, Hansen JA (2005) Alignment error models and ensemble-based data assimilation. Mon Weather Rev 133:1687–1709

Meng Z, Zhang F (2008) Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part IV: Comparison with 3DVAR in a month-long experiment. Mon Weather Rev 136:3671–3682

Snyder C, Bengtsson T, Anderson J (2008) Obstacles to high-dimensional particle filtering. Mon Weather Rev 136:4629–4640

Szunyogh I, Kostelich EJ, Gyarmati G, Kalnay E, Hunt BR, Ott E, Satterfield E, Yorke JA (2008) A local ensemble transform Kalman filter data assimilation system for the NCEP global model. Tellus A 60:113–130

Torn RD, Hakim GJ (2008) Performance characteristics of a pseudo-operational ensemble Kalman filter. Mon Weather Rev 136:3947–3963

Tippet MK, Anderson JL, Bishop CH, Hamill TM, Whitaker JS (2003) Ensemble square root filters. Mon Weather Rev 131:1485–1490 van Leeuwen PJ (2009) Particle filtering in geophysical systems. Mon Weather Rev 137:4089–4114

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

Zupanski M (2005) Maximum likelihood ensemble filter: theoretical aspects. Mon Weather Rev 133:1710–1726

Chapter 8
Background Error Correlation Modeling with Diffusion Operators

Max Yaremchuk, Matthew Carrier, Scott Smith, and Gregg Jacobs

Abstract Many background error correlation (BEC) models in data assimilation are formulated in terms of a positive-definite smoothing operator $\mathbf{B}$ that is employed to simulate the action of correlation matrix on a vector in state space. In this chapter, a general procedure for constructing a BEC model as a rational function of the diffusion operator $\mathbf{D}$ is presented and analytic expressions for the respective correlation functions in the homogeneous case are obtained. It is shown that this class of BEC models can describe multi-scale stochastic fields whose characteristic scales can be expressed in terms of the polynomial coefficients of the model. In particular, the connection between the inverse binomial model and the well-known Gaussian model $\mathbf{B}_g = \exp \mathbf{D}$ is established and the relationships between the respective decorrelation scales are derived.

By its definition, the BEC operator has to have a unit diagonal and requires appropriate renormalization by rescaling. The exact computation of the rescaling factors (diagonal elements of $\mathbf{B}$) is a computationally expensive procedure, therefore an efficient numerical approximation is needed. Under the assumption of local homogeneity of $\mathbf{D}$, a heuristic method for computing the diagonal elements of $\mathbf{B}$ is proposed. It is shown that the method is sufficiently accurate for realistic applications, and requires 10^2 times less computational resources than other methods of diagonal estimation that do not take into account prior information on the structure of $\mathbf{B}$.

M. Yaremchuk (✉) · M. Carrier · S. Smith · G. Jacobs
Naval Research Laboratory, Stennis Space Center, MS 39529, USA
e-mail: max.yaremchuk@nrlssc.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_8,

8.1 Introduction

In recent years, heuristic background error correlation (BEC) modelling has become an area of active research in geophysical data assimilation. Of particular interest are the BEC models constructed with positive functions of the diffusion operator,

$$\mathbf{D} = \nabla \boldsymbol{\nu} \nabla \tag{8.1}$$

where $\boldsymbol{\nu}$ is the spatially varying positive-definite diffusion tensor. This type of BEC model is attractive for several reasons: (a) it guarantees positive definiteness of the resulting correlation functions (CFs), (b) it is computationally inexpensive in most practical applications, and (c) it allows straightforward control of inhomogeneity and anisotropy via the diffusion tensor. In the traditional approach of correlation modeling where spatial correlations are specified by prescribed analytical functions, care should be taken to maintain positive definiteness of the respective correlation operator, especially in anisotropic and/or inhomogeneous cases (Gaspari et al. 2006; Gregori et al. 2008).

Among the most popular operators $\mathbf{B}$ used in practical BEC modeling are those using the exponential and the inverse binomial functions of $\mathbf{D}$:

$$\mathbf{B}_g = \exp(a^2\mathbf{D}); \qquad \mathbf{B}_m = \left(\mathbf{I} - \frac{a^2\mathbf{D}}{m}\right)^{-m} \tag{8.2}$$

where $\mathbf{I}$ is the identity operator, a is a scaling parameter and m is a positive integer. Since $\mathbf{D}$ has a non-positive spectrum whose larger eigenvalues correspond to the smaller-scale eigenvectors, the operators $\mathbf{B}_g$ and $\mathbf{B}_m$ are positive-definite and suppress small-scale variability. Both types of BEC models (8.2) are extensively used in geophysical applications. Numerically, they are implemented by integration of the diffusion equation using either explicit (in the case of $\mathbf{B}_g$ Derber and Rosati 1989; Egbert et al. 1994; Weaver et al. 2003) or implicit (in the case of $\mathbf{B}_m$ Ngodock et al. 2000; Di Lorenzo et al. 2007) integration schemes.

A disadvantage of the BEC models (8.2) is that there is a limited freedom in the shape of local CFs, which have either the shape of the Gaussian bell ($\mathbf{B}_g$) or provide its mth-order strictly positive approximations ($\mathbf{B}_m$) (Xu 2005; Yaremchuk and Smith 2011). In order to allow negative correlations, one has to consider operators generated by the arbitrary polynomials in $\mathbf{D}$. The quadratic polynomial case was studied recently by Hristopulos and Elogne (2007, 2009) and Yaremchuk and Smith (2011), who obtained analytic representations of the CFs and derived relationships between the polynomial coefficients and the spectral parameters of $\mathbf{B}$ in the homogeneous case.

In a more realistic inhomogeneous setting, the diffusion tensor varies in space, making analytic methods inapplicable. Nevertheless, they can still give a reasonable guidance for quick estimation of the diagonal elements of $\mathbf{B}$ (normalization factors), whose values are crucial for constructing the BEC models. The importance of accurately computing diag$\mathbf{B}$ is evident from the fact that the operators $\mathbf{B}$ under consideration are formulated numerically as multiplication algorithms by the matrices, whose elements are not explicitly known. On the other hand, since the BEC operator $\mathbf{C}$ is represented numerically by the correlation matrix, it must have a unit diagonal and, therefore, knowledge of the diagonal elements of $\mathbf{B}$ is required for renormalization:

$$\mathbf{C} = (\text{diag}\mathbf{B})^{-1/2}\mathbf{B}(\text{diag}\mathbf{B})^{-1/2} \tag{8.3}$$

Equation (8.3) shows that the considered BEC models involve two separate algorithms: one for computing the action of $\mathbf{B}$ and another for estimating the normalization factors (diag$\mathbf{B}$) that are necessary for computing the action of (diag$\mathbf{B})^{-1/2}$.

Purser with coauthors (Purser et al. 2003; Purser 2008a,b) were among the first to employ analytic methods for estimating the normalization factors for the Gaussian operator $\mathbf{B}_g$ in geophysical applications. Somewhat earlier, an asymptotic technique was developed for estimating the diagonal of the Gaussian kernel in Riemannian spaces to study quantum effects in general relativity (e.g., Gusynin and Kushnir 1991; Avramidi 1999). These ideas can be utilized to derive a useful algorithm for estimating the normalization factors.

In this chapter, we first give an overview of the recent developments in constructing the $\mathbf{D}$-operator BEC models, and illustrate their major features with the examples in the homogeneous case $\boldsymbol{\nu} = const.$ In particular, in Sect. 8.2.2, the relationships between the scaling parameters for the Gaussian model and its mth-order approximation (8.2) are obtained and the respective CFs are given. In Sect. 8.2.3 the inverse binomial model is extended to an arbitrary polynomial of $\mathbf{D}$: Expressions for the CFs and normalization factors are derived, and relationships are established between the structure of the BEC spectrum and the polynomial coefficients. In Sect. 8.3, after a brief overview of the diagonal estimation methods, a heuristic formula for computing diag$\mathbf{B}_g$ is derived (Sect. 8.3.2) and then tested numerically against other methods in a set of realistic oceanographic applications (Sects. 8.3.3–8.3.5). Results of similar tests with the $\mathbf{B}_m$ model are also presented. Summary and discussion of the prospects for the $\mathbf{D}$-operator BEC modeling complete the chapter.

8.2 Diffusion Operator and Covariance Modeling

The convenience of the diffusion operator (8.1) for constructing the BEC models can be explained by the non-negative spectrum of $-\mathbf{D}$: An operator that is generated by a positive rational function F of $-\mathbf{D}$ whose eigenvalues tend to zero at large wavenumbers, is positive-definite and has a smoothing property, i.e. tends to suppress high-frequency components of the solution. In this section we consider two types of such functions: Those that are generated by the mth-order binomials (Sect. 8.2.2) and the others by the inverse of a positive polynomial (Sect. 8.2.3). To allow analytical treatment, anisotropic homogeneous case in the boundless domain is considered.

The benefit of analytical consideration is its ability to reveal local correlation structure and therefore provide a reasonable guidance to construction of more general operators $\mathbf{B}$. In addition, as it has been shown recently, good approximations to diag$\mathbf{B}$ can be obtained by using analytical results obtained with the homogeneous versions of $\mathbf{B}$ (e.g., Purser et al. 2003; Mirouze and Weaver 2010; Yaremchuk and Carrier 2012). Therefore, analytical formulas describing homogeneous BEC operators are of significant practical interest. The analytical results may facilitate practical design of the cost functions in variational data assimilation problems, because they give explicit relationships between the shape of the local CFs and the structure of the corresponding BEC operator.

8.2.1 Correlation Functions and Normalization

Consider an anisotropic, homogeneous diffusion operator (8.1) in $\mathbb{R}^n, n = 1, \dots, 3$, with $\boldsymbol{x} \in \mathbb{R}^n$ representing points in the physical space. By using the coordinate transformation $\boldsymbol{x}' = \boldsymbol{\nu}^{-1/2}\boldsymbol{x}$, the problem can be reduced to considering isotropic operators of the form

$$\mathbf{B} = F(-\Delta), \tag{8.4}$$

where Δ is the Laplacian (e.g., Xu 2005; Hristopulos and Elogne 2007) and F is an arbitrary positive function. In the case of an inhomogeneous diffusion ($\boldsymbol{\nu} \neq const$) the global transformation cannot be found. Transformations of this type, however, can be used locally for constructing $\mathbf{B}$ and the normalization factors (Sect. 8.3). All of the formulas that are written below are assumed to be in the transformed coordinates $\boldsymbol{x}'$ with primes omitted to simplify the notation.

The operator (8.4) is diagonalized with the Fourier transform, and the diagonal elements are $B(\boldsymbol{k}) = F(\boldsymbol{k}^2)$ where $\boldsymbol{k}$ is the Fourier coordinate (wavenumber). Because of homogeneity, the matrix elements of $\mathbf{B}$ in the $\boldsymbol{x}$-representation depend only on the distance $r = |\boldsymbol{x}|$ from the diagonal. They can be computed by applying the inverse Fourier transform to $B(\boldsymbol{k})$:

$$B^n(\boldsymbol{x}) = (2\pi)^{-n} \int_{\mathbb{R}^n} B(\boldsymbol{k}) \exp(-i\boldsymbol{k}\boldsymbol{x}) d\boldsymbol{k}. \tag{8.5}$$

By integrating over the directions in $\mathbb{R}^n$ (Appendix 1), (8.5) can be reduced to

$$B^n(r) = (2\pi)^{-n/2} \int_0^\infty B(k) k^{n-1} (kr)^s J_{-s}(kr) dk \tag{8.6}$$

where $k \equiv |\boldsymbol{k}|$, J denotes the Bessel function of the first kind, and $s = 1 - n/2$. The respective matrix elements of the correlation operator (CFs) are obtained by normalization:

$$C^n(r) = B^n(r)/B^n(0) \tag{8.7}$$

In practical applications, the diffusion operator is not homogeneous, and the analytic representations (8.6) and (8.7) cannot be obtained. However, the action of $\mathbf{B}$ on a state vector can be computed numerically at a relatively low cost. The major problem with such modelling is the efficient estimation of the diagonal elements

$$\mathbf{B}^n(\boldsymbol{x},\boldsymbol{x}) \equiv \int\limits_{\mathbb{R}^n} \mathbf{B}^n(\boldsymbol{x},\boldsymbol{y})\delta(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \tag{8.8}$$

which are necessary to rescale $\mathbf{B}$ to have its diagonal elements equal to unity. In practice, the rescaling factors $N^n(\boldsymbol{x})$ are defined as reciprocals of $\mathbf{B}^n(\boldsymbol{x},\boldsymbol{x})$.

Computing the integral (8.8) numerically is expensive, because the convolutions with the δ-functions have to be performed at all points $\boldsymbol{x}$ of the numerical grid. However, reasonable approximations (Purser et al. 2003; Yaremchuk and Carrier 2012) for $N^n(\boldsymbol{x})$ can be obtained by using asymptotic expansions of (8.8) under the assumption of weak inhomogeneity (see Sect. 8.3).

8.2.2 The Gaussian Model and Its Binomial Approximations

The Gaussian-shaped correlation model is widely used in geophysical applications. Numerically, it is implemented by approximating $\exp(a^2\mathbf{D}/2)$ with the binomial:

$$\mathbf{B}_g(\mathbf{D}) = \exp(\frac{a^2\mathbf{D}}{2}) \approx \left(\mathbf{I} + \frac{a^2\mathbf{D}}{2m}\right)^m, \tag{8.9}$$

where m is a large positive integer. This numerical approach is often referred to as "integration of the diffusion equation" and has been used in practice for several decades (Derber and Rosati 1989; Egbert et al. 1994; Weaver et al. 2003; Di Lorenzo et al. 2007). There is, however, a certain disadvantage associated with the numerical stability of the integration: The number of "integration time steps" m has to be large enough for the eigenvalues of the binomial operator in the rhs of (8.9) to be less than 1 in the absolute value. This constraint may limit m from below by a large value, which can make the computation rather expensive.

Another option is to use a different approximation in (8.9):

$$\mathbf{B}_m(\mathbf{D}) = \left(\mathbf{I} - \frac{a^2\mathbf{D}}{2m}\right)^{-m}. \tag{8.10}$$

The eigenvalues of the operator in the rhs of (8.10) do not exceed 1, and the "integration procedure" is unconditionally stable. This approach is often referred to as "implicit integration of the diffusion equation" (see Appendix 2). and has been used in many practical applications as well (Ngodock et al. 2000; Di Lorenzo et al. 2007; Carrier and Ngodock 2010).

In the Fourier representation both models (8.9) and (8.10) approximate the same Gaussian function of k:

$$\mathbf{B}_e^n(k) = \left[1 - \frac{a^2k^2}{2m}\right]^m \approx \exp(-\frac{a^2k^2}{2}) \tag{8.11}$$

$$\mathbf{B}_m^n(k) = \left[1 + \frac{a^2k^2}{2m}\right]^{-m} \approx \exp(-\frac{a^2k^2}{2}) \tag{8.12}$$

Since the value of m in (8.11) is fairly large in practice, the resulting CF is hardly distinguishable from a Gaussian-shaped curve with a half-width a.

Substituting (8.12) into (8.6), integrating over k, and normalizing the result by $\mathbf{B}_m^n(0)$ yields the CFs of the Matern family (Stein 1999) enumerated by $\mathrm{s} = m - n/2$ and scaled by $a_* = a/\sqrt{2m}$:

$$C_m^n(\rho) = \frac{\rho^{\mathrm{s}} K_{\mathrm{s}}(\rho)}{2^{\mathrm{s}-1}\Gamma(\mathrm{s})}, \tag{8.13}$$

where $\rho = r/a_*$, Γ is the gamma-function and K stands for the modified Bessel function of the second kind (Abramowitz and Stegun 1972). The respective normalization factors are

$$N_m^n = \frac{\sqrt{\pi}\,\Gamma(m)}{\Gamma(m-1/2)} \omega_n a_*^n \tag{8.14}$$

where $\omega_1 = 2$, $\omega_2 = 2\pi$, and $\omega_3 = 4\pi$. In the limiting case of $m \to \infty$, the CFs (8.13) take the Gaussian form:

$$C_\infty^n = \exp(-r^2/2a^2); \quad n = 1, .. \tag{8.15}$$

Consecutive approximations of the Gaussian CF by (8.13) are shown in Fig. 8.1. It is remarkable that when $m = 1$, the CFs (8.13) have singularities at $\rho = 0$ in both two and three dimensions (see also Table 8.1). This means that *in the continuous case* the first-order approximations become invalid when $n > 1$. Numerically, however, the CFs do exist for $n > 1$, but their decorrelation scale is limited by the grid size δ (the corresponding CF is shown by the dotted line in the left panel of Fig. 8.1). This occurs because the numerical analogue of the δ-function is never singular, but has a finite amplitude inversely proportional to the volume of a grid cell, therefore, resulting in a finite value of the convolution (8.8) even if it is infinite in the continuous case. After normalization by that finite value, the CF is 1 at $r = 0$, but its effective decorrelation scale remains proportional to the local grid size.

The left panel in Fig. 8.1 shows that low-order binomial approximations (8.13) underestimate the decorrelation scale a of the target Gaussian function. This unpleasant property can be corrected by optimizing the value of a in (8.10) to obtain the best fit with the Gaussian CF. Since the Gaussian and its approximating functions

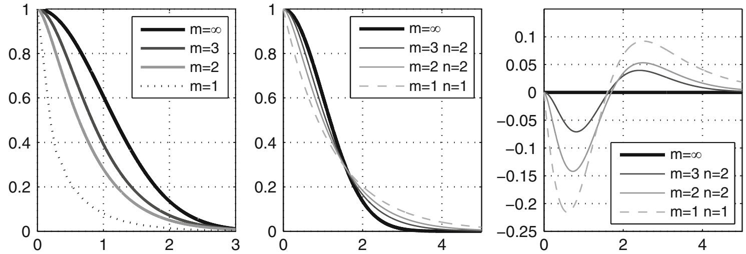


Fig. 8.1 *Left*: Binomial approximations (8.13) of the Gaussian CF in two dimensions ($n = 2$). The CF for $m = 1$ is shown by the *dotted line* for the numerical realization with the grid step $\delta = a/4$. *Middle*: Same approximations, but with optimally adjusted correlation radii for various combinations of m and n. *Right*: Differences between the Gaussian CF and its approximations shown in the *middle panel*. The *horizontal axes* are scaled by a

are both positive and have similar shapes, a reasonable optimization criterion is to set their integral decorrelation scales equal to each other:

$$\int_0^\infty C_m^n(\rho)dr \equiv \frac{a_{opt}}{\sqrt{2m}}\int_0^\infty C_m^n(y)dy = \int_0^\infty \exp(-\frac{r^2}{2a^2})dr = \frac{\sqrt{\pi}a}{\sqrt{2}}. \tag{8.16}$$

Expression (8.16) shows that $a_{opt} = \xi_m^n a$, where the rescaling coefficient ξ_m^n is defined as:

$$\xi_m^n = \sqrt{\pi m}\left[\int_0^\infty C_m^n(y)dy\right]^{-1} = \frac{\Gamma(s)}{\Gamma(s+1/2)}\sqrt{m}. \tag{8.17}$$

The values of ξ_m^n for $m, n < 4$ and their respective approximation errors

$$e_m^n = \int_0^\infty |C_m^n - C_\infty|dr/[\int_0^\infty |C_\infty|dr]$$

are assembled in Table 8.1.

The coefficients ξ_m^n along with relationship (8.12) provide an expression for estimating the scaling parameter in the binomial model (8.10) which approximates the Gaussian-shaped CF with a given radius a:

$$a_{binom} = \xi_m^n a/\sqrt{2m} \tag{8.18}$$

Table 8.1 Correlation functions associated with the power approximations (8.10) of the Gaussian CF in n dimensions. The CFs for $n = 1$ and 3 are rewritten in terms of elementary functions for convenience. The correlation radius adjustment coefficients ξ_m^n are shown below the formulas together with the corresponding relative errors e_m^n in approximation of the Gaussian CF (bold numbers)

	$n = 1$	$n = 2$	$n = 3$
$m = 1$	$\exp(-\rho)$	$K_0(\rho)$	$\exp(-\rho)/\rho$
	$\sqrt{\pi}$ **0.33**	—	—
$m = 2$	$(1+\rho)\exp(-\rho)$	$\rho K_1(\rho)$	$\exp(-\rho)$
	$\sqrt{\pi/2}$ **0.13**	$\sqrt{8/\pi}$ **0.19**	$\sqrt{2\pi}$ **0.33**
$m = 3$	$(1+\rho+\rho^2/3)\exp(-\rho)$	$\rho^2 K_2(\rho)/2$	$(1+\rho)\exp(-\rho)$
	$\sqrt{27\pi}/8$ **0.08**	$\sqrt{16/3\pi}$ **0.10**	$\sqrt{3\pi/4}$ **0.13**

8.2.3 The Inverse Polynomial Model

A certain disadvantage of the binomial models (8.9) and (8.10) is their inability to represent oscillating CFs whose spectra may have multiple maxima. This issue can be overcome by considering the BEC models of the form:

$$\mathbf{B} = \left[\mathbf{I} + \sum_{j=1}^{J} a_j \mathbf{D}^j\right]^{-1} \tag{8.19}$$

Here a_j are the real numbers, constrained by the positive definiteness requirement of $\mathbf{B}$. In the Fourier representation, the operator (8.19) acts as multiplication by the inverse of the polynomial in k^2, and the positive-definiteness property translates into the requirement that the spectral polynomial

$$B^{-1}(k^2) = 1 + \sum_{j=1}^{J} a_j (-k^2)^j \tag{8.20}$$

should be positive for all $k^2 > 0$. This constraint is equivalent to the statement that the rhs of (8.20) must not have real positive roots. Therefore, $B^{-1}(k^2)$ can also be represented in the form

$$B^{-1}(k^2) = \frac{1}{Z}\prod_{m=1}^{M}(k^2 + z_m^2)(k^2 + \bar{z}_m^2), \tag{8.21}$$

where $M = J/2$,

$$Z = \prod_m |z_m^2|^2, \tag{8.22}$$

the overline denotes the complex conjugate, and $z_m = a_m + ib_m$ are arbitrary complex numbers with $\mathrm{Im}(z_m^2) \neq 0$. In its general form, the polynomial (8.21) is additionally multiplied by the product of the arbitrary number of real negative roots ($b_m = 0$). The ensuing analysis of (8.21) will be simplified by omitting the product (summation) limits over m and assuming there are no real negative or multiple roots. The latter requirement is not restrictive in practice, because location of the roots is never known exactly, and the BEC spectrum can always be well approximated by (8.21) (Yaremchuk and Sentchev 2012).

It is instructive to note that the polynomial (8.21) can also be rewritten as

$$B^{-1}(k^2) = \frac{1}{Z}\prod_{m=1}^{M}(a_m^2 + (k - b_m)^2)(a_m^2 + (k + b_m)^2), \tag{8.23}$$

Compared to the spectral representation (8.20), representation (8.23) has the advantage that its free parameters are not constrained by the positive-definiteness requirement, and they have a sensible meaning of the scales (b^{-1}) and "energies" (a^{-1}) of the modes forming the spectrum.

Using (8.6), the matrix elements of **B** can now be written as

$$B^n(r) = \frac{Zr^{-s}}{(2\pi)^{\frac{n}{2}}}\int_0^\infty \frac{k^{s+1}J_s(kr)dk}{\prod_m (k^2 + z_m^2)(k^2 + \bar{z}_m^2)}, \tag{8.24}$$

where $s = n/2 - 1$. The integral in (8.24) can be taken by decomposing

$$B(k) = \frac{Z}{\prod_m (k^2 + z_m^2)(k^2 + \bar{z}_m^2)} \tag{8.25}$$

into elementary fractions:

$$B(k) = \sum_m \left[\frac{q_m}{k^2 + z_m^2} + \frac{\bar{q}_m}{k^2 + \bar{z}_m^2}\right], \tag{8.26}$$

where

$$q_m = \frac{Z}{(\bar{z}_m^2 - z_m^2)\prod_{j\neq m}(z_m^2 - z_j^2)(z_m^2 - \bar{z}_j^2)} \tag{8.27}$$

After substitution of (8.26) into (8.24), the integral is reduced to the sum of Hankel-Nicholson type integrals (Abramowitz and Stegun 1972) and can be taken explicitly, yielding

$$B^n(r) = \frac{2r^{2-n}}{(2\pi)^{\frac{n}{2}}}\sum_m \langle q_m \rho_m^s K_s(\rho_m)\rangle \tag{8.28}$$

where $\rho_m = z_m r$, and angular brackets denote taking the real part (cf. (8.13)).

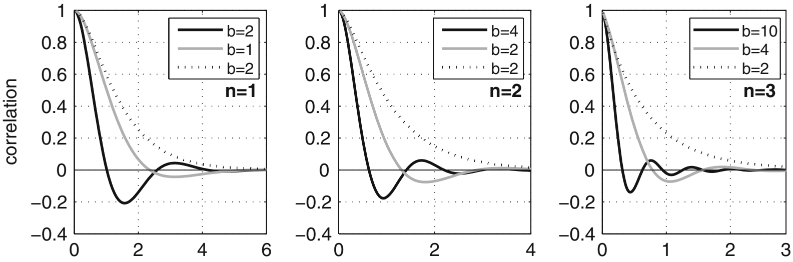


Fig. 8.2 Two-parameter CFs corresponding to the inverse BEC (8.21) with $a = 1, M = 1$. The *horizontal axis* is scaled by a. *Dotted lines* show CFs corresponding to the special case with two negative roots $k_1^2 = -a, \quad k_2^2 = -b$ not described by the spectral polynomial (8.21)

The corresponding correlation functions $C^n(r)$ are obtained through normalizing (8.28) by $B^n(0)$. The first three values at $r = 0$ are

$$B^1(0) = \sum_m \langle q_m \bar{z}_m \rangle |z_m|^{-2} \tag{8.29}$$

$$B^2(0) = -\frac{1}{\pi} \sum_m \langle q_m \log z_m \rangle \tag{8.30}$$

$$B^3(0) = -\frac{1}{2\pi} \sum_m \langle q_m z_m \rangle \tag{8.31}$$

The normalization factors can be found by integrating $C^n(r)$ over $\mathbb{R}^n$:

$$N^n = \frac{2}{B^n(0)} \sum_m \frac{\langle q_m \bar{z}_m^2 \rangle}{|z_m|^4} \tag{8.32}$$

Relationships (8.28)–(8.32) provide analytical expressions for the CFs and the normalization factors.

In the important case of the quadratic polynomial ($M = 1$) the BEC model is defined by two parameters a, b (Fig. 8.2). Expressions for the respective CFs in 1- and 3-dimensional cases can be rewritten in terms of the elementary functions (Yaremchuk and Smith 2011; Yaremchuk and Sentchev 2012)

$$C^1(a,b,r) = \frac{\sqrt{a^2+b^2}}{b} \exp(-ar) \cos(br - \arctan\frac{a}{b}) \tag{8.33}$$

$$C^3(a,b,r) = \exp(-ar) \frac{\sin(br)}{br} \tag{8.34}$$

and the normalization factors are given by

$$N^1 = \frac{4a}{a^2+b^2}; \quad N^2 = \frac{8\pi ab}{2(a^2+b^2)^2 \arctan(b/a)}; \quad N^3 = \frac{8\pi a}{(a^2+b^2)^2} \tag{8.35}$$

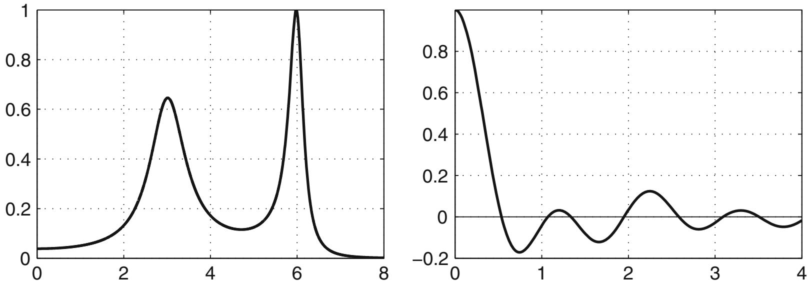


Fig. 8.3 An example of the normalized spectrum (*left*) and the respective correlation function (*right*) for the fourth-order polynomial (8.26) in two dimensions ($M = 2; z_1 = .5 + 3i; z_2 = .2 + 6i$)

In practical applications, a BEC model is often constructed by fitting the spectral (8.25) or correlation (8.28) functions to those derived from experimental data. These functions are characterized by $2m$ parameters which give enough freedom for approximating complex spectra. The approximation procedure can be formulated as a least squares problem in $2m$ dimensions, which may be rather difficult to solve due to the non-linearity of B with respect to the fitting parameters a_m and b_m. Therefore, it is useful to have guidance on how the BEC model parameters are related to the scales and amplitudes of the physical modes that contribute to the experimental spectrum (Fig. 8.3).

The contribution of the mth mode to the spectrum can be assessed by integrating the right hand side of (8.26):

$$E_m = \int_0^{\infty} \left[\frac{q_m}{k^2 + z_m^2} + \frac{\bar{q}_m}{k^2 + \bar{z}_m^2} \right] dk = \frac{\pi \langle q_m \bar{z}_m \rangle}{|z_m|^2} \tag{8.36}$$

In the limit when distances $|b_l - b_m|$ between the spectral peaks of $\mathbf{B}$ are much larger than their half-widths a_m, (i.e. $a_m/b_m \ll 0$ in particular), (8.36) can be simplified using the asymptotic approximations

$$z_m \approx i b_m; \quad q_m \approx \frac{b_m^3}{4i a_m \Pi_m}; \quad \Pi_m \equiv \prod_{j \neq m} (1 - b_m^2/b_j^2)^2$$

to yield

$$E_m \approx \frac{\pi b_m^2}{4 a_m \Pi_m}. \tag{8.37}$$

Asymptotic values of the spectral density at the peaks are respectively

$$B(b_m) \approx \frac{b_m^2}{4 a_m^2 \Pi_m} = \frac{E_m}{\pi a_m}, \tag{8.38}$$

i.e. the peak amplitudes are inversely proportional to b_m^2 and to the square of the mode scale a_m^{-1}. Expressions (8.36)–(8.38) can be useful in generating the first guess values for z_m to initialize an iterative procedure of approximating experimental data.

After the model parameters are established, the action of $\mathbf{B}^{-1}$ can be computed recursively (cf. (8.21) and (8.22)):

$$\mathbf{B}^{-1} = \prod_m \left[\mathbf{I} - |z_m^2|^{-2}\mathbf{D}(2\langle z_m^2\rangle\mathbf{I} - \mathbf{D})\right] \tag{8.39}$$

The inverse BEC model (8.39) can then be employed to compute either the action of $\mathbf{B}$ with an iterative inversion algorithm or to directly compute the gradient of a 3dVar cost function involving the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}$, where $\mathbf{x}$ is the state vector.

The above analysis gives an insight on the shape of the local CFs and provides a direct connection between the scales described by $\mathbf{B}$ and the polynomial coefficients of the considered BEC models (8.9), (8.10), (8.25) or (8.39). The second important ingredient in constructing the BEC operator $\mathbf{C}$ (8.3) is estimating the diagonal elements of $\mathbf{B}$, which is a more technical but equally important problem.

8.3 Diagonal Estimation

8.3.1 Stochastic Methods

In the last few decades a large family of stochastic algorithms were developed for estimating elements and traces of extra-large matrices emerging from numerical soluitons of the PDEs in applied physics (e.g., Girard 1987; Dong and Liu 1994; Hutchison 1989). Weaver and Courtier (2001) were among the first to use this approach in geophysical applications for estimating the diagonal of the operator (8.9).

The underlying idea is to define an ensemble of K random vectors $\mathbf{s}_k$ on the model grid and perform componentwise averaging of the products $\tilde{\mathbf{s}} = \mathbf{B}\mathbf{s}$ according to the formula:

$$\tilde{\mathbf{d}}(x) = \overline{\mathbf{s} \odot \tilde{\mathbf{s}}} \oslash \overline{\mathbf{s} \odot \mathbf{s}}, \tag{8.40}$$

where the overline denotes averaging over the ensemble and $\odot$, $\oslash$ stand for the componentwise multiplication and division of the vectors respectively. Simple considerations show that when all the components of $\mathbf{s}$ have identical δ-correlated distributions with zero mean, the contributions to $\tilde{\mathbf{d}}$ from the off-diagonal elements tend to cancel out, and $\tilde{\mathbf{d}}$ converges to $\mathbf{d} = \mathrm{diag}\mathbf{B}$ as $K \to \infty$. More accurately, the squared relative approximation error

$$\varepsilon^2(x) = (\tilde{\mathbf{d}} - \mathbf{d})^2/\mathbf{d}^2 \tag{8.41}$$

is inversely proportional to the ensemble size K. In other words, one may expect to achieve 10 % accuracy at the expense of approximately 100 multiplications by $\mathbf{B}$ if the first ensemble member gives a 100 % error. This estimate may seem acceptable since in geophysical applications the BE variances are usually known with limited precision and approximating the diagonal with 5–10 % error seems satisfactory.

The above described Monte-Carlo (MC) technique was developed further by Bekas et al. (2007), who noticed that the method may converge to $\mathbf{d}$ in the *finite* number of iterations that equals to the matrix dimension N if the ensemble vectors are mutually orthogonal. An easy way to construct such an ensemble is to draw the vectors $\mathbf{s}_k$ from the columns of the $N \times N$ Hadamard matrix (HM), which span the model's state space (see Appendix 3).

In the numerical experiments below we use MC and HM techniques as testbeds for the diagonal estimation methods which can be derived from analytical considerations and take into account prior knowledge of the structure of $\mathbf{B}$.

8.3.2 Locally Homogeneous Approximations

Consider homogeneous ($\boldsymbol{\nu} = const$) operators (8.2) with $a^2 = 1/2$ and assume that the coordinate axes are aligned along the eigenvectors of the diffusion tensor, whose (positive) eigenvalues are $\lambda_i^2, i = 1, .., n$. Then the matrix elements of $\mathbf{B}_{g,m}$ can be written down explicitly as

$$\mathbf{B}_g(\boldsymbol{x}, \boldsymbol{y}) \quad = \exp(\mathbf{D}/2) \quad = d \ \exp\left[\frac{-\rho^2}{2}\right] \tag{8.42}$$

$$\mathbf{B}_m(\boldsymbol{x}, \boldsymbol{y}) = (\mathbf{I} - \mathbf{D}/2m)^{-m} = d \ \frac{\bar{\rho}^{\mathrm{s}} K_{\mathrm{s}}(\bar{\rho})}{2^{\mathrm{s}-1}\Gamma(\mathrm{s})} \tag{8.43}$$

where

$$\rho = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^{\mathsf{T}} \boldsymbol{\nu}^{-1} (\boldsymbol{x} - \boldsymbol{y})}$$

is the distance between the correlated points (measured in terms of the smoothing scales λ_i), $d = (2\pi)^{-n/2}\Omega^{-1}$ are the (constant) diagonal elements of $\mathbf{B}_{g,m}$, $\Omega = \Pi \lambda_i = \sqrt{\det \boldsymbol{\nu}}$ is the diffusion volume element, and $\bar{\rho} = \sqrt{2m}\rho$.

When $\boldsymbol{\nu}$ varies in space, (8.42) and (8.43) are no longer valid, and the diagonal elements $\mathbf{d}$ depend on $\boldsymbol{x}$ and the type of the $\mathbf{B}$ operator. However, if we assume that $\boldsymbol{\nu}$ is locally homogeneous (LH), i.e. varies in space on a typical scale L which is much larger than λ_i, the diagonal elements $\mathbf{d}(\boldsymbol{x})$ can be expanded in the powers of the small parameter $\epsilon = \bar{\lambda}/L$, where $\bar{\lambda}$ is the mean eigenvalue of $\sqrt{\boldsymbol{\nu}}$. The zeroth-order LH approximation term (LH0) is apparently

$$\mathbf{d}^0(\boldsymbol{x}) = (2\pi)^{-n/2}\Omega(\boldsymbol{x})^{-1} \tag{8.44}$$

because for infinitely slow variations of $\boldsymbol{\nu}$ ($L \to \infty$), the normalization factors must converge to the above expression for the constant diagonal elements d. It is noteworthy that the formula (8.44) is found to be useful even in the case of strong inhomogeneity $\epsilon \geq 1$. In particular, numerical experiments of Mirouze and Weaver (2010) have shown that such an approximation provided 10 % errors in a simplified 1d case.

The accuracy of (8.44) can formally be increased by considering the next term in the expansion of the diagonal elements of $\mathbf{B}_{g,m}$. The technique of such asymptotics has been well developed for the diagonal of the Gaussian kernel (8.42) in Riemannian spaces (e.g., Gusynin and Kushnir 1991; Avramidi 1999). More recently, the approach was considered by Purser (2008a,b) in the atmospheric data assimilation context. The application of this technique to the diffusion operator (8.1) in flat space yields the following asymptotic expression for the diagonal elements of $\mathbf{B}_g$ in the local coordinate system where $\boldsymbol{\nu}(\boldsymbol{x})$ is equal to the identity matrix, and $\mathbf{D}$ takes the form of the Laplacian operator:

$$\mathbf{B}_g(\boldsymbol{x},\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}}\left[1 - \frac{1}{2}\mathrm{tr}\boldsymbol{h} - \frac{1}{12}\left(\frac{\Delta}{2}\mathrm{tr}\boldsymbol{h} + \nabla\cdot\,\mathrm{div}\boldsymbol{h}\right)\right] + O(\epsilon^5) \tag{8.45}$$

Here $\boldsymbol{h}$ is a small ($|\boldsymbol{h}| \sim \epsilon$) correction to $\boldsymbol{\nu}$ within the vicinity of $\boldsymbol{x}$. Note that the terms in the parentheses have the order $O(\epsilon^3)$, because each spatial differentiation adds an extra power of ϵ.

The asymptotic estimate (8.45) involves second derivatives which tend to amplify errors in practical applications when ϵ may not be small. Therefore, using (8.45) in its original form could be inaccurate even at a moderately small value of ϵ. To increase the computational efficiency, it is also desirable to formulate the first-order approximation as a linear operator, which acts on $\mathbf{d}^0(\boldsymbol{x})$. Keeping in mind that $|\boldsymbol{h}| \sim \epsilon$, and utilizing the relationships:

$$\mathbf{d}^0(\boldsymbol{x}) = (2\pi)^{-n/2}\Omega(\boldsymbol{x})^{-1} \approx (2\pi)^{-n/2}\left(1 - \frac{1}{2}\mathrm{tr}\boldsymbol{h}\right) \tag{8.46}$$

$$\exp(\Delta/2) \approx \mathbf{I} + \frac{1}{2}\Delta, \tag{8.47}$$

the second term in the parentheses of (8.45) can be represented as follows:

$$\nabla\cdot\,\mathrm{div}\boldsymbol{h} = \frac{1}{n}\Delta\mathrm{tr}\boldsymbol{h} + \nabla\cdot\,\mathrm{div}\boldsymbol{h}' \tag{8.48}$$

where $\boldsymbol{h}'$ is the traceless part of $\boldsymbol{h}$. On the other hand, if the divergence of $\boldsymbol{h}'$ is neglected, the (8.45) can be rewritten in the form

$$\mathbf{B}_g(\boldsymbol{x},\boldsymbol{x}) \approx \frac{1}{(2\pi)^{n/2}}\left(1 + \gamma_n\frac{\Delta}{2}\right)\left(1 - \frac{1}{2}\mathrm{tr}\boldsymbol{h}\right) \tag{8.49}$$

where

$$\gamma_n = \frac{1}{6} + \frac{1}{3n}. \tag{8.50}$$

Taking (8.46) and (8.47) into account and replacing Δ by $\mathbf{D}$, the desired ansatz for the first-order approximation (LH1) of the diagonal elements is obtained:

$$\mathbf{d}_g^1 = \exp\left(\gamma_n \frac{\mathbf{D}}{2}\right) \mathbf{d}_g^0 \tag{8.51}$$

The relationship (8.51) was derived by Purser et al. (2003) for the one-dimensional case ($\gamma_1 = 0.5$) and tested by Mirouze and Weaver (2010), who reported a significant (2–4 times) improvement of the accuracy in 1d simulations.

An estimate similar to (8.51) can also be obtained for $\mathbf{B}_m$, possibly with a different coefficient $\tilde{\gamma}_n$. It is assumed, however, that $\tilde{\gamma}_n$ may not differ too much from γ_n given similarity in the shapes (Fig. 8.1) of the correlation functions (8.42) and (8.43). Furthermore, because of the approximate nature of (8.51), the best representation of $\mathbf{d}(\boldsymbol{x})$ in realistic applications may be achieved with a value of γ_n that ts significantly different from the one given by (8.50). For this reason, a more general form of (8.51) was adopted in the numerical experiments, assuming

$$\mathbf{d}_g^1(\boldsymbol{x}) \approx \exp\left[\gamma \mathbf{D}/2\right] \mathbf{d}_g^0(\boldsymbol{x}); \qquad \mathbf{d}_2^1(\boldsymbol{x}) \approx \left[\mathbf{I} - \gamma \mathbf{D}/4\right]^{-2} \mathbf{d}_2^0(\boldsymbol{x}) \tag{8.52}$$

for the Gaussian model and its second-order ($m = 2$) spline approximation (8.10).

The following experiments investigate the dependence of the respective approximation errors $\langle \varepsilon_{g,2} \rangle$ on the free parameter γ.

8.3.3 Numerical Results

To assess the efficiency of the methods outlined in Sects. 8.3.1 and 8.3.2, two series of numerical experiments with realistically inhomogeneous BEC models are performed. In the first series the methods were tested in the 2d case with the state vector having a dimension of several thousand. In the second series, the LH0 and LH1 techniques are examined in a realistic 3d setting with a state space dimension of $N \sim 10^6$.

8.3.3.1 Experimental Setting in 2d

The state space is described by scalar functions defined on the orthogonal curvilinear grid of the Navy Coastal Ocean Model (NCOM) (Martin et al. 2008) set up in the Monterrey Bay (Fig. 8.4). The number N of grid points (dimension of the state space) was 3,438. A vector field $\boldsymbol{u}(\boldsymbol{x})$ was used to generate the diffusion tensor as follows. The smaller principal axis λ_2 of $\sqrt{\boldsymbol{\nu}}$ is set to be orthogonal

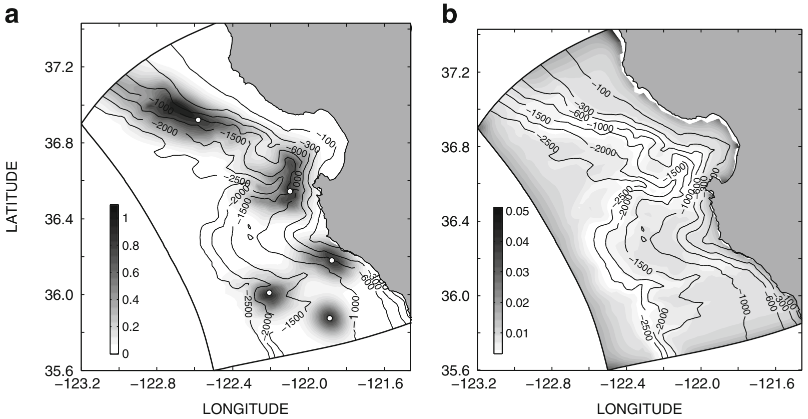


Fig. 8.4 *Left*: A composite map of five columns of the $\mathbf{B}_g$ operator. *White circles* denote locations of the diagonal elements of the corresponding correlation matrices. *Right panel* shows the map of the non-normalized diagonal elements of $\mathbf{B}_g$. Depth contours are in meters

to $\boldsymbol{u}$ with the corresponding "background" length scale $\lambda_2 = 3\delta$, where $\delta(\boldsymbol{x})$ is the spatially varying grid step. The length of the larger axis λ_1 is set to be equal to $\max(1, \sqrt{|\boldsymbol{u}|/u})\lambda_2$, where u is a prescribed threshold value of $|\boldsymbol{u}|$. If $\boldsymbol{u}$ is a velocity field, then a structure like this simulates enhanced diffusive transport of model errors in the regions of strong currents on the background of isotropic error diffusion with the decorrelation scale λ_2.

In the 2d experiments, the vector field $\boldsymbol{u}$ is generated by treating bottom topography $h(\boldsymbol{x})$ (Fig. 8.4) as a stream function. The threshold value v was taken to be one-fifth of the rms variation of $|\nabla h|$ over the domain.

All the experiments described in the next two sections are performed using the BEC models (8.42) and (8.43) with the parameters $n = m = 2$. A composite map of five columns of $\mathbf{B}_g$ is shown in Fig. 8.4a. The diffusion operator (8.1) is constrained to have zero normal derivative at the open and rigid boundaries of the domain in both 2d and 3d experiments.

Numerically, the action of $\mathbf{B}_g$ on a state vector $\mathbf{y}_0$ was evaluated by explicitly integrating the corresponding diffusion equation $\mathbf{y}_t = \mathbf{D}/2\mathbf{y}$ for the virtual "time period" defined by $\boldsymbol{v}$, starting from the "initial condition" $\mathbf{y}_0$. The minimum number of "time steps" required for the scheme's stability in such a setting was 5,256. The action of $\mathbf{B}_2$ was computed by solving the system of equations $(\mathbf{I} - \mathbf{D}/4)^2\mathbf{y} = \mathbf{y}_0$ with a conjugate gradient method. The number of iterations, required for obtaining a solution, varied within 2,000–2,500. To make the shapes of the $\mathbf{B}_g$ and $\mathbf{B}_2$ compatible (Fig. 8.1), the diffusion tensor in $\mathbf{B}_2$ was multiplied by $8/\pi$ (see Table 8.1).

The exact values $\mathbf{d}(\boldsymbol{x})$ of the diagonal elements are shown in Fig. 8.4b. Their magnitude appears to be lower in the regions of "strong currents" (large $\boldsymbol{u}$), as the corresponding δ-functions are dispersed over larger areas by diffusion. $\mathbf{d}(\boldsymbol{x})$

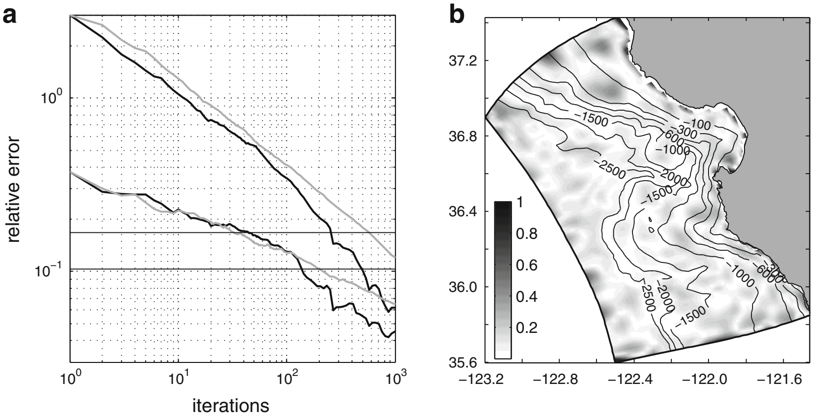


Fig. 8.5 (**a**) reduction of the domain-averaged diagonal estimation error $\langle\varepsilon\rangle$ with iterations for the HM (*black*) and MC (*gray*) methods for the $\mathbf{B}_2$ model. The *lower curves* are obtained after optimal smoothing of the estimates. The *thin horizontal lines* show the error levels that are provided by the asymptotic zeroth- ($\langle\varepsilon\rangle = 0.17$) and first-order ($\langle\varepsilon\rangle = 0.10$) methods which do not require iterative schemes. (**b**) Horizontal distribution of $\epsilon(\mathbf{B}_2)$ after 60 iterations of the HM method with smoothing

are higher near the boundaries because part of the domain available for dispersion is screened by the condition that prescribes zero flux across either open or rigid boundaries.

8.3.3.2 Statistical Methods

The MC method is implemented in two ways: In the first series of experiments, the components of $\mathbf{s}_k$ are taken to be either 1 or -1 with equal probability. In the second series they are drawn from the white noise on the interval $[-1, 1]$. The residual error ε is computed using (8.41). In both series, the rates of reduction of ε with iteration k are similar and closely follow the $\sqrt{k}$ law (upper gray line in Fig. 8.5a).

To improve the accuracy, the MC estimates are low-pass filtered with the corresponding $\mathbf{B}$-operators at every iteration (Fig. 8.5b). To optimize the filter, the diffusion operators in $\mathbf{B}_{g,2}$ are multiplied by the tunable parameter γ, which effectively reduced the mean decorrelation (smoothing) scale $\gamma^{-1/2}$ times. The lower lines in Fig. 8.5a demonstrate the result of such optimal smoothing: this procedure resulted in an almost four-fold reduction of the domain-averaged error $\langle\varepsilon\rangle$ to 0.16 after performing 60 iterations (averaging over 60 ensemble members).

Experiments with the HM method (black curves in Fig 8.5a) show that horizontal smoothing significantly improves the accuracy of the estimates, especially after the first few dozens of iterations. Comparison with the MC method (gray curves in Fig. 8.5a) demonstrates a noticeable advantage of the HM technique (black curves), which remains visible at higher iterations $k > 100$ even after smoothing (lower curves). This advantage increases with increasing iterations for two reasons: The

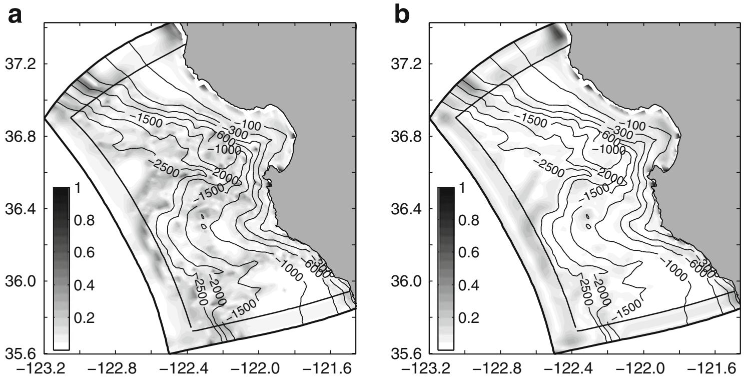


Fig. 8.6 Diagonal approximation errors under the zeroth-order (**a**), and first-order (**b**) LH methods for the $\mathbf{B}_g$ model. The *thin black line* inside the boundaries shows the domain of error averaging

HM method converges faster than $k^{-1/2}$ by its nature, whereas the efficiency of smoothing (targeted at removing the small-scale error constituents) degrades as the signal-to-noise ratio of the diagonal estimates increases with the iteration number k.

From the practical point of view, it is not reasonable to do more than several hundred iterations, as $\langle\varepsilon\rangle$ drops to the value of a few per cent (Fig. 8.5a), which is much smaller than the accuracy in the determination of the background error variances. It can therefore be concluded that it is advantageous to use the HM technique, when making more than a 100 iterations is computationally affordable.

8.3.3.3 Asymptotic Expansion Method

Since the principal axes of the diffusion tensor at every point are defined by construction, computation of the zeroth-order approximation (8.44) to the normalization factors is not expensive. Near the boundaries, however, the factors described by (8.44) have to be adjusted by taking into account the geometric constraints imposed on the diffusion. This adjustment was computed for points located closer than $3\lambda_1$ from the boundary and it was assumed that the boundary had negligible impact on the shape of the diffused δ-function (Yaremchuk and Carrier 2012).

Figure 8.6 demonstrates the horizontal distribution of the error $\varepsilon(\boldsymbol{x})$ obtained by approximating the diagonal elements of $\mathbf{B}_g$ with (8.44) (zeroth-order LH method, or LH0) and with (8.51), (the first-order LH method LH1). Despite an apparent violation of the LH assumption in many regions (e.g., λ_1 changes from 20δ to the background value of 3δ at distances $L \sim 5 - 6\delta < \lambda_1$ across the shelf break), the mean approximation error of the diagonal elements appears to be relatively small (19 %) for the LH0 method, with most of the maxima confined to the regions of strong inhomogeneity (Fig. 8.6a). The next approximation (Fig. 8.6b) reduces $\langle\varepsilon\rangle$ to 9 %. Numerical experiments with the $\mathbf{B}_2$ model have shown similar results (16 and 10 % errors).

Table 8.2 Relative CPU times required by the MC and HM methods to achieve the accuracies $\langle \varepsilon \rangle$ of the LH0 and LH1 methods (shown in brackets)

	MC/LH0	MC/LH1	HM/LH0	HM/LH1
$\mathbf{B}_g$	755 (0.19)	1205 (0.09)	680 (0.19)	520 (0.09)
$\mathbf{B}_2$	780 (0.17)	490 (0.10)	850 (0.17)	330 (0.10)

Another series of experiments are performed with the varying scaling parameter γ to find an optimal fit to $\mathbf{d}$. Computations were made for $0 \leq \gamma \leq 1$. The best result for $\mathbf{B}_g$ was obtained for $\gamma_2 = 0.30$ which is fairly consistent with the value ($\gamma_2 = 0.33$) given by (8.50). In the case of the $\mathbf{B}_2$ operator, the optimal value is $\gamma_2 = 0.24$, still in a reasonable agreement with (8.50), given the strong inhomogeneity of $\boldsymbol{\nu}$ and deviation of the $\mathbf{B}_2$ operator from the Gaussian form. A somewhat smaller value of $\gamma_2(\mathbf{B}_2)$ can be explained by the sharper shape of the respective correlation function at the origin (Fig. 8.1), which renders $\mathbf{d}^0$ to be less dependent on the inhomogeneities in the distribution of $\boldsymbol{\nu}$, and, therefore, requires less smoothing in the next approximation.

8.3.4 Numerical Efficiency

Table 8.2 provides an overview of the performance for the tested methods. For comparison purposes we show CPU requirements by the smoothed MC and HM methods after they achieve the accuracies of the LH0 and LH1 methods. It is seen that both MC and HM methods are 300–1,000 times more computationally expensive than the LH technique. In fact, for the 2d case considered, the computational cost of the stochastic diagonal estimation method is similar to the cost of the 3dvar analysis itself, which required several hundred iterations. The remarkable CPU savings are due to the fact that the LH methods explicitly take into account information on the local structure of $\mathbf{B}$ which can be derived by analytical methods. Comparison of the spatial distributions of the approximation error $\langle \varepsilon \rangle(\boldsymbol{x})$ (Figs. 8.5b and 8.6b) favor the LH methods as well: They show significantly less small-scale variations and may have a potential for further improvement. Comparing Figs. 8.5b and 8.6b also shows that, in contrast to the statistical methods, LH0 errors tend to increase in the regions of strong inhomogeneity, but decrease substantially after smoothing by the LH1 algorithm. At the same time, the LH1 errors tend to have relatively higher values near the boundaries. The effect is less visible in the HM pattern (Fig. 8.5b). This feature can be partly attributed to certain inaccuracy in estimation of the near-boundary elements. However, there is certainly room for further improvement with the issue.

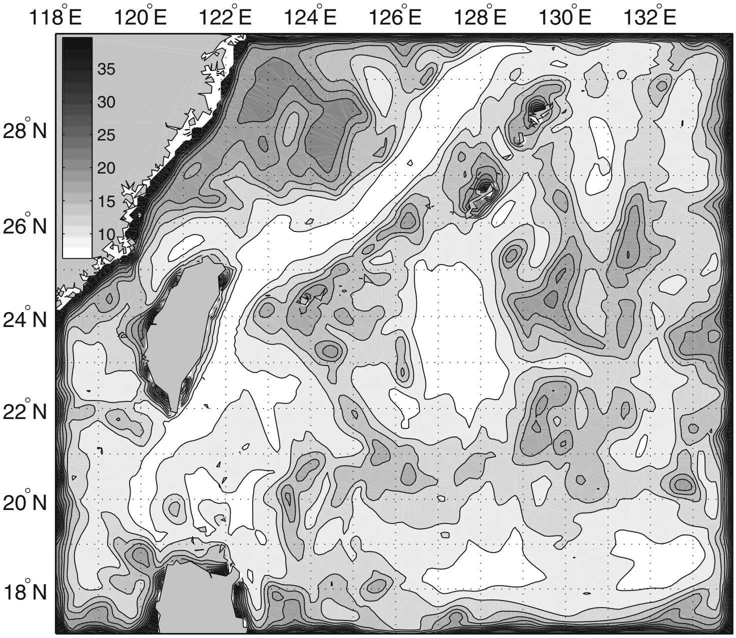


Fig. 8.7 Diagonal elements of $\mathbf{B}_g$ in the Okinawa region at $z = 20$ m. The actual values are multiplied by 10^4

8.3.5 LH Experiments in 3d Setting

To check the performance of the LH0 and LH1 methods further, a larger 3d NCOM domain is set up in the Okinawa region (Fig. 8.7) with horizontal resolution of 10 km and 45 vertical levels. The state vector dimension N (total number of the grid points) in this setting was 862,992.

Because of the large N, it is computationally unfeasible to directly compute all the diagonal elements of the BEC matrix. Therefore, accuracy checks are performed on a subset of 10,000 points, randomly distributed over the domain and the value of $\langle \varepsilon \rangle$ is estimated by averaging over these points.

The diffusion tensor is constructed in the same way that is described in Sect. 8.3.1, but the generating field $\boldsymbol{u}(\boldsymbol{x})$ is taken to be the horizontal velocity field from an NCOM run. The value of λ_3 (in the vertical direction) is independent of horizontal coordinates, but varies in the vertical as $3\delta_z$, where δ_z is the vertical grid step. Figure 8.7 illustrates spatial variability of the $\mathbf{B}_g$ diagonal elements at $z = 20$ m. The smallest values are observed in the regions of the Kuroshio and the North Equatorial Current, where the largest velocities are observed, and the $\Omega = \sqrt{\det \boldsymbol{\nu}}$ reaches its largest values (8.44). To better test the algorithm, a relatively small threshold value of $v = 0.02$ m/s is prescribed, so that diffusion is anisotropic in more than 90 % of the grid points.

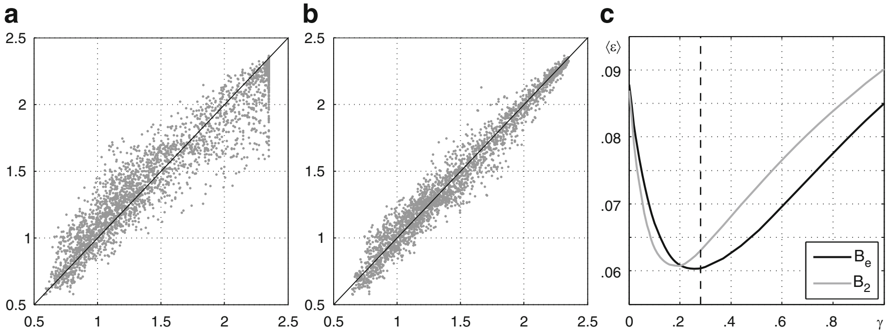


Fig. 8.8 Scatter plots of the true diagonal elements of $\mathbf{B}_g$ (*vertical axis*) versus their approximations by LH0 (**a**) and LH1 (**b**) algorithms. The actual values are multiplied by 10^3. Near-boundary points are excluded. *Right*: Diagonal approximation errors as a function of γ for the $\mathbf{B}_g$ (*black*) and $\mathbf{B}_2$ (*gray*) models. *Dashed line* shows the value of γ_3^e given by (8.50)

Figure 8.8 demonstrates the accuracy of LH0 and LH1 methods in such setting: the LH0 method provides an accuracy of 9 % which is further improved to 6 % by the LH1 scheme. The major improvement occurs in the regions where points with highly anisotropic $\boldsymbol{\nu}$ neighbor isotropic points and reduce the diagonal elements in the latter. The effect is reflected by the negative bias of the scatter plot at high values of $\mathbf{d}^0$, which reaches its maximum of 0.0237 in the points with isotropic $\boldsymbol{\nu}$ (Fig. 8.8a).

Figure 8.8c shows the dependence of approximation error ε on the value of γ_3 for both correlation models. The best approximation is obtained at $\gamma_3 = 0.26$, a value somewhat lower than suggested by the heuristic formula ($\gamma_3 = 5/18 = 0.28$, dashed line). Similarly to the 2d case, the optimal value of $\gamma_3(\mathbf{B}_2) = 0.21$ is less than $\gamma_3(\mathbf{B}_g)$, which is in agreement with the more rapid off-diagonal decay of the $\mathbf{B}_2$ matrix elements.

In general, it appears that the relationship (8.50) provides a reasonable guidance to the estimation of the smoothing parameter in the LH1 method. For the $\mathbf{B}_g$ model, the operator acting on $\mathbf{d}_g^0$ can be implemented by either reducing the number of "time steps" in integration of the diffusion equation γ^{-1} times, or by $\gamma^{-1/2}$-fold reduction of the decorrelation radius. For the $\mathbf{B}_2$ model only the second option is applicable: it also reduces the number of iterations required for computing the action of the $\mathbf{B}_2$ due to the decrease of the condition number.

8.4 Summary and Discussion

BEC modeling with the diffusion operator is an efficient and flexible tool for evaluating matrix-vector products of large dimension which emerge in minimization algorithms of variational data assimilation. In this chapter, we discussed two major issues associated with this type of models: construction of a positive-definte smoothing operator $\mathbf{B}$ as a rational function of $\mathbf{D}$ and the estimation of diag$\mathbf{B}$.

In Sect. 8.2 analytic relationships between the polynomial coefficients of $\mathbf{B}$ and the parameters controlling the shape of correlation functions were derived. Although only homogeneous operators in boundless domains were considered, these relationships provide reasonable guidance to constructing more realistic BEC operators, especially in cases when the typical scale of variability of the diffusion tensor is much larger than the local decorrelation scale ρ_c and/or most of the observations are separated from the boundaries by distances, exceeding ρ_c. In a similar way, weak inhomogeneity can be introduced by variable coefficients $z_m(\boldsymbol{x})$, and the local CF shapes can be assessed using (8.13) and (8.28)–(8.31).

Similar issues have been recently studied by many authors (e.g., Xu 2005; Hristopulos and Elogne 2007, 2009; Mirouze and Weaver 2010). In particular, analytic formulas analogous to (8.33)–(8.35), were derived in somewhat different setting by Hristopulos and Elogne (2007, 2009) who considered quadratic polynomials of similar structure. Xu (2005) analyzed Taylor expansions of exp $\mathbf{B}$ and obtained recursive relations for the polynomial coefficients associated with an arbitrary CF. Mirouze and Weaver (2010) demonstrated a possibility to generate oscillating CFs using higher-order polynomials in one dimension.

Relationships (8.28)–(8.32) generalize these results for the polynomial model of an arbitrary order M. We assume, however, that the inverse quadratic model ($M = 1$) is of major practical interest for two reasons. First, the BEC operators that are encountered in GFD applications are rarely homogeneous and observational statistics are usually insufficient to capture the details of the spatial variability of the CFs. Therefore, experimental estimates of the BECs are either limited to low-rank ensemble estimates or have to rely on the very rough assumption of homogeneity. Needless to say, that in the latter case the structure of a sample CF should be elaborated with sufficiently low detailization and be well accounted for by a two-parameter BEC model (Fig. 8.2). The second reason is that the use of higher-order polynomials considerably degrades the conditioning of the linear systems that are being solved in the assimilation process and, therefore, may require sophisticated preconditioners.

The second equally important aspect of the $\mathbf{D}$-operator BEC modeling is the computational efficiency of estimating the diagonal elements of $\mathbf{B}$. Two types of the BEC operators were considered: with the Gaussian-shaped kernel $\mathbf{B}_g$ and with the kernel generated by the second-order binomial approximation to $\mathbf{B}_g$. The tested techniques include the “stochastic” MC and HM methods, which retrieve diag$\mathbf{B}$ iteratively from its action on a sequence of model state vectors, and the “deterministic” scheme based on the analytic diagonal expansion under the assumption of local homogeneity of the diffusion tensor. The deterministic scheme was tested in two regimes: the zeroth (LH0) and the first-order (LH1) approximations.

Numerical experiments conducted with realistic diffusion tensor models show that: (a) the HM technique proves to be superior in efficiency compared to the MC technique when accuracies of less than 10 % error ($k > 100$) are required; (b) both stochastic methods require 300–1,000 times more CPU time to achieve the accuracy, compatible with the most efficient LH1 method; (c) with the Gaussian model, the LH1 method demonstrates the best performance with the value of the smoothing parameter γ compatible with the one given by the relationship (8.50) derived from the asymptotic approximation of the Gaussian kernel diagonal. In deriving the ansatz (8.51) for the LH1 model, we followed the approach of Purser et al. (2003), who proposed to smooth the zeroth-order diagonal by the square-root of the BEC operator in the one-dimensional case. Using the asymptotic technique for the heat kernel expansion, we obtained a formula for higher dimensions, and tested its validity by numerical experimentation.

It should be noted that the formal asymptotic expansion (8.45) is local by nature and tends to diverge in practical applications, where spatial variations of the diffusion tensor may occur at distances L comparable with the typical decorrelation scale $\bar{\lambda}$. To effectively immunize the expansion from the ill-effects of the abrupt changes in $\boldsymbol{\nu}$, we utilized a non-local empirical modification, still fully consistent with the original expansion in the limit $\bar{\lambda}/L \to 0$, but sufficiently robust with respect to the numerical errors related to the high-order derivatives of $\boldsymbol{\nu}$. A similar technique was developed by Purser (Purser et al. 2003; Purser 2008a), who used empirical saturation functions to stabilize higher-order approximations of the $\mathbf{B}_g$.

In general, results of our experiments show high computational efficiency of the LH1 scheme, whose total CPU requirements is just a fraction of the CPU time required by the convolution with the BEC operator – a negligible amount compared to the cost of a 3dVar analysis. Therefore, LH1 approximations to the BEC diagonal may serve as an efficient tool for renormalization of the correlation operators in variational data assimilation, as they are capable of reducing the error to 3–10 % in realistically inhomogeneous BEC models.

A separate question, that requires further investigation, is the accurate treatment of the boundary conditions. In the present study we assumed that boundaries affect only the magnitude of the corresponding columns of $\mathbf{B}$, but not their structure. This approximation is only partly consistent with the zero normal flux conditions for $\mathbf{D}$, but can be avoided if one uses “transparent” boundary conditions (e.g. Mirouze and Weaver 2010) which do not require computation of the adjustment factors. On the other hand, it might be beneficial to keep physical (no-flux) boundary conditions in the formulation of $\mathbf{D}$, as they are likely to bring more realism to the dynamics of the BE field.

Another important issue is parameterization of $\boldsymbol{\nu}(\boldsymbol{x})$ using the background fields and their statistics. In the simple diffusion tensor model used in the experiments, anisotropic BE propagation is governed by the background velocity field and superimposed on the small-scale isotropic BE diffusion, which takes place at scales that are not well resolved by the grid (less than 3δ). More sophisticated parameterizations of $\boldsymbol{\nu}(\boldsymbol{x})$ are surely possible and require further studies. In particular, recent studies have shown that since $\boldsymbol{\nu}(\boldsymbol{x})$ has only $n(n+1)/2$ independent components, it can be estimated from ensembles of moderate ($\sim 100n$) size with reasonable accuracy (Belo-Pereira and Berre 2006; Pannekoucke and Massart 2008; Pannekoucke et al. 2008; Berre and Desroziers 2010). Finally, the considered BEC models could also be effectively used for adaptive/flow-dependent covariance localization (Bishop and Hodyss 2007, 2011; Yaremchuk and Nechaev 2013), which is an issue of crucial importance in improving the forecast skill of the state-of-the-art data assimilation systems.

Acknowledgements This study was supported by the Office of Naval Research (Program element 0602435N) as part of the project "Exploring error covariances in variational data assimilation".

Appendix 1

Let θ be the angle between $\boldsymbol{x}$ and $\boldsymbol{k}$ in $\mathbb{R}^n$ and $n > 2$. Then the integral (8.5) can be rewritten in spherical coordinates as

$$B^n(r) = (2\pi)^{-n}\int\limits_0^\infty B(k)\int\limits_{\Omega_{n-1}}\exp(-ikr\cos\theta)k^{n-1}dk\ d\Omega_{n-1}, \tag{8.53}$$

where $d\Omega_{n-1}$ is the element of the surface area of the unit sphere. Since $\cos\theta$ changes symmetrically within the limits of integration, the imaginary part of the exponent vanishes. Furthermore, using the identity $d\Omega_{n-1} = d\Omega_{n-2}\cdot\sin^{n-2}\theta d\theta$, the integral (8.53) can be rewritten as

$$B^n(r) = (2\pi)^{-n}\int\limits_0^\infty B(k)k^{n-1}dk\int\limits_{\Omega_{n-2}} d\Omega_{n-2}\int\limits_0^\pi \cos(kr\cos\theta)\sin^{n-2}\theta d\theta \tag{8.54}$$

Integration over θ and substitution of the formula for the surface of $(n-2)$-dimensional unit sphere into (8.54) yields (8.6).

The general relationship (8.6) also holds for $n = 1, 2$ although these cases require a special (less complicated) treatment.

Appendix 2

In practice, the matrix elements of the operator (8.10) are never calculated explicitly due to the immense cost of such a computation. Instead, the result $\hat{\mathbf{x}}^m(\boldsymbol{x})$ of the action by $\mathbf{B}$ on a (discrete) model state vector $\hat{\mathbf{x}}^0(\boldsymbol{x})$ is calculated by solving the linear system of equations

$$\left(\mathbf{I} - \hat{\mathbf{D}}/2m\right)^m \hat{\mathbf{x}}^m = \hat{\mathbf{x}}^0, \tag{8.55}$$

where $\hat{\mathbf{D}}$ denotes the discretized diffusion operator. If $\hat{\mathbf{x}}^0(\boldsymbol{x})$ represents the "initial state" and the "time step" δt is prescribed such that the "integration time" is $m\delta t = 1$, then action of the operator (8.55) can be identified as a result of a discrete-time integration of the diffusion equation $\partial_t\mathbf{x} = \mathbf{D}/2\mathbf{x}$ with the implicit scheme

$$\hat{\mathbf{x}}^i - \hat{\mathbf{x}}^{i-1} = \frac{1}{2}\delta t\ \hat{\mathbf{D}}\ \hat{\mathbf{x}}^i, \quad i = 1, \ldots, m \tag{8.56}$$

starting from the initial state $\hat{\mathbf{x}}^0$. Here i denotes the time step number.

Similarly, the action of $\exp(\mathbf{D}/2)$ is never computed by convolving a state vector $\hat{\mathbf{x}}^0$ with the discretized kernel (8.42), but rather by the discrete-time integration of the diffusion equation with the explicit numerical scheme

$$\hat{\mathbf{x}}^i - \hat{\mathbf{x}}^{i-1} = \frac{1}{2}\delta t\ \hat{\mathbf{D}}\ \hat{\mathbf{x}}^{i-1}, \quad i = 1, \ldots, m \tag{8.57}$$

such that

$$\hat{\mathbf{x}}^m = \left(\mathbf{I} + \hat{\mathbf{D}}/2m\right)^m \mathbf{x}^0 \tag{8.58}$$

in correspondence with the asymptotic relation (8.9) for the Gaussian kernel $\mathbf{B}_g$.

Appendix 3

By definition, a Hadamard matrix (HM) is a square matrix whose entries are either 1 or -1 and whose columns are mutually orthogonal. The simplest way to construct HMs is the recursive Sylvester algorithm which is based on the obvious property: if H_N is an $N \times N$ Hadamard matrix, then

$$H_{2N} = \begin{bmatrix} H_N & H_N \\ H_N & -H_N \end{bmatrix}$$

is also an HM. Starting from $H_2 = [1 \quad 1;\ 1\ -1]$, the HMs with order $N = 2^n,\ n = 1, 2 \ldots$ can be easily constructed. HMs with $N = 12, 20$ were constructed "manually" more than a century ago. A more general HM construction algorithm, which employs the Galois fields theory, was found in 1933. In the present study we used the MatLab software that only handles the cases when $M/12$, or $M/20$ is a power of 2. Despite this restriction, the available values of M were sufficient for purposes of this chapter.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions with formulas, graphs and mathematical tables. Dover Publications, New York

Avramidi IG (1999) Covariant techniques for computation of the heat kernel. Rev Math Phys 11:947–980

Bekas CF, Kokiopoulou E, Saad Y (2007) An estimator for the diagonal of a matrix. Appl Numer Math 57:1214–1229

Belo-Pereira M, Berre L (2006) The use of ensemble approach to study the background error covariances in a global NWP model. Mon Weather Rev 134:2466

Berre L, Desroziers G (2010) Filtering of background error variances and correlations by local spatial averraging: a review. Mon Weather Rev 138:3693

Bishop C, Hodyss D (2007) Flow adaptive moderation of spurious ensemble correlations and its use in ensemble data assimilation. Q J R Meteorol Soc 133:2029

Bishop C, Hodyss D (2011) Adaptive ensemble covariance localization in ensemble 4D-VAR estimation. Mon Weather Rev 139:1241

Carrier M, Ngodock H (2010) Background error correlation model based on the implicit solution of a diffusion equation. Ocean Model 35:45–53

Derber J, Rosati A (1989) A global oceanic data assimilation system. J Phys Oceanogr 19:1333

Di Lorenzo E, Moore AM, Arango HG, Cornuelle BD, Miller AJ, Powell BS, Chua BS, Bennett AF (2007) Weak and strong constraint data assimilation in the Inverse Ocean Modelling System (ROMS): development and application for a baroclinic coastal upwelling system. Ocean Model 16:160

Dong S-J, Liu K-F (1994) Stochastic estimation with Z_2 noise. Phys Lett B 328:130–136

Egbert GD, Bennett AF, Foreman MGG (1994) Topex/Poseidon tides estimated using a global inverse model. J Geophys Res 99:24821

Gaspari G, Cohn SE, Guo J, Pawson S (2006) Construction and application of covariance functions with variable length-fields. Q J R Meteorol Soc 132:1815

Girard DF (1987) Un algorithme simple et rapide pour la validation croissee generalisee sur des problemes de grande taillee. RR 669-M, Grenoble, France: Informatique et Mathématiques Appliquées de Grenoble

Gregori P, Porcu E, Mateu J, Sasvari Z (2008) On potentially negative space time covariances obtained as sum of products of marginal ones. Ann Inst Stat Math 60:865

Gusynin VP, Kushnir VA (1991) On-diagonal heat kernel expansion in covariant derivatives in curved space. Class Quantum Gravity 8:279–285

Hristopulos DT, Elogne SN (2007) Analytic properties and covariance functions of a new class of generalized Gibbs random fields. IEEE Trans Inf Theory 53:4467–4679

Hristopulos DT, Elogne SN (2009) Computationally efficient spatial interpolators based on Spartan spatial random fields. IEEE Trans Signal Process 57:3475–3487

Hutchison MF (1989) A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. J Commun Stat Simul 18:1059–1076

Martin P, Barron C, Smedstad L, Wallcraft A, Rhodes R, Campbell T, Rowley C (2008) Software design description for the navy coastal ocean model (NCOM) Ver. 4.0. Naval Research Laboratory Report NRL/MR/7320-08-9149, p 151

Mirouze I, Weaver A (2010) Representation of the correlation functions in variational data assimilation using an implicit diffusion operator. Q J R Meteorol Soc 136:1421

Ngodock HE, Chua BS, Bennett (2000) Generalized inversion of a reduced gravity primitive equation ocean model and tropical atmosphere ocean data. Mon Weather Rev 128:1757

Pannekoucke O, Massart S (2008) Estimation of the local diffusion tensor and normalization for heterogeneous correlation modelling using a diffusion equation. Q J R Meteorol Soc 134:1425

Pannekoucke O, Berre L, Desroziers G (2008) Background-error correlation length-scale estimates and their sampling statistics. Q J R Meteorol Soc 134:497

Purser RJ (2008a) Normalization of the diffusive filters that represent the inhomogeneous covariance operators of variational assimilation, using asymptotic expansions and the techniques of non-euclidean geometry: part I: analytic solutions for symmetrical configurations and the validation of practical algorithms. NOAA/NCEP Office Note 456, p 48

Purser RJ (2008b) Normalization of the diffusive filters that represent the inhomogeneous covariance operators of variational assimilation, using asymptotic expansions and the techniques of non-euclidean geometry: part II: Riemannian geometry and the generic parametrix expansion method. NOAA/NCEP Office Note 457, p 55

Purser RJ, Wu W, Parrish DF, Roberts NM (2003) Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: spatially inhomogeneous and anisotropic general covariances. Mon Weather Rev 131:1536–1548

Stein ML (1999) Interpolation of spatial data. Some theory for kriging. Springer, New York

Weaver A, Courtier P (2001) Correlation modelling on the sphere using a generalized diffusion equation. Q J R Meteorol Soc 127:1815

Weaver AT, Vialard J, Anderson DLT (2003) Three and four-dimensional variational assimilation with a general circulation model of the Tropical Pacific Ocean. Part I: formulation, internal diagnostics and consistency checks. Mon Weather Rev 131:1360

Xu Q (2005) Representations of inverse covariances by differential operators. Adv Atmos Sci 22(2):181

Yaremchuk M, Carrier M (2012) On the renormalizaiton of the covariance operators. Mon Weather Rev 140:639–647

Yaremchuk M, Nechaev D (2013) Covariance localization with diffusion-based correlation models. Mon Weather Rev 141:848–860

Yaremchuk M, Sentchev A (2012) Multi-scale correlation functions associated with the polynomials of the diffusion operator. Q J R Meteorol Soc 138:1948–1953

Yaremchuk M, Smith S (2011) On the correlation functions associated with polynomials of the diffusion operator. Q J R Meteorol Soc 137:1927–1932

Chapter 9
The Adjoint Sensitivity Guidance to Diagnosis and Tuning of Error Covariance Parameters

Dacian N. Daescu and Rolf H. Langland

Abstract Adjoint techniques are effective tools for the analysis and optimization of the observation performance on reducing the errors in the forecasts produced by atmospheric data assimilation systems (DASs). This chapter provides a detailed exposure of the equations that allow the extension of the adjoint-DAS applications from observation sensitivity and forecast impact assessment to diagnosis and tuning of parameters in the observation and background error covariance representation. The error covariance sensitivity analysis allows the identification of those parameters of potentially large impact on the forecast error reduction and provides a first-order diagnostic to parameter specification. A proof-of-concept is presented together with comparative results of observation impact assessment and sensitivity analysis obtained with the adjoint versions of the Naval Research Laboratory Atmospheric Variational Data Assimilation System – Accelerated Representer (NAVDAS-AR) and the Navy Operational Global Atmospheric Prediction System (NOGAPS).

9.1 Introduction

Advanced measurement capabilities and algorithms for operational retrieval of atmospheric parameters from data acquired by remote sensing instruments have increased at a fast pace the amount of information provided by the global observing system (Thépaut and Andersson 2010; Lahoz 2010). As the data volume has

D.N. Daescu (✉)
Portland State University, Portland, OR, USA
e-mail: daescu@pdx.edu

R.H. Langland
Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA
e-mail: Rolf.Langland@nrlmry.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_9,

increased tremendously, there is a growing gap between the ability to collect information and the ability to optimally ingest it into numerical weather prediction (NWP) models through data assimilation techniques. The high model resolution and data density must be matched by an improved representation in the data assimilation system (DAS) of the statistical properties of the errors in the prior state estimate, model, and observations. Suboptimal information weighting poses a fundamental limitation on the DAS performance and the development of structured error covariance models for NWP applications and of computationally efficient techniques for tuning of error covariance parameters are areas of active research (Gaspari and Cohn 1999; Dee and Da Silva 1999; Lorenc 2003; Desroziers et al. 2005; Buehner et al. 2005; Chapnik et al. 2006; Bannister 2008a,b; Li et al. 2009; Frehlich 2011).

Adjoint-data assimilation system (adjoint-DAS) techniques provide effective tools for the analysis and optimization of the observation impact on reducing the forecast errors. The adjoint-DAS evaluation of the observation sensitivity has been introduced in NWP by Baker and Daley (2000) for the analysis and design of observation targeting strategies. Practical applications include monitoring the impact of data provided by the global observing system to reduce short-range forecast errors, data quality diagnostics and guidance to optimal satellite channel selection, and adaptive observation targeting (Langland and Baker 2004; Langland 2005; Cardinali 2009; Baker and Langland 2009; Gelaro and Zhu 2009; Gelaro et al. 2010; Cardinali and Prates 2011; Lupu et al. 2011).

The assessment of the forecast impact as a result of variations in the specification of observation and background error covariance parameters has been mainly performed through observing system experiments (Zhang and Anderson 2003; Joiner et al. 2007) and recently, the extension of the adjoint-DAS approach has been formulated to include the forecast sensitivity to the specification of error covariance parameters (Daescu 2008; Daescu and Todling 2010). This chapter presents a detailed exposure of the equations to evaluate the sensitivity of a forecast error aspect with respect to parameters in the observation and background error covariance representation and recent results obtained with the adjoint versions of the Naval Research Laboratory Atmospheric Variational Data Assimilation System – Accelerated Representer (NAVDAS-AR) (Xu et al. 2005; Rosmond and Xu 2006) and the Navy Operational Global Atmospheric Prediction System (NOGAPS) (Hogan and Rosmond 1991). The chapter is organized as follows below. Section 9.2 includes a brief review of the adjoint-DAS approach to observation impact assessment in variational data assimilation. A simple scalar example of statistical estimation illustrates the suboptimal observation performance in the presence of misspecified information error statistics. In Sect. 9.3 we present the theoretical basis to adjoint-DAS forecast sensitivity and first order impact estimation for observation and background error covariance parameters. A proof-of-concept to error covariance diagnosis is provided with the Lorenz 40-variable model. Section 9.4 presents results of observation impact and forecast sensitivity to error covariance weight parameters obtained with NAVDAS-AR/NOGAPS and their adjoint versions. The sensitivity analysis provides guidance on the parameter adjustments that are necessary to improve the DAS performance. A summary and further research perspectives are in Sect. 9.5. The notational convenience adopted in this work and some useful elements of matrix calculus are in the appendix.

9.2 The Analysis Equation

Variational data assimilation (Kalnay 2002) provides an analysis $\mathbf{x}^a \in \mathscr{R}^n$ to the true state $\mathbf{x}^t$ of the atmosphere by minimizing the cost functional

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}\left[\mathbf{h}(\mathbf{x}) - \mathbf{y}\right]^{\mathrm{T}}\mathbf{R}^{-1}\left[\mathbf{h}(\mathbf{x}) - \mathbf{y}\right] \tag{9.1}$$

where $\mathbf{x}^b \in \mathscr{R}^n$ is a prior (background) state estimate, $\mathbf{y} \in \mathscr{R}^p$ is the vector of observational data, and $\mathbf{h} : \mathscr{R}^n \to \mathscr{R}^p$ is the observation operator that maps the state into observations. In a four-dimensional variational (4D-Var) DAS the operator $\mathbf{h}$ incorporates the nonlinear forecast model and evolves the initial state to the observation time. Statistical information on the background error $\boldsymbol{\epsilon}^b = \mathbf{x}^b - \mathbf{x}^t$ and observational error $\boldsymbol{\epsilon}^o = \mathbf{y} - \mathbf{h}(\mathbf{x}^t)$ is used to specify the weighting matrices $\mathbf{B} \in \mathscr{R}^{n\times n}$ and $\mathbf{R} \in \mathscr{R}^{p\times p}$ that are representations in the DAS of the background and observation error covariances $\mathbf{B}_t = E(\boldsymbol{\epsilon}^b {\boldsymbol{\epsilon}^b}^{\mathrm{T}})$ and $\mathbf{R}_t = E(\boldsymbol{\epsilon}^o {\boldsymbol{\epsilon}^o}^{\mathrm{T}})$ respectively, where $E(\cdot)$ denotes the statistical expectation operator.

In practice, an approximate solution to the nonlinear minimization problem (9.1) is obtained using a linearization of the observation operator (Courtier et al. 1994),

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{h}(\mathbf{x}^b) + \mathbf{H}(\mathbf{x} - \mathbf{x}^b) \tag{9.2}$$

where

$$\mathbf{H} = \left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right]_{|\mathbf{x}=\mathbf{x}^b} \in \mathscr{R}^{p\times n} \tag{9.3}$$

is the Jacobian matrix of the observation operator $\mathbf{h}$ evaluated at $\mathbf{x}^b$. In this study we consider a single outer loop iteration such that the analysis state is expressed as

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \tag{9.4}$$

where the gain matrix $\mathbf{K}$ is defined as

$$\mathbf{K} = \left[\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1} \tag{9.5}$$

The observation-space evaluation of the analysis (9.4) is a two-stage process consisting of solving the linear system

$$\left[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]\mathbf{z} = \mathbf{y} - \mathbf{h}(\mathbf{x}^b) \tag{9.6}$$

for the vector $\mathbf{z} \in \mathscr{R}^p$ and followed by a post-multiplication operation

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{z} \tag{9.7}$$

In NAVDAS-AR the computational steps (9.6) and (9.7) are performed using a matrix-free implementation (Xu et al. 2005; Rosmond and Xu 2006).

9.2.1 Adjoint-DAS Observation Impact Estimation

Adjoint techniques are currently implemented as an effective approach (all-at-once) to estimate the impact of any data subset in the DAS on reducing the forecast errors. The forecast score is typically defined as a short-range forecast error measure

$$e(\mathbf{x}) = (\mathbf{x}_f - \mathbf{x}_f^v)^{\mathrm{T}}\mathbf{E}(\mathbf{x}_f - \mathbf{x}_f^v) \tag{9.8}$$

where $\mathbf{x}_f = \mathscr{M}_{t_0,t_f}(\mathbf{x})$ is the model forecast at verification time t_f initiated at t_0 from $\mathbf{x}$, $\mathbf{x}_f^v$ is the verifying analysis at t_f and serves as a proxy to the true state $\mathbf{x}_f^t$, and $\mathbf{E}$ is a diagonal matrix of weights that gives (9.8) units of energy per unit mass.

The adjoint approach to observation impact (OBSI) estimation relies on the *adjoint-DAS* operator $\mathbf{K}^{\mathrm{T}}$ to obtain an observation-space estimation of the change in the model forecast due to the assimilation of all data in the DAS

$$e(\mathbf{x}^a) - e(\mathbf{x}^b) \approx \left\langle \mathbf{g}, \mathbf{x}^a - \mathbf{x}^b \right\rangle_{\mathscr{R}^n} = \left\langle \mathbf{K}^{\mathrm{T}}\mathbf{g}, \mathbf{y} - \mathbf{h}(\mathbf{x}^b) \right\rangle_{\mathscr{R}^p} \tag{9.9}$$

The order of the approximation (9.9) is determined by the specification of the vector $\mathbf{g} \in \mathscr{R}^n$ (Gelaro et al. 2007; Daescu and Todling 2009).

In NAVDAS-AR the OBSI assessment is performed based on a second-order accurate approximation introduced by Langland and Baker (2004) with $\mathbf{g}$ defined as the average of two forecast gradients that are evaluated with adjoint model integrations along the analysis and background trajectories

$$\mathbf{g} = \left[\frac{1}{2}\frac{\partial e}{\partial \mathbf{x}}(\mathbf{x}^a) + \frac{1}{2}\frac{\partial e}{\partial \mathbf{x}}(\mathbf{x}^b)\right] = [\mathbf{M}^a_{t_o,t_f}]^{\mathrm{T}}\mathbf{E}(\mathbf{x}^a_f - \mathbf{x}^v_f) + [\mathbf{M}^b_{t_o,t_f}]^{\mathrm{T}}\mathbf{E}(\mathbf{x}^b_f - \mathbf{x}^v_f) \tag{9.10}$$

where $\mathbf{M}_{t_0,t_f}$ denotes the tangent linear model from t_0 to t_f. A measure of the contribution of individual data components in the assimilation scheme to the forecast error reduction, per observation type, instrument type, and data location, is obtained as

$$OBSI(\mathbf{y}_i) = \left\langle \{\mathbf{K}^{\mathrm{T}}\mathbf{g}\}_i, \{\mathbf{y} - \mathbf{h}(\mathbf{x}^b)\}_i \right\rangle_{\mathscr{R}^{p_i}} \tag{9.11}$$

where $\mathbf{y}_i \in \mathscr{R}^{p_i}$ is the data component whose impact is being evaluated. Data components for which $OBSI(\mathbf{y}_i) < 0$ contribute to the forecast error reduction (improve the forecast), whereas data components with $OBSI(\mathbf{y}_i) > 0$ increase the forecast error (degrade the forecast). The second-order approximation (9.9) and (9.10) has been found to provide satisfactory results for OBSI estimates associated to short-range forecast error measures (Baker and Langland 2009; Cardinali 2009). An intercomparison study on OBSI assessment at various NWP centers is provided by Gelaro et al. (2010).

9.2.2 Suboptimal Observation Performance: A Scalar Example

A simple scalar example of statistical estimation is used to illustrate the suboptimal observation performance for misspecified information error statistics. Consider a prior estimate $x_b = x_t + \epsilon_b$ and a measurement $y = x_t + \epsilon_o$ to the true value x_t and let assume that the errors ϵ_b and ϵ_o are unbiased $E(\epsilon_b) = 0, E(\epsilon_o) = 0$, and uncorrelated $E(\epsilon_b \epsilon_o) = 0$. The error variances $E(\epsilon_b^2) = \sigma_{b,t}^2$ and $E(\epsilon_o^2) = \sigma_{o,t}^2$ are assumed to be unknown and are specified in the analysis equation as σ_b^2 and σ_o^2, respectively. A suboptimal analysis estimate to x_t is obtained as

$$x_a = \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2} x_b + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2} y = \frac{\mu}{1+\mu} x_b + \frac{1}{1+\mu} y \tag{9.12}$$

where μ denotes the ratio $\mu = \sigma_o^2/\sigma_b^2$. The observation performance on improving the prior estimate x_b is investigated in terms of the specification μ versus the optimal ratio $\mu_t = \sigma_{o,t}^2/\sigma_{b,t}^2$. The analysis variance is

$$\sigma_a^2 = \left(\frac{\mu}{1+\mu}\right)^2 \sigma_{b,t}^2 + \left(\frac{1}{1+\mu}\right)^2 \sigma_{o,t}^2 \tag{9.13}$$

and the ratio

$$\frac{\sigma_a^2}{\sigma_{b,t}^2} = \left(\frac{\mu}{1+\mu}\right)^2 + \left(\frac{1}{1+\mu}\right)^2 \mu_t \tag{9.14}$$

provides a measure of the statistical quality of the analysis as compared with the prior estimate x_b. The minimum value of the ratio (9.14) as a function of μ is achieved at $\mu = \mu_t$ and corresponds to the *optimal analysis* $x_{a,t}$,

$$\frac{\sigma_{a,t}^2}{\sigma_{b,t}^2} = \frac{\mu_t}{1+\mu_t} < 1 \tag{9.15}$$

In the practical situation when the specification of the information error statistics is such that $\mu \neq \mu_t$, the observation performance is suboptimal and, in certain situations, the assimilation of data y may provide an estimate x_a of lower quality as compared to x_b. Contours $\sigma_a^2/\sigma_{b,t}^2 = const$ of the ratio (9.14) as a function of the two variables (μ, μ_t) are shown in Fig. 9.1. A threshold value to the μ specification is obtained when $\sigma_a^2/\sigma_{b,t}^2 = 1$,

$$\frac{\sigma_a^2}{\sigma_{b,t}^2} = 1 \Leftrightarrow \mu_t = 2\mu + 1 \tag{9.16}$$

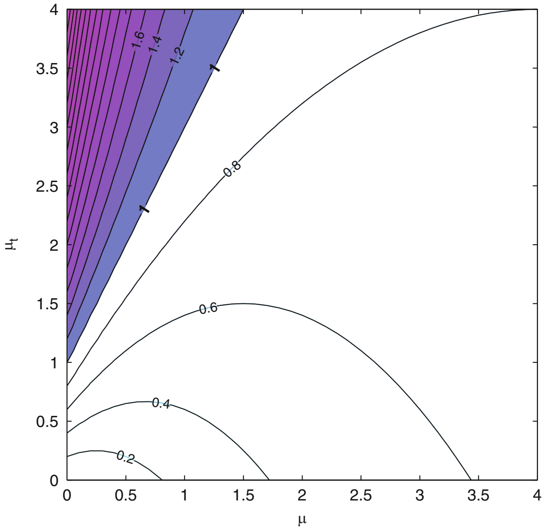


Fig. 9.1 Contours of the ratio $\sigma_a^2/\sigma_{b,t}^2$ as a function of the two variables (μ, μ_t). The contour interval is of 0.2. For values of $\mu_t > 1$, misspecification of the observation and/or background error variances may result in a detrimental observation impact (*shaded region*)

The line $\mu_t = 2\mu + 1$ divides the positive quadrant of the (μ, μ_t) plane in a region $\mu_t < 2\mu+1$ of *benefic observation impact*, $\sigma_a^2/\sigma_{b,t}^2 < 1$, and a region $\mu_t > 2\mu+1$ of *detrimental observation impact*, $\sigma_a^2/\sigma_{b,t}^2 > 1$. Since μ depends on the specification of both σ_o^2 and σ_b^2, this is a simple illustration that the observing system performance (observation "value") is closely determined by the representation in the DAS of both observation and background error statistics.

9.3 Adjoint-DAS Sensitivity Analysis

The adjoint-DAS sensitivity analysis aims to provide an assessment of the response of a functional $e(\mathbf{x}^a)$ to variations in the DAS input parameters. The first order variation $\delta e(\mathbf{x}^a)$ induced by the analysis variation $\delta \mathbf{x}^a$ is defined as

$$\delta e = \left\langle \frac{\partial e}{\partial \mathbf{x}^a}, \delta \mathbf{x}^a \right\rangle_{\mathscr{R}^n} = (\delta \mathbf{x}^a)^{\mathrm{T}} \frac{\partial e}{\partial \mathbf{x}^a} \tag{9.17}$$

For the functional (9.8), the forecast sensitivity to analysis, $\partial e/\partial \mathbf{x}^a \in \mathscr{R}^n$, is evaluated using a backward adjoint model integration from t_f to t_0 along the analysis trajectory

$$\frac{\partial e}{\partial \mathbf{x}^a} = 2[\mathbf{M}_{t_0,t_f}^a]^{\mathrm{T}} \mathbf{E}(\mathbf{x}_f^a - \mathbf{x}_f^v) \tag{9.18}$$

where $\mathbf{M}_{t_0,t_f}^a$ denotes the tangent linear model from t_0 to t_f.

Equations (9.4) and (9.5) establish the relationship $\mathbf{x}^a = \mathbf{x}^a(\mathbf{y}, \mathbf{x}^b, \mathbf{R}, \mathbf{B})$ and allow the expression of the first order analysis variation $\delta\mathbf{x}^a$ in (9.17) in terms of variations in the DAS input components $\mathbf{y}$, $\mathbf{x}^b$, $\mathbf{R}$, and $\mathbf{B}$.

9.3.1 Sensitivity to Observations and Background

Baker and Daley (2000) derived the equations of the forecast sensitivity with respect to observations and background

$$\frac{\partial e}{\partial \mathbf{y}} = \mathbf{K}^{\mathrm{T}} \frac{\partial e}{\partial \mathbf{x}^a} \tag{9.19}$$

$$\frac{\partial e}{\partial \mathbf{x}^b} = [\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}] \frac{\partial e}{\partial \mathbf{x}^a} = \frac{\partial e}{\partial \mathbf{x}^a} - \mathbf{H}^{\mathrm{T}} \frac{\partial e}{\partial \mathbf{y}} \tag{9.20}$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix. It is noticed that the $\mathbf{x}^b$-sensitivity equation (9.20) is formally valid only for a linear observation operator, $\mathbf{h}(\mathbf{x}) = \mathbf{H}\mathbf{x}$, since it neglects the dependence of the linearized observation operator (9.3) on $\mathbf{x}^b$. For the purpose of estimating the forecast sensitivity to background error covariance parameters we will simply interpret (9.20) as a vector notation. Second order derivative information or additional approximations are also necessary to evaluate the observation sensitivity in a variational DAS with multiple outer loop iterations (Daescu 2008; Trémolet 2008).

The relationship

$$\left\langle \frac{\partial e}{\partial \mathbf{x}^b}, \mathbf{x}^a - \mathbf{x}^b \right\rangle_{\mathscr{R}^n} = \left\langle \frac{\partial e}{\partial \mathbf{y}}, \mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b) \right\rangle_{\mathscr{R}^p} \tag{9.21}$$

may be established from (9.19), (9.20), and the analysis equations (9.4), (9.5) and its significance to the parametric error covariance sensitivity is explained in Sect. 9.3.3.

The evaluation of the observation sensitivity is currently integrated in the routine activities at NWP centers to monitor the observing system performance using OBSI measures such as (9.9) and (9.10). As shown below, the vectors (9.19) and (9.20) are also key ingredients to obtain information on the forecast $\mathbf{R}$- and $\mathbf{B}$-sensitivity, respectively.

9.3.2 Forecast R- and B-Sensitivity and Impact Estimation

The forecast $\mathbf{R}$- and $\mathbf{B}$-sensitivity and impact estimation identifies those error covariance parameters of potentially high forecast impact and provides guidance on the forecast benefit that may be achieved from adjusting the error covariance

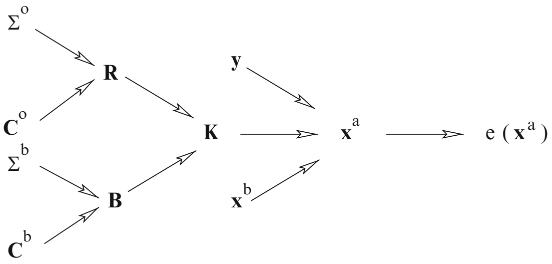


Fig. 9.2 Flow chart of the relationship between the forecast aspect and various input components of the data assimilation system

parameters. The DAS operator **K** incorporates both **R** and **B** models and from (9.4) the analysis variation induced by a perturbation in the **K** operator is expressed as

$$\delta \mathbf{x}^a = \delta \mathbf{K}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \tag{9.22}$$

By replacing (9.22) in (9.17), the first order forecast variation δe is expressed in terms of $\delta \mathbf{K}$ as

$$\delta e = \left\langle \frac{\partial e}{\partial \mathbf{x}^a}, \delta \mathbf{K}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \right\rangle_{\mathscr{R}^n} \tag{9.23}$$

From (9.23) and (9.62), the forecast sensitivity to the **K** operator is the rank-one matrix

$$\frac{\partial e}{\partial \mathbf{K}} = \frac{\partial e}{\partial \mathbf{x}^a}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)]^{\mathrm{T}} \in \mathscr{R}^{n \times p} \tag{9.24}$$

To obtain the forecast sensitivity to the error covariance specification it is necessary to analyze how the **K** operator responds to variations in **R** and **B**. Additionally, each covariance model incorporates a diagonal matrix $\boldsymbol{\Sigma}$ whose entries are the values assigned to the error standard deviation and an error correlation model **C**,

$$\mathbf{R} = \boldsymbol{\Sigma}^o \mathbf{C}^o \boldsymbol{\Sigma}^o, \qquad \mathbf{B} = \boldsymbol{\Sigma}^b \mathbf{C}^b \boldsymbol{\Sigma}^b \tag{9.25}$$

A flow chart of the functional dependence of the forecast aspect on various DAS input components is illustrated in Fig. 9.2 and the extension of the adjoint-DAS applications to parameters in the error covariance specification is discussed next.

9.3.2.1 Forecast R-Sensitivity and Impact Estimation

From (9.5) and (9.65), the first order variation in the **K** operator induced by a variation $\delta \mathbf{R}$ in the observation error covariance model **R** is expressed as

$$\delta \mathbf{K} = -\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}\delta \mathbf{R}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1} = -\mathbf{K}\delta \mathbf{R}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1} \tag{9.26}$$

and the relationship between δe and $\delta\mathbf{R}$ is obtained by replacing (9.26) in (9.23)

$$\delta e = -\left\langle \frac{\partial e}{\partial \mathbf{x}^a}, \mathbf{K}\delta\mathbf{R}[\mathbf{HBH}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \right\rangle_{\mathscr{R}^n} \tag{9.27}$$

An observation-space formulation to (9.27) is obtained by using the adjoint-DAS operator $\mathbf{K}^{\mathrm{T}}$ and Eqs. (9.6) and (9.19)

$$\delta e = -\left\langle \mathbf{K}^{\mathrm{T}}\frac{\partial e}{\partial \mathbf{x}^a}, \delta\mathbf{R}[\mathbf{HBH}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \right\rangle_{\mathscr{R}^p} = -\left\langle \frac{\partial e}{\partial \mathbf{y}}, \delta\mathbf{R}\mathbf{z} \right\rangle_{\mathscr{R}^p} \tag{9.28}$$

From (9.28) and (9.62), the forecast $\mathbf{R}$-sensitivity is the rank-one matrix

$$\frac{\partial e}{\partial \mathbf{R}} = -\frac{\partial e}{\partial \mathbf{y}}\mathbf{z}^{\mathrm{T}} \in \mathscr{R}^{p\times p} \tag{9.29}$$

In an observation-space DAS, the evaluation of the vector $\mathbf{z}$ is performed in the intermediate stage (9.6) of the analysis. An $\mathbf{R}$-sensitivity formulation that is equivalent to (9.29) and may be used in both observation-space and analysis-space data assimilation systems is obtained by expressing $\mathbf{z}$ from (9.6) and (9.7) as

$$\mathbf{z} = \mathbf{R}^{-1}[\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)] \tag{9.30}$$

If the observation error correlations are not modeled in the DAS then $\mathbf{R}$ is a diagonal matrix, $\mathbf{R} = diag(\boldsymbol{\sigma}_o^2)$, and the forecast sensitivity to the specification of the observation error variance is expressed from (9.29) as

$$\frac{\partial e}{\partial \boldsymbol{\sigma}_{o,i}^2} = -\frac{\partial e}{\partial y_i}z_i,\ i = 1 : p \tag{9.31}$$

A first order assessment of the forecast performance of a new covariance model $\hat{\mathbf{R}}$, as compared with the model $\mathbf{R}$ in the DAS, may be obtained by setting $\delta\mathbf{R} = \hat{\mathbf{R}} - \mathbf{R}$ in (9.28) and requires only the additional ability to provide the matrix/vector product $[\delta\mathbf{R}]\mathbf{z}$,

$$e[\mathbf{x}^a(\hat{\mathbf{R}})] - e[\mathbf{x}^a(\mathbf{R})] \approx -\{[\delta\mathbf{R}]\mathbf{z}\}^{\mathrm{T}}\frac{\partial e}{\partial \mathbf{y}} \tag{9.32}$$

In particular, if the observation error covariance models $\mathbf{R}$ and $\hat{\mathbf{R}}$ are specified as diagonal matrices, $\mathbf{R} = diag(\sigma_{o,i}^2)$ and $\hat{\mathbf{R}} = diag(\hat{\sigma}_{o,i}^2)$, then (9.32) is expressed as

$$e[\mathbf{x}^a(\hat{\boldsymbol{\sigma}}_o^2)] - e[\mathbf{x}^a(\boldsymbol{\sigma}_o^2)] \approx -\sum_{i=1}^{p}\left(\delta\sigma_{o,i}^2\right)\left(\frac{\partial e}{\partial y_i}z_i\right) \tag{9.33}$$

The right side of (9.33) provides an *all-at-once* first order assessment to the forecast impact of each individual variation $\delta\sigma_{o,i}^2 = \hat{\sigma}_{o,i}^2 - \sigma_{o,i}^2$. The impact estimates (9.32) and (9.33) may be evaluated prior to the actual implementation of the model $\hat{\mathbf{R}}$ in the DAS and provide insight on the potential forecast gain at a reduced computational effort.

9.3.2.2 Forecast B-Sensitivity and Impact Estimation

From (9.5) and (9.65), the first order variation in the **K** operator associated with a variation $\delta\mathbf{B}$ in the background error covariance model **B** is expressed as

$$\begin{aligned}\delta\mathbf{K} &= \delta\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1} - \mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1}\mathbf{H}\delta\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1}\\ &= [\mathbf{I}-\mathbf{K}\mathbf{H}]\delta\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1}\end{aligned} \tag{9.34}$$

An explicit relationship between δe and $\delta\mathbf{B}$ is obtained by replacing (9.34) in (9.23),

$$\begin{aligned}\delta e &= \left\langle \frac{\partial e}{\partial \mathbf{x}^a}, [\mathbf{I}-\mathbf{K}\mathbf{H}]\delta\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1}[\mathbf{y}-\mathbf{h}(\mathbf{x}^b)]\right\rangle_{\mathscr{R}^n}\\ &= \left\langle [\mathbf{I}-\mathbf{K}\mathbf{H}]^{\mathrm{T}}\frac{\partial e}{\partial \mathbf{x}^a}, \delta\mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R}]^{-1}[\mathbf{y}-\mathbf{h}(\mathbf{x}^b)]\right\rangle_{\mathscr{R}^n}\end{aligned} \tag{9.35}$$

With the aid of (9.6) and (9.20), (9.35) may be expressed as

$$\delta e = \left\langle \frac{\partial e}{\partial \mathbf{x}^b}, \delta\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{z}\right\rangle_{\mathscr{R}^n} \tag{9.36}$$

From (9.36) and (9.62), the forecast **B**-sensitivity is the rank-one matrix

$$\frac{\partial e}{\partial \mathbf{B}} = \frac{\partial e}{\partial \mathbf{x}^b}\left(\mathbf{H}^{\mathrm{T}}\mathbf{z}\right)^{\mathrm{T}} \in \mathscr{R}^{n\times n} \tag{9.37}$$

A first order assessment of the forecast performance of a new background error covariance model $\hat{\mathbf{B}}$, as compared with the model **B** in the DAS, may be obtained by setting $\delta\mathbf{B} = \hat{\mathbf{B}} - \mathbf{B}$ in (9.36),

$$e[\mathbf{x}^a(\hat{\mathbf{B}})] - e[\mathbf{x}^a(\mathbf{B})] \approx \left([\delta\mathbf{B}]\mathbf{H}^{\mathrm{T}}\mathbf{z}\right)^{\mathrm{T}}\frac{\partial e}{\partial \mathbf{x}^b} \tag{9.38}$$

Having available the $\mathbf{x}^b$-sensitivity vector defined as in (9.20), the evaluation of the first order impact estimate (9.38) may be performed prior to the actual implementation of the model $\hat{\mathbf{B}}$ in the DAS at a computational cost roughly equivalent to the cost of a post-multiplication operation (9.7).

9.3.3 Forecast Sensitivity to Error Covariance Parameters

While the explicit evaluation and storage of the **R**- and **B**-sensitivity matrices is not feasible in an operational system, from (9.29) and (9.37) it is noticed that evaluation and storage of only a few vectors are necessary to capture the information content of the sensitivity matrices. Of practical significance is the ability to evaluate directional derivatives associated with perturbations $(\delta\mathbf{R}, \delta\mathbf{B})$ and to obtain sensitivities to key parameters used to model the error covariances. The observation sensitivity vector (9.19) is a key ingredient to both **R**- and **B**-sensitivity estimation and techniques to observation sensitivity and impact estimation in an ensemble-based DAS have been also formulated (Liu and Kalnay 2008; Liu et al. 2009). The **R**- and **B**-sensitivity equations provided in this work are thus of relevance to both variational and ensemble-based data assimilation systems and their use to perform parameter sensitivity analysis is presented below.

9.3.3.1 Sensitivity to Multiplicative Error Covariance Parameters

A practical approach to perform error covariance tuning relies on the parametric representation

$$\mathbf{B}(s^b) = s^b\mathbf{B}, \qquad \mathbf{R}_i(s_i^o) = s_i^o\mathbf{R}_i,\ i \in I \tag{9.39}$$

where $s^b > 0$ and $s_i^o > 0$ are scalar coefficients used to adjust the weight given in the DAS to the background information and to the information provided by the observing system component $\mathbf{y}_i, i \in I$, respectively (Chapnik et al. 2006; Desroziers et al. 2009). In the formulation (9.39) it is assumed that $\{\mathbf{y}_i, i \in I\}$ is a partition of the observations consisting of data subsets $\mathbf{y}_i \in \mathscr{R}^{p_i}, i \in I$, with uncorrelated observation errors such that the model **R** is structured as a block diagonal matrix

$$\mathbf{R} = diag(\mathbf{R}_i),\ \ \mathbf{R}_i \in \mathscr{R}^{p_i \times p_i} \tag{9.40}$$

The covariance specification $(\mathbf{R}, \mathbf{B})$ in the reference DAS corresponds to all weight parameters in set to 1 i.e., $s_i^o = 1,\ i \in I$ and $s^b = 1$. From (9.39), the covariance variations induced by perturbations δs_i^o and δs^b in the weight coefficients are expressed respectively, as

$$\delta\mathbf{R}_i = \delta s_i^o\mathbf{R}_i,\ \ i \in I \tag{9.41}$$

$$\delta\mathbf{B} = \delta s^b\mathbf{B} \tag{9.42}$$

By replacing (9.41) in (9.28) and with the aid of (9.30), the first order variation in the forecast aspect $e(\mathbf{x}^a)$ is expressed in terms of δs_i^o as

$$\delta e = -\sum_{i \in I} \delta s_i^o \left\langle \frac{\partial e}{\partial \mathbf{y}_i}, [\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)]_i \right\rangle_{\mathscr{R}^{p_i}} \tag{9.43}$$

Equation (9.43) provides the forecast sensitivity to each observation error covariance weight coefficient,

$$\begin{aligned}\frac{\partial e}{\partial s_i^o} &= -\left\langle \frac{\partial e}{\partial \mathbf{y}_i}, [\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)]_i \right\rangle_{\mathscr{R}^{p_i}} \\ &= -[\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)]_i^{\mathrm{T}} \frac{\partial e}{\partial \mathbf{y}_i}, \ i \in I \end{aligned} \tag{9.44}$$

By replacing (9.42) in (9.36) and with the aid of the analysis equation (9.7), the first order variation in the forecast aspect $e(\mathbf{x}^a)$ is expressed in terms of δs^b as

$$\delta e = \delta s^b \left\langle \frac{\partial e}{\partial \mathbf{x}^b}, \mathbf{x}^a - \mathbf{x}^b \right\rangle_{\mathscr{R}^n} \tag{9.45}$$

Equation (9.45) provides the forecast sensitivity to the background error covariance weight coefficient,

$$\frac{\partial e}{\partial s^b} = \left\langle \frac{\partial e}{\partial \mathbf{x}^b}, \mathbf{x}^a - \mathbf{x}^b \right\rangle_{\mathscr{R}^n} \tag{9.46}$$

and the identity (9.21) allows the observation-space evaluation of the s^b-sensitivity as

$$\frac{\partial e}{\partial s^b} = [\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)]^{\mathrm{T}} \frac{\partial e}{\partial \mathbf{y}} \tag{9.47}$$

From (9.44) and (9.47) it is noticed that the identity (9.21) is formally equivalent to

$$\frac{\partial e}{\partial s^b} + \sum_{i \in I} \frac{\partial e}{\partial s_i^o} = 0 \tag{9.48}$$

and reflects an intrinsic property of the optimization problem (9.1) in variational data assimilation: multiplication of both **R** and **B** matrices by the same positive constant has no impact on the analysis.

9.3.3.2 Sensitivity to the Observation Error Correlation Specification

The standard practice in operational data assimilation and forecast systems is to neglect the statistical correlation of the observation errors and tuning of the assigned observation error variance parameters is used to compensate for unrepresented error correlations. Recent diagnostic studies have shown evidence of both spatial and inter-channel error correlations in the radiance data provided by the atmospheric sounders (Garand et al. 2007; Bormann et al. 2010, 2011) and research to assess the potential gain that may be achieved from modeling the observation error correlations is becoming increasingly important as the next generation of hyperspectral instruments will further increase the data density.

By replacing

$$\delta \mathbf{R} = \boldsymbol{\Sigma}^o \delta \mathbf{C}^o \boldsymbol{\Sigma}^o \tag{9.49}$$

in (9.28), the first order forecast variation δe induced by a $\delta \mathbf{C}^o$-perturbation in the observation error correlation model is expressed as

$$\delta e = -\left\langle \frac{\partial e}{\partial \mathbf{y}}, \boldsymbol{\Sigma}^o \delta \mathbf{C}^o \boldsymbol{\Sigma}^o \mathbf{z} \right\rangle_{\mathscr{R}^p} = -\left\langle \boldsymbol{\Sigma}^o \frac{\partial e}{\partial \mathbf{y}}, \delta \mathbf{C}^o \boldsymbol{\Sigma}^o \mathbf{z} \right\rangle_{\mathscr{R}^p} \tag{9.50}$$

From (9.50), the forecast $\mathbf{C}^o$-sensitivity is the rank-one matrix

$$\frac{\partial e}{\partial \mathbf{C}^o} = -\left(\boldsymbol{\Sigma}^o \frac{\partial e}{\partial \mathbf{y}} \right) (\boldsymbol{\Sigma}^o \mathbf{z})^{\mathrm{T}} \in \mathscr{R}^{p \times p} \tag{9.51}$$

and may be expressed using the elementwise vector product (9.60) as

$$\frac{\partial e}{\partial \mathbf{C}^o} = -\left(\boldsymbol{\sigma}_o \circ \frac{\partial e}{\partial \mathbf{y}} \right) (\boldsymbol{\sigma}_o \circ \mathbf{z})^{\mathrm{T}} \in \mathscr{R}^{p \times p} \tag{9.52}$$

where $\boldsymbol{\sigma}_o \in \mathscr{R}^p$ denotes the vector of values assigned in the DAS to the observation error standard deviation, $\boldsymbol{\Sigma}^o = diag(\boldsymbol{\sigma}_o)$.

9.3.3.3 Sensitivity to the Background Error Correlation Specification

The specification of the background error correlations is a key ingredient of the data assimilation system and ongoing research at NWP centers is focused on the development of flow-dependent background error covariance models (Buehner 2005; Bannister 2008a,b; Brousseau et al. 2011). By replacing

$$\delta \mathbf{B} = \boldsymbol{\Sigma}^b \delta \mathbf{C}^b \boldsymbol{\Sigma}^b \tag{9.53}$$

in (9.36), the first order forecast variation δe induced by a perturbation $\delta \mathbf{C}^b$ in the background error correlation model is expressed as

$$\delta e = \left\langle \frac{\partial e}{\partial \mathbf{x}^b}, \boldsymbol{\Sigma}^b \delta \mathbf{C}^b \boldsymbol{\Sigma}^b \mathbf{H}^{\mathrm{T}} \mathbf{z} \right\rangle_{\mathscr{R}^n} = \left\langle \boldsymbol{\Sigma}^b \frac{\partial e}{\partial \mathbf{x}^b}, \delta \mathbf{C}^b \boldsymbol{\Sigma}^b \mathbf{H}^{\mathrm{T}} \mathbf{z} \right\rangle_{\mathscr{R}^n} \tag{9.54}$$

From (9.54), the forecast $\mathbf{C}^b$-sensitivity is the rank-one matrix

$$\frac{\partial e}{\partial \mathbf{C}^b} = \left(\boldsymbol{\Sigma}^b \frac{\partial e}{\partial \mathbf{x}^b} \right) \left(\boldsymbol{\Sigma}^b \mathbf{H}^{\mathrm{T}} \mathbf{z} \right)^{\mathrm{T}} \in \mathscr{R}^{n \times n} \tag{9.55}$$

Table 9.1 Forecast sensitivity to various input parameters of a data assimilation system with a single outer loop iteration

Parameter	Significance	Dimension	Sensitivity equation
$\mathbf{y}$	Observation vector	$\mathscr{R}^p$	$\mathbf{K}^{\mathrm{T}}\dfrac{\partial e}{\partial \mathbf{x}^a}$
$\mathbf{R}$	Observation error covariance model	$\mathscr{R}^{p\times p}$	$-\dfrac{\partial e}{\partial \mathbf{y}}\mathbf{z}^{\mathrm{T}}$
$\mathbf{C}^o$	Observation error correlation model	$\mathscr{R}^{p\times p}$	$-\left(\boldsymbol{\sigma}_o\circ\dfrac{\partial e}{\partial \mathbf{y}}\right)(\boldsymbol{\sigma}_o\circ\mathbf{z})^{\mathrm{T}}$
σ_o	Observation error standard deviation	$\mathscr{R}^p$	$-\sigma_o^{-1}\circ\left[\dfrac{\partial e}{\partial \mathbf{y}}\circ(\mathbf{Rz})+\left(\mathbf{R}\dfrac{\partial e}{\partial \mathbf{y}}\right)\circ\mathbf{z}\right]$
s_i^o	Observation error covariance weight	$\mathscr{R}^1$	$-[\mathbf{y}-\mathbf{h}(\mathbf{x}^b)-\mathbf{H}(\mathbf{x}^a-\mathbf{x}^b)]_i^{\mathrm{T}}\dfrac{\partial e}{\partial \mathbf{y}_i}$
$\mathbf{x}^b$	Background state vector[a]	$\mathscr{R}^n$	$\dfrac{\partial e}{\partial \mathbf{x}^a}-\mathbf{H}^{\mathrm{T}}\dfrac{\partial e}{\partial \mathbf{y}}$
$\mathbf{B}$	Background error covariance model	$\mathscr{R}^{n\times n}$	$\dfrac{\partial e}{\partial \mathbf{x}^b}(\mathbf{H}^{\mathrm{T}}\mathbf{z})^{\mathrm{T}}$
$\mathbf{C}^b$	Background error correlation model	$\mathscr{R}^{n\times n}$	$\left(\boldsymbol{\sigma}_b\circ\dfrac{\partial e}{\partial \mathbf{x}^b}\right)\left[\boldsymbol{\sigma}_b\circ(\mathbf{H}^{\mathrm{T}}\mathbf{z})\right]^{\mathrm{T}}$
$\boldsymbol{\sigma}_b$	Background error standard deviation	$\mathscr{R}^n$	$\boldsymbol{\sigma}_b^{-1}\circ\left[\dfrac{\partial e}{\partial \mathbf{x}^b}\circ(\mathbf{x}^a-\mathbf{x}^b)+\left(\mathbf{B}\dfrac{\partial e}{\partial \mathbf{x}^b}\right)\circ(\mathbf{H}^{\mathrm{T}}\mathbf{z})\right]$
s^b	Background error covariance weight	$\mathscr{R}^1$	$[\mathbf{y}-\mathbf{h}(\mathbf{x}^b)-\mathbf{H}(\mathbf{x}^a-\mathbf{x}^b)]^{\mathrm{T}}\dfrac{\partial e}{\partial \mathbf{y}}$

[a] See Sect. 9.3.1 on the interpretation of the $\mathbf{x}^b$-sensitivity equation

and may be expressed using the elementwise vector product (9.60) as

$$\frac{\partial e}{\partial \mathbf{C}^b}=\left(\boldsymbol{\sigma}_b\circ\frac{\partial e}{\partial \mathbf{x}^b}\right)\left[\boldsymbol{\sigma}_b\circ\left(\mathbf{H}^{\mathrm{T}}\mathbf{z}\right)\right]^{\mathrm{T}}\in\mathscr{R}^{n\times n} \tag{9.56}$$

where $\boldsymbol{\sigma}_b\in\mathscr{R}^n$ denotes the vector of values assigned in the DAS to the background error standard deviation, $\boldsymbol{\Sigma}^b=diag(\boldsymbol{\sigma}_b)$.

A similar reasoning strategy may be used to derive the equations of the forecast sensitivity with respect to the specification of the observation and background error standard deviation vectors $\boldsymbol{\sigma}_o$ and $\boldsymbol{\sigma}_b$ respectively, in the covariance representation (9.25). A summary of equations to evaluate the forecast sensitivity with respect to various input parameters of a data assimilation system with a single outer loop iteration is provided in Table 9.1.

9.3.4 The Adjoint Sensitivity Guidance: A Proof-of-Concept

The derivative information obtained through adjoint-DAS techniques provides guidance on the local behavior of the forecast aspect as a function of various parameters in the DAS. For a generic parameter $\mathbf{u}$, the steepest descent direction

$$\mathbf{d} = -\frac{\partial e}{\partial \mathbf{u}} \tag{9.57}$$

identifies the direction of small parameter variations $\delta\mathbf{u}$, from the current DAS configuration, that will be of largest forecast benefit and provides a first-order optimality diagnostic. For applications to parameter tuning an additional search must be performed to determine an optimal step length along the descent direction. Daescu and Todling (2010) provided an illustration of iterative gradient-based tuning of observation error variances. Whereas the optimal parameter values may not be inferred from the derivative information alone, valuable insight may be gained by monitoring the forecast error sensitivity to the specification of the error covariance parameters.

A proof-of-concept is given with the Lorenz 40-variable model (Lorenz and Emanuel 1998)

$$\frac{\mathrm{d}\,x_j}{\mathrm{d}\,t} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F,\ \ j = 1:n \tag{9.58}$$

where $n = 40, x_{n+j} = x_j$, and the forcing constant is specified as $F = 8$. The system (9.58) is integrated with the standard fourth-order explicit Runge-Kutta method and a constant time step $\varDelta t = 0.05$ that is identified with a 6-h time period to produce a reference trajectory ("the truth") $\mathbf{x}^t$.

The adjoint-DAS sensitivity guidance to diagnosis and tuning of error covariance parameters is illustrated using an idealized data assimilation system (DAS-I) and a suboptimal data assimilation system (DAS-II). Observations are assumed to be available at each grid point, with unbiased and uncorrelated observation errors taken from a normal distribution with the standard deviation of $\sigma_{o,t} = 0.5$ at locations 1–20 and of $\sigma_{o,t} = 1$ at locations 21–40. In both DAS-I and DAS-II the assigned σ_o values are consistent with the true observation error statistics, $\sigma_o = \sigma_{o,t}$, and distinction between the experiments is made through the **B**-matrix specification. DAS-I provides an optimal analysis by implementing a full Extended Kalman Filter (EKF) to update in time the background error covariance. Figure 9.3a shows the average over $N = 7{,}200$ analysis cycles ($\sim$ a 5-year period) of the background error covariance in DAS-I,

$$\tilde{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{B}(t_i) \tag{9.59}$$

Deficiencies are introduced in DAS-II by ignoring the background error correlations and the flow dependence of the **B** matrix. In DAS-II, the background error covariance is specified as a diagonal matrix ($\mathbf{C}^b = \mathbf{I}$), frozen in time, with the diagonal entries σ_b^2 taken from (9.59), as shown in Fig. 9.3b.

The **B**-sensitivity associated to the Euclidean norm of the 24-h forecast errors is monitored in each of DAS-I and DAS-II and time averaged results are displayed in Fig. 9.3c and in Fig. 9.3d, respectively. In the optimal system DAS-I, the **B**-sensitivity matrix displays no particular structure and this is an indication that no systematic deficiency in the **B** matrix specification has been identified. In the

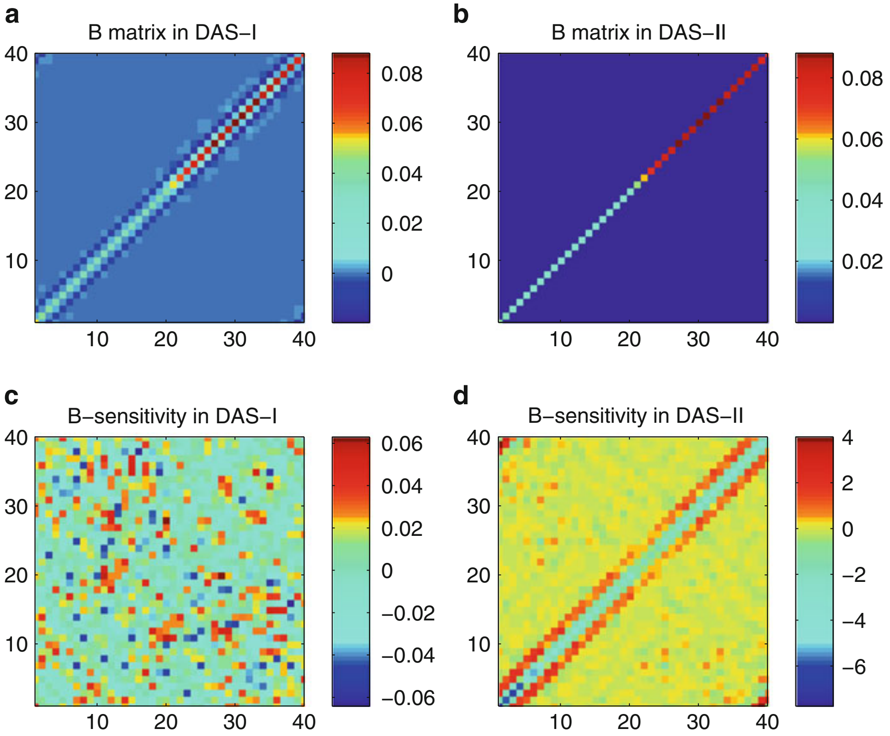


Fig. 9.3 (**a**) Time-averaged background error covariance in DAS-I. (**b**) The specification of the background error covariance in DAS-II. (**c**) Time-averaged forecast error **B**-sensitivity in DAS-I. (**d**) Time-averaged forecast error **B**-sensitivity in DAS-II. Displayed is the symmetric part of the **B**-sensitivity matrices

suboptimal system DAS-II, the **B**-sensitivity matrix distinguishes a pronounced off-diagonal structure from the noisy entries and provides an indication that the background error correlations are misspecified in the model **B**.

The time-averaged structure of the background error correlation matrix $\mathbf{C}^b$ in DAS-I is shown in Fig. 9.4a. The time-averaged structure of the forecast error sensitivity to the background error correlation model in DAS-II ($\mathbf{C}^b$-sensitivity) is shown in Fig. 9.4b. The diagonal entries of the correlation model $\mathbf{C}^b$ are constrained to $\mathbf{C}^b_{ii} = 1, i = 1 : n$. For visual display purposes only and to emphasize the cross-correlation structure, the diagonal entries of the $\mathbf{C}^b$ and $\mathbf{C}^b$-sensitivity matrices in Fig. 9.4a, b respectively, were set to zero. Negative values of the entries in the $\mathbf{C}^b$-sensitivity matrix in DAS-II are associated with positive values of the background error correlations in DAS-I (and vice versa) which is in agreement with the guidance provided by the steepest descent direction (9.57). The $\mathbf{C}^b$-sensitivity in Fig. 9.4b provides a first order guidance on the $\delta\mathbf{C}^b$ update that is necessary to correct the correlation model in DAS-II from its current specification of $\mathbf{C}^b = \mathbf{I}$ toward the correlation structure of the background errors in DAS-I (Fig. 9.4a).

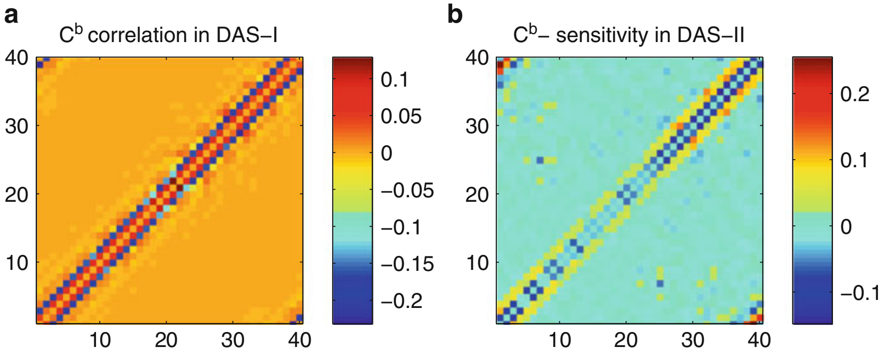


Fig. 9.4 (**a**) The time-averaged structure of the background error correlations in DAS-I. (**b**) Time-averaged forecast error sensitivity to the background error correlation model in DAS-II. Displayed is the symmetric part of the $\mathbf{C}^b$-sensitivity matrix

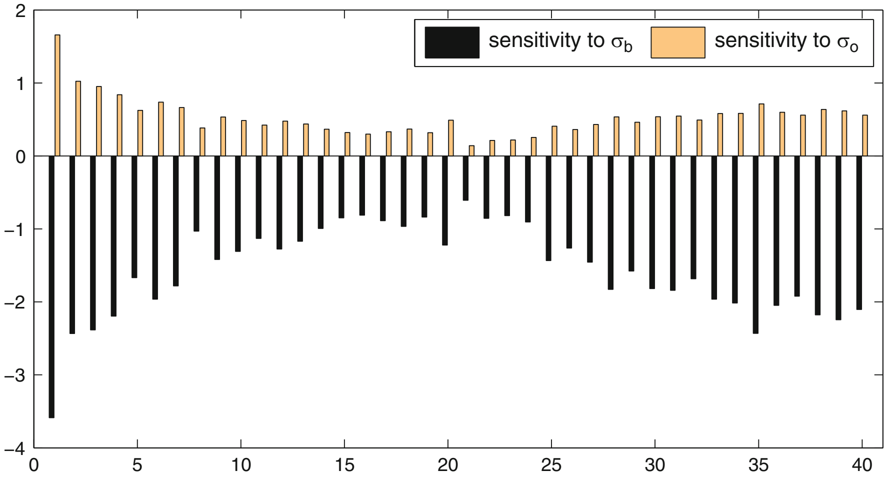


Fig. 9.5 Time-averaged forecast error sensitivity to the σ_b and σ_o specification in DAS-II

A suboptimal specification of the error covariance parameters in a DAS input component will be reflected in the sensitivities with respect to other input components. To illustrate this aspect, Fig. 9.5 displays time-averaged values of the forecast sensitivity to the specification of the background error standard deviation σ_b and to the specification of the observation error standard deviation σ_o in DAS-II. The negative values associated with the σ_b-sensitivity indicate that in average, the background information is overweighted in the current configuration of DAS-II and that an improved performance may be obtained by increasing the assigned σ_b values (variance inflation). At the same time, positive values associated with the σ_o-sensitivity indicate that the information provided by the observing system is underweighted and that an improved performance of DAS-II may be also obtained

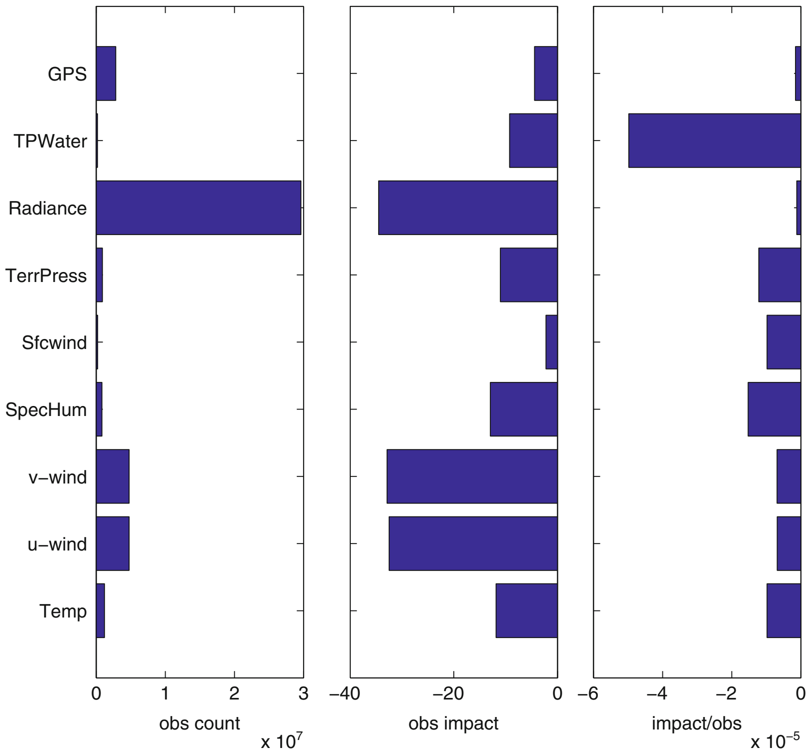


Fig. 9.6 The total number of observations assimilated in NAVDAS-AR to produce the 00 UTC analyses over the time period of study, the total observation impact on the 24-h forecast error reduction (J Kg^{-1}), and the average observation impact/observation (J Kg^{-1})

by *artificially reducing* the assigned σ_o values from the current specification of $\sigma_o = \sigma_{o,t}$ (optimal estimate). In this context, the evaluation of the forecast error sensitivity with respect to a selected parameter in the DAS merely provides guidance on the parameter variations that are necessary to compensate for (unknown) deficiencies in the error covariance specification associated with other components of the DAS.

9.4 Results with the Adjoint NAVDAS-AR/NOGAPS

The guidance derived from the error covariance sensitivity analysis and its relevance to the observation impact estimates are presented with the NRL NAVDAS-AR/NOGAPS and their adjoint versions. The forecast error measure (9.8) is defined as the moist total energy norm over the global domain and the verification state $\mathbf{x}^v$ is the analysis valid at the forecast time. The results are valid for the 24-h NOGAPS forecasts associated with the 00 UTC NAVDAS-AR analyses produced

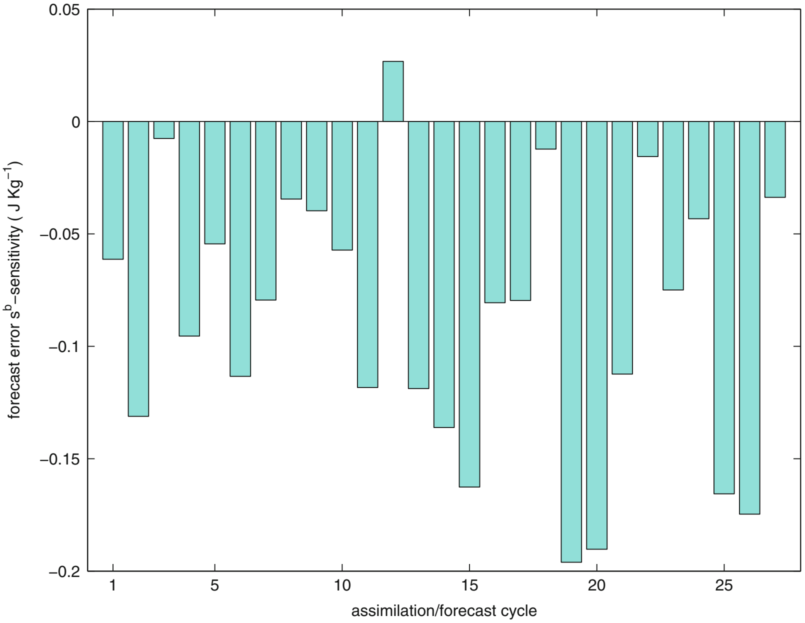


Fig. 9.7 The sensitivity of the 24-h forecast error to the background error covariance weight coefficient s^b during each assimilation/forecast cycle

in a 6-h 4D-Var assimilation interval during the time period of 2010 September 29–October 26 (data for 2010 October 15 was not incorporated in this study). For each observed parameter, the total number of observations assimilated in NAVDAS-AR to produce the 00 UTC analyses over the period of study (27 data sets) is shown in Fig. 9.6 together with the OBSI (9.10) and (9.11). It is noticed that each observed parameter had a benefic OBSI on the forecast error reduction and that radiance and wind speed observations had the largest overall impact. The total precipitable water results are for profiles through an entire atmospheric column, which may explain their relatively high value of impact per observation.

First guidance derived from the error covariance sensitivity analysis is on the proper weighting in the DAS between the information provided by the background estimate and the observing system as a whole (covariance parameterization through a single parameter). This is obtained by systematically monitoring the forecast error sensitivity to the background error covariance weight coefficient s^b (9.47). The forecast s^b-sensitivity values for each assimilation/forecast cycle are shown in Fig. 9.7. The sensitivity magnitude is closely determined by the forecast episode and negative derivative values indicate that, in general, inflation of the background error covariance is of potential benefit to the forecasts (at weight $s^b = 1$ the forecast aspect is a locally decreasing function of the s^b parameter). An alternative interpretation of the s^b-sensitivity guidance is that, in general, the information provided by the observing system as a whole is underweighted in the DAS.

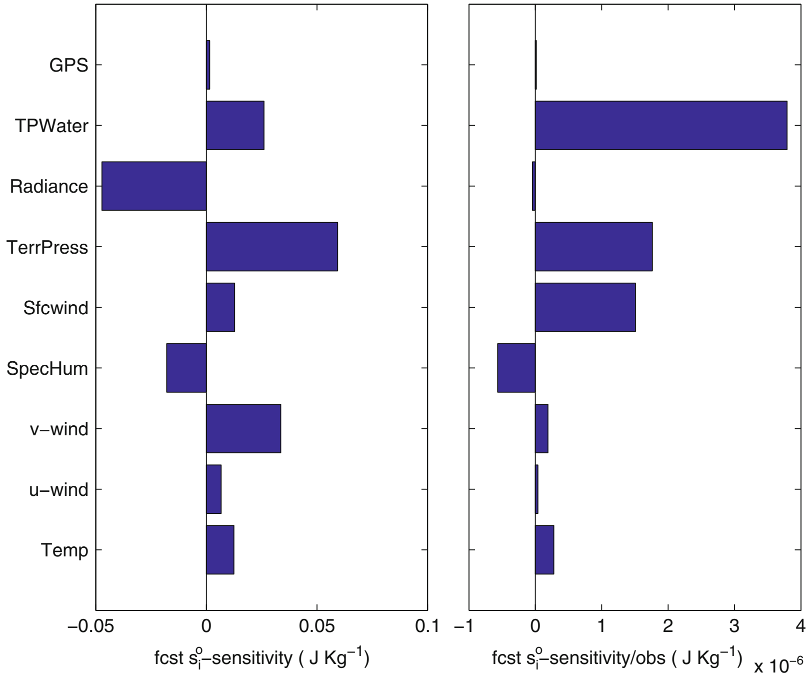


Fig. 9.8 The sensitivity of the 24-h forecast error to the observation error covariance weight coefficient s_i^o. Displayed are average values per assimilation/forecast cycle (*left side* graphic) and average values per observation (*right side* graphic)

By analogy with the OBSI estimates, the adjoint approach provides all-at-once **R**-sensitivity information for each data type, instrument, and observation location in the time-space domain. For each observed parameter, the forecast error sensitivity to the observation error covariance weight coefficient s_i^o (9.44) is shown in Fig. 9.8.

The comparison is facilitated by the fact that all s_i^o-sensitivities have units of J/Kg since all the weight coefficients are non-dimensional scalar parameters. Positive values identify those data types whose reduced σ_o^2 values will be of potential benefit to the forecasts and it is noticed that total precipitable water observations exhibit the largest sensitivity/observation. Radiances and specific humidity are identified in Fig. 9.8 as data types whose negative sensitivity values point toward σ_o^2-inflation. The presence of both positive and negative s_i^o-sensitivity values indicate that an optimal weighting between the information provided by the background and observations may not be achieved by adjusting a single scalar covariance coefficient (e.g., s^b-inflation) and that a systematic analysis of each data type and instrument is necessary to optimize the DAS performance. To further illustrate this aspect, results of forecast sensitivity to the σ_o^2-weight coefficient for the Special Sensor Microwave Imager (SSMI) total precipitable water and for the radiosonde specific humidities are contrasted in Fig. 9.9 for each assimilation/forecast episode.

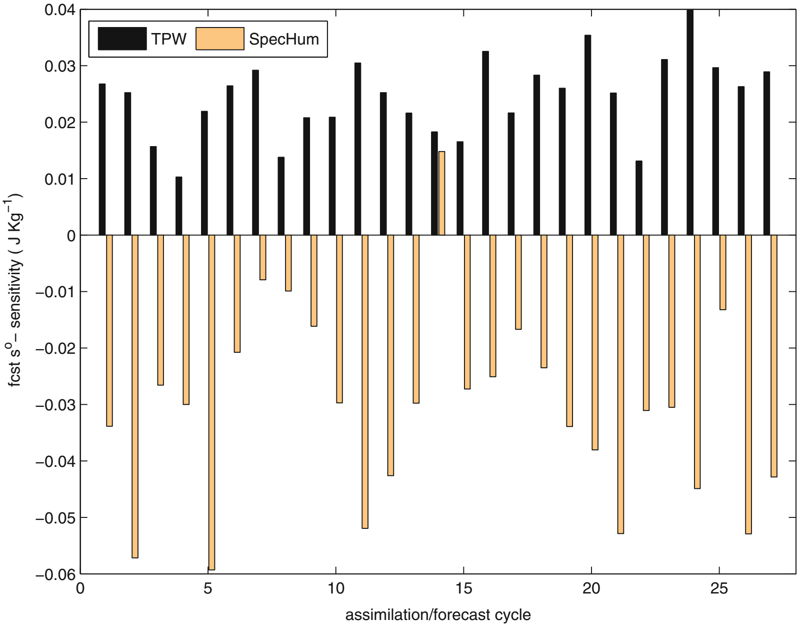


Fig. 9.9 Forecast error sensitivity (J Kg^{-1}) to the observation error variance weight coefficient s^o for SSMI total precipitable water observations (TPW) and for the radiosonde specific humidity observations (SpecHum) during each assimilation/forecast cycle

The sensitivity results obtained for the radiosondes indicate that, in general, the information provided by temperature measurements is underweighted whereas the information provided by specific humidity measurements is overweighted in the DAS. The geographical distribution of the forecast error sensitivity to the assigned σ_o^2-weight coefficient for radiosonde temperature and specific humidity measurements is shown in Figs. 9.10 and 9.11, respectively. These maps display the cumulative values of the sensitivities over the time period of study, vertically integrated, and at a horizontal bin resolution of 2.5°.

The sensitivity analysis of the radiance data distinguishes the Advanced Microwave Sounding Unit (AMSU)-A from the Infrared Atmospheric Sounding Interferometer (IASI), the Atmospheric Infrared Sounder (AIRS), and the Special Sensor Microwave Imager Sounder (SSMIS) instruments. For each instrument, the forecast error sensitivity to the assigned observation error variance weight coefficient was monitored during each assimilation/forecast episode and the results are displayed in Fig. 9.12. Positive s^o-sensitivities associated with AMSU-A indicate a conservative use of information and that reducing the σ_o^2 values assigned to this instrument is of potential benefit to the forecasts. Systematic negative s^o-sensitivities are noticed for IASI and provide an indication that further inflation of the σ_o^2 values assigned to this instrument is of potential benefit to the forecasts. Negative sensitivity values are also noticed in the majority of the analysis/forecast episodes for AIRS and SSMIS instruments.

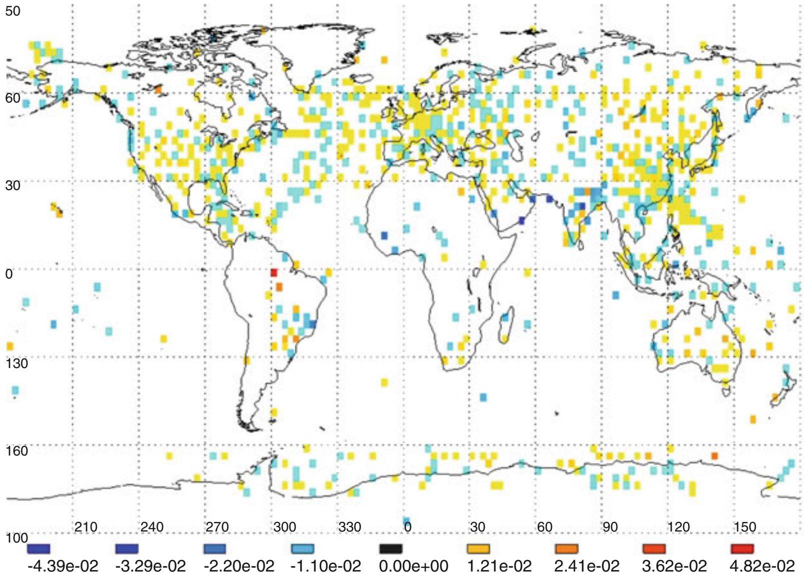


Fig. 9.10 Forecast error sensitivity (J Kg^{-1}) to the observation error variance weight coefficient s^o for radiosonde temperatures

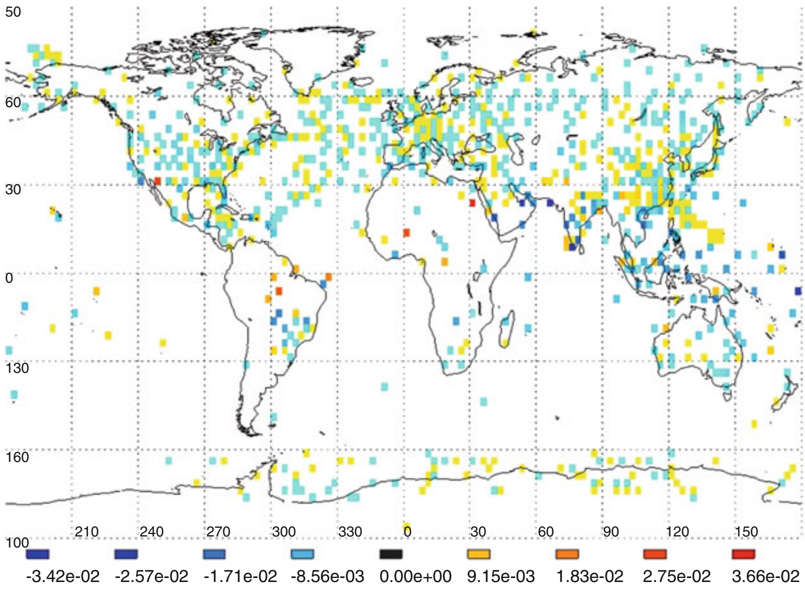


Fig. 9.11 Forecast error sensitivity (J Kg^{-1}) to the observation error variance weight coefficient s^o for radiosonde specific humidities

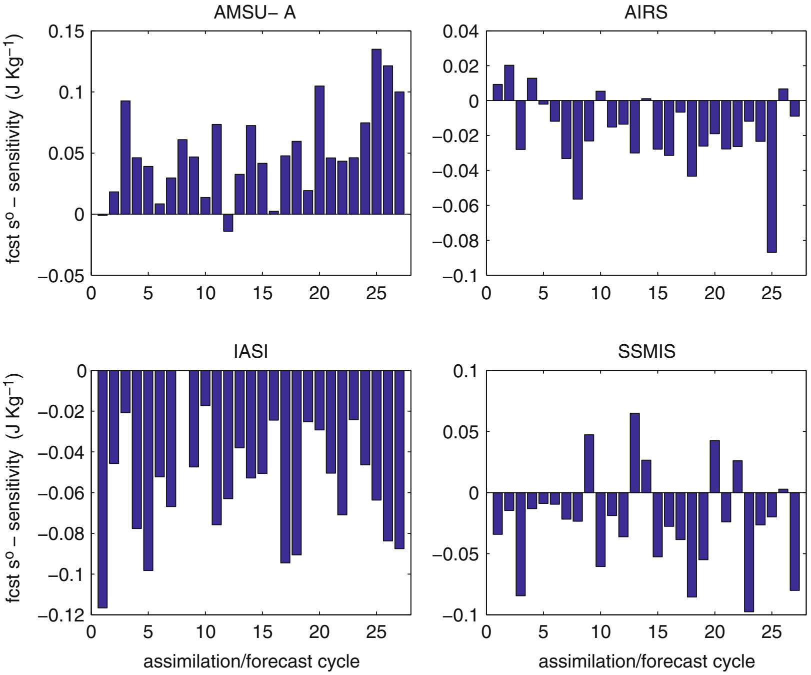


Fig. 9.12 Forecast error sensitivity (J Kg^{-1}) to the observation error variance weight coefficient s^o for the atmospheric sounders AMSU-A, AIRS, IASI, and SSMIS during each assimilation/forecast cycle

Bormann and Bauer (2010) and Bormann et al. (2010, 2011) implemented various diagnostics to estimate observation error statistics in the European Centre for Medium-Range Weather Forecasts (ECMWF) assimilation system. Their findings suggest that observation-error estimates for sounder radiances are significantly lower than the values assigned in the operational systems and indicate a too-conservative use of the AMSU-A instrument. In the above-mentioned studies it was also found that AMSU-A shows little spatial and interchannel error correlations and that error correlations of larger magnitude are present for the IASI, AIRS, and SSMIS instruments. In this context, the sensitivity guidance may be interpreted as an attempt to compensate for unrepresented observation error correlations in the DAS through artificial inflation of the assigned error variances. Caution must be exercised in the interpretation of the sensitivity analysis for tuning DAS error covariance parameters and, as explained in Sect. 9.3.4, the derivative information only allows the identification of a descent direction in the parameter space and without providing the optimal parameter values. This information may be considered in conjunction with other diagnostic tools such as the methods of Desroziers and Ivanov (2001) and Desroziers et al. (2005).

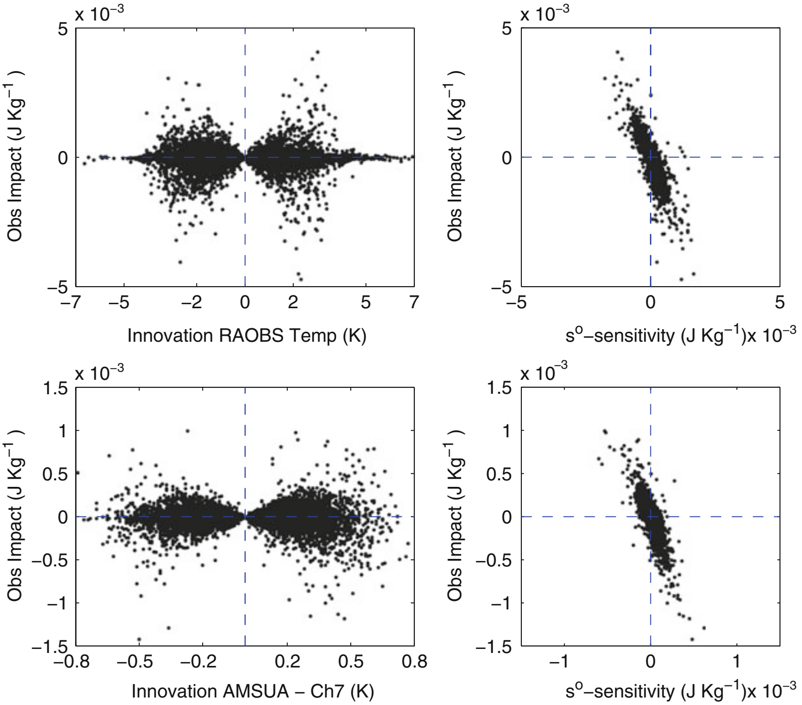


Fig. 9.13 Scatter diagram of innovation vs. observation impact (*left side* graphics) and of forecast sensitivity to the error variance weight coefficient vs. observation impact (*right side* graphics). Results are for the radiosonde temperatures (*top*) and AMSU-A channel 7 radiances (*bottom*) assimilated in NAVDAS-AR to produce the analysis valid at 00 UTC 2010 September 29

We conclude this section with an illustration of the statistical correlation between the observation impact and the forecast sensitivity to the observation error variance weight coefficient. Gelaro et al. (2010) noticed that in general, in a given assimilation/forecast episode the percentage of observations in the DAS that have a benefic forecast impact is in the range of 50–54 %. Their study also investigates the relation of observation impact to innovation value, $\mathbf{y} - \mathbf{h}(\mathbf{x}^b)$. In Fig. 9.13, scatter diagrams are used to additionally illustrate the correlation between the observation impact and the forecast sensitivity to the error variance weight coefficient associated with the observation. For practical reasons, results are shown for a single 24-h forecast initiated at 00 UTC 2010 September 29 for the radiosonde temperatures and for the radiances from the AMSU-A channel 7. It is noticed that for the majority of observations a negative (benefic) observation impact is associated with a positive s^o-sensitivity (guidance is to reduce the assigned σ_o^2) and a positive (detrimental) observation impact is associated with a negative s^o-sensitivity (guidance is to increase the assigned σ_o^2). In each quadrant of the (s^o-sensitivity, OBSI)-plane (listed in counterclockwise order, with Q-I being the positive quadrant) the percentage distribution of the number of observations was found to be as follows.

For radiosonde temperatures, of 27,063 observations analyzed: 10 % in Q-I, 38.5 % in Q2, 12 % in Q-III, and 39.5 % in Q-IV. For AMSU-A channel 7, of 22,215 observations analyzed: 13 % in Q-I, 36 % in Q-II, 14 % in Q-III, and 37 % in Q-IV. A similar distribution was noticed in the analysis of data for other days of the study period and a throughout investigation of the global observing system remains to be performed. These results also illustrate that in a given assimilation/forecast episode tuning an instrument through a single covariance weight coefficient is suboptimal. For practical applications, a potential use of the combined information derived from sensitivity and impact estimates is to identify data components and to provide guidance on the adjustment in the corresponding covariance weight parameters that are necessary to reduce the errors in a specified forecast aspect.

9.5 Summary and Research Perspectives

The value added by observations to a data assimilation and forecast system is closely determined by the weight assigned in the DAS to the information provided by the prior state estimate and measurements. The adjoint-DAS methodology offers a computationally feasible approach to assess the significance of each DAS input component to a selected forecast aspect. The evaluation of the observation sensitivity, observation impact, and forecast **R**- and **B**-sensitivity share the same adjoint-DAS tools and may be performed simultaneously to obtain complementary information on the DAS performance. The necessary software for these calculations is currently in place or it is being developed at various NWP centers and new practical applications remain to be investigated. Valuable insight to the design of observing system experiments and implementation of parameter tuning procedures that are effective in reducing the forecast errors may be gained by systematically monitoring the forecast sensitivity to parameters in the observation and background error covariance representation. Observation sensitivity calculations provide guidance for observation-space targeting and the practical ability to obtain **R**- and **B**-sensitivity information establishes a basis for extending the traditional targeting approach from the observation space to the error covariance space.

Acknowledgements The work of D.N. Daescu was supported by the Naval Research Laboratory Atmospheric Effects, Analysis, and Prediction BAA #75-09-01 under award N00173-10-1-G032 and by the National Science Foundation under award DMS-0914937. Support for the second author from the sponsor ONR PE-0602435N is gratefully acknowledged.

Appendix

All vectors are represented in column format and the superscript $^{\mathrm{T}}$ denotes the transposition operator. The elementwise (Hadamard) product of two vectors $\mathbf{u} \in \mathscr{R}^n$ and $\mathbf{v} \in \mathscr{R}^n$ is denoted $\mathbf{u} \circ \mathbf{v}$ and is the vector $\mathbf{w} \in \mathscr{R}^n$ with entries defined as

$$\mathbf{w} = \mathbf{u} \circ \mathbf{v},\ w_i = u_i v_i,\ i = 1 : n \tag{9.60}$$

For two matrices of the same order $\mathbf{X}, \mathbf{Y} \in \mathscr{R}^{n\times m}$

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathscr{R}^{n\times m}} = Tr\left(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right) = Tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right) \tag{9.61}$$

denotes the Frobenius inner product that is expressed in terms of the matrix trace operator Tr. Given the vectors $\mathbf{u} \in \mathscr{R}^n$, $\mathbf{v} \in \mathscr{R}^m$, and the matrix $\mathbf{X} \in \mathscr{R}^{n\times m}$,

$$\langle \mathbf{u}, \mathbf{X}\mathbf{v} \rangle_{\mathscr{R}^n} = \mathbf{u}^{\mathrm{T}}\mathbf{X}\mathbf{v} = \langle \mathbf{u}\mathbf{v}^{\mathrm{T}}, \mathbf{X} \rangle_{\mathscr{R}^{n\times m}} \tag{9.62}$$

Given a functional $e : \mathscr{R}^{n\times m} \rightarrow \mathscr{R}$ of matrix argument $\mathbf{X} \in \mathscr{R}^{n\times m}$, the sensitivity of e with respect to $\mathbf{X}$ is the matrix of the first order partial derivatives denoted as

$$\frac{\partial e}{\partial \mathbf{X}} = \left[\frac{\partial e}{\partial X_{i,j}}\right]_{i=1,n;j=1,m} \in \mathscr{R}^{n\times m} \tag{9.63}$$

The first order variation δe induced by a variation $\delta\mathbf{X}$ is expressed as

$$\delta e = \left\langle \frac{\partial e}{\partial \mathbf{X}}, \delta\mathbf{X} \right\rangle_{\mathscr{R}^{n\times m}} = Tr\left[\frac{\partial e}{\partial \mathbf{X}}(\delta\mathbf{X})^{\mathrm{T}}\right] \tag{9.64}$$

For a nonsingular matrix $\mathbf{X} \in \mathscr{R}^{n\times n}$, the first order variation $\delta\mathbf{X}^{-1}$ in the inverse matrix $\mathbf{X}^{-1}$ induced by a variation $\delta\mathbf{X}$ is expressed as

$$\delta\mathbf{X}^{-1} = -\mathbf{X}^{-1}\delta\mathbf{X}\mathbf{X}^{-1} \tag{9.65}$$

References

Baker NL, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation-targeting problem. Q J R Meteorol Soc 126:1431–1454

Baker NL, Langland RH (2009) Diagnostics for evaluating the impact of satellite observations. In Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications. Springer, Berlin, pp 177–196

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

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 134:1971–1996

Bormann N, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: methods and applications to ATOVS data. Q J R Meteorol Soc 136:1036–1050

Bormann N, Collard A, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: applications to AIRS and IASI data. Q J R Meteorol Soc 136:1051–1063

Bormann N, Geer AJ, Bauer P (2011) Estimates of observation-error characteristics in clear and cloudy regions for microwave imager radiances from numerical weather prediction. Q J R Meteorol Soc 137:2014–2023.

Brousseau P, Berre L, Bouttier F, Desroziers G (2011) Flow-dependent background-error covariances for a convective-scale data assimilation system. Q J R Meteorol Soc. doi: 10.1002/qj.920

Buehner M (2005) Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. Q J R Meteorol Soc 131:1013–1043

Buehner M, Gauthier P, Liu Z (2005) Evaluation of new estimates of background- and observation-error covariances for variational assimilation. Q J R Meteorol Soc 131:3373–3383

Cardinali C (2009) Monitoring the observation impact on the short-range forecast. Q J R Meteorol Soc 135:239–250

Cardinali C, Prates F (2011) Performance measurement with advanced diagnostic tools of all-sky microwave imager radiances in 4D-Var. Q J R Meteorol Soc 137:2038–2046

Chapnik B, Desroziers G, Rabier F, Talagrand O (2006) Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. Q J R Meteorol Soc 132:543–565

Courtier P, Thépaut JN, Hollingsworth A (1994) A strategy of operational implementation of 4D-Var using an incremental approach. Q J R Meteorol Soc 120:1367–1388

Daescu DN (2008) On the sensitivity equations of four-dimensional variational (4D-Var) data assimilation. Mon Weather Rev 136:3050–3065

Daescu DN, Todling R (2009) Adjoint estimation of the variation in model functional output due to the assimilation of data. Mon Weather Rev 137:1705–1716

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

Dee DP, Da Silva AM (1999) Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: methodology. Mon Weather Rev 127:1822–1834

Desroziers G, Ivanov S (2001) Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. Q J R Meteorol Soc 127:1433–1452

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

Desroziers G, Berre L, Chabot V, Chapnik B (2009) A posteriori diagnostics in an ensemble of perturbed analyses. Mon Weather Rev 137:3420–3436

Frehlich R (2011) The definition of ‘truth’ for numerical weather prediction error statistics. Q J R Meteorol Soc 137:84–98

Garand L, Heilliette S, Buehner M (2007) Interchannel error correlation associated with AIRS radiance observations: inference and impact in data assimilation. J Appl Meteorol 46:714–725

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J R Meteorol Soc 125:723–757

Gelaro R, Zhu Y (2009) Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. Tellus 61A:179–193

Gelaro R, Zhu Y, Errico RM (2007) Examination of various-order adjoint-based approximations of observation impact. Meteorol Z 16:685–692

Gelaro R, Langland RH, Pellerin S, Todling R (2010) The THORPEX observation impact intercomparison experiment. Mon Weather Rev 138:4009–4025

Hogan TF, Rosmond TE (1991) The description of the Navy Operational Global Atmospheric Prediction System’s spectral forecast model. Mon Weather Rev 119:1786–1815

Joiner J, Brin E, Treadon R, Derber J, Van Delst P, Da Silva A, Le Marshall J, Poli P, Atlas R, Bungato D, Cruz C (2007) Effects of data selection and error specification on the assimilation of AIRS data. Q J R Meteorol Soc 133:181–196

Kalnay, E. (2002) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge

Lahoz W (2010) Research satellites. In: Lahoz W, Khattatov B, Ménard R (eds) Data assimilation: making sense of observations. Springer, Heidleberg/London, pp 301–321

Langland RH (2005) Issues in targeted observing. Q J R Meteorol Soc 131:3409–3425

Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201

Li H, Kalnay E, Miyoshi T (2009) Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Q J R Meteorol Soc 135:523–533

Liu J, Kalnay E (2008) Estimating observation impact without adjoint model in an ensemble Kalman filter. Q J R Meteorol Soc 134:1327–1335

Liu J, Kalnay E, Miyoshi T, Cardinali C (2009) Analysis sensitivity calculation in an ensemble Kalman filter. Q J R Meteorol Soc 135:1842–1851

Lorenc AC (2003) Modelling of error covariances by 4D-Var data assimilation. Q J R Meteorol Soc 129:3167–3182

Lorenz EN, Emanuel KA (1998) Optimal sites for supplementary weather observations: simulation with a small model. J Atmos Sci 55:399–414

Lupu C, Gauthier P, Laroche S (2011) Evaluation of the impact of observations on analyses in 3D- and 4D-Var based on information content. Mon Weather Rev 139:726–737

Rosmond T, Xu L (2006) Development of NAVDAS-AR: non-linear formulation and outer loop tests. Tellus 58A:45–58

Thépaut JN, Andersson E (2010) The global observing system. In Lahoz W, Khattatov B, Ménard R (eds) Data assimilation: making sense of observations. Springer, Heidleberg/London, pp 263–281

Trémolet Y (2008) Computation of observation sensitivity and observation impact in incremental variational data assimilation. Tellus 60A:964–978

Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR: formulation and initial tests of the linear problem. Tellus 57A:546–559

Zhang S, Anderson JL (2003) Impact of spatially and temporally varying estimates of error covariance on assimilation in a simple atmospheric model. Tellus 55A:126–147

Chapter 10
Treating Nonlinearities in Data-Space Variational Assimilation

Brian S. Powell

Abstract One goal of four-dimensional variational (4D-Var) state estimation is to utilize the longest time window that maximizes the observational constraints to improve predictive skill; unfortunately, nonlinearities are present in geophysical flows and limit the time in which the linear approximation is valid. For weakly nonlinear flows, updating the background trajectory, relinearizing, and repeating the minimization is a way to lengthen the time window. This so called "outer-loop" requires special consideration when minimizing the solution in data-space. This discussion provides a review of the relevant theory and presents two data-space cost functions: the standard cost-function that becomes unconstrained during additional outer-loops and a modified function that preserves the original constraint. Experiments with the Lorenz (J Atmos Sci 20:130–141, 1963) model show that unconstrained outer-loops perform similarly to sequentially applied 3D-Var assimilations by overfitting the observations and producing state estimates with poor predictive skill. Evaluating the *posterior* error covariances, the analysis error, and minimum cost function illustrate how overfitting degrades the solution. This is an important lesson for assimilation schemes: minimizing the model data residuals without proper constraint does not provide the optimal solution. By properly constraining the data-space outer-loop, adjoint-based methods will be well constrained over time windows that are longer than those required by linearity.

B.S. Powell (✉)
Department of Oceanography, University of Hawai'i at Mānoa, Marine Sciences Building,
1000 Pope Road, Honolulu, HI 96822, USA
e-mail: powellb@hawaii.edu

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_10,

10.1 Introduction

Data assimilation has become an important component of numerical modeling combining numerical models with observational data to obtain an improved estimate of the circulation. Variational methods aim to minimize the residual difference between the model and observations via least-squares. Three-dimensional variational assimilation (3D-Var) holds time constant and is valid for synoptic observations. Four-dimensional variational assimilation (4D-Var) is constrained by the physics of the model to preserve the dynamical relationships between the observations during a time window. Much of the theoretical 4D-Var work is described in Le Dimet and Talagrand (1986), Talagrand and Courtier (1987), and Courtier et al. (1993, 1994). Lorenc (2006) provides a comparison of 3D-Var and 4D-Var.

The problem of minimizing the residuals can be accomplished in either the space of the model or in the space defined by the observations referred to as model-space and data-space, respectively. The focus of this work is on the data-space methods that are well described in Courtier (1997), Bennett (2002), Chua and Bennett (2001), Bennett et al. (2008), El Akkraoui et al. (2008), and El Akkraoui and Gauthier (2010).

Variational methods make assumptions of linearity for the time-scales over which the assimilation occurs. For geophysical circulations that are nonlinear, a choice must be made: limit the time window over which the assimilation is performed, or occasionally update the nonlinear trajectory during the assimilation procedure. For some applications, reducing the time window to ensure linearity would collapse the problem to 3D-Var (in the limit as the assimilation time-window converges to the model time-step, 4D-Var becomes 3D-Var). For many applications, it is unacceptable to consider time-dependent observations synoptic. In 4D-Var, the goal is to use the longest possible time window, incorporating as many observations as are available; however, the growth of nonlinearities in the flow may disrupt the convergence of the iteration scheme.

The purpose of this discussion is to examine how data-space variational methods in weakly nonlinear regimes are properly constrained to prevent overfitting of noisy observations. With a longer time window and more observational constraints, a better estimate and prediction are produced. Furthermore, long time-windows provide dynamically consistent circulations without frequent initialization shocks. If the prior estimate is not the fixed reference as is often done in sequential 3D-Var and the standard 4D-Var cost function, the residuals are minimized, but the model structure is no longer physically consistent.

This discussion presents how outer-loops in variational data-space methods can be updated to use a longer assimilation window to better estimate the state. First, the standard data-space cost-function is derived and shown to be inappropriate when applying multiple outer-loops. Two outer-loop methods (one constrained and one unconstrained) and sequential 3D-Var are used to illustrate the insidious effects of overfitting the observations. A number of posterior diagnostics are presented to examine the consistency of the solution before concluding.

10.2 Background

An abbreviated description of variational assimilation is presented, but detailed descriptions may be found in Talagrand and Courtier (1987), Le Dimet and Talagrand (1986), Courtier et al. (1994), Courtier (1997), Chua and Bennett (2001), and Moore et al. (2011b) among others. Assuming a linearly-forced, nonlinear model, the time-step integration is expressed as:

$$\mathbf{x}(t_{i+1}) \;= \mathscr{M}\,(\mathbf{x}(t_i)) \;+\; \mathscr{L}\,(\mathbf{f}(t_i)) \;+\; \mathbf{q}(t_i), \tag{10.1}$$

where $\mathbf{x}(t_i)$ is the model state vector at time t_i, $\mathscr{M}(\mathbf{x}(t_i))$ is the forward nonlinear integration of state $\mathbf{x}(t_i)$ to state $\mathbf{x}(t_{i+1})$, $\mathbf{f}(t_i)$ is the forcing and boundary condition parameter vector, $\mathscr{L}(\cdot)$ transforms the given vector to model forcing and boundary condition influence, and $\mathbf{q}(t_i)$ represents the model errors. Using a model guess for the initial conditions, $\mathbf{x}(0)$, and forcing, $\mathbf{f}(t)$, the model can be integrated to produce a reference or "background" trajectory $\mathbf{x}_b(t)$.

The residuals between the observations and the model are given by the innovation vector,

$$\mathbf{d}_i \;=\; \mathbf{y}_i - \mathscr{H}_i\,(\mathbf{x}_b(t_i))\,, \tag{10.2}$$

for N observations, where $\mathscr{H}_i(\cdot)$ maps data from the model to a given observation $\mathbf{y}_i$ location in time and space. The goal of any assimilation scheme is to reduce these residuals.

Assuming that perturbations, $\delta\mathbf{x}(t_i)$, $\delta\mathbf{f}(t_i)$, and $\delta\mathbf{q}(t_i)$ to the background are within the realm of linearity, then these perturbations evolve by the tangent-linear model linearized around the background trajectory, $\mathbf{x}_b$, with a forcing function, $\mathbf{L}$, linearized about $\mathscr{L}$.

The residuals between the perturbed model solution and the observations are given by

$$\mathbf{r}_i \;=\; \mathbf{y}_i - (\mathscr{H}_i\,(\mathbf{x}_b(t_i)) \;+\; \mathbf{H}_i\mathbf{M}_i\mathbf{z}(t_i)) \;=\; \mathbf{d}_i \;-\; \mathbf{G}_i\mathbf{z}(t_i), \tag{10.3}$$

where $\mathbf{H}_i$ is the linearized sampling matrix, $\mathbf{M}_i$ represents the integration of the tangent-linear model over the interval (t_0, t_i), $\mathbf{G}_i = \mathbf{H}_i\mathbf{M}_i$, and $\mathbf{z}$ is a vector that comprises the perturbations to the initial state, forcing, and model error fields: $\mathbf{z} = (\delta\mathbf{x}, \delta\mathbf{f}, \delta\mathbf{q})^T$.

In 4D-Var, our goal is to solve for $\mathbf{z}$ that minimizes (10.3) via least-squares by employing a quadratic cost function,

$$\mathscr{J} \;=\; \frac{1}{2}\sum_{i=1}^{N}\mathbf{r}_i^T\mathbf{R}^{-1}\mathbf{r}_i \;+\; \frac{1}{2}\mathbf{z}^T\mathbf{P}^{-1}\mathbf{z}, \tag{10.4}$$

where $\mathbf{P}$ is the covariance of uncertainty in the model initial state, forcing, and model error $(\mathbf{x}_b, \mathbf{f}, \mathbf{q})$ and $\mathbf{R}$ is the covariance of uncertainty in the residuals, $\mathbf{r}$.

Substituting for $\mathbf{r}_i$ from (10.3), using vector notation rather than i indices, such that $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_N)^T$, the cost function is given by

$$\mathscr{J} = \frac{1}{2}(\mathbf{d} - \mathbf{Gz})^T \mathbf{R}^{-1} (\mathbf{d} - \mathbf{Gz}) + \frac{1}{2}\mathbf{z}^T \mathbf{P}^{-1} \mathbf{z}. \tag{10.5}$$

The minimal solution for (10.5) is the analysis increment, $\mathbf{z}^a$, that yields $\partial \mathscr{J} / \partial \mathbf{z} = 0$, given by

$$\mathbf{z}^a = \left(\mathbf{G}^T \mathbf{R}^{-1} \mathbf{G} + \mathbf{P}^{-1}\right)^{-1} \mathbf{G}^T \mathbf{R}^{-1} \mathbf{d}, \tag{10.6}$$

where the transpose to the tangent-linear integration, $\mathbf{G}^T$, is the adjoint model integrated backwards over (t_i, t_0). Equation (10.6) is the solution to the 4D-Var problem in model space. The analysis increment is the perturbation applied to the initial conditions and forcing such that the residuals (10.3) are minimized. This is the form used in the incremental scheme shown by Courtier et al. (1994) and used by the European Centre for Medium-range Weather Forecasting (ECMWF) as well as in the ocean as shown in studies such as Weaver et al. (2003), Powell et al. (2008), and Broquet et al. (2009). Simplifying (10.6) by replacing $\mathbf{G}$ and $\mathbf{G}^T$ with $\mathbf{H}$ and $\mathbf{H}^T$, respectively, results in the solution to 3D-Var, which ignores the time-dependent dynamics of the system.

The solution (10.6) may be rearranged using the Woodbury Identity (Golub and Van Loan 1989) such that the minimization is performed in the data space to yield

$$\mathbf{z}^a = \mathbf{PG}^T \left(\mathbf{GPG}^T + \mathbf{R}\right)^{-1} \mathbf{d}. \tag{10.7}$$

This is used by the Physical-space Statistical Analysis System (PSAS) (Courtier 1997) and "Representer" (Chua and Bennett 2001; Bennett 2002) methods. These data-space methods have been used successfully for research in both atmospheric (Chua et al. 2009) and oceanic applications (Di Lorenzo et al. 2007; Muccino et al. 2008; Kurapov et al. 2007). For this discussion,

$$\mathbf{K} = \mathbf{PG}^T \left(\mathbf{GPG}^T + \mathbf{R}\right)^{-1}, \tag{10.8}$$

will be referred to as the "Kalman Gain Matrix." The total solution to the data-space minimization procedure is given by $\mathbf{x}^a = \mathbf{x}_b + \mathbf{Kd} = \mathbf{x}_b + \mathbf{z}^a$.

Due to the large size of the typical geophysical problem, one cannot solve for $\mathbf{z}^a$ explicitly, and the solution is found iteratively via a conjugate-gradient algorithm. This minimization iteration to find an approximation to (10.7) is referred to as the "inner-loop." The estimate of $\mathbf{z}^a$ is revealed at the conclusion (determined by convergence or reaching a maximum number of inner-loops) of the inner-loops. If the problem is linear, then the global minimum is reached when the inner-loops converge.

If the particular problem is sufficiently nonlinear that the cost function is not well represented by the linear model, it is desirable to update $\mathbf{x}_b$ with $\mathbf{z}^a$, re-linearize, and find a new update $\mathbf{z}^a$. This iterative procedure is referred to as the "outer-loop."

In many 4D-Var applications, the problem is sufficiently linear that multiple outer-loops are not used. When it is truly linear, the increments would not change the subsequent linearization about the perturbed $\mathbf{x}_b$. However, any nonlinearities may result in differing cost functions between the nonlinear and linear spaces. To deal with this, one may choose to update the nonlinear trajectory with the increments to avoid local minima in the nonlinear cost function. During these additional outer-loops, consideration must be given that the increments remain constrained to the original background, $\mathbf{x}_b$; otherwise, the iterates will overfit the data at the expense of the background covariance. Solving the problem in model-space (see Eq. 10.6) provides an implicit constraint to the initial-state and additional outer-loops are not an issue (Tshimanga et al. 2008). This is not the case in data-space and during additional outer-loops, the cost-function must be modified to add an additional constraint.

Upon completing the inner-loops, an estimate of $\mathbf{z}_k^a$ is found, where k signifies the outer-loop iteration (beginning with $k = 1$). The operators $\mathbf{G}_{k-1}$ and $\mathbf{G}_{k-1}^T$ are linearized about the prior trajectory $\mathbf{x}_{k-1}$ (where $\mathbf{x}_0 = \mathbf{x}_b$ and $\mathbf{z}_0 = 0$). If the problem were perfectly linear, $\mathbf{G}_k = \mathbf{G}_{k-1}$ and there would be no need for additional outer-loops. The next iteration of the outer-loop ($k + 1$) requires linearization about the new prior, ($\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{z}_k^a$) integrated by (10.1). The question becomes how to keep the original background constraint, $\mathbf{x}_b$, active when further iterating in the data-space methods. The typical approach is to simply reapply (10.7) such that

$$\mathbf{z}_k = \mathbf{P}\mathbf{G}_{k-1}^T \left(\mathbf{G}_{k-1}\mathbf{P}\mathbf{G}_{k-1}^T + \mathbf{R}\right)^{-1} \mathbf{d}_{k-1}, \tag{10.9}$$

where $(\mathbf{d}_{k-1})_i = \mathbf{y}_i - \mathscr{H}_i \mathbf{x}_{k-1}(t_i)$ as was used in Zaron et al. (2011). This constrains $\mathbf{z}_k$ to be small; however, the total increment, $\sum_{j=1}^{k} \mathbf{z}_j$, is required to be small. Hence (10.9) constrains the increments only against the prior nonlinear circulation ($\mathbf{x}_{k-1}$), which allows the increments to deviate from the background trajectory.

To rectify, the total constraint must be incorporated to yield a new cost function,

$$\begin{aligned}\mathscr{J}_k = {} & \frac{1}{2}\left(\mathbf{d}_{k-1} - \mathbf{G}_{k-1}\mathbf{z}_k\right)^T \mathbf{R}^{-1} \left(\mathbf{d}_{k-1} - \mathbf{G}_{k-1}\mathbf{z}_k\right) \\ & + \frac{1}{2}\left(\mathbf{z}_k + \sum_{j=1}^{k-1} \mathbf{z}_j\right)^T \mathbf{P}^{-1} \left(\mathbf{z}_k + \sum_{j=1}^{k-1} \mathbf{z}_j^a\right).\end{aligned} \tag{10.10}$$

The gradient at its minimum is given by

$$\frac{\partial \mathscr{J}_k}{\partial \mathbf{z}_k} = -\mathbf{G}_{k-1}^T \mathbf{R}^{-1} \left(\mathbf{d}_{k-1} - \mathbf{G}_{k-1}\mathbf{z}_k\right) + \mathbf{P}^{-1} \left(\mathbf{z}_k + \sum_{j=1}^{k-1} \mathbf{z}_j^a\right) = 0. \tag{10.11}$$

Solving for the analysis increment reveals

$$\mathbf{z}_k^a = \left(\mathbf{G}_{k-1}^T\mathbf{R}^{-1}\mathbf{G}_{k-1} + \mathbf{P}^{-1}\right)^{-1}\mathbf{G}_{k-1}^T\mathbf{R}^{-1}\mathbf{d}_{k-1} - \mathbf{P}^{-1}\sum_{j=1}^{k-1}\mathbf{z}_j^a. \qquad (10.12)$$

The constraint to minimize the total increment acts as an additional forcing on the model-space solution as compared to (10.6). The additional background constraint forcing from $\sum_{j=1}^{k-1}\mathbf{z}_j^a$ opposes the residuals force and prevents overfitting the observations. Once updating the prior nonlinear circulation no longer provides additional increments, the right hand side of (10.12) vanishes and convergence is reached. However, (10.12) remains in model-space, and the goal is to understand how the additional constraint impacts data-space methods. This requires the additional forcing term to be applied within data-space.

The solution is found by El Akkraoui et al. (2008), such that the analysis increment for outer-loop k is given by

$$\mathbf{z}_k^a = \mathbf{P}\mathbf{G}_{k-1}^T\left(\mathbf{G}_{k-1}\mathbf{P}\mathbf{G}_{k-1}^T + \mathbf{R}\right)^{-1}\left(\mathbf{d}_{k-1} + \mathbf{G}_{k-1}\mathbf{z}_{k-1}\right). \qquad (10.13)$$

This provides the constraint that each data-space outer-loop is constrained by its prior loop. The prior increment is propagated through the tangent-linear model that is now linearized about the previous outer-loop trajectory. It should be noted that if the problem is fully linear, this method is identical to the typical approach because $\mathbf{G}_1 = \mathbf{G}_0$, and the additional term in (10.13) is negated.

This procedure can be carried forward for as many outer-loops as necessary to approximate nonlinearities without violating the initial increment constraint. It is important to note that with each outer-loop, a new minimization descent begins from approximately zero using a new linearization, which will have negative consequences on preconditioning schemes that improve their estimates during subsequent outer-loop runs.

10.3 Experiments

As presented, there are two potential methods for handling multiple outer-loops in data-space methods. The typical cost-function formulation (10.9) will be referred to as the "overfit" method because it forgets the background constraint after more than one outer-loop giving full weight to the observations. Overfitting observations is never the goal of assimilation and should always be avoided, and the insidious effects of overfitting are found in the results. The method derived by El Akkraoui et al. (2008) is designated the "constrained" method as it respects the total increment constraint. Because overfitting observations leads to solutions with highly variable structure and poor prediction skill, the 4D-Var solutions are compared along with sequential 3D-Var during both assimilation and prediction phases. The strong-constraint of all methods is used in both data poor and data rich regimes to examine the efficacy. To accomplish these experiments, a system that is weakly nonlinear during the assimilation window is required. The Lorenz (1963) (hereafter referred to as Lorenz63) system provides a nonlinear, dynamic system with particular sensitivity to the initial state, $\mathbf{x}(t_0)$. The Lorenz63 system is a well used test case in dynamical systems due to its highly nonlinear but basic structure. Prediction is a difficult problem with its strong sensitivity to changes in the initial conditions.

One of the first adjoint-based assimilation experiments using Lorenz63 was performed by Gauthier (1992). This work was followed by other experiments using both adjoint and Kalman filter techniques by Evensen and Fario (1997) and Miller et al. (1994). More recently, Ngodock et al. (2007) examined how well the weakly-constrained Representer method can be used to approximate strongly nonlinear flows in an observation-rich environment.

The Lorenz63 system is a simplified model of convective atmospheric dynamics. It is unforced $(\mathbf{f}(t_i) = 0)$ and uses a three-dimensional state vector (x, y, z) to describe the convective motion intensity, temperature difference between vertical currents, and vertical temperature deviation from linearity, respectively. The model parameters (σ, r, b) describe the Prandtl number, ratio of the Rayleigh number to criticality, and convective period. Parameter values of $(10, 28, 8/3)$ are chosen to provide a strongly nonlinear flow as in Gauthier (1992). The forward integration of the model is performed with the standard fourth-order Runge-Kutta method. The tangent-linear operator $\mathbf{M}$ is implemented with the tangent-linearization of both the nonlinear model and Runge-Kutta integrator. Likewise, the adjoint model is represented by $\mathbf{M}^T$.

Gauthier (1992) examined two regimes of the Lorenz63 system: "regular," in which the system remains within a single attractor during the time window and is weakly nonlinear, and the "transition" case that changes attractor for the other during the time window and is strongly nonlinear. For each case, Gauthier (1992) integrated from $t = [0, 8]$; however, these periods are too long for linear methods. As shown by Gauthier (1992) and Evensen and Fario (1997), the local minima (due to nonlinearity) in the transition period limit the effectiveness of the assimilation. Miller et al. (1994) showed that these issues were due to the length of time window used.

Determining the length of the window to keep the system weakly nonlinear requires a metric to evaluate any differences between the linear and nonlinear solutions. A simple cost-function to compare a perturbed trajectory against the unperturbed is given by $\mathscr{A} = (\mathbf{x} - \mathbf{x}_t)^T\,\mathbf{Q}^{-1}\,(\mathbf{x} - \mathbf{x}_t)$, where $\mathbf{x}$ is a perturbed model trajectory sampled at every timestep, $\mathbf{x}_t$ is the unperturbed model trajectory, and $\mathbf{Q}$ is the prescribed measure of estimate error chosen to be diagonal with values of 2. Perturbations were randomly chosen with a variance of $\mathbf{Q}$ and integrated through the tangent-linear model (linearized about $\mathbf{x}_t$). These perturbations were also added to $\mathbf{x}_t$ and integrated through the nonlinear model (10.1). Any differences between the nonlinear and linear cost functions, $\mathbf{A}$, are due to unresolved nonlinearity. Figure 10.1 shows the dramatic difference between two different time windows. In Fig. 10.1a, the system is weakly nonlinear over 1.9 time units; however, over the

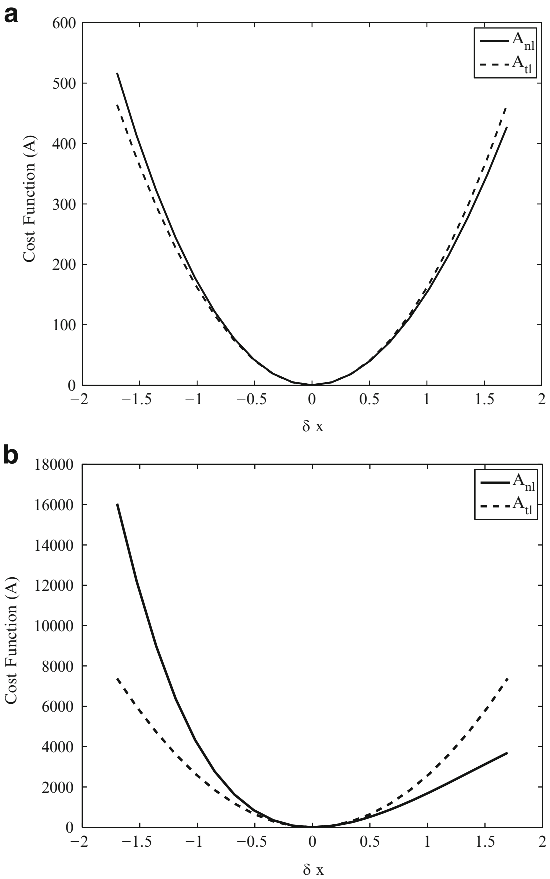


Fig. 10.1 Cost functions resulting from various sized perturbations (x-axis) for both the nonlinear (*solid*) and linear (*dashed*) models integrated for (**a**) 1.9 and (**b**) 8 time units. If the system were within the linear regime, both lines would be identical

8 time units (Fig. 10.1b) as used by Gauthier (1992), the system is highly nonlinear (note the difference in scales). Perturbations are only shown for x and results (not shown) are similar for y and z. The situation is far worse during transition periods. The system is weakly nonlinear only over 0.5 time units, but strongly nonlinear over 8 time units. In fact, multiple minima are present in the nonlinear cost-function; however, the linear cost function is nearly five orders of magnitude greater.

It is clear that the degree of nonlinearity is a function of time window length and size of the initial perturbation. In order to determine a range of valid time windows that remain weakly nonlinear, numerous perturbations were integrated through both the linear and nonlinear models and the differences were normalized by the initial perturbation. The time at which the ratio achieves one is determined as the maximum possible time window length to be considered. The results for the regular case in Fig. 10.2 show that with perturbations near zero, the maximum window length is 14 time units. As the perturbations in x and y (perturbations in z are not shown, but are similar) increase, the maximum window size decreases significantly. Differences

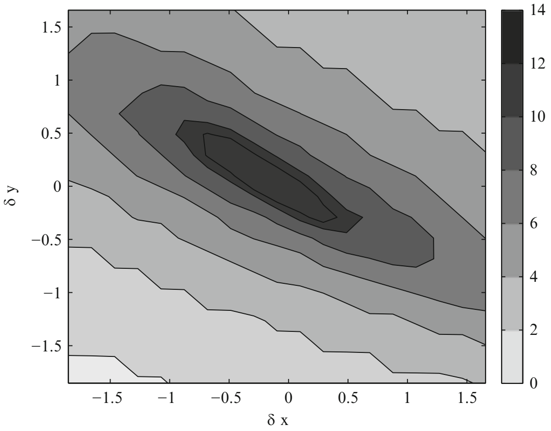


Fig. 10.2 The maximum time before linearity breaks down for the "regular" regime using perturbations in x or y of varying sizes

equal to the initial error are too great to be considered weakly nonlinear; however, these time window estimates provide an upper time bound for the experiments.

As is shown in Figs. 10.1 and 10.2, the prescribed initial error is significant to determining the time window of the system. No matter the growth of the nonlinearities, a longer period can be used with smaller initial error as the error growth rate can be approximated by $\mathbf{P}^{\frac{1}{2}}e^{\lambda t}$. If the *prior* error, $\mathbf{P}$, is small, long time-windows are possible. Figure 10.1b shows that despite the large growth of linear error, long window solutions are possible if $\mathbf{P}$ is small. The linear and nonlinear cost functions deviate illustrating that this system is not fully linear over the periods of interest, and it provides an ideal configuration to examine the role of constraints in the outer-loops.

Ensembles of twin experiments are created to compare the two separate cost functions. The regular and transition initial conditions were integrated for the predetermined time window to generate truth trajectories then sampled to generate the observations. Observations were equally spaced over the time window avoiding the starting and ending times and each state variable was sampled an equal number of times, such that at each observation time, the entire state is observed. An ensemble of 500 members was created by randomly perturbing the initial conditions with variance consistent with $\mathbf{P}$. For each ensemble member, random error consistent with $\mathbf{R}$ was added to each observation. Initially, $\mathbf{P}$ and $\mathbf{R}$ were chosen as diagonal matrices with elements set to 2. Each member assimilated the randomly perturbed observations using four outer-loops with both the overfit and constrained data-space methods as well as with sequentially applied 3D-Var. The 3D-Var was applied at each time of the full state observations during the time window of interest. After applying the 3D-Var, the system was integrated to the next set of observations, and the 3D-Var is repeated with the same $\mathbf{P}$ and $\mathbf{R}$ and using the current state as $\mathbf{x}_b$. Unlike the 4D-Var solutions, the 3D-Var solutions are discontinuous during each of the examined time windows and *posterior* statistics discussed below are not readily available. The time window was taken as

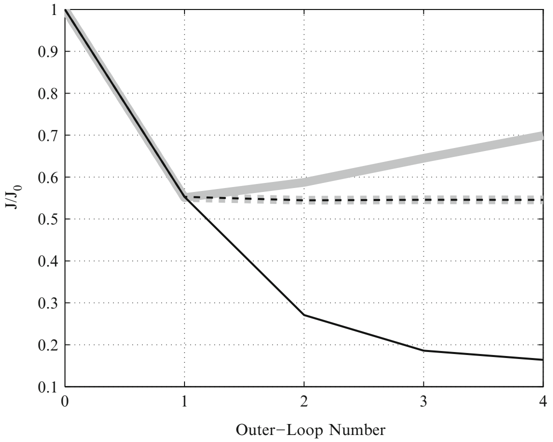


Fig. 10.3 The normalized total cost function, $\mathscr{J}$, from the nonlinear (*grey*) and linear (*black*) models after each other-loop for the regular case with the overfit (*solid*) and the constrained (*dashed*) methods

one-half of the values found in Fig. 10.2 as a conservative measure of the time window ($t = [0, 1.9]$ for the regular case). Because the Lorenz63 system state contains only three state variables, the Kalman matrix (10.8) was computed directly rather than employ a gradient-descent inner-loop.

For each ensemble member, the linearity is estimated similar to that shown in Fig. 10.1, but scaled by the observational error, $\mathbf{R}$, because this provides an estimate of the errors expected in the fit (including nonlinearity). Any member with a linearity error greater than 10 % of $\mathbf{R}$ is thrown out. This insures that only weakly nonlinear ensemble members are compared. For all cases, at least 200 members passed this criterion. Because the distribution of the ensemble member trajectories is not Gaussian, for all remaining discussion (unless noted) the median of the valid ensemble members is used to represent the entire ensemble. Ensembles using 3, 6, 9, 12, 15, 18, and 21 observations were created.

The cost functions (Eq. 10.4) for each ensemble member upon completion of each outer-loop, normalized by the initial guess are computed to compare the behavior of each outer-loop. The ensemble median cost-functions for the three observation case are shown in Fig. 10.3. Because the overfit method constrains only the residuals and ignores the background constraint, it significantly under-estimates the cost function. In both the regular and transition (not shown as it is similar) cases, the overfit method significantly reduces the linear cost function through each outer-loop. After the second inner-loop, most further reduction is accomplished against the background constraint, $\mathscr{J}_b$, term. The overfit method considers only the current increment weighted by $\mathbf{P}$ rather than the total increments over all inner-loops. The constrained case shows little improvement after the first outer-loop, and after the second outer-loop, it has converged.

It is expected that overfitting the observations would degrade the forecasting ability of the model. Using the true (unperturbed) trajectory, a new cost function $\mathscr{A} = (\mathbf{x} - \mathbf{x}_t)^T\, \mathbf{S}^{-1}\, (\mathbf{x} - \mathbf{x}_t)$ is computed every 0.05 time units of the window. The diagonal error matrix $\mathbf{S}$ is composed of the prescribed random observational error.

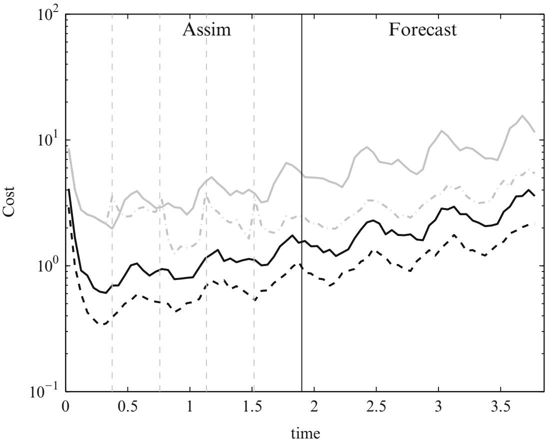


Fig. 10.4 The temporal cost function between the truth and four cases: background trajectory (*grey*), 3D-Var (*dashed grey*), "overfit" (*solid*), and "constrained" (*dashed black*) methods

Figure 10.4 shows the ensemble median temporal cost function for the assimilation and forecast periods for the regular case (transition case is similar) using 12 observations when solved with sequential 3D-Var (at four time steps, marked with grey, vertical line in the figure) and the two 4D-Var outer-loop schemes. All methods improve upon the initial guess; however, the forecasts from the 3D-Var and overfit methods suffer. Once the final set of observations are accounted for, the 3D-Var and the overfit method follow similar trajectories.

A skill ratio between the total cost functions of the two cost-functions is given by $1 - \mathscr{A}_c/\mathscr{A}_o$, where $\mathscr{A}_o$ and $\mathscr{A}_c$ are the overfit and constrained cost functions, respectively is shown in Fig. 10.5. Furthermore, the skill ratio between the 3D-Var and constrained method $(1 - \mathscr{A}_c/\mathscr{A}_3)$ is shown in grey. For both cases, the assimilation and forecast periods are shown as a function of the number of observations assimilated.

In the limit as the cost of $\mathscr{A}_o$ or $\mathscr{A}_3$ increases, the ratio goes to zero meaning that the errors of the overfit or 3D-Var methods are significantly higher than the constrained. So, as the skill reaches higher values shows, overfitting penalizes the solution. As before, the overfit method improves only when significantly increasing the number of observational constraints. This is an important consideration for geophysical assimilation where in situ data are temporally and spatially sparse. As more noisy data are assimilated into the sequential 3D-Var solutions, overfitting becomes a significant issue because it is a series of individual time solutions unconstrained by the original background. These series of independent fits lose dynamical consistency as more and more observations are included through time.

The cost functions reveal that the overfit procedure in data-space assimilation violates the fundamental constraints and produces worse assimilation and forecasting results. Sequential 3D-Var was found to perform similarly for these number of observations; however, it worsened as the number of observations increased. No matter the choice of values of **P** and **R** (not shown) or the regular or transition cases, the overfit method consistently undervalues the cost function and produces

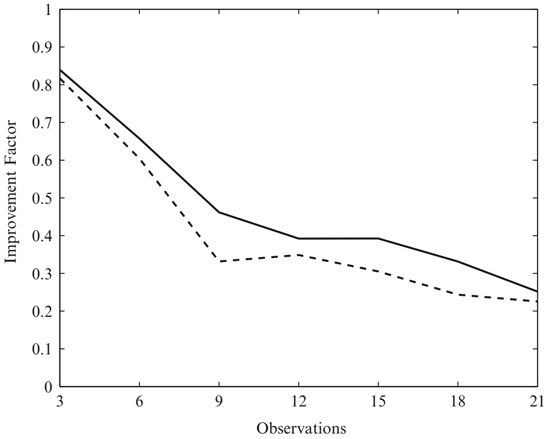


Fig. 10.5 The total cost function improvement factor of the constrained versus overfit (*black*) methods and constrained versus 3D-Var (*grey*) as a function of number of observations. The assimilation (*dashed*) and forecast (*solid*) periods are shown. Values near 1 show a significant error in the overfit or 3D-Var cases

poor forecast results. These two methods may exhibit extremely small residuals at individual times; however, this comes at the expense of adding noisy model structure with a loss of dynamical consistency throughout the time window. For the remainder of the discussion, *posterior* analysis of the two 4D-Var cases are examined.

10.4 Posterior Statistics

The overfit method exhibits worse predictive skill for every case examined. By ignoring the total increment constraint, it emphasizes the noisy observations. There are a number of posterior diagnostics available to quantify the performance of each outer-loop method, including: the final analysis error, the consistency of the *prior* and *posterior* errors, and the true minimum cost functions.

The analysis error ($\mathbf{E}_a$) of the assimilation is provided by the inverse term in (10.6). For most geophysical applications, direct calculation of this matrix is not possible due to the size; however, it can be computed directly for the Lorenz63 system. Because these are twin experiments, the true error between the analysis and the truth is computed and compared against the diagonal of $\mathbf{E}_a$. The ratio between the true error, $\mathbf{E}_t$, and the analysis error, (diag[$\mathbf{E}_a$]), is compared to examine how well each cost-function evaluates the true statistics. By ignoring the background error, it is expected that the overfit method will underestimate the analysis error. For all valid ensemble members, the mean $\mathbf{E}_t/(\text{diag}[\mathbf{E}_a])$ was 1.075 for the overfit method. This is an underestimate of the true error by 7.5 %. For the constrained method, the mean ratio was 0.984, which is an overestimate of the true analysis error by 1.6 %.

The initial background state and analysis state are compared to the true initial state for all valid ensemble members tested and the overfit method increases the initial error by 8.8 % on average, while the constrained method improves the initial error by 3.2 %. Without a fixed constraint, the overfit method tends to push away from the true initial conditions in order to better represent the noisy observations. In fact, comparing the true observations without random error perturbations to the final analysis trajectories, the overfit method only reduces the innovation residuals (Eq. 10.3) between the model and true observations by 37 %, while the constrained method reduces it by 88 %. Without a proper background constraint, the overfit method attempts to best fit the noisy observations. The constrained method attempts to best represent the noisy observations while decreasing the initial error. This initial constraint actually provides for a better observational fit against the true observations.

Another important measure is the consistency between the prescribed *prior* background ($\mathbf{P}$) and observational ($\mathbf{R}$) errors with the *posterior* estimates from the analysis. As shown in Desroziers et al. (2005) and used in Moore et al. (2011a), two relationships are found. First, two vectors are defined: $(\mathbf{d}_b^k)_i = \mathscr{H}(\mathbf{x}_k(t_i)) - \mathscr{H}(\mathbf{x}_b(t_i))$, where $\mathbf{x}_k$ is the integrated solution of the kth outer-loop and provides the difference between the analysis and background at the observation locations; and, $(\mathbf{d}_o^k)_i = \mathbf{y}_i - \mathscr{H}(\mathbf{x}_k(t_i))$ provides the difference between the analysis and the observations. Desroziers et al. (2005) shows that the *posterior* error estimates are given by,

$$E\left[\mathbf{d}_b^k \mathbf{d}^T\right] = \mathbf{G}_k \mathbf{P} \mathbf{G}_k^T \tag{10.14}$$

$$E\left[\mathbf{d}_o^k \mathbf{d}^T\right] = \mathbf{R}. \tag{10.15}$$

Because each term is computed directly in the Lorenz63 problem, the specified *prior* error values are compared with (10.14) and (10.15). First, the *posterior* error estimate from the analysis, $\mathbf{P}_a = \mathbf{d}_b^k \mathbf{d}^T$, is compared with the diagonal of the *prior* error, $\mathbf{P}_p = \mathbf{G}_k \mathbf{P} \mathbf{G}_k^T$. Because the diagonal values of $\mathbf{P}$ are prescribed, one would expect consistency with the *posterior* error; however, the true error, $\mathbf{P}_t$, is less than $\mathbf{P}$ because ensemble members were selected that did not violate weak nonlinearity. This created a selection bias that decreases the true initial error of the ensemble. For all of the cases performed, the $\mathbf{P}_t$ averages 30 % less than the specified value, $\mathbf{P}$, regardless of the method.

With this in mind, the ratio of the *posterior* error, $\mathbf{P}_a$, to the *prior*, $\mathbf{P}_p$, is compared. The inner-product of the sampled observation locations should be equivalent to the initial error projected into data-space and the $\mathbf{P}_a$ is highly dependent upon the number of observations as expected by (10.14). As shown in Fig. 10.6, both methods tend to underestimate the actual background error. Interestingly, the ratio depends only upon the choice of $\mathbf{R}$. As $\mathbf{R}$ decreases, both assimilation methods grossly underestimate the error in the background. As $\mathbf{R}$ increases, the overfit method significantly overestimates the error in the background as it relies solely on the observations, while the constrained method becomes more consistent. The results are from the regular case, but are consistent with the transition case.

Likewise, (10.15) is used to compute the *posterior* $\mathbf{R}_a$ for comparison with the prescribed *prior* $\mathbf{R}$. Figure 10.7 shows that no matter the number of observations or selection of $\mathbf{P}$, the overfit method always underestimates the actual error in

Fig. 10.6 Ratio ($\mathbf{P}_a/\mathbf{P}_p$) of the *posterior* error to the *prior* projected in data-space. The overfit (*black*) method generates a higher error estimate compared to the constrained (*grey*) method across all choices of background and observational errors as well as over varying numbers of observations

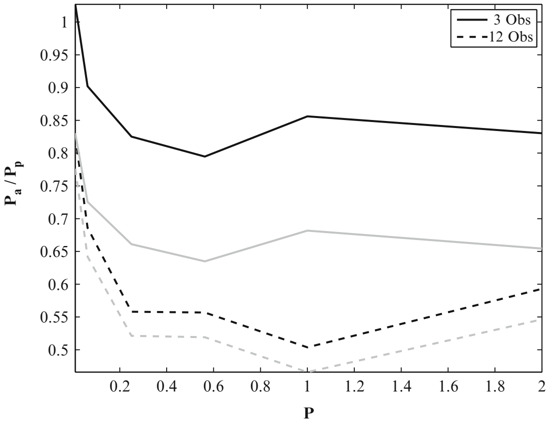
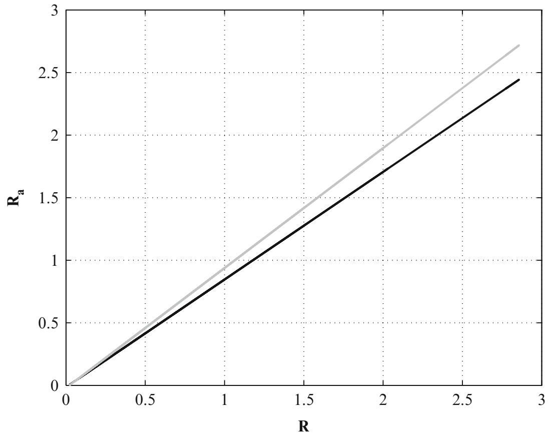


Fig. 10.7 Comparison of *prior* **R** and *posterior* $\mathbf{R}_a$ observational error for overfit (*black*) and constrained (*grey*) methods



the observations. The constrained method slightly underestimates the observational error, but is consistent with the *prior*. This exposes the overfit method for placing too much emphasis on the observations.

As a final comparison, the theoretical cost-function minimum is compared to the minimized cost-function at the end of each outer-loop. Because the overfit method is unconstrained by **P** after the first outer-loop, it was shown that it will minimize $\mathscr{J}_o$ at the expense of $\mathscr{J}_b$.

Bennett (2002) showed that for correctly specified **P** and **R**, the minimum value of the cost-function is $\mathscr{J}_{min} = N_{obs}/2$. This measure has been used as a useful diagnostic by Weaver et al. (2003) and Powell et al. (2008). Unfortunately, this measure does not quantify the contribution of each component of the cost function. Moore et al. (2011a) provides a concise review of the work of Talagrand (1999), Chapnik et al. (2006), and Desroziers et al. (2009) along with the derivation for determining the minimum theoretical values of each cost-function component. The

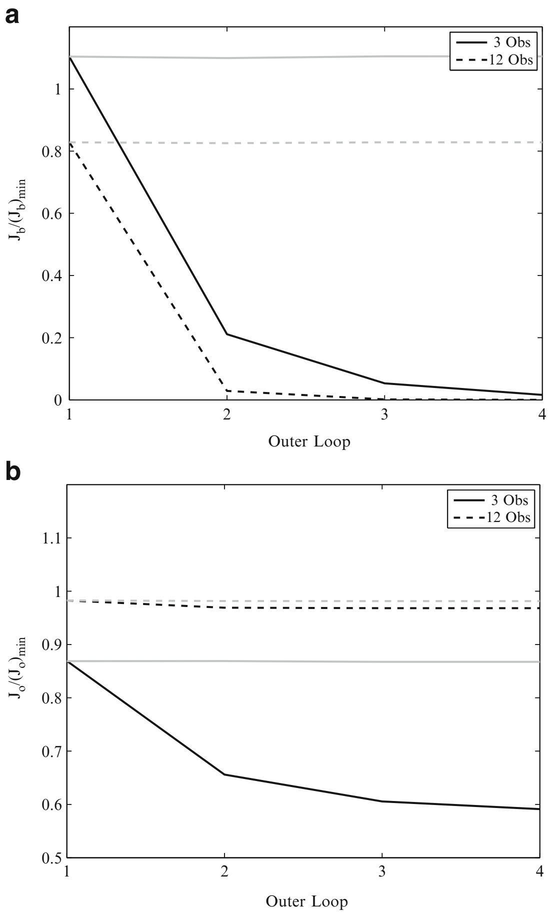


Fig. 10.8 Ratio of background (**a**) and observational (**b**) cost functions versus the theoretical minimum for overfit (*black*) and constrained (*grey*) methods. The *prior* observational error controls level

cost-function minima are functions of the trace of the Kalman gain matrix, and are given by:

$$\left(\mathscr{J}_b\right)_{min} = \frac{1}{2}\left(N_{obs} - \mathrm{Tr}\left[\mathbf{G}_{k-1}\mathbf{K}_k\right]\right) \tag{10.16}$$

$$\left(\mathscr{J}_o\right)_{min} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{G}_{k-1}\mathbf{K}_k\right] \tag{10.17}$$

Normally, to compute the trace of the matrices involved would be prohibitively expensive; however, for the Lorenz63 system, the Kalman Gain matrix is computed explicitly to solve (10.16) and (10.17).

Both methods tend to underestimate the background cost because the actual background error is less than the prescribed; however, the overfit method increasingly underestimates the background cost with each subsequent outer-loop. This is well illustrated in Fig. 10.8 (other results are very similar and are not shown).

The constrained method achieves the minimum cost function precisely when the prescribed observational errors are equal to the background error. When the observational errors are greater than the background error, the background cost function is overestimated. The observational cost function shows that the overfit method continues to reduce the observational cost far below the minimum level, which is a clear indication of strong overfitting. After the second outer-loop, the constrained method no longer reduces the cost-function as it has reached the minimum that balances the two gradients.

Cardinali et al. (2004) showed that $\left(\mathscr{J}_o\right)_{min}$ is a measure of the degrees of freedom in the system and that $\left(\mathscr{J}_b\right)_{min}$ is a measure of the degrees of freedom in the observations. The overfit method reduces $\mathscr{J}_b$ to zero, thus placing all weight onto the observations; hence, the degrees of freedom to the observations is zero. Only the constrained method maintains the proper relationship between the constraints of the system.

10.5 Discussion

Unconstrained outer-loops in data-space and sequential 3D-Var degrades the solution and overfits the data. With the proper formulation, weakly nonlinear problems can be solved with longer time-windows to improve both the number of observational constraints and length of the trajectory. For problems that are purely linear (such as when employing the "Representer" method of Bennett (2002)), there is no need for additional outer-loops because the problem is solved when the inner-loops converge. For many geophysical applications that are weakly nonlinear, multiple outer-loops are advantageous. It is important to note that in methods that give greater weight to the observations (3D-Var, multi-variate optimal interpolation, etc.), careful consideration must be paid to prevent increments from adding unrealistic structure to the model in order to fit the observations because it was shown that this structure leads to severely handicapped predictive skill.

Using a number of quantifiable measures of the assimilation framework to compare the *posterior* errors of each method, the constrained method preserves the *prior* error and does not underestimate the true error. The analysis error of the constrained method assimilation was consistent with the true error. Although the overfit method tended to worsen the background initial conditions, it still underestimated the true analysis error by an average of 7.5 %. Not only does the overfit method provide an improper minimization, the *posterior* analysis statistics are invalid. In addition, the overfit method further underestimated the error in the observations, which is expected as it gave higher consideration to the noisy observations. The constrained method was consistent with the true *posterior* statistics, and as shown by the cost metrics, provided significant estimate to the degrees of freedom in the system.

When assimilating data in geophysical models, a long time window constrained by the model dynamics with as many available observations provides a quantifiably superior analysis circulation. This circulation is dynamically consistent for the longer time window, avoiding the frequent discontinuities present in 3D-Var or short window 4D-Var solutions. To achieve similar results, the overfitting methods would require nearly error free observations at a frequency that approaches the time-step of the model. For geophysical flows with temporally sparse data, properly constrained data-space methods provide an ideal configuration for assimilation.

Acknowledgements The author would like to thank Dr. Bruce Cornuelle for his thoughts and discussions. Dr. Powell was supported by the Office of Naval Research contract #N00014-09-10939.

References

Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, Cambridge/New York

Bennett AF, Chua BS, Pflaum BL, Erwig M, Fu Z, Loft RD, Muccino JC (2008) The inverse ocean modeling system. Part I: implementation. J Atmos Ocean Technol 25:1608–1622

Broquet G, Edwards CA, Moore A, Powell BS, Veneziani M, Doyle JD (2009) Application of 4D-Variational data assimilation to the California Current System. Dyn Atmos Oceans 48:69–92

Cardinali C, Pezzulli S, Andersson E (2004) Influence-matrix diagnostic of a data assimilation system. Q J R Meteorol Soc 130:2767–2786

Chapnik B, Desroziers G, Talagrand O (2006) Diagnosis and tuning of observational error statistics in a quasi-operational data assimilation setting. Q J R Meteorol Soc 132:543–565

Chua BS, Bennett AF (2001) An inverse ocean modeling system. Ocean Model 3:137–165

Chua BS, Xu L, Rosmond T, Zaron ED (2009) Preconditioning representer-based variational data assimilation systems: application to NAVDAS-AR. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications. Springer, Berlin/Heidelberg, pp 307–319. doi:10.1007/978-3-540-71056-1

Courtier P (1997) Dual formulation of four-dimensional variational assimilation. Q J R Meteorol Soc 123:2449–2461

Courtier P, Andersson E, Heckley WA, Kelly G, Pailleux J, Rabier F, Thépaut JN, Undén P, Vasiljević D, Cardinali C, Eyre J, Hamrud M, Haseler J, Hollingsworth A, McNally AP, Stoffelen A (1993) Variational assimilation at ECMWF. Technical report 194, European Centre for Medium-Range Weather Forecasts

Courtier P, Thépaut JN, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. Q J R Meteorol Soc 120:1367–1387

Desroziers GL, 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

Desroziers G, Berre L, Chabot V, Chapnik B (2009) A posteriori diagnostics in an ensemble of perturbed analyses. Mon Weather Rev 137:3420–3436

Di Lorenzo E, Moore AM, Arango HG, Cornuelle BD, Miller AJ, Powell BS, Chua BS, Bennett AF (2007) Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): development and application for a baroclinic coastal upwelling system. Ocean Model 16:160–187

El Akkraoui A, Gauthier P (2010) Convergence properties of the primal and dual forms of variational data assimilation. Q J R Meteorol Soc 136:107–115

El Akkraoui A, Gauthier P, Pellerin S, Buis S (2008) Intercomparison of the primal and dual formulations of variational data assimilation. Q J R Meteorol Soc 134:1015–1025

Evensen G, Fario N (1997) Solving for the generalized inverse of the Lorenz model. J Meteorol Soc Jpn 75:229–243

Gauthier P (1992) Chaos and quadric-dimensional data assimilation: a study based on the Lorenz model. Tellus 44A:2–17

Golub GH, Van Loan CF (1989) Matrix computations. Johns Hopkins University Press, Baltimore

Kurapov AL, Egbert GD, Allen JS, Miller RN (2007) Representer-based variational data assimilation in a nonlinear model of nearshore circulation. J Geophys Res 112:C11019

Le Dimet F, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus 38A:97–110

Lorenc A (2006) Why does 4D-Var beat 3D-Var? Q J R Meteorol Soc 131:3247–3257

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

Miller R, Ghil M, Gauthiez F (1994) Advanced data assimilation in strongly nonlinear dynamical systems. J Atmos Sci 51:1037–1056

Moore AM, Arango HG, Broquet G, Edwards C, Veneziani M, Powell BS, Foley D, Doyle J, Costa D, Robinson P (2011a) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: part II – performance and application to the California Current System. Prog Oceanogr 91:50–73. doi:10.1016/j.pocean.2011.05.003

Moore AM, Arango HG, Broquet G, Powell BS, Zavala-Garay J, Weaver AT (2011b) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: part I – system overview and formulation. Prog Oceanogr 91:34–49. doi:10.1016/j.pocean.2011.05.004

Muccino JC, Luo H, Arango HG, Haidvogel D, Levin JC, Bennett AF, Chua BS, Egbert GD, Cornuelle BD, Miller AJ, Di Lorenzo E, Moore AM, Zaron ED (2008) The inverse ocean modeling system. Part II: applications. J Atmos Ocean Technol 25:1623–1637

Ngodock HE, Smith SR, Jacobs GA (2007) Cycling the representer algorithm for variational data assimilation with the Lorenz attractor. Mon Weather Rev 135:373–386

Powell BS, Arango HG, Moore AM, Di Lorenzo E, Milliff RF, Foley D (2008) 4DVAR data assimilation in the Intra-Americas Sea with the Regional Ocean Modeling System (ROMS). Ocean Model 25:173–188

Talagrand O (1999) A posteriori verification of analysis and assimilation algorithms. In: Proceedings of ECMWF Workshop on Diagnostics of Data Assimilation Systems, pp 17–28. Shineld Park, Reading, UK

Talagrand O, Courtier P (1987) Variational assimilation of meteorological observations with the adjoint vorticity equation. I: theory. Q J R Meteorol Soc 113:1311–1328

Tshimanga J, Gratton S, Weaver AT, Sartenaer A (2008) Limited-memory preconditioners with application to incremental four-dimensional variational data assimilation. Q J R Meteorol Soc 134:751–769

Weaver AT, Vialard J, Anderson DLT (2003) Three- and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part I: formulation, internal diagnostics, and consistency checks. Mon Weather Rev 131:1360–1378

Zaron ED, Pradal MA, Miller PD, Blumberg AF, Georgas N, Li W, Cornuelle JM (2011) Bottom topography mapping via nonlinear data assimilation. J Atmos Ocean Technol 28:1606–1623

Chapter 11
Linearized Physics for Data Assimilation at ECMWF

Marta Janisková and Philippe Lopez

Abstract A comprehensive set of linearized physical parameterizations has been developed for the global ECMWF Integrated Forecasting System. Implications of the linearity constraint for any parametrization scheme, such as the need for simplification and regularization, are discussed. The description of the methodology to develop linearized parameterizations highlights the complexity of obtaining a physics package that can be efficiently used in practical applications. The impact of the different physical processes on the tangent-linear approximation and adjoint sensitivities, as well as their performance in data assimilation are demonstrated.

11.1 Introduction

Adjoint models have several applications in numerical weather prediction (NWP). In variational data assimilation (DA) for instance, they are used to efficiently determine optimal initial conditions. Another application of the adjoint technique is the computation of the fastest growing modes (i.e. singular vectors) over a finite time interval, which can be used in Ensemble Prediction Systems (EPS). Adjoint models can also be used for sensitivity studies since they enable the computation of the gradient of a selected output parameter from a numerical model with respect to all its input parameters. In practice, this is often used to obtain the sensitivity of the analysis to model parameters, sensitivities of one aspect of the forecast to initial conditions or sensitivities of the analysis to observations.

M. Janisková (✉) · P. Lopez
European Centre for Medium–Range Weather Forecasts, Shinfield Park, Reading, Berks, RG2 9AX, UK
e-mail: Marta.Janiskova@ecmwf.int; philippe.lopez@ecmwf.int

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_11,

Initially, only the adiabatic linearized models were used in NWP. However, the significant role played by physical processes in various large-scale and mesoscale phenomena was soon recognized. Physical processes are particularly important in the tropics, near the surface, in the planetary boundary layer or the stratosphere, where the description of the atmospheric processes is controlled by both physics and dynamics. Therefore a lot of effort was devoted to include physical parameterizations in adjoint models. Several studies aimed at including physical parameterizations in adjoint models (Zou et al. 1993; Zupanski and Mesinger 1995; Tsuyuki 1996; Errico and Reader 1999; Janisková et al. 1999; Mahfouf 1999; Janisková et al. 2002; Laroche et al. 2002; Lopez 2002; Tompkins and Janisková 2004; Lopez and Moreau 2005; Mahfouf 2005) with encouraging results. However, these studies also showed that the linearization of physical parametrization schemes is not straightforward because of the non-linear and on/off nature of physical processes. Strong non-linearities that could lead to noise problems had to be removed from the models in order to be able to benefit from the inclusion of physical processes in the linearized model.

In recent years, four-dimensional variational (4D-Var) data assimilation became a powerful tool for exploiting information from irregularly distributed observations for initial conditions of a numerical forecast model. 4D-Var minimizes the distance between a model trajectory and observations spread over a given time interval, using the adjoint equations of the model to compute the gradient of the cost function with respect to the model state at the beginning of the assimilation period. The mismatch between model solution and observations can remain large if the imperfect adiabatic adjoint model would only be used in the minimization. Many satellite observations, such as radiances, rainfall and cloud measurements, cannot be directly assimilated with such overly simple adjoint models. Therefore it is crucial to represent physical processes in the assimilating models. Parametrization schemes for adjoint models started from very simple ones, such as Buizza (1994), which aimed at removing very strong increments produced by the adiabatic adjoint models. More complex, but still incomplete schemes were developed by Zou et al. (1993), Zupanski and Mesinger (1995), Janisková et al. (1999), Mahfouf (1999, 2005), and Laroche et al. (2002). More recently, comprehensive schemes were implemented, which describe the whole set of physical processes and interactions between them almost as in the non-linear model, just slightly simplified and/or regularized (e.g. Janisková et al. 2002; Tompkins and Janisková 2004; Lopez and Moreau 2005).

In this paper, a comprehensive set of physical parameterizations developed for the linearized version of the global ECMWF model is described together with its applications in sensitivity studies and data assimilation. A description of the current package, which is unique because of its complexity, has never been published in the literature. Readers would only be able to find summaries of old parametrization schemes (Mahfouf 1999) from which hardly anything is left in the current operational model. Some information about updated versions of the schemes for shortwave radiation (Janisková et al. 2002) and moist processes (Tompkins and Janisková 2004; Lopez and Moreau 2005) is available, but is no longer up-to-date. In Sect. 11.2, the reasons for using physics in variational data assimilation are explained. The implications of the linear constraint for any parametrization schemes, such as simplification and regularization are described in Sect. 11.3. The methodology for the development of linearized simplified parameterizations is provided in Sect. 11.4. To fully appreciate the achieved level of sophistication of the linearized physical parametrization schemes used at ECMWF, which can still be integrated even over 48 h on the global scale without producing spurious noise, each of them is described in Sect. 11.5. The impact of different physical processes on the tangent-linear approximation, adjoint sensitivity, as well as the performance in data assimilation are demonstrated in Sect. 11.6. Finally, conclusions and perspectives are given in Sect. 11.7.

11.2 The Need for Physics in Variational Data Assimilation

Two main reasons can justify the need for linearized physical parameterizations in variational data assimilation.

The first one lies in the necessity to compute model-observation departures at a given time, so that the variational cost function can be minimized. For instance, if satellite microwave brightness temperatures are to be assimilated, one must be able to translate the model control variables (typically temperature, humidity, wind and surface pressure) into some equivalent simulated brightness temperatures. In this example, this can be achieved by applying moist physics parameterizations to simulate cloud and precipitation fields first, and then a radiative transfer model to obtain the desired microwave brightness temperatures, as seen by the model. The goal of data assimilation is to define the atmospheric state such that the mismatch between the model and observations (or cost function, J) is minimum. To minimize the cost function for obtaining the optimal increments in each model state vector component, its gradient with respect to model variables needs to be assessed. In the chosen example of microwave brightness temperatures, this would be achieved by applying the adjoint of the radiative transfer model followed by the adjoint of the moist physical parameterizations to the gradient of J in observation space. The adjoint of a given operator is simply the transpose of its Jacobian matrix with respect to its input variables.

Secondly, in the particular context of 4D-Var data assimilation, the model state needs to be compared to each available observation at the time the latter was performed. It is therefore necessary to evolve the model state from the beginning of the 4D-Var assimilation window (time 0) to the time of the observation (time i). This is achieved by integrating the full non-linear (NL) forecast model, M, from time 0 to i. Again, the minimization of the 4D-Var cost function, J, which measures the total distance between the model and all observations available throughout the assimilation window, requires the computation of its gradient, $\nabla_{\mathbf{x}(t_0)} J$ with respect to the model state at the beginning of the 4D-Var assimilation window, $\mathbf{x}(t_0)$. To achieve this, the gradient of the observation term, J_o, of the cost function in observation space can be first computed through simple differentiation as

$$\nabla_{\tilde{\mathbf{y}}_i} J_o = \sum_{i=0}^{n} \mathbf{R}_i^{-1}(\tilde{\mathbf{y}}_i - \mathbf{y}_i^o) \tag{11.1}$$

where $\tilde{\mathbf{y}}_i = H(\mathbf{x}_i)$ is the model observed equivalent, $\mathbf{y}_i^o$ is the vector of available observations and $\mathbf{R}_i$ is the observation error covariance matrix. Using the adjoint of the observation operator, $\mathbf{H}_i^T$, one can then calculate the gradient of J_o with respect to the model state at observation time, $\mathbf{x}(t_i)$,

$$\nabla_{\mathbf{x}_i} J_o = \sum_{i=0}^{n} \mathbf{H}_i^T \mathbf{R}_i^{-1}(H_i[\mathbf{x}(t_i)] - \mathbf{y}_i^o) \tag{11.2}$$

Finally, the gradient of J_o with respect to the model state at time 0 can be obtained by applying the adjoint (AD) of the forecast model, $\mathbf{M}^T(t_i, t_0)$,

$$\nabla_{\mathbf{x}(t_0)} J_o = \sum_{i=0}^{n} \mathbf{M}^T(t_i, t_0)\mathbf{H}_i^T \mathbf{R}_i^{-1}(H_i[\mathbf{x}(t_i)] - \mathbf{y}_i^o) \tag{11.3}$$

Again, since the adjoint version of the forecast model can be seen as the transpose of its Jacobian matrix, the forecast model first needs to be differentiated with respect to its inputs, yielding the so-called tangent-linear (TL) model, $\mathbf{M}$.

In contrast with the full non-linear model, the tangent-linear model works on perturbations of the input variables rather than on full model fields and is fully linear by construction. The adjoint is therefore a fully linear operator as well and, in the case of 4D-Var, its inputs are the components of $\nabla_{\mathbf{x}(t_i)} J$. As a consequence, solving the 4D-Var minimization requires the linearization of the forecast model's physical parameterizations (e.g. vertical diffusion, radiation, convection, large-scale moist processes) so that their TL and AD versions can be used to describe the (forward, respectively backward) time evolution of the model state during the minimization as seen from (11.3).

11.3 Implication of the Linearity Constraint

The minimization of the 4D-Var cost function is solved with an iterative algorithm and is therefore computationally rather demanding. Even though the minimization is usually performed at a much lower resolution (T159/T255[1] in current ECMWF's operations) than in the standard forecast model (T1279[2] at ECMWF), the several tens of iterations required to obtain the optimal model state means that the linearized

[1] T159/T255 corresponding approximately to 130/80 km

[2] T1279 corresponding approximately to 16 km

physics package must be as cheap as possible. To reduce computational cost, it is therefore often necessary to simplify the set of linearized parameterizations by retaining only physical processes that dominate in the full forecast model. Linearity considerations can also influence this choice: if a given process is known to be highly non-linear (e.g. thresholds, switches), this process should be discarded from the linearized code since this might otherwise lead to instabilities during TL and AD integrations. However, some of those instabilities can be overcome through adequate modifications of the code. At the same time, though simplified, parametrization schemes used in the linearized model must remain realistic enough to keep the description of atmospheric processes physically sound.

11.3.1 Simplification

For important practical applications (incremental approach of 4D-Var – Courtier et al. 1994, adjoint based sensitivities, initial perturbations of EPS), the linearized version of the forecast model is run at a lower resolution than the non-linear model. In this case, since the dynamics is already simplified through the reduction in horizontal resolution, the linearized physics does not necessarily need to be exactly tangent to the full physics. In principle, physical parameterizations can already behave differently between non-linear and tangent-linear models due to the change in resolution. Consequently, some freedom exists in the development of a simplified physics package, as long as the parameterizations can represent general feedbacks of physical phenomena present in the atmosphere. Simplified approaches can allow the progressive inclusion of physical processes in the tangent-linear and adjoint models. This strategy has been used, for instance, in the operational 4D-Var systems of ECMWF (Mahfouf 1999; Mahfouf and Rabier 2000; Rabier et al. 2000; Janisková et al. 2002; Janisková 2003; Tompkins and Janisková 2004; Lopez and Moreau 2005) and at Météo-France (Janisková et al. 1999; Geleyn et al. 2001).

11.3.2 Regularization

As already mentioned, physical processes are often characterized by thresholds. These can be:

- Discontinuities of some functions themselves describing the physical processes or some on/off processes (for instance produced by saturation, changes between liquid and solid phase);
- Some discontinuities in the derivative of a continuous function (i.e. the derivative can go towards infinity at some points);
- Some strong non-linearities (such as those created by the transition from unstable to stable regimes in the planetary boundary layer).

In each of these situations, an estimation of the derivative close to the discontinuity point will be different between the non-linear model (in terms of finite differences) and the TL model. All of this makes the tangent linear approximation less valid when the linearized model includes physical parameterizations compared to the adiabatic version only. To treat the described problems, it is important to regularize, i.e. to smooth the parameterized discontinuities in order to make the scheme as much differentiable as possible. One should recognize that it is often quite difficult to achieve a tradeoff between a physically sound description of atmospheric processes and a well-behaved linear physical parametrization. However, without a proper treatment of the most significant thresholds, the TL model can quickly become too inaccurate to be useful. Therefore a lot of effort was devoted by a number of investigators to deal with discontinuities present in parameterized physical processes (e.g. Zou et al. 1993; Zupanski and Mesinger 1995; Tsuyuki 1996; Errico and Reader 1999; Janisková et al. 1999; Mahfouf 1999; Laroche et al. 2002; Tompkins and Janisková 2004; Lopez and Moreau 2005).

To illustrate a potential source of problem in the linearized model, the rain production function, describing which portion of the cloud water is converted into precipitation, is shown in Fig. 11.1. An increase of cloud water mixing ratio by a small amount dx (Fig. 11.1a) leads to a small change in the precipitation amount dy_{NL} in the case of the non-linear (NL) model, but to a much larger change (dy_{TL}) in the case of the TL model. As a possible solution, one can modify the function to make it less steep (dotted line on Fig. 11.1b). In this case, the resulting TL increment will be significantly smaller (dy_{TL_2}). However, the required modification can be substantial and it can deteriorate the overall quality of the physical parametrization itself. Therefore one must always be careful to keep the right balance between linearity and realism of the parametrization schemes. In the future, the better the non-linear forecast model will become, the smaller 4D-Var analysis increments and hence the hope to have less difficulties with using linearized physical processes will be.

11.4 Methodology for the Development of Linearized Simplified Parameterizations

There are several problems with including physics in adjoint models. The development requires substantial resources and it is technically very demanding. The validation must be very thorough and it must be done for the non-linear, tangent-linear and adjoint versions of the physical parametrization schemes. The computational cost of the model with physical processes can be very high despite some possible simplifications applied. One must be also very careful with the non-linear and threshold nature of physical processes which can affect the range of validity of the tangent-linear approximation as mentioned above.

The development of a new linearized physical parametrization can be divided into four main stages:

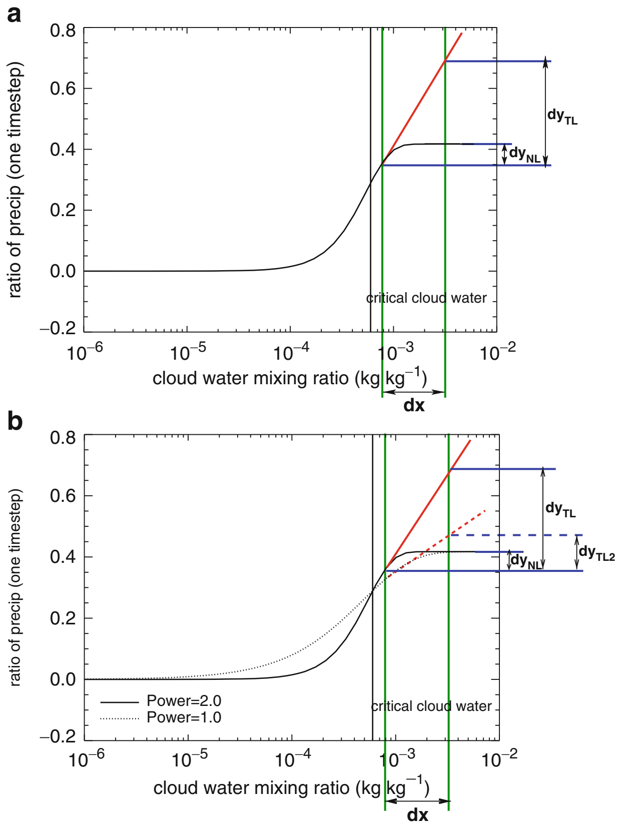


Fig. 11.1 Autoconversion function of cloud water into precipitation (*black solid line*) based on Sundqvist et al. (1989). A change in the cloud water, dx, results in a change of precipitation, dy_{NL}, in the case of non-linear (NL) model. dy_{TL} is the corresponding change in precipitation given by the tangent-linear (TL) model. (**b**) Describes the modified function which is less steep and helps to reduce the TL increments to dy_{TL_2} (closer to dy_{NL})

1. Simplified non-linear forward model design, coding, tuning and validation.
2. Tangent-linear coding and testing.
3. Adjoint coding and testing.
4. Performance assessment in data assimilation and other applications (see examples in Sect. 11.6).

11.4.1 Simplified Non-linear Version

In the first stage, the non-linear version of the new simplified physical parametrization needs to be designed. This can be achieved through either an "upward" or "downward" approach, both relying on the prioritization of the various processes represented in the full NL code used in standard forecasts. With the upward technique, the simplified code is obtained by keeping only the most relevant processes found in the full NL version. In the downward approach, the simplified code is built by ignoring the least significant processes from the full NL code. Ideally, both approaches should converge to more or less similar simplified codes, which should be computationally cheaper than the full NL code, and contain fewer discontinuities but are still able to provide realistic forecasts. Once the simplified code has been written, it is thus necessary to tune and validate it in traditional forecasts over periods at least equal to the maximum length of the expected applications. At ECMWF for instance, this period corresponds to 12 h for 4D-Var DA or to 24 h for singular vector computations involved in the ensemble prediction system. It is particularly crucial to ensure that the new simplified NL code does not depart too much from its full NL counterpart over this period of time. Verification in much longer integrations (up to climate timescales), although not essential, is also recommended to make sure that the new simplified scheme is stable and behaves reasonably well.

11.4.2 Linearization Techniques

Once the NL version of the simplified scheme is deemed adequate, efforts are devoted to the development of the TL code, first, and then of the AD code. In practice, linearization can be achieved using either a manual line-by-line approach or an automatic coding software (e.g. Giering and Kaminski 1998; Araya-Polo and Hascoët 2004). However attractive automatic coding may sound, the manual technique is usually more suitable as soon as one has to deal with the large amounts of complex code used in modern NWP systems. Until now, in our own experience, the code produced through automatic differentiation and adjoining often turned out to be computationally very expensive (no optimization) and sometimes not bug-free. This is the reason why so far only manual line-by-line TL and AD coding has been applied to derive and update ECMWF's full set of linearized physical parameterizations. In the future this strategy might be revisited if automatic softwares become more efficient and reliable.

11.4.3 Tangent-Linear Version

An estimation of sensitivity of model output with respect to input required by many studies can be efficiently done by using the adjoint. For atmospheric models evolving in time, this backward integration requires to have the tangent linear model acting forward in time. To build the TL model, the linearization is performed with respect to the local tangent of the model trajectory.

If M is the model describing the time evolution of the model state $\mathbf{x}$ as:

$$\mathbf{x}(t_{i+1}) = M[\mathbf{x}(t_i)] \tag{11.4}$$

then the time evolution of a small perturbation $\delta\mathbf{x}$ can be estimated to the first order approximating by the tangent linear model $\mathbf{M}$ (derived from the NL model M):

$$\begin{aligned} \delta\mathbf{x}(t_{i+1}) &= \mathbf{M}[\mathbf{x}(t_i)]\delta\mathbf{x}(t_i) \\ \delta\mathbf{x}(t_{i+1}) &= \frac{\partial M[\mathbf{x}(t_i)]}{\partial\mathbf{x}}\delta\mathbf{x}(t_i) \end{aligned} \tag{11.5}$$

The verification of the correctness of the TL model is first performed through the classical Taylor formula:

$$\lim_{\lambda\to 0} \frac{M(\mathbf{x}+\lambda\delta\mathbf{x}) - M(\mathbf{x})}{\mathbf{M}(\lambda\delta\mathbf{x})} = 1 \tag{11.6}$$

This examination of asymptotic behaviour, using perturbations the size of which becomes infinitesimally small, is performed to check the numerical correctness of the TL code.

For practical applications, it is also important to investigate the accuracy of TL models for finite-amplitude perturbations (typically perturbations of the size of analysis increments). The results from applications of tangent-linear and adjoint models are only useful when the linearized approximation is valid for such perturbations. Therefore, for the validation of the tangent-linear approximation, the accuracy of the linearization of a parametrization scheme is studied with respect to pairs of non-linear results. The difference between two non-linear integrations (one starting from a background field, $\mathbf{x}^b$, and the other from an analysis, $\mathbf{x}^a$) run with the full NL model, M, is compared to time evolution of the analysis increments ($\mathbf{x}^a - \mathbf{x}^b$) obtained by integrating the TL model, $\mathbf{M}$, with the trajectory taken from the background field.

For a quantitative evaluation of the impact of linearized schemes, their relative importance is evaluated using mean absolute errors between tangent-linear and non-linear perturbations as:

$$\varepsilon = \overline{\left|\mathbf{M}(\mathbf{x}^a - \mathbf{x}^b) - \left[M(\mathbf{x}^a) - M(\mathbf{x}^b)\right]\right|} \tag{11.7}$$

As a reference for the comparisons, an absolute mean error for the TL model without physics, ε_{ref}, is taken. If ε_{exp} is defined as the absolute mean error of the TL model with the different physical schemes included, then an improvement coming from the inclusion of more physics in the TL model is expressed as $\varepsilon_{exp} < \varepsilon_{ref}$. The relative errors, r_{er}, and relative improvements, η, are also computed as:

$$r_{er} = \frac{\left|\mathbf{M}(\mathbf{x}^a - \mathbf{x}^b) - \left[M(\mathbf{x}^a) - M(\mathbf{x}^b)\right]\right|}{\left|M(\mathbf{x}^a) - M(\mathbf{x}^b)\right|} \tag{11.8}$$

$$\eta = \frac{\varepsilon_{exp} - \varepsilon_{ref}}{\varepsilon_{ref}} \tag{11.9}$$

Validity tests of the tangent-linear approximations are mostly performed over the time period and at the resolution at which adjoint models will be applied in practice: resolution and time length of an inner-loop integration of 4D-Var system (e.g. 12 h, T255 and 91 vertical levels at ECMWF) or longer time periods for singular vectors and sensitivity applications (e.g. 24 h at ECMWF). An example of the results from such TL approximation assessment will be given in Sect. 11.6.1.

11.4.4 Adjoint Version

The adjoint of a linearized operator, $\mathbf{M}$, is the linear operator, $\mathbf{M}^*$, such that:

$$\forall \mathbf{x}, \forall \mathbf{y} \qquad < \mathbf{M}.\mathbf{x}, \mathbf{y} > = < \mathbf{x}, \mathbf{M}^*.\mathbf{y} > \tag{11.10}$$

where $<,>$ denotes the inner product and $\mathbf{x}$ and $\mathbf{y}$ are input vectors.

Besides, the adjoint model $\mathbf{M}^*$ can provide the gradient of any objective function, J, with respect to $\mathbf{x}(t_i)$ from the gradient of the objective function with respect to $\mathbf{x}(t_{i+1})$

$$\frac{\partial J}{\partial \mathbf{x}(t_i)} = \mathbf{M}^* \left(\frac{\partial J}{\partial \mathbf{x}(t_{i+1})} \right) \tag{11.11}$$

The integration of the AD forecast model works backward in time. One should remember that, M being non-linear, $\mathbf{M}$ as well as $\mathbf{M}^*$ depend on the particular state of the atmosphere, $\mathbf{x}$, about which the linearization is performed. The adjoint operator, for the simplest canonical scalar product $<,>$ (11.10), is actually the transpose of the tangent linear operator, $\mathbf{M}^T$ (not its inverse).

For the practical verification of the adjoint code, one must test the identity described in (11.10). It should be emphasized that it is absolutely essential to ensure that the TL and AD codes verify (11.10) to the level of machine precision, even when vectors $\mathbf{x}$ and $\mathbf{y}$ are global 3D atmospheric states and even for time integrations up to 12 or 24 h. Note that a correct adjoint test does not imply the correctness of tangent-linear code.

11.4.5 Singular Vectors

Besides the verification of the numerical correctness of TL and AD versions of the model and the examination of the validity of TL approximation, singular vectors can be used to find out whether the new schemes do not lead to a growth of spurious unstable modes. Such modes would indicate the existence of strong non-linearities and threshold processes in the model and would have a negative impact on the usefulness of the linearized model.

11.5 ECMWF's Linearized Physics Package

11.5.1 Description

The set of ECMWF physical parameterizations used in the linearized model (called simplified or linearized parameterizations) comprises six different schemes: radiation, vertical diffusion, orographic gravity wave drag, moist convection, large-scale condensation/precipitation and non-orographic gravity wave activity, sequentially called in this order. The current linearized physics package is therefore quite sophisticated and is believed to be the most comprehensive one used in operational global data assimilation. Each physical parametrization scheme of this package is described below starting with dry processes.

11.5.1.1 Radiation

The radiation scheme solves the radiative transfer equation in two distinct spectral regions. The computations for the longwave (LW) radiation are performed over the spectrum from 0 to 2,820 cm^{-1} ($\sim$100 to 3.5 μm). The shortwave (SW) part of the scheme integrates the fluxes over the whole shortwave spectrum between 0.2 and 4.0 μm. The scheme used for data assimilation purposes must be computationally efficient to be called at full spatial resolution to improve the description of cloud-radiation interactions during the assimilation period (Janisková et al. 2002).

The Shortwave Radiation Scheme

The linearized code for the SW radiation scheme has been derived from ECMWF's original non-linear scheme developed by Fouquart and Bonnel (1980) and revised by Morcrette (1991). In this scheme, previously used in the operational forecast model, the photon-path-distribution method is applied to separate the parametrization of scattering processes from that of molecular absorption. Upward $F_{\text{sw}}^{\uparrow}$ and downward

$F^{\downarrow}_{\mathrm{sw}}$ fluxes at a given level j are obtained from the reflectance and transmittance of the atmospheric layers as

$$F^{\downarrow}_{\mathrm{sw}}(j) = F_0 \prod_{k=j}^{N} T_{\mathrm{bot}}(k) \tag{11.12}$$

$$F^{\uparrow}_{\mathrm{sw}}(j) = F^{\downarrow}_{\mathrm{sw}}(j) R_{\mathrm{top}}(j-1) \tag{11.13}$$

Computations of the transmittance at the bottom of a layer, T_{bot}, start at the top of atmosphere and work downwards. Those of the reflectance at the top of the same layer, R_{top}, start at the surface and work upwards. In the presence of cloud in the layer, the final fluxes are computed as a weighted average of the fluxes in the clear-sky and in the cloudy fractions of the column (depending on the cloud-overlap assumption).

The non-linear scheme is reasonably fast for application in 4D-Var and has therefore been linearized without a-priori changes (Janisková et al. 2002). The only modification with respect to the non-linear version (used operationaly until June 2007; since then Rapid Radiation Transfer model for SW radiation is used – Mlawer and Clough 1997), is the use of two spectral intervals, instead of six intervals. This is meant to reduce the computational cost.

The Longwave Radiation Scheme

The LW radiation scheme, used in the ECMWF full NL forecast model is the Rapid Radiation Transfer Model (RRTM; Mlawer et al. 1997; Morcrette et al. 2001). The complexity of the RRTM scheme for the LW part of the spectrum makes accurate computations expensive. In the variational assimilation framework, the older operational scheme of Morcrette (1989) was linearized. In this scheme, the LW spectrum from 0 to 2,820 cm^{-1} is divided into six spectral regions. The transmission functions for water vapour and carbon dioxide over those spectral intervals are fitted using Padé approximations on narrow-band transmissions obtained with statistical band models (Morcrette et al. 1986). Integration of the radiation transfer equation over wavenumber ν within the particular spectral regions yields the upward and downward fluxes.

The inclusion of cloud effects on the LW fluxes follows the treatment discussed by Washington and Williamson (1997). The scheme first calculates upward and downward fluxes ($F_0^{\uparrow}(i)$ and $F_0^{\downarrow}(i)$) for a clear-sky atmosphere. In any cloudy layer, the scheme evaluates the fluxes assuming a unique overcast cloud of emissivity unity, i.e. $F_n^{\uparrow}(i)$ and $F_n^{\downarrow}(i)$ for a cloud present in the nth layer of the atmosphere. The fluxes for the actual atmosphere are derived from a linear combination of the fluxes calculated in the previous steps with some cloud overlap assumption (see below) in the case of clouds spreading over several layers. If N is the number of model layers starting from the top of atmosphere to the bottom, C_i the fractional cloud cover in layer i, the cloudy upward $F_{lw}^{\uparrow}$ and downward $F_{lw}^{\downarrow}$ fluxes are then expressed as:

$$F_{lw}^{\uparrow}(i) = (1 - CC_{N,i})F_0^{\uparrow}(i) + \sum_{k=i}^{N}(CC_{i,k+1} - CC_{i,k})F_k^{\uparrow}(i) \tag{11.14}$$

$$F_{lw}^{\downarrow}(i) = (1 - CC_{i-1,0})F_0^{\downarrow}(i) + \sum_{k=1}^{i-1}(CC_{i,k+1} - CC_{i,k})F_k^{\downarrow}(i) \tag{11.15}$$

where $CC_{i,j}$ is the cloudiness encountered between any two levels i and j in the atmosphere computed using the overlap assumption described below.

In the case of semi-transparent clouds, the fractional cloudiness entering the calculations is an effective cloud cover equal to the product of the emissivity due to condensed water and gases in the layer by the horizontal coverage of the cloud cover. This is the so called effective emissivity approach.

To reduce a computational cost of the linearized LW radiation for data assimilation, the scheme is not called at each time step. Furthermore, the transmission functions are only computed for H_2O and CO_2 absorbers (though the version taking into account the whole spectrum of absorbers is also coded for aerosols and other gases). The cloud effects on LW radiation are only computed up to cloud top.

Cloud Overlap Assumptions

Cloud overlap assumptions must be made in atmospheric models in order to organize the cloud distribution used for radiation. This is is necessary to account for the fact that clouds often do not fill the whole grid box. The maximum-random overlap assumption (originally introduced in Geleyn and Hollingsworth 1997) is used operationally (Morcrette and Jakob 2000). Adjacent cloudy layers are combined by assuming maximum overlap to form a contiguous cloud and discrete layers separated by clear-sky are combined randomly.

Cloud Optical Properties

When one considers cloud-radiation interactions, it is not only the cloud fraction or cloud volume, but also cloud optical properties that matter. In the case of SW radiation, cloud radiative calculations depend on three different parameters: the optical thickness, the asymmetry factor and the single scattering albedo. They are derived from Fouquart (1987) for water clouds, and Ebert and Curry (1992) for ice clouds. They are functions of cloud condensate and a specified effective radius.

Cloud LW optical properties are represented by the emissivity, related to the condensed water amount, and by the condensed-water mass absorption coefficient obtained from Smith and Shi (1992) for water clouds and Ebert and Curry (1992) for ice clouds.

11.5.1.2 Vertical Diffusion

Vertical diffusion applies to wind components, dry static energy and specific humidity. The exchange coefficients in the planetary boundary layer and the drag coefficients in the surface layer are expressed as functions of the local Richardson number, Ri, (Louis et al. 1982). They differ from the formulation of the operational forecast model (i.e. full non-linear scheme) where for the unstable regime ($Ri < 0$) the Monin-Obukhov (M-O) formulation is used, together with a K-profile approach for convectively mixed layer in the case of unstable surface conditions. In the linearized model, the exchange coefficients are computed according to Louis et al. (1982). For the stable regime ($Ri > 0$), diffusion coefficients according to the Louis scheme are used close to the surface and above 300 m, then they tend asymptotically to the M-O formulation. A mixed layer parametrization is also included. This is consistent with the full non-linear model.

Analytical expressions are generalized for the situation with different roughness lengths for heat and momentum transfer. For any conservative variable ψ (wind vector components, u and v; dry static energy, s; specific humidity, q), the tendency produced by vertical diffusion is

$$\frac{\partial \psi}{\partial t} = \frac{1}{\rho}\frac{\partial}{\partial z}\left(K(Ri)\frac{\partial \psi}{\partial z}\right) \tag{11.16}$$

where ρ is the air density. The exchange coefficient K for heat and momentum transfer is given by

$$K = l^2 \left\| \frac{\partial \mathbf{U}}{\partial z} \right\| f(Ri) \tag{11.17}$$

where $\mathbf{U}$ is the wind vector and $f(Ri)$ represents the coefficient accounting for the dependence of vertical turbulent diffusion on the local Richardson number, either computed according to Louis et al. (1982), $f_L(Ri)$, or to the Monin-Obukhov formulation, $f_{MO}(Ri)$. l is the mixing length profile based on the formulation of Blackadar (1962) with a reduction in the free atmosphere.

A continuous transition between Louis coefficients near the surface to about 300 m and M-O coefficients above is computed as

$$\frac{1}{l\sqrt{f(Ri)}} = \frac{1}{\kappa z\sqrt{f_L(Ri)}} + \frac{1}{\lambda\sqrt{f_{MO}(Ri)}} \tag{11.18}$$

where κ is the Von Karman's constant, z is the height and λ is the asymptotic mixing length.

To parameterize turbulent fluxes at the surface, the drag coefficient, C_{sf}, (i.e. the exchange coefficient between the surface and the lowest model level) is computed as

$$C_{sf} = g_{sf}(Ri)\, C_{\mathrm{N}} \tag{11.19}$$

where C_{N} is the neutral drag coefficient, which is a function of the roughness length, and $g_{sf}(Ri)$ is a function of the local Richardson number. Different formulations of C_{N} and $g_{sf}(Ri)$ are used for momentum and heat, according to Louis et al. (1982).

Regularization

In earlier versions of the model, perturbations of the exchange coefficients were simply neglected ($K' = 0$), in order to prevent spurious unstable perturbations from growing in the linearized version of the scheme (Mahfouf 1999). Later, some regularization of exchange coefficients was introduced at upper model levels to allow partial perturbations of these coefficients. This consists in the perturbations being more significantly reduced around the neutral state (i.e. Ri close to zero) where both the function of Ri itself and its derivative exhibit a significant rate of change. The reduction is eased exponentially away from the neutral state.

11.5.1.3 Subgrid Scale Orographic Effects

Only the low-level blocking part of the operational non-linear scheme developed by Lott and Miller (1997) is taken into account in TL and AD calculations. The deflection of the low-level flow around orographic obstacles is supposed to occur below an altitude Z_{blk} such that

$$\int_{Z_{\mathrm{blk}}}^{3\mu} \frac{N}{|\mathbf{U}|}\,\mathrm{d}z \geq H_{n_{\mathrm{crit}}} \tag{11.20}$$

where $H_{n_{\mathrm{crit}}}$ is a critical non-dimensional mountain height, μ is the standard deviation of subgrid-scale orography and N is the Brunt-Väisälä frequency.

The deceleration of wind due to low-level blocking is given by

$$\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{\mathrm{blk}} = -C_{\mathrm{d}} \max\left(2 - \frac{1}{r}, 0\right) \frac{\sigma}{2\mu} \sqrt{\frac{Z_{\mathrm{blk}} - z}{z + \mu}} (B \cos^2 \alpha + C \sin^2 \alpha) \frac{\mathbf{U}|\mathbf{U}|}{2} \tag{11.21}$$

where C_{d} is the low-level drag coefficient, σ is the mean slope of the subgrid-scale orography, and α is the angle between the low-level wind and the principal axis of orography. r is determined as $r = (\cos^2 \alpha + \gamma \sin^2 \alpha)/(\gamma \cos^2 \alpha + \sin^2 \alpha)$, where γ is the anisotropy of the subgrid-scale orography. The functions B, C are written as (Phillips 1984)

$$B = 1 - 0.18\gamma - 0.04\gamma^2 \qquad \text{and} \qquad C = 0.48\gamma + 0.3\gamma^2.$$

The final wind tendency produced by the low-level drag parametrization is then obtained from the following partially implicit discretization of (11.21)

$$\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{\text{blk}} = \frac{\mathbf{U}^{n+1} - \mathbf{U}^{n}}{\Delta t} = -A|\mathbf{U}^{n}|\mathbf{U}^{n+1} = -\frac{\beta}{1+\beta}\frac{\mathbf{U}^{n}}{\Delta t} \tag{11.22}$$

where $\beta = A|\mathbf{U}^n|\Delta t$ and $\mathbf{U}^{n+1} = \mathbf{U}^n/(1+\beta)$.

11.5.1.4 Non-orographic Gravity Wave Drag

The tangent-linear and adjoint versions of the non-linear scheme for non-orographic gravity waves (in details described by Orr et al. 2010) were developed in order to reduce discrepancies between the full NL and linearized versions of the model, especially in the stratosphere. The parametrization scheme used in the NL model is based on Scinocca (2003). This is a spectral scheme that follows from the Warner and McIntyre (1996) scheme representing the three basic mechanisms that are conservative propagation, critical level filtering, and non-linear dissipation. Since the full nonhydrostatic and rotational wave dynamics considered by Warner and McIntyre (1996) is too costly for operational models, only hydrostatic and non-rotational wave dynamics are employed.

The dispersion relation for a hydrostatic gravity wave in the absence of rotation is

$$m^2 = \frac{k^2 N^2}{\tilde{\omega}^2} = \frac{N^2}{\tilde{c}^2} \tag{11.23}$$

where k, m are horizontal and vertical wavenumbers, while $\tilde{\omega} = \omega - kU$ and $\tilde{c} = c - U$ are the intrinsic frequency and phase speed (with c being the ground based phase speed and U the background wind speed in the direction of propagation), respectively.

The launch spectrum, which is globally uniform and constant, is given by the total wave energy per unit mass in each azimuth angle ϕ following Fritts and VanZandt (1993) as

$$\tilde{E}(m,\tilde{\omega},\phi) = B\left(\frac{m}{m_*}\right)^s \frac{N^2\tilde{\omega}^{-d}}{1-\left(\frac{m}{m_*}\right)^{s+3}} \tag{11.24}$$

where B, s and d are constants, and $m_* = 2\pi L$ is a transitional wavelength. Instead of the total wave energy $\tilde{E}(m,\tilde{\omega},\phi)$, the momentum flux spectral density $\rho\tilde{F}(m,\widetilde{\omega,\phi})$ is required. This is obtained through the group velocity rule. In order to have the momentum flux conserved in the absence of dissipative processes as the spectrum propagates vertically through height-varying background wind and buoyancy frequency, the coordinate framework (k,ω) is used instead of $(m,\tilde{\omega})$ as shown in Scinocca (2003).

The dissipative mechanisms applied to the wave field in each azimuthal direction and on each model level are critical level filtering and non-linear dissipation. After application of them, the resulting momentum flux profiles are used to derive the net eastward $\rho\bar{F}_E$ and northward $\rho\bar{F}_N$ fluxes. The tendencies for the $(u,\, v)$ wind components are then given by the divergence of those fluxes obtained through summation of the total momentum flux (i.e. integrated over all phase speed bins) in each azimuth ϕ_i projected onto the east and north directions, respectively:

$$\frac{\partial u}{\partial t} = g\frac{\partial(\rho \bar{F}_E)}{\partial p} \tag{11.25}$$

$$\frac{\partial v}{\partial t} = g\frac{\partial(\rho \bar{F}_N)}{\partial p} \tag{11.26}$$

where g is the gravitational acceleration and p is pressure.

Regularization

In order to eliminate the spurious noise in TL computations caused by the introduction of this scheme, it was necessary to implement some regularizations. These include rewriting buoyancy frequency (N) computations in the NL scheme in height coordinates instead of pressure coordinates (as used in the original code) and setting increments for momentum flux in the last three spectral elements (high phase speed) of the launch spectrum to zero.

11.5.1.5 Moist Convection

The original version of the simplified mass-flux convection scheme currently used in the minimization of 4D-Var is described in Lopez and Moreau (2005). Through time, the original scheme has been updated so as to gradually converge towards the full convection scheme used in high-resolution 10-day forecasts (Bechtold et al. 2008). The transport of tracers by convection is also implemented.

The physical tendencies produced by convection on any conservative variable ψ (dry static energy, wind components, specific humidity, cloud liquid water) can be written in mass-flux form following Betts (1997)

$$\frac{\partial \psi}{\partial t} = \frac{1}{\rho}\left[(M_{\mathrm{u}} + M_{\mathrm{d}})\frac{\partial \psi}{\partial z} + D_{\mathrm{u}}(\psi_{\mathrm{u}} - \psi) + D_{\mathrm{d}}(\psi_{\mathrm{d}} - \psi)\right] \tag{11.27}$$

The first term on the right hand side represents the compensating subsidence induced by cumulus convection on the environment through the mass flux, M. The other terms account for the detrainment of cloud properties in the environment with a detrainment rate, D. Subscripts u and d refer to the updraughts and downdraughts properties, respectively. Evaporation of cloud water and precipitation should also be added in (11.27) for dry static energy, $s = c_pT + gz$, and specific humidity.

Equations for Updraught and Downdraught

The equations describing the evolution with height of the convective updraught and downdraught mass fluxes, M_{u} and M_{d}, respectively, are

$$\frac{\partial M_{\mathrm{u}}}{\partial z} = (\epsilon_{\mathrm{u}} - \delta_{\mathrm{u}})M_{\mathrm{u}} \tag{11.28}$$

$$\frac{\partial M_{\mathrm{d}}}{\partial z} = -(\epsilon_{\mathrm{d}} - \delta_{\mathrm{d}})M_{\mathrm{d}} \tag{11.29}$$

where ϵ and δ respectively denote the entrainment and detrainment rates. A second set of equations is used to describe the evolution with height of any other characteristic, ψ, of the updraught or downdraught, namely

$$\frac{\partial \psi_{\mathrm{u}}}{\partial z} = -\epsilon_{\mathrm{u}}(\psi_{\mathrm{u}} - \overline{\psi}) \tag{11.30}$$

$$\frac{\partial \psi_{\mathrm{d}}}{\partial z} = \epsilon_{\mathrm{d}}(\psi_{\mathrm{d}} - \overline{\psi}) \tag{11.31}$$

where $\overline{\psi}$ is the value of ψ in the large-scale environment.

In practice, (11.28) and (11.29) are solved in terms of $\mu = M/M_{\mathrm{u}}^{\mathrm{base}}$, where $M_{\mathrm{u}}^{\mathrm{base}}$ is the mass flux at cloud base (determined from the closure assumption as described further down).

Triggering of Moist Convection

The determination of the occurrence of moist convection in the model is based on whether a positively buoyant test parcel starting at each model level (iteratively from the surface and upwards) can rise high enough to produce a convective cloud and possibly precipitation. For an updraught starting from the lowest model level, its initial temperature and moisture departures with respect to the environment and its initial vertical velocity depend on surface sensible and latent heat fluxes, following Jakob and Siebesma (2003). When starting from higher model levels, the ascent is initially set to 1 m s^{-1} and its initial thermodynamic characteristics are assumed to be representative of a few hundred metre deep mixed-layer, with typical excesses of 0.2 K for temperature and 1×10^{-4} kg kg^{-1} for moisture. A 200 hPa threshold for cloud depth is prescribed to distinguish between shallow and deep convection. Mid-level convection is treated as deep convection.

Entrainment and Detrainment

Entrainment rate in the updraught (ϵ_{u}) is split into turbulent and organized components, which are both modulated by humidity conditions in the environment. Detrainment in the updraught (δ_{u}) is assumed to occur inside the convective cloud only where the updraught vertical gradient of kinetic energy and buoyancy are negative, that is usually in the upper part of the convective cloud.

Entrainment in downdraughts (ϵ_{d}) is assumed to occur only between the level of free sinking and the top of the 60 hPa atmospheric layer just above the surface. Inside this layer, it is set to a constant value. Detrainment (δ_{d}) is defined such as to ensure a downward linear decrease of downdraught mass flux to zero at the surface.

Precipitation Processes

The formation of precipitation from the cloud water contained in the updraught is parameterized according to Sundqvist et al. (1989) and a simple representation of precipitation evaporation is included. Precipitation formed from cloud liquid water at temperatures below the freezing point is assumed to freeze instantly, which affects the dry static energy tendency.

Closure Assumptions

One needs to formulate so-called closure assumptions to relate the convective updraught mass-flux at cloud base, $M_{\mathrm{u}}^{\mathrm{base}}$, to quantities that are explicitly resolved by the model. For deep convection, the closure is based on the balance between the convective available potential energy in the subgrid-scale updraught and the total heat release in the resolved larger-scale environment. The cloud base mass flux is expressed as the ratio of the latter two quantities, modulated with an adjustment timescale. This timescale depends on the updraught vertical velocity averaged over its depth and on spectral truncation. For shallow convection, the closure assumption links the moist energy excess at cloud base to the moist energy convergence inside the sub-cloud layer. The ratio of these two quantities yields the cloud base mass flux for shallow convection.

Regularization

Various regularizations need to be applied in the TL and AD code of the convection scheme to avoid spurious instabilities. These include reducing or setting to zero the perturbations of some terms that directly depend on the updraught vertical velocity as well as reducing updraught buoyancy and cloud base mass flux perturbations.

11.5.1.6 Large-Scale Condensation and Precipitation

The original version of the simplified diagnostic large-scale clouds and precipitation scheme currently used in the minimization of 4D-Var is described in Tompkins and Janisková (2004). Only a summary of its main features is given here.

The physical tendencies of temperature and specific humidity produced by moist processes on the large-scale can be written as

$$\frac{\partial q}{\partial t} = -C_{\mathrm{ce}} + E_{\mathrm{prec}} + D_{\mathrm{conv}} \tag{11.32}$$

$$\frac{\partial T}{\partial t} = L(C_{\mathrm{ce}} - E_{\mathrm{prec}} - D_{\mathrm{conv}}) + L_f(F_{\mathrm{rain}} - M_{\mathrm{snow}}) \tag{11.33}$$

where C_{ce} denotes large-scale condensation/evaporation, E_{prec} is the moistening due to the evaporation of precipitation and D_{conv} is the detrainment of cloud water from convective clouds. F_{rain} and M_{snow} correspond to the freezing of rain and melting of snow, respectively. L and L_{f} are the latent heats of vaporisation/sublimation and fusion, respectively.

Condensation

The subgrid-scale variability of humidity is assumed to be represented by a uniform distribution. This allows the estimation of the stratiform cloud fraction, C_{strat}, and cloud condensate amount, $q_{\mathrm{c}}^{\mathrm{strat}}$, from the grid-box relative humidity, RH, as

$$C_{\mathrm{strat}} = 1 - \sqrt{\frac{1 - RH}{1 - RH_{\mathrm{crit}} - \kappa(RH - RH_{\mathrm{crit}})}} \tag{11.34}$$

$$q_{\mathrm{c}}^{\mathrm{strat}} = q_{\mathrm{sat}} C_{\mathrm{strat}}^2 \{\kappa(1 - RH) + (1-\kappa)(1 - RH_{\mathrm{crit}})\} \tag{11.35}$$

where q_{sat} is the saturation specific humidity. The critical relative humidity threshold, RH_{crit}, and the coefficient κ are specified as in Tompkins and Janisková (2004). A simple diagnostic partitioning based on temperature is used to separate cloud condensate into liquid and ice.

The impact of convective activity on large-scale clouds, which is particularly important in the tropics and mid-latitude summers, is accounted for through the detrainment rate produced by the convection scheme. This detrainment term is used to compute the additional cloud cover and cloud condensate resulting from convection (i.e. convection called before condensation).

Precipitation Processes

The formation of precipitation from cloud condensate, q_c, is parameterized according to Sundqvist et al. (1989), but the Bergeron-Findeisen mechanism and collection processes are currently disregarded. Precipitation formed from cloud liquid water at temperatures below the freezing point is assumed to freeze instantly, which corresponds to term F_{rain} in (11.33). On the other hand, precipitation evaporation is estimated from the overlap of precipitation with the uniformly distributed subgrid fluctuations of humidity inside the clear-sky fraction of the grid box.

Regularization

Perturbations of C_{strat} were found to cause spurious instabilities in TL and AD integrations and are therefore artificially reduced according to the value of C_{strat} in the trajectory. A reduction of perturbations in the autoconversion of cloud condensate to precipitation is also needed.

11.5.2 A Few Remarks

The set of physical parametrization schemes developed for the ECMWF linearized model was described in Sect. 11.5.1. Although there are some simplifications and regularizations applied in the different parametrization schemes, the whole package is comprehensive and its non-linear form is able to provide up to 3 days forecasts that show a degree of realism which does not depart too much from that of the non-linear physics. Different levels of simplification of the schemes have been driven either by the requirement to decrease computational cost for operational applications or the necessity to avoid unrealistic perturbations in the linearized version of the scheme. The applied regularizations and simplifications allow global integrations of the linearized model with elaborated physical parametrization schemes even up to 48 h without producing spurious noise.

Overall, the presented package is a result of compromise between realism, linearity and computational cost while at the same time the level of complexity for the parametrization schemes is also influenced by the required applications. It is a constant challenge to maintain the best tradeoff between all those requirements.

11.5.3 Benefits of Regularization

The validity of the tangent-linear approximation can be highly degraded due to the non-linear and discontinuous nature of physical processes (see Sect. 11.3.2). If the

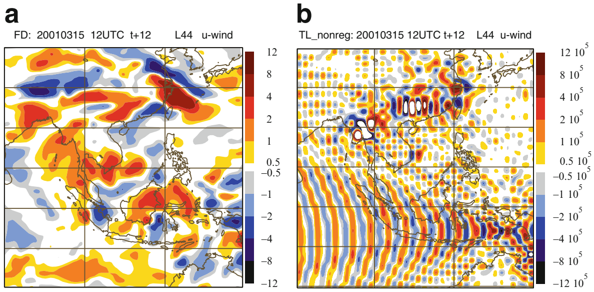


Fig. 11.2 Zonal wind increments around 700 hPa after 12-h evolution: (**a**) finite-differences (FD), (**b**) tangent-linear (TL) model without any regularization and (**c**) TL model with applied regularization in the cloud parametrization scheme

derivatives of model outputs with respect to the model state variables become very large, the linearization will become useless.

As an illustration of how strong nonlinearities can lead to erroneous behaviour of the tangent linear model, Fig. 11.2b shows the evolution of zonal wind increments at the model level around 700 hPa when using the TL model without any regularization in the cloud parametrization scheme. When compared to the finite differences (differences between two non-linear integrations) in Fig. 11.2a, one can notice that strong spurious noise develops in the tangent linear model. This noise comes from the autoconversion function (Fig. 11.1) describing the conversion of cloud water to precipitation. When regularization is applied to this function, TL increments (Fig. 11.2c) agree well with the finite differences (Fig. 11.2a).

11.6 Performance of the Linearized Physics

11.6.1 TL Approximation

Once discontinuities in the different physical parametrization schemes have been properly smoothed, the TL model describing the evolution of perturbations with the simplified physical parameterizations generally fits the finite differences between two non-linear forecasts much better than an adiabatic TL model.

To demonstrate the impact of the different physical processes in the TL model, experiments have been performed at a horizontal resolution equal to T255 and 91 levels in vertical (L91) using the linearized physics of ECMWF described in Sect. 11.5. The impact of the different physical processes on the tangent-linear evolution of temperature and zonal wind increments is shown in Fig. 11.3. Results are presented in terms of zonal mean of error difference as in (11.7) (i.e. the fit to the non-linear model with full physics) between the TL model including the whole set of linearized parametrization schemes and the adiabatic TL model (i.e.

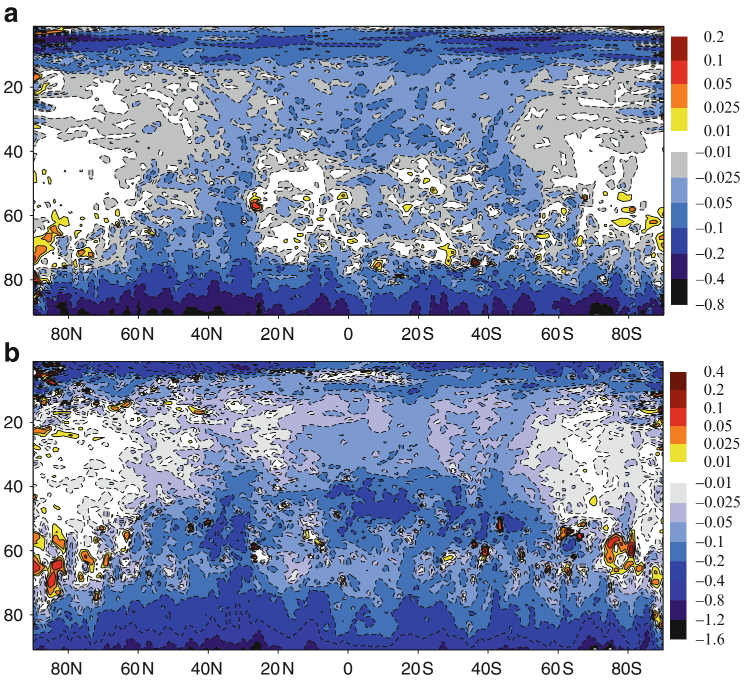


Fig. 11.3 Impact of the ECMWF operational linearized physics on the evaluation of (**a**) temperature and (**b**) u-wind increments in zonal mean. Results are presented as the error differences (in terms of fit to the non-linear model with full physics) between the TL model with full linearized physics and the purely adiabatic TL model

$\varepsilon_{exp} - \varepsilon_{adiab}$). Negative values are associated with an improvement of the model using the parametrization schemes with respect to the adiabatic TL model since they correspond to a reduction of the errors. The improvement is observed over most of the atmosphere, and is maximum in the lower troposphere.

Figure 11.4 shows the global relative improvement (see (11.9)) coming from including (a) dry physical parametrization schemes (i.e. vertical diffusion, gravity wave drag, non-orographic gravity wave and radiation) alone and (b) in combination with the moist processes (i.e. convection and cloud with large-scale parametrization schemes) in the linearized model compared to adiabatic tangent linear model for temperature, wind and specific humidity. The additional improvement due to the inclusion of the moist parametrization schemes is not only coming from these schemes, but also from cloud-radiation interactions.

The relative error of the TL model with respect to the finite differences using the full non-linear physical parameterizations is also used for evaluation. Figure 11.5 shows vertical profiles of global relative error reduction (therefore negative values) of the TL model using different physical parametrization schemes with respect to the adiabatic TL model. The error reduction becomes larger by gradually including

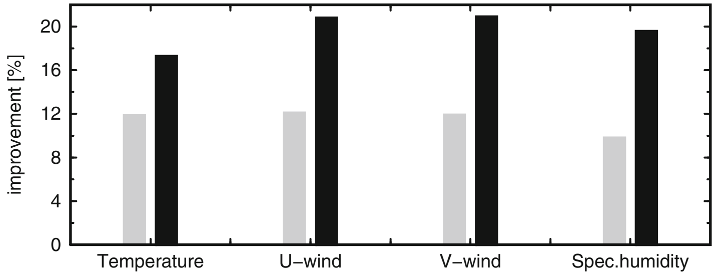


Fig. 11.4 Global relative improvement [%] of the tangent-linear approximation for temperature, wind and specific humidity coming from including: (a) dry physical parametrization schemes (i.e. vertical diffusion, gravity wave drag, non-orographic gravity wave and radiation) alone (*grey bars*) and (b) in combination with moist processes (i.e. convection and large-scale cloud parametrization schemes – *black bars*) into the linearized model compared to the purely adiabatic tangent linear model

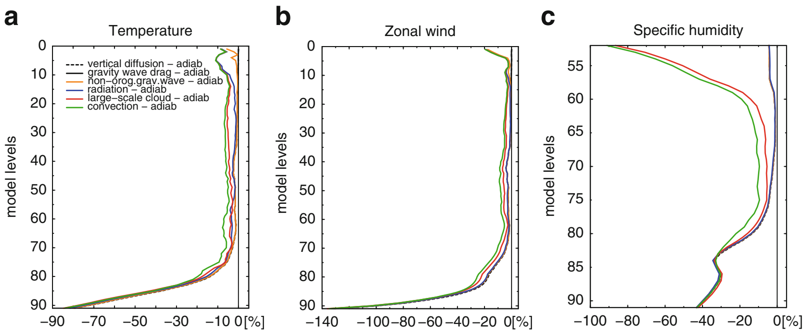


Fig. 11.5 Global relative error reduction of the tangent-linear (TL) model obtained from the gradual inclusion of physical parameterizations as shown in legend with respect to adiabatic TL model for (**a**) temperature, (**b**) zonal wind and (**c**) specific humidity. Relative errors of TL model are computed with respect to finite differences using the full non-linear physics

parametrization schemes, e.g. by including moist processes on top of the dry physical parametrization schemes.

11.6.2 Adjoint Sensitivities

Adjoint models can also be used for sensitivity studies since they allow to compute the gradient of one output parameter of a numerical model with respect to all input parameters. This property of adjoint allows to study, for instance, the sensitivity of a physical parametrization scheme to its input parameters (e.g. Li and Navon 1998; Janisková and Morcrette 2005). It is a more effective method compared with other standard approaches of repetitively using the direct schemes to obtain the sensitivity of all outputs by modifying in turn each input variables. More generally, an adjoint can be applied to analyze the sensitivity of a forecast aspect to initial conditions as proposed, for instance, by Errico and Vukicevic (1992) or Rabier et al. (1996). The adjoint method can also be used to measure the sensitivity with respect to any parameter of importance of the data assimilation system. In recent years, adjoint-based observation sensitivity techniques have been used as a diagnostic tool to monitor the observation impact on short-range forecasts (e.g. Langland and Baker 2004; Cardinali and Buizza 2004; Zhu and Gelaro 2008; Cardinali 2009). Such technique is restricted by the tangent-linear assumption and its validity. The better the tangent-linear approximation, the more realistic and useful the sensitivity patterns. Results obtained through the adjoint integration when using a too simplified adjoint model with large inaccuracies or adjoint models without a proper treatment of nonlinearities and discontinuities, can be incorrect.

The adjoint($\mathbf{F}^T$) of the linear operator $\mathbf{F}$ can provide the gradient of any objective function, J, with respect to $\mathbf{x}$ (input variables) given the gradient of J with respect to $\mathbf{y}$ (output variables) as:

$$\frac{\partial J}{\partial \mathbf{x}} = \mathbf{F}^T . \frac{\partial J}{\partial \mathbf{y}} \tag{11.36}$$

As an example, Fig. 11.6 displays the adjoint sensitivity of the 24-h forecast error to the initial conditions, i.e. to the analysis $\frac{\partial J}{\partial \mathbf{x}}$, where J is a measure of the forecast error (Rabier et al. 1996; Cardinali 2009). The sensitivity with respect to specific humidity and temperature at the lowest model level are shown for the situation on 28 August 2010 at 21:00 UTC from the run at T255L91 resolution. The results are presented for two different experiments, the first one run with only the dry parametrization schemes (i.e. vertical diffusion, gravity wave drag, non-orographic gravity wave and radiation) included in the adjoint model (Fig. 11.6a, b) and the second one with moist processes also taken into account (Fig. 11.6c, d). With only dry parametrization schemes, sensitivity to specific humidity is quite small and localized in areas of strongest dynamical activity. Even for temperature, it is obvious that some sensitivities are quite weak, especially in convective regions. Adding moist processes in the adjoint model brings additional structures to the sensitivity in areas affected by large-scale condensation/evaporation and convection. Therefore, using a more sophisticated adjoint model also provides more flow-dependent and more realistic sensitivities.

Another example of adjoint sensitivity computations using the adjoint version of the linearized physics package is given here, where the cost function was defined as the 3-h precipitation averaged over the core of a mid-latitude winter storm over northwestern Europe. One should emphasize that this kind of computation is only possible if the adjoint of moist physics parameterizations is available. Figure 11.7 shows the field of 3-h precipitation accumulation used for the evaluation of the precipitation cost function inside the black box at 0000 UTC 10 February 2009. As an illustration, Fig. 11.8 displays the adjoint sensitivities of the precipitation cost function with respect to 500 hPa temperature at 0000 UTC 9 February 2009 (i.e. 24 h beforehand and computed at T159L91 resolution). In other words, Fig. 11.8

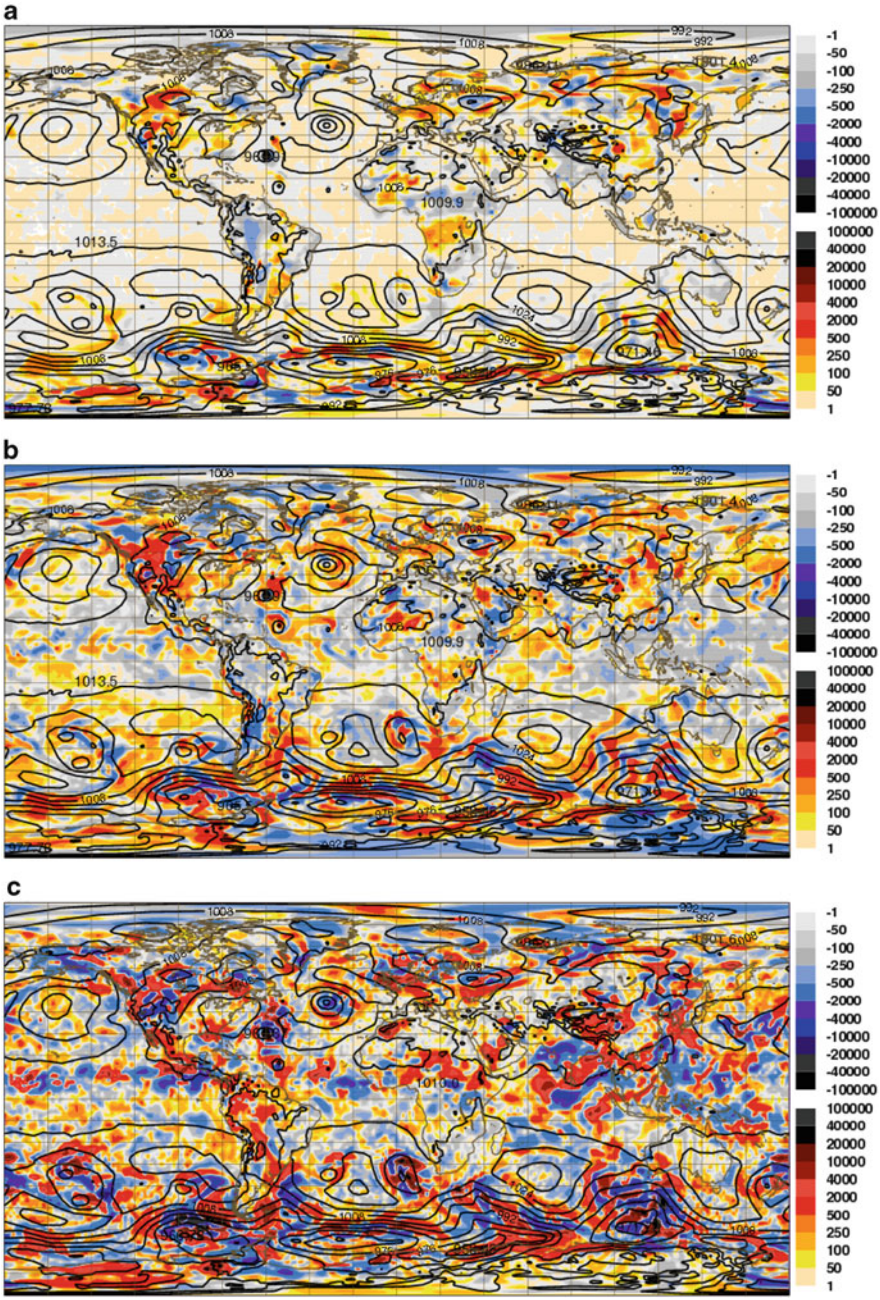


Fig. 11.6 (continued)

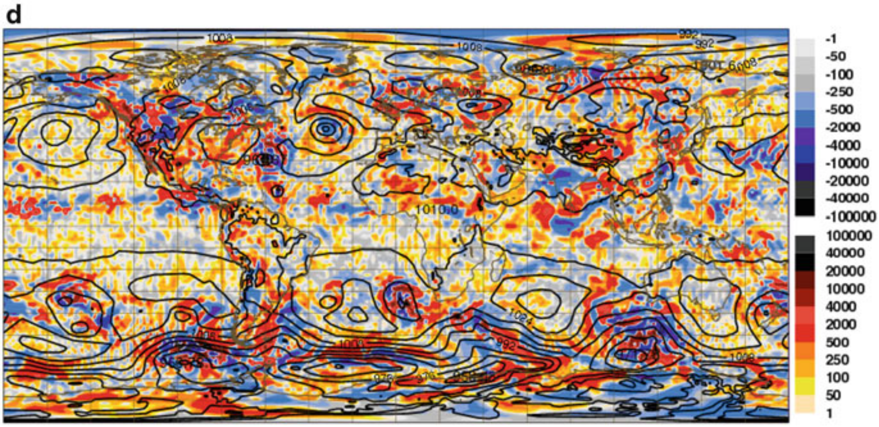


Fig. 11.6 Adjoint sensitivity of the 24-h forecast error to initial conditions in (**a**, **c**) specific humidity (J kg^{-1}/(g kg^{-1})) and (**b**, **d**) temperature (J kg^{-1}/K) at the lowest model level for the situation on 28 August 2010 at 21:00 UTC. The results are presented for (**a**, **b**) an experiment with dry parametrization schemes (i.e. vertical diffusion, gravity wave drag, non-orographic gravity wave and radiation) used in the adjoint model and (**c**, **d**) with moist processes also included. Sensitivities are shown with colour shading. *Black isolines* represent mean-sea-level pressure (hPa)

points out the regions where temperature ought to be modified in order to change precipitation inside the target box, 24 h later. In Fig. 11.8, the region of maximum sensitivity is found in the vicinity of the cold front associated with the 990 hPa low pressure system located at 19°W/47°N. The dipolar pattern of sensitivities indicates that a strengthening of the cross-frontal temperature gradient would result in a precipitation increase inside the black box, 24 h later.

Of course, it would also be possible to plot sensitivities with respect to moisture, wind and surface pressure fields for this case (not shown). In fact, sensitivities can be computed with respect to any variable which is part of the control vector of the adjoint model. However, one should also keep in mind that the relevance and usefulness of adjoint sensitivities can be limited by the degradation of the linearity assumption over time.

11.6.3 Data Assimilation

Experiments have been performed over July–September 2011 in order to compare two versions of the ECMWF 4D-Var system at resolution T511[3]L91: the first one including the linearized physics described above and the second one without it. Actually, in the version without the described linearized physics, a simple linear

[3] T511 corresponding approximately to 40 km

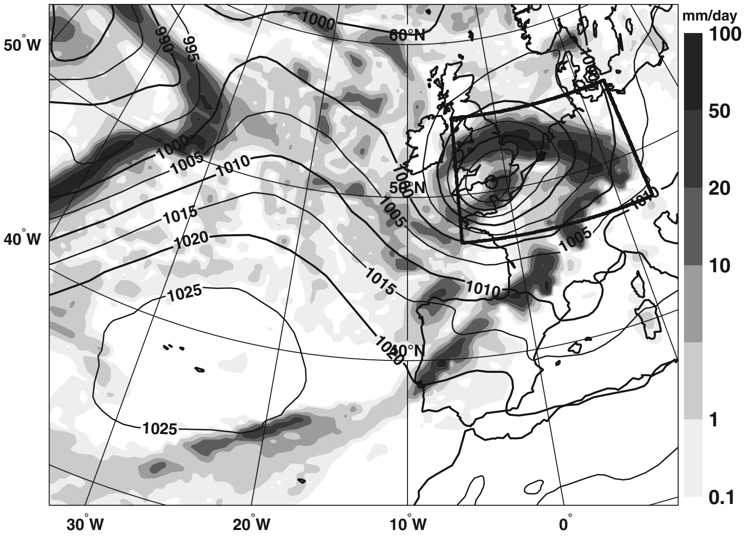


Fig. 11.7 Map of 3-h precipitation accumulations ending at 0000 UTC 10 February 2009 and used for computing the precipitation cost function inside the target black box over northwestern Europe. *Grey shading* shows precipitation (in mm day^{-1}), while *black isolines* of mean-sea-level pressure are also plotted (in hPa)

vertical diffusion (dry and acting mainly close to surface) and surface drag scheme (Buizza 1994) had to be used to avoid strong wind increments close to the surface. Precipitation and cloud related observations have not been taken assimilated in order to use the same type of observations in both experiments. Indeed, without the linearized moist physics in 4D-Var, cloud and precipitation observations cannot be assimilated since no observation equivalent can be produced from the model.

Including physical processes in the linear model of 4D-Var not only decreases the background cost function (measuring the distance between the initial state of the model and the background), but also brings model closer to observations as indicated by the general decreased observation cost function (measuring the distance between the model trajectory and corresponding observations) as seen in Fig. 11.9. Thus the distance between the model and the observations is better optimized when the linearized physics is used in the 4D-Var minimization.

The significance of the impact coming from including the linearized physical parametrization schemes in the 4D-Var system on the subsequent forecast is illustrated in Fig. 11.10 for the period of July–September 2011. The forecasts are scored against operational analyses in terms of anomaly correlation. A systematic and significant improvement for all plotted parameters, levels and regions is clearly obvious. Close to analysis time (where obviously the impact of the linearized physics in 4D-Var should be the largest), the biggest improvement is found in the middle and upper troposphere (e.g. 200 hPa wind vector scores) and overall in the

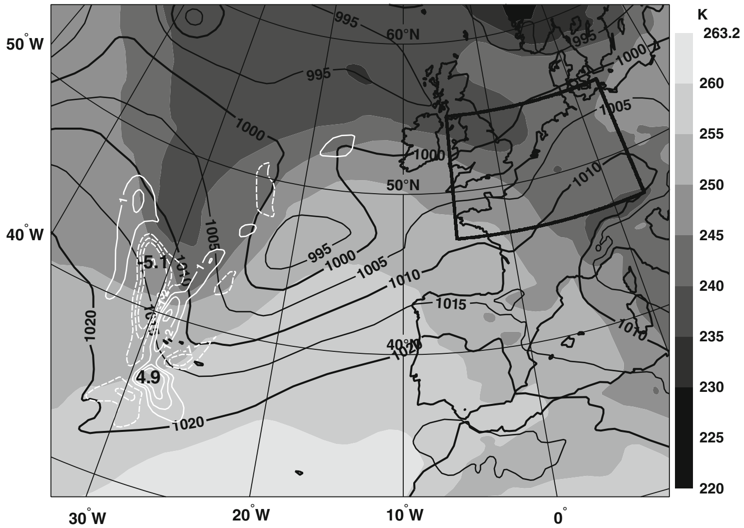


Fig. 11.8 Adjoint sensitivities of the precipitation cost function defined over the black box (see Fig. 11.7) with respect to 500-hPa temperature at 0000 UTC 9 February 2009, i.e. with a lead time of 24 h. Sensitivities are shown with white isolines (*solid* for positive, *dash* for negative) and are expressed in units of 10^{-4} (mm day^{-1}) K^{-1}. The background 500 hPa temperature field valid at 0000 UTC 9 February 2009 is displayed using *grey shading* and *black isolines* of mean-sea-level pressure are also plotted (in hPa)

Tropics. The positive impact is also generally remarkable in the lower troposphere (e.g. 700 hPa temperature scores or 700 hPa relative humidity scores).

The results presented above only show which impact the linearized physical parameterizations have on the evolution of the model state from the beginning of the 4D-Var assimilation window to the time of observations. However, including physical processes in the linearized model also allows to assimilate observations that are directly related to the physical processes, such as cloud and precipitation observations. Therefore further improvement in producing more realistic initial atmospheric states can be achieved. Since the late 1990s, significant efforts have been devoted to the assimilation of such observations. This is also the case at ECMWF, where a 1D + 4D-Var technique has been first used operationally for the assimilation of precipitation-related observations using microwave brightness temperatures from SSM/I (Bauer et al. 2006) from June 2005 until March 2009. This was then replaced by direct 4D-Var assimilation unifying the treatment of clear-sky, cloudy and precipitation situations, leading to an all-sky approach (Bauer et al. 2010; Geer et al. 2010). Direct 4D-Var of rain- and cloud-affected observations allows a physically consistent adjustment of model dynamics with temperature and humidity increments, due to the sensitivity of radiance observations to the atmospheric state through the combined radiative transfer model and the moist-

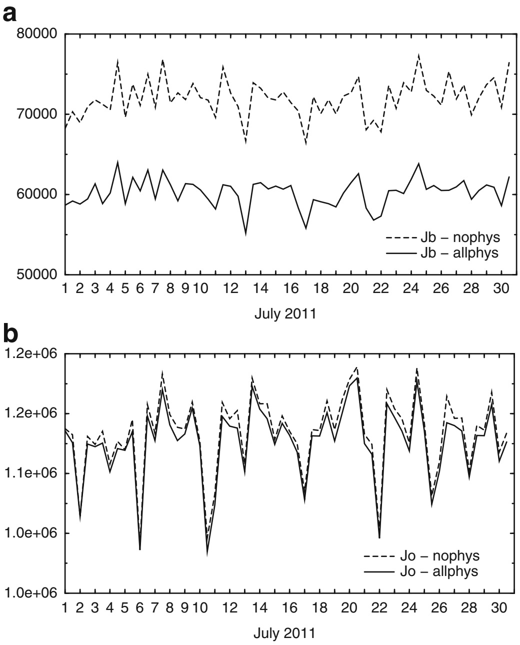


Fig. 11.9 Global values of (**a**) background cost function and (**b**) observation cost function for 4D-Var assimilation experiments run with all linearized physical parametrization schemes included (*solid line*) and using only very simple vertical diffusion of Buizza (1994) (*dashed line*). Statistics are shown over July 2011

physics parametrization. Furthermore, direct 4D-Var of surface rain data from ground-based NCEP Stage IV rain radars and gauges over the Eastern USA recently became operational in ECMWF global forecasting system (Lopez 2011) providing the clear improvement of short-range precipitation forecasts over the region. In the longer term, one could consider the assimilation of more radar networks (e.g. Europe, China, Canada, ...) once problems of data availability and quality are solved.

Experimental studies for assimilation of other observations related to the physical processes which may be considered for the future operational assimilation and therefore requiring parametrization schemes being able to provide a realistic counterpart to these observations were also performed at ECMWF. Experiments were conducted to assimilate spaceborne cloud optical depths (from MODIS, Benedetti and Janisková 2008), precipitation radar reflectivities (from TRMM precipitation radar, Benedetti et al. 2005) and cloud radar data (from CloudSat, Janisková et al. 2011). More recently, the potential benefits of directly assimilating synoptic station (SYNOP) rain gauge observations in 4D-Var were investigated (Lopez 2012) in both, a high resolution operations-like context and a lower-resolution data-sparse reanalysis-like framework.

The results from all above mentioned studies are not shown here, since they are well documented in the literature.

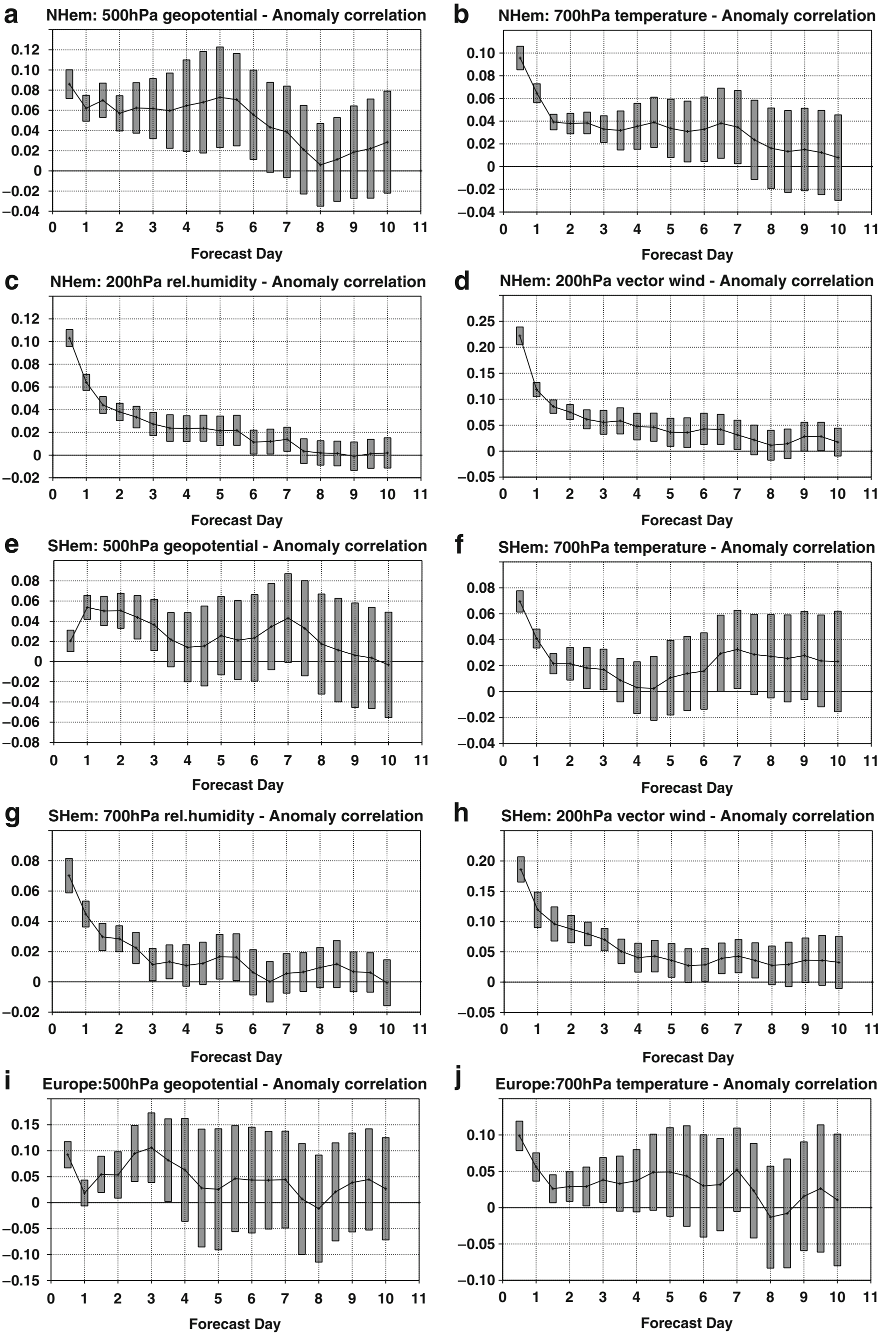


Fig. 11.10 (continued)

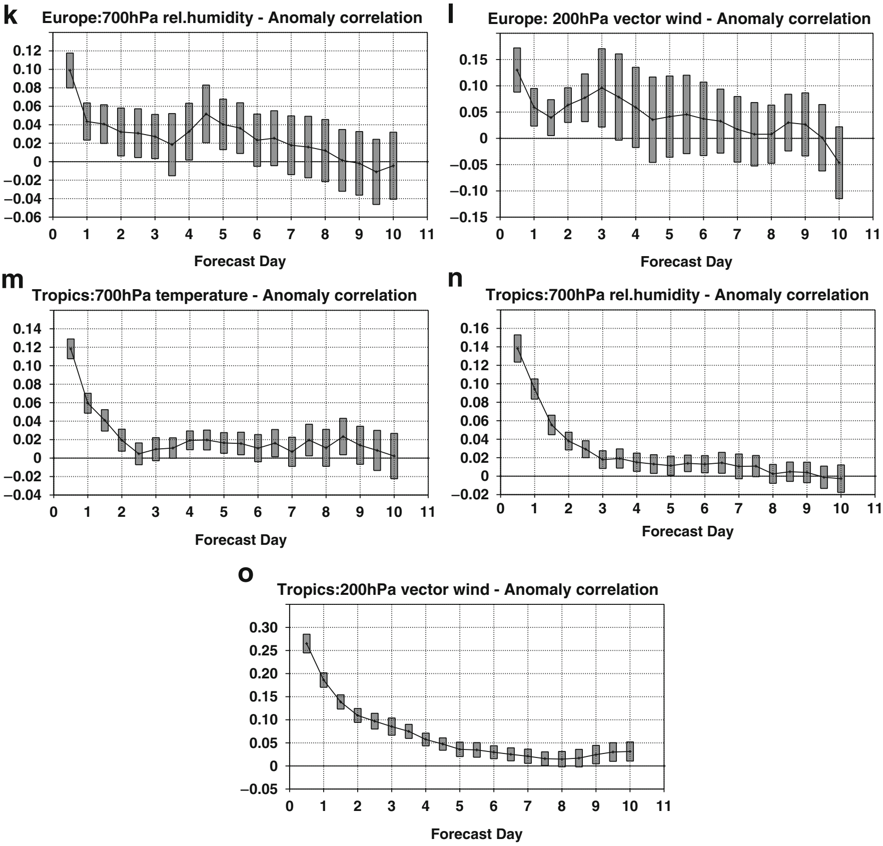


Fig. 11.10 Relative impact from the inclusion of the linearized physical parametrization schemes in ECMWF's 4D-Var system. Forecast scores against operational analysis are shown in terms of anomaly correlations for ranges up to 10 days. Score change is normalized by the control and positive values correspond to an improvement. *Grey bars* indicate significance at the 95 % confidence level. Results are shown for: 500 hPa geopotential, 700 hPa temperature, 700 hPa relative humidity, 200 hPa vector wind and for the different regions: (**a**)–(**d**) Northern extratropics, (**e**)–(**h**) Southern extratropics, (**i**)–(**l**) Europe and (**m**)–(**o**) tropics. Statistics are valid for the period of July–September 2011

11.7 Conclusions and Prospects

Past experimentation and operational implementation in ECMWF's Integrated Forecasting System have clearly demonstrated the benefits of including linearized physical parameterization schemes in the data assimilation process. Linearized physics can also be beneficial to singular vector computations for the Ensemble Prediction System, leading to more realistic initial perturbations. It can be useful to diagnose short-range forecast sensitivities to observations. Furthermore, employing linearized moist physics parameterizations in the 4D-Var minimizations has permitted the assimilation of the ever-increasing number of satellite and ground-based observations that are sensitive to clouds and/or precipitation.

However, the development of efficient and well-behaved TL and AD codes is made difficult by many obstacles and is therefore time consuming and often tedious, if not sometimes rather frustrating. In particular, a substantial amount of work is required to simplify and regularize the code or, in other words, to eliminate or smooth out the discontinuities and non-linearities that often characterize physical processes. The behaviour of the linearized physics package also needs to be constantly and thoroughly monitored in a wide range of potential applications (e.g. data assimilation, singular vectors computations, sensitivity experiments). In particular, every time one of the physical parameterizations is modified in the non-linear forecast model (which in practice occurs at every new model release), it is necessary to verify that the tangent-linear approximation is not degraded. If it is, appropriate updates have to be made to the TL and AD code so as to avoid a likely degradation of the 4D-Var operational performance. Eventually, a delicate compromise must constantly be achieved between linearity, computational efficiency and realism, to ensure that the best analysis and (above all) forecast performance are obtained.

With the continual trend towards higher and higher resolutions (both in the horizontal and the vertical), maintaining a well-behaved linearized physics package is bound to become more and more challenging. Currently, the minimizations involved in ECMWF's 4D-Var are still run at a relatively coarse resolution of roughly 80 km, even though trajectories and final analyses are computed at 16 km resolution. When minimization resolution is increased, the ability to represent smaller-scale and often noisier processes (such as convection) is likely to make it more difficult to fulfil the TL hypothesis. However it should be mentioned that preliminary TL approximation tests were recently performed with a global resolution of 25 km and over 12 h, with no sign of a degradation. One of the major uncertainty for the future is whether it will remain possible to make linearized physics to work when the resolution of the non-linear forecast model reaches a few kilometres, while the resolution remains well above 10 km in the 4D-Var minimizations. At this stage, the paradox of explicitly resolving convection in the trajectory but still needing to parameterize it in the minimization could be very challenging, and the current 4D-Var approach might need to be modified so as not to include the smaller scales in the entire analysis process (e.g. through trajectory smoothing).

There should nonetheless be some even greater concern about the growing complexity of the physical parameterizations used in the non-linear forecast model. Over the years, the increasing level of detail added to the representation of physical processes has been synonymous for enhanced and more numerous sources of non-linearity, which by construction cannot be included in the linearized physics package. There is a risk that if nothing is done to keep this trend under control, it will become impossible to make the linearized physics follow its non-linear counterpart closely enough, in which case 4D-Var as we know it may not be sustainable.

Even though there is some hope that future configurations of data assimilation based on weak-constraint 4D-Var might provide some ways to slightly relax the linearity constraint in time, it is paramount that non-linear model developers always remember that 4D-Var can only deliver good analyses if the linearity assumption remains valid over the entire assimilation window.

Acknowledgements The authors would like to thank Peter Bauer, Anton Beljaars and Jean-Noël Thépaut for their useful reviews of this paper. The anonymous reviewer is also acknowledged for reviewing the manuscript.

References

Araya-Polo M, Hascoët L (2004) Data flow algorithms in the Tapenade tool for automatic differentiation. In: Proceedings of 4th European congresson computational methods, ECCOMAS'2004, Jyvaskyla, Finland

Bauer P, Lopez P, Salmond D, Benedetti A, Saarinen S, Bonazzola M (2006) Implementation of 1D+4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. II: 4D-Var. Q J R Meteorol Soc 132:2307–2332

Bauer P, Geer AJ, Lopez P, Salmond D (2010) Direct 4D-Var assimilation of all-sky radiances. I: implementation. Q J R Meteorol Soc 136:1868–1885

Blackadar AK (1962) The vertical distribution of wind and turbulent exchange in a neutral atmosphere. J Geophys Res 67:3095–3102

Bechtold P, Köhler M, Jung T, Leutbecher M, Rodwell M, Vitart F, Balsamo G (2008) Advances in simulating atmospheric variability with the ECMWF model: from synoptic to decadal timescales. Q J R Meteorol Soc 134:1337–1351

Benedetti A, Janisková M (2008) Assimilation of MODIS cloud optical depths in the ECMWF model. Mon Weather Rev 136:1727–1746

Benedetti A, Lopez P, Bauer P, Moreau E (2005) Experimental use of TRMM precipitation radar observations in 1D+4D-Var assimilation. Q J R Meteorol Soc 131:2473–2495

Betts A (1997) The parametrization of deep convection: a review. In: Proceedings ECMWF workshop on new insights and approaches to convective parametrization, Reading, 4–7 Nov 1996, pp 166–188

Buizza R (1994) Impact of simple vertical diffusion and of the optimisation time on optimal unstable structures. ECMWF technical memorandum, vol 192. ECMWF, Reading, 25p

Cardinali C (2009) Monitoring the observation impact on the short-range forecast. Q J R Meteorol Soc 135:239–250

Cardinali C, Buizza R (2004) Observation sensitivity to the analysis and the forecast: a case study during ATreC targeting campaign. In: Proceedings of the first THORPEX international science symposium, Montreal, Canada, 6–10 Dec 2004. WMO/TD-1237, WWRP/THORPEX No. 6

Courtier P, Thépaut JN, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var using an incremental approach. Q J R Meteorol Soc 120:1367–1387

Errico RM, Reader KD (1999) An examination of the accuracy of the linearization of a mesoscale model with moist physics. Q J R Meteorol Soc 125:169–196

Errico RM, Vukicevic T (1992) Sensitivity analysis using an adjoint of the PSU/NCAR mesoscale model. Mon Weather Rev 120:1644–1660

Ebert EE, Curry JA (1992) A parametrization of ice optical properties for climate models. J Geophys Res 97D:3831–3836

Fouquart Y (1987) Radiative transfer in climate models. In: Schlesinger ME (eds) Physically based modelling and simulation of climate and climate changes. Kluwer Academic Publishers, Dordrecht/Boston, pp 223–284

Fouquart Y, Bonnel B (1980) Computations of solar heating of the earth's atmosphere: a new parametrization. Beitr Phys Atmos 53:35–62

Fritts DC, VanZandt TE (1993) Spectral estimates of gravity wave energy and momentum fluxes. Part I: energy dissipation, acceleration, and constraints. J Atmos Sci 50:3685–3694

Geer AJ, Bauer P, Lopez P (2010) Direct 4D-Var assimilation of all-sky radiances. Part II: assessment. Q J R Meteorol Soc 136:1886–1905

Geleyn J-F, Hollingsworth A (1997) An economical and analytical method for the interactions between scattering and line absorption of radiation. Contrib Atmos Phys 52:1–16

Geleyn J-F, Banciu D, Bellus M, El Khatib R, Moll P, Saez P, Thépaut J-N (2001) The operational 4D-Var data assimilation system of Météo-France: specific characteristics and behaviour in the special case of the 99 Xmas storms over France. In: Proceedings of 18th conference on weather analysis and forecasting and 14th conference on numerical weather prediction. Ft. Lauderdale, Florida, USA

Giering R, Kaminski T (1998) Recipes for adjoint code construction. ACM Trans Math Softw 24(4):437–474

Jakob C, Siebesma AP (2003) A new subcloud model for mass flux convection schemes. Influence on triggering, updraught properties and model climate. Mon Weather Rev 131:2765–2778

Janisková M (2003) Physical processes in adjoint models: potential pitfalls and benefits. In: Proceedings of ECMWF seminar on recent developments in data assimilation for Atmosphere and Ocean, Reading, UK, 8–12 Sept 2003, pp 179–191

Janisková M, Morcrette J-J (2005) Investigation of the sensitivity of the ECMWF radiation scheme to input parameters using adjoint technique. Q J R Meteorol Soc 131:1975–1995

Janisková M, Thépaut J-N., Geleyn J-F (1999) Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. Mon Weather Rev 127:26–45

Janisková M, Mahfouf J-F, Morcrette J-J, Chevallier F (2002) Linearized radiation and cloud schemes in the ECMWF model: development and evaluation. Q J R Meteorol Soc 128: 1505–1527

Janisková M, Lopez P, Bauer P (2012) Experimental 1D+4D-Var assimilation of CloudSat observations. Q J R Meteorol Soc 138:1196–1220 doi:10.1002/qj.988

Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201

Laroche S, Tanguay M, Delage Y (2002) Linearization of a simplified planetary boundary layer parametrization. Mon Weather Rev 130:2074–2087

Li Z, Navon IM (1998) Adjoin sensitivity of the Earth's radiation budget in the NCEP medium-range forecasting model. J Geophys Res 103(D4):3801–3814

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: 229–257

Lopez P (2011) Direct 4D-Var Assimilation of NCEP Stage IV Radar and Gauge precipitation data at ECMWF. Mon Weather Rev 139:2098–2116

Lopez P (2012) Experimental 4D-Var assimilation of SYNOP rain gauge data at ECMWF. ECMWF technical memorandum, vol 661. European Centre for Medium-Range Weather Forecasts, Reading, 25p

Lopez P, Moreau E (2005) A convection scheme for data assimilation: description and initial tests. Q J R Meteorol Soc 131:409–436

Lott F, Miller MJ (1997) A new subgrid-scale orographic drag parametrization: its formulation and testing. Q J R Meteorol Soc 123:101–127

Louis J-F, Tiedtke M, Geleyn J-F (1982) A short history of the PBL parametrization at ECMWF. In: Proceedings ECMWF workshop on planetary boundary layer parameterization, Reading, 25–27 Nov 1981, pp 59–80

Mahfouf J-F (1999) Influence of physical processes on the tangent-linear approximation. Tellus 51:147–166

Mahfouf J-F (2005) Linearization of a simple moist convection for large-scale NWP models. Mon Weather Rev 133:1655–1670

Mahfouf J-F, Rabier F (2000) The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Part II: experimental results with improved physics. Q J R Meteorol Soc 126:1171–1190

Mlawer E, Clough SA (1997) Shortwave and longwave enhancements in the rapid radiative transfer model. In: Proceedings of 7th atmospheric radiation measurement (ARM) science team meeting, U.S. Department of Energy, CONF-9603149, available from http://www.arm.gov/publications/proceedings/conf07/title.stm/mlaw-97.pdf

Mlawer E, Taubman SJ, Brown PD, Ianoco M, Clough SA (1997) Radiative transfer for inhomogeneous atmospheres: RRTM a validated correlated-k model for the longwave. J Geophys Res 102:16663–16682

Morcrette J-J (1989) Description of the radiation scheme in the ECMWF operational weather forecast model. ECMWF technical memorandum, vol 165 ECMWF, Reading, UK

Morcrette J-J (1991) Radiation and cloud radiative properties in the ECMWF operational forecast model. J Geophys Res 96D:9121–9132

Morcrette J-J, Jakob C (2000) The response of the ECMWF model changes in the cloud overlap assumption. Mon Weather Rev 128:876–887

Morcrette J-J, Smith L, Fouquart Y (1986) Pressure and temperature dependence of absorption in longwave radiation parametrizations. Beitr Phys Atmos 59:455–469

Morcrette J-J, Mlawer E, Iacono M, Clough S (2001) Impact of the radiation transfer scheme RRTM in the ECMWF forecating system. ECMWF newsletter No. 91, ECMWF, Reading, UK pp 2–9

Orr A, Bechtold P, Scinoccia J, Ern M, Janisková M (2010) Improved middle atmosphere climate and analysis in the ECMWF forecasting system through a non-orographic gravity wave parametrization. J Climate 23:5905–5926

Phillips SP (1984) Analytical surface pressure and drag for linear hydrostatic flow over three-dimensional elliptical mountains. J Atmos Sci 41:1073–1084

Rabier F, Klinker E, Courtier P, Hollingsworth A (1996) Sensitivity of forecast errors to initial conditions. Q J R Meteorol Soc 122:121–150

Rabier F, Järvinen H, Klinker E, Mahfouf J-F, Simmons A (2000) The ECMWF operational implementation of four-dimensional variational assimilation. Part I: experimental results with simplified physics. Q J R Meteorol Soc 126:1143–1170

Scinocca JF (2003) An accurate spectral nonorographic gravity wave drag parameterization for general circulation models. J Atmos Sci 60:667–682

Smith EA, Shi L (1992) Surface forcing of the infrared cooling profile over the Tibetan plateau. Part I: influence of relative longwave radiative heating at high altitude. J Atmos Sci 49:805–822

Sundqvist H, Berge E, Kristjansson JE (1989) Condensation and cloud parametrization studies with mesoscale numerical weather prediction model. Mon Weather Rev 117:1641–1657

Tompkins AM, Janisková M (2004) A cloud scheme for data assimilation: description and initial tests. Q J R Meteorol Soc 130:2495–2517

Tsuyuki T (1996) Variational data assimilation in the tropics using precipitation data. Part II: 3-D model. Mon Weather Rev 124:2545–2561

Warner CD, McIntyre ME (1996) An ultra-simple spectral parametrization for non-orographic gravity waves. J Atmos Sci 58:1837–1857

Washington WM, Williamson DL (1977) A description of the NCAR GCMs. GCMs of the atmosphere. In: Chang J (ed) Methods in computational physics, vol 17. Academy Press, New York, pp 111–172

Zhu Y, Gelaro R (2008) Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. Mon Weather Rev 136:335–351

Zou X, Navon IM, Sela JG (1993) Variational data assimilation with moist threshold processes using the NMC spectral model. Tellus 45A:370–387

Zupanski D, Mesinger F (1995) Four-dimensional variational assimilation of precipitation data. Mon Weather Rev 123:1112–1127

Chapter 12
Recent Applications in Representer-Based Variational Data Assimilation

Boon S. Chua, Edward D. Zaron, Liang Xu, Nancy L. Baker, and Tom Rosmond

Abstract Data assimilation with representer-based algorithms (also called “dual space” algorithms) are currently being used for weak-constraint four-dimensional variational data assimilation (W4D-Var) atmospheric prediction, distributed parameter estimation, and other hydrodynamic data assimilation problems. The iterative linear solvers at the core of these systems may display non-monotonic convergence in the norm defined by the primal objective function, and this behavior makes problematic the development of practical stopping criteria. One approach to this problem is described, namely an implementation of the inner solver using the generalized conjugate residual(GCR) algorithm. Additional elements of data assimilation systems are error model for the background, model forcings, and observations. An implementation of a posterior analysis method for diagnosing the error variances is described, and representative results from an atmospheric data assimilation systems are shown.

B.S. Chua (✉) · T. Rosmond
SAIC, Monterey, CA, USA

Marine Meteorology Division, Naval Research Laboratory Monterey, CA, USA
e-mail: boon.chua@nrlmry.navy.mil; tom.rosmond@nrlmry.navy.mil

E.D. Zaron
Department of Civil and Environmental Engineering, Portland State University, Portland, OR, USA
e-mail: zaron@cecs.pdx.edu

L. Xu · N.L. Baker
Marine Meteorology Division, Naval Research Laboratory Monterey, CA, USA
e-mail: liang.xu@nrlmry.navy.mil; nancy.baker@nrlmry.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_12,

12.1 Introduction

Four-dimensional variational data assimilation (4D-Var) is an estimation technique which finds a model state $\mathbf{x}(t_0)$, at initial time t_0, that minimizes a quadratic objective function, the sum of the distance between the initial state $\mathbf{x}(t_0) \in R^n$ and a prior estimate (the so-called background field) $\mathbf{x}^b \in R^n$, and the distance between a real-valued vector of observations $\mathbf{y} \in R^m$ and measurements $\mathcal{H}(\mathbf{x})$ of the trajectory $\mathbf{x}(t)$ obtained by integration of a dynamical model from $\mathbf{x}(t_0)$. The objective function $\mathcal{J}$ is written

$$\begin{aligned} \mathcal{J}[\mathbf{x}(t_0)] = {} & (\mathbf{x}(t_0) - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}(t_0) - \mathbf{x}^b) \\ & + [\mathbf{y} - \mathcal{H}(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y} - \mathcal{H}(\mathbf{x})], \end{aligned} \tag{12.1}$$

where $\mathbf{B}$ and $\mathbf{R}$ are estimates of the background and observation error covariance matrices, respectively, and the observations, $\mathbf{y} = \{y_i\}_{i=1}^m$, are nonlinear functions of the initial state,

$$y_i = \mathcal{H}_i[\mathcal{M}(t_i, t_0)\mathbf{x}(t_0)] + \delta_i. \tag{12.2}$$

Here we assume that $\mathcal{M}(t_i, t_0)$ propagates the model state from t_0 to t_i, $\mathcal{H}_i$ is the $i-$th observation operator, and δ_i is the observation error. Note that if the initial condition and observation errors are Gaussian distributed with covariances $\mathbf{B}$ and $\mathbf{R}$, if the observation errors are unbiased, and if the background field $\mathbf{x}^b$ is equal to the statistical mean of $\mathbf{x}(t_0)$, then the minimizer of $\mathcal{J}$ is the maximum likelihood estimate of $\mathbf{x}(t_0)$.

In addition to errors in the initial conditions, it is clear that oceanic and atmospheric models contain other sources of error which must be considered. Specifically, there are errors in model inhomogeneities such as boundary conditions and radiative forcing. Weak-constraint four-dimensional variational data assimilation (W4D-Var) is a generalization of 4D-Var which permits one to estimate these additional inhomogeneities, denoted $\mathbf{f}$. Assuming that prior or background values of the forcing fields are available, $\mathbf{f}^b$, then the above objective function naturally generalizes to

$$\begin{aligned} \mathcal{J}[\mathbf{x}(t_0), \mathbf{f}] = {} & (\mathbf{f} - \mathbf{f}^b)^T \mathbf{F}^{-1} (\mathbf{f} - \mathbf{f}^b) \\ & + (\mathbf{x}(t_0) - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}(t_0) - \mathbf{x}^b) \\ & + [(\mathbf{y} - \mathcal{H}(\mathbf{x})]^T \mathbf{R}^{-1} [(\mathbf{y} - \mathcal{H}(\mathbf{x})], \end{aligned} \tag{12.3}$$

where it should be understood that the model propagator $\mathcal{M}$ now depends on both the space-time-dependent inhomogeneities, $\mathbf{f}$, and the initial conditions, $\mathbf{x}(t_0)$.

In the incremental formulation (Courtier et al. 1994), the dynamics and measurement operators are linearized around a background trajectory $\bar{\mathbf{x}}$, and an incremental objective function is defined in terms of $\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$. Of course, if the model dynamics and observation operator are linear, the extremum of the incremental objective function corresponds to an extremum of the original objective function. When nonlinearity is present, the incremental objective function is used to build an iterative solver for the original, nonlinear, data assimilation problem. In this article we assume that some linearization strategy has been selected, e.g., the tangent linearization proposed in Courtier et al. (1994) or the bounded iterate strategy of Bennett and Thorburn (1992), so that the so-called *inner loop* solver must minimize a strictly quadratic objective function. Henceforth, we shall restrict our attention to the objective function,

$$\begin{aligned}\mathcal{J}[\mathbf{x}(t_0),\mathbf{f}] = {} & (\mathbf{f}-\mathbf{f}^b)^T\mathbf{F}^{-1}(\mathbf{f}-\mathbf{f}^b)\\ & + (\mathbf{x}(t_0)-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}(t_0)-\mathbf{x}^b)\\ & + (\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}),\end{aligned} \tag{12.4}$$

where the matrix $\mathbf{H} \in R^{m\times n}$ is a linear approximation to the operator $\mathcal{H}$, and inhomogeneities resulting from the linearization have been absorbed into $\mathbf{x}^b$, $\mathbf{f}^b$, and $\mathbf{y}$.

There are practical considerations which make the implementation of W4D-Var considerably more complex than 4D-Var for realistic models. The first issue is the dimensionality of the unknown vectors, which has consequences for the design and implementation of solvers for minimizing $\mathcal{J}$. Assuming the state vector $\mathbf{x}(t)$ is of dimension n, then the model forcing $\mathbf{f}$ may be as large as $T \times n$, where T is the cardinality of the time interval under consideration. The dimension of the space-time covariance matrix $\mathbf{F}$ is formally the square of this. The second key issue is scientific, and relates to the determination of the error covariances $\mathbf{B}$ and $\mathbf{F}$. Quantitative estimation of these objects requires vast amounts of data which are rarely available; in practice they are often parameterized in terms of a spatially- or temporally-varying variance function, and a set of correlation scales for the orthogonal coordinate directions.

Here we review recent developments associated with the application of representer-based solvers (Bennett 1992) to 4D-Var and W4D-Var problems, an approach which is the foundation for the so-called dual form of variational data assimilation (Courtier 1997). Recall that the minimizer of the objective function is the solution to $\frac{1}{2}\nabla\mathcal{J}(\mathbf{x}) = 0$; applied it to (12.1) yields,

$$(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{x} = \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} + \mathbf{B}^{-1}\mathbf{x}^b, \tag{12.5}$$

where uniqueness is assured provided that $\mathbf{B}$ is of full rank. Equivalently, the solution can be expressed as the sum of the background and a linear combination of representer functions $\mathbf{x} = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T\hat{\mathbf{x}}$, yielding the equation for the dual variables $\hat{\mathbf{x}}$,

$$(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})\hat{\mathbf{x}} = \mathbf{y} - \mathbf{H}\mathbf{x}^b. \tag{12.6}$$

In this dual formulation the unknown vector $\hat{\mathbf{x}}$ lies in R^m, whereas $\mathbf{x}$ lies in R^n. Also, the expansion in terms of representer functions is valid even in the continuum limit of the discretized dynamics, in which case (12.5) become the Euler-Lagrange equations for the extremum of the objective functional. The columns of the $\mathbf{BH}^T$ matrix, which are approximations to the representer functions in the continuum limit, span the space of observable increments; i.e., they are exactly the m degrees of freedom which are determined by the measurements (Bennett 1992).

The dual formulation and representer expansion have by now been utilized in many data assimilative modeling studies of the ocean and atmosphere. Because the dimension of the vector of unknowns is m in either case of 4D-Var or W4D-Var, there is no intrinsic limitation of the method in the latter case. In order to fix the notation so that a single system describes both 4D-Var and W4D-Var, consider the following augmented vectors and covariance matrices:

$$\mathbf{x}' = \begin{pmatrix} \mathbf{x}(t_0) \\ \mathbf{f} \end{pmatrix}, \quad \mathbf{B}' = \begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{F} \end{pmatrix}, \quad \mathbf{H}' = \begin{pmatrix} \mathbf{H} \\ 0 \end{pmatrix}, \quad \mathbf{R}' = \mathbf{R}, \quad \mathbf{y}' = \mathbf{y}. \tag{12.7}$$

Henceforth, we drop primes and simply write the objective function as

$$\begin{aligned} \mathcal{J}[\mathbf{x}] =& (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) \\ &+ (\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}), \end{aligned} \tag{12.8}$$

noting that the extremal conditions (12.5) and dual formulation (12.6) are formally unchanged.

Recent advances for representer-based variational assimilation have been connected with technologies for solving (12.6), e.g., preconditioners and iterative solvers, and with developing justifiable error models for the background and model forcing errors, $\mathbf{B}$ and $\mathbf{F}$.

In the next section, recent technological developments for solving (12.6) are discussed, and we share our experience concerning the primal and dual forms of the variational data assimilation algorithms, as has been the focus of recent papers (El Akkraoui and Gauthier 2010; El Akkraoui et al. 2008; Gratton and Tshimanga 2009). Following that, recent work on covariance modeling is described. The latter developments are not unique to representer-based approaches.

12.2 Solver Improvements

Several considerations have led to improvements in representer-based solvers for variational data assimilation.

First, it has been noted that iterative solvers for (12.6) may yield a non-monotonic sequence of $\mathcal{J}(\mathbf{x}_p)$ values, where $\mathbf{x}_p$ represents the approximate solution at step p of the iterative solver (El Akkraoui et al. 2008). This phenomenon has been observed

Fig. 12.1 The GCR algorithm for solving $\mathbf{Ax} = \mathbf{b}$

```
initialize $\mathbf{x}_0, \epsilon$;
$\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$;
$i = 0$;
while $(\mathbf{r}_i^T \mathbf{r}_i)^{1/2} > \epsilon$, do
        $i = i + 1$;
        $\mathbf{u}_i = \mathbf{r}_{i-1}$;
        $\mathbf{c}_i = \mathbf{Au}_i$;
        for $k = 1, i - 1$, do
            $\alpha_k = \mathbf{c}_i^T \mathbf{c}_k$;
            $\mathbf{c}_i = \mathbf{c}_i - \alpha_k \mathbf{c}_k$;
            $\mathbf{u}_i = \mathbf{u}_i - \alpha_k \mathbf{u}_k$;
        end;
        $\mathbf{c}_i = \mathbf{c}_i / (\mathbf{c}_i^T \mathbf{c}_i)^{1/2}$;
        $\mathbf{u}_i = \mathbf{u}_i / (\mathbf{c}_i^T \mathbf{c}_i)^{1/2}$;
        $\mathbf{x}_i = \mathbf{x}_{i-1} + (\mathbf{c}_i^T \mathbf{r}_{i-1}) \mathbf{u}_i$;
        $\mathbf{r}_i = \mathbf{r}_{i-1} - (\mathbf{c}_i^T \mathbf{r}_{i-1}) \mathbf{c}_i$;
end
```

with the Physical-space Statistical Analysis System (PSAS, Cohn et al. 1998), which employs the conjugate-gradient algorithm applied to (12.6) using $\mathbf{R}^{-1/2}$ as preconditioner, and it was also displayed in Zaron (2006) with a non-preconditioned solver. The non-monotonic reduction in the value of the objective function makes it problematic to establish an acceptable stopping criteria for the iterative solver. In spite of the fact that $m << n$, data sets are frequently large enough that executing full set of m iterations, the worst-case iteration count for conjugate-gradient-type linear solvers in exact arithmetic, is prohibitive.

Another issue which arises in practice is that the huge condition number of the covariance matrices and asymmetry of the linearized model and its approximate adjoint may cause $\mathbf{R} + \mathbf{HBH}^T$ to be non-positive-definite symmetric. Experience with idealized problems, where the operators can be explicitly constructed as matrices, shows that the lack of monotonic convergence discussed in the previous paragraph is exacerbated by symmetry errors and lack of positive-definiteness in the $\mathbf{HBH}^T$ matrix.

A final consideration in the development of new solvers is the availability of diagnostic data to assess the progress of the iteration or to evaluate the quality of the state variable which is obtained.

Recent experience has shown that the generalized conjugate residual (GCR) method (de Sturler 1994, 1996) addresses all the above-mentioned points. GCR is a general-purpose Krylov method for solving non-symmetric systems, $\mathbf{Ax} = \mathbf{b}$, which builds matrices $\mathbf{U}$ and $\mathbf{C}$ in $R^{p \times m}$ such that $\mathbf{AU} = \mathbf{C}$. The columns of both $\mathbf{U}$ and $\mathbf{C}$ are in the span of the Krylov subspace $K = Span\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{p-1}\mathbf{b}\}$, and $\mathbf{C}$ is orthogonal, such that $\mathbf{C}^T\mathbf{C} = \mathbf{I}$. The GCR algorithm shown in Fig. 12.1 computes $\mathbf{x}_p \in K$ to minimize $\| \mathbf{Ax}_p - \mathbf{b} \|_2$, which is similar to the *minimum residual* algorithm suggested by El Akkraoui and Gauthier (2010). Although the GCR algorithm can fail when either the residual is orthogonal to the Krylov subspace or when $\mathbf{b}$ is an eigenvector of $\mathbf{A}^p$, neither of these situations has occurred in practice.

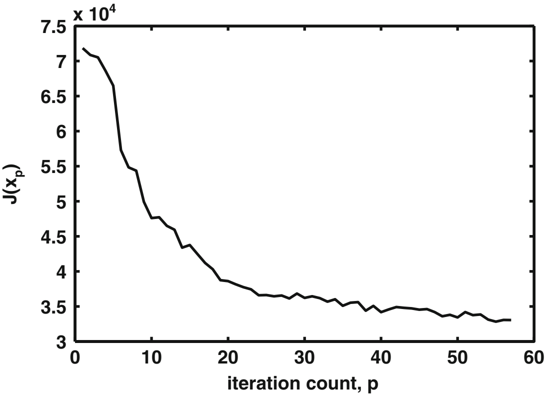


Fig. 12.2 Reduction of $\mathcal{J}(\mathbf{x})$ using GCR. The performance of the GCR solver as measured by the value of the objective function for an ocean data assimilation problem is shown. $\mathcal{J}(\mathbf{x}_p)$ is computed using (12.14) and (12.15) in the text. The application involves the assimilation of satellite altimetry data into a three dimensional primitive equations ocean model encompassing the Hawaiian Ridge, with the goal of estimating the tidal circulation around the Ridge

Figure 12.2 shows the progress of $\mathcal{J}(\mathbf{x}_p)$ for a data-assimilative three-dimensional ocean model with approximately $n = 400 \times 300 \times 30 \times 5 = 18 \times 10^6$ state variables and $m = 17 \times 10^4$ observations (see Zaron et al. 2009 for a similar application in a smaller computational domain). The figure shows that the decrease in cost function is not monotonic, and increases can occur. This behavior does not occur in smaller, exactly symmetric problems, and the working hypothesis is that the non-monotonicity is caused by asymmetry or lack of positive-definiteness in either the adjoint model or background covariance. Pointwise tests of the symmetry of $\mathbf{B}$ and $\mathbf{HBH}^T$ indicate that the former is symmetric to machine precision, while the latter contains symmetry errors of 10 % of the diagonal elements. The computational cost of evaluating $\mathbf{Ax}$ is approximately 100 cpu-hours, so there is a substantial need for computational efficiency.

Further diagnostic information is available from the GCR iterates as well. Qualitative assessment of the solution in the state space is available since the solution $\mathbf{x}_p$ is computed at each iterate. Because $\mathbf{AU} = \mathbf{C}$, with $\mathbf{C}$ orthogonal, the singular values $\lambda(\mathbf{U})$ of $\mathbf{U}$ approximate the singular values of $\mathbf{A}^{-1}$ (Golub and Van Loan 1989). Knowledge of the singular spectrum and orthogonal decomposition of $\mathbf{U}$ may be used to better precondition subsequent outer iterations (Giraud et al. 2006; Parks et al. 2006).

Assuming the observation error is uncorrelated and constant, $\mathbf{R} = \sigma\mathbf{I}$, one can approximate the singular spectrum of the so-called representer matrix $\mathcal{R} = \mathbf{HBH}^T$ (Bennett 1992) with $\lambda(\mathcal{R}) \approx \lambda(\mathbf{U})^{-1} - \sigma$. Here the notation $\lambda(\mathbf{U}) = \{\lambda_i(\mathbf{U})\}_{i=1}^{p}$ denotes the ordered singular spectrum, the set of nonzero singular values of the matrix $\mathbf{U} \in R^{m \times p}$, where $\lambda_{i+1}(\mathbf{U}) \leq \lambda_i(\mathbf{U})$ and $p \leq m$ are assumed, and the inverse of the singular spectrum $\lambda(\mathbf{U})^{-1}$ is defined as the set of reciprocals of the singular values. This singular spectrum is useful when assessing the observing array or covariance model, since it establishes a criterion for counting the number of degrees of freedom effectively constrained by the data (Bennett 1985, 1992). When the observation error is not a constant it is advantageous to transform with the change of variables, $\hat{\mathbf{v}} = \mathbf{R}^{-1/2}\hat{\mathbf{x}}$.

The singular spectrum can be used to develop a stopping criterion for the iterative solver in terms of the predicted percent of variance explained. Recall that the representer matrix $\mathcal{R}$ can be interpreted as a covariance matrix, the trace of which is the total amount of variance expected in the observations exclusive of measurement noise (Bennett 2002). Recall also, that the degrees of freedom associated with singular vectors may be classified as either smoothed or interpolated by the data assimilation, according to whether $\lambda_i(\mathcal{R}) < \sigma$ or $\lambda_i(\mathcal{R}) > \sigma$, respectively (Bennett 2002). Let k denote the mode number with the singular value comparable to the measurement error, e.g., $\lambda_k(\mathcal{R}) > \sigma \geq \lambda_{k+1}(\mathcal{R})$, then

$$S = \sum_{i=1}^{k} \lambda_i(\mathcal{R}) \tag{12.9}$$

is the expected total observed variance explainable by the given data assimilation system. In practice $\lambda(\mathcal{R})$ is not known exactly, but its approximation $\hat{\lambda}(\mathcal{R}) = \lambda(\mathbf{U})^{-1} - \sigma$ is available from the orthogonal decomposition of $\mathbf{U}$. An approximation to S can be made by extrapolating $\hat{\lambda}(\mathcal{R})$ out to $i = k$. Letting $\hat{\lambda}^e(\mathcal{R})$ denote this approximate spectrum, then the fraction of S explained by stopping at iterate p may be estimated as

$$f = \left(\sum_{i=1}^{p} \hat{\lambda}_i(\mathcal{R})\right)\left(\sum_{i=1}^{k} \hat{\lambda}_i^e(\mathcal{R})\right)^{-1}. \tag{12.10}$$

Figure 12.3 shows an application of these ideas with the data-assimilative ocean model described in Zaron et al. (2009). The estimated spectrum $\hat{\lambda}(\mathcal{R})$ is computed for iterates $p = 10, 20, 40$ (gray) and for the final iterate $p = 58$ (black). The extrapolated spectrum $\hat{\lambda}^e(\mathcal{R})$ is computed from a power-law fit to the middle 50 % of the singular values, and one sees that the extrapolated spectrum and data error variance intersect at approximately $k = 200$; thus, one expects approximately 142 additional iterates would be necessary to minimize $\mathcal{J}(\mathbf{x})$. Applying (12.10) to compute the fraction of variance explained, one finds $f = 88\,\%$. In other words, the solution obtained by stopping the solver at $p = 58$ accounts for 88 of the explainable observed variance. Note that the variance associated with modes $p > k$ is un-explainable with the covariance model $\mathbf{B}$, and it is ascribed to observation error. While the details are certainly problem-dependent, we have found that $\hat{\lambda}(\mathcal{R})$ adequately approximates the true spectrum when judged against the uncertainty in $\mathbf{B}$. Experience with idealized, low-dimensional, data assimilation problems suggests that these methods are applicable in realistic systems, where complete knowledge of the spectra cannot be obtained.

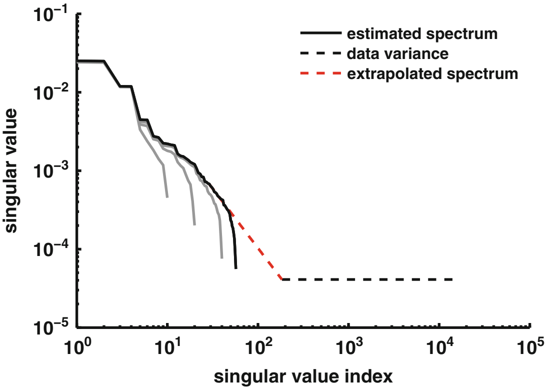


Fig. 12.3 Spectral Diagnostics from GCR. The estimated spectrum $\hat{\lambda}(\mathcal{R})$ of the representer matrix $\mathcal{R} = \mathbf{HBH}^T$ is shown by the *dark solid line* corresponding to the last GCR iterate ($p = 58$) in Fig. 12.2. *Solid gray lines* show $\hat{\lambda}(\mathcal{R})$ based on iterates $p = 10, 20$, and 40, for comparison. The data variance is σ, where $\mathbf{R} = \sigma\mathbf{I}$. The extrapolated spectrum is computed from a linear fit to $(log(i), log(\lambda_i(\mathcal{R})))$ in the range $p/4 \leq i \leq 3p/4$

Finally, the two components of $\mathcal{J}(\mathbf{x}_p)$ due to the background and observations may be obtained as diagnostic information from the GCR iterates. Substituting $\mathbf{x}_p = \mathbf{BH}^T\hat{\mathbf{x}}_p$ in (12.4), one obtains

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{x}}_p) &= \mathcal{J}^B(\hat{\mathbf{x}}_p) + \mathcal{J}^R(\hat{\mathbf{x}}_p)\\ &= \hat{\mathbf{x}}_p^T\mathbf{HBH}^T\hat{\mathbf{x}}_p\\ &\quad + (\mathbf{HBH}^T\hat{\mathbf{x}}_p - \mathbf{y})^T\mathbf{R}^{-1}(\mathbf{HBH}^T\hat{\mathbf{x}}_p - \mathbf{y}).\end{aligned} \tag{12.11}$$

Because the GCR solver computes the residual $\mathbf{r}_p$ at each iterate, one has

$$(\mathbf{HBH}^T + \mathbf{R})\hat{\mathbf{x}}_p = \mathbf{y} - \mathbf{r}_p. \tag{12.12}$$

Assuming that $\mathbf{R}\hat{\mathbf{x}}_p$ can be computed on demand, then

$$\mathbf{HBH}^T\hat{\mathbf{x}}_p = \mathbf{y} - \mathbf{r}_p - \mathbf{R}\hat{\mathbf{x}}_p, \tag{12.13}$$

and all terms in the expression for the objective function are computable. The contribution from the background term is

$$\mathcal{J}^B(\mathbf{x}_p) = (\hat{\mathbf{x}}_p)^T(\mathbf{y} - \mathbf{r}_p - \mathbf{R}\hat{\mathbf{x}}_p), \tag{12.14}$$

while the contribution from the observations is

$$\mathcal{J}^R(\mathbf{x}_p) = (\mathbf{r}_p + \mathbf{R}\hat{\mathbf{x}}_p)^T\mathbf{R}^{-1}(\mathbf{r}_p + \mathbf{R}\hat{\mathbf{x}}_p). \tag{12.15}$$

In summary, the GCR algorithm has been found useful for data assimilation solvers based on the representer expansion. Being applicable to non-symmetric linear systems, the solver is more tolerant of symmetry errors in the adjoint model, such as are present when the continuous adjoint equations are discretized. The GCR solver is currently being used for a variety of weak-constraint ocean data assimilation problems, and it has been implemented within the IOM data assimilation software system (Bennett et al. 2008; Muccino et al. 2008).

12.3 Diagnosis of Error Variances

The preceding analysis of the solver performance and interpretation in terms of explained variance is contingent upon having correct descriptions of the model and observation error covariances. Validation of **B** and **R** is thus of paramount importance. This section outlines the *posterior diagnosis* strategy of Desroziers and Ivanov (2001) for validating the errors **B** and **R**, with application to a large-scale operational weather analysis system, the Naval Research Laboratory Atmospheric Variational Data Assimilation System-Accelerated Representer, or (NAVDAS-AR; Xu et al. 2005; Rosmond and Xu 2006).

12.3.1 Notation and Background Materials

First, recall some established results using the notation employed here. It may be shown (Lorenc 1986) that the *analysis* $\mathbf{x}^a$, the minimizer of the objective function (12.8), is given by

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{12.16}$$

where **K** denotes the so-called *Kalman gain*,

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}. \tag{12.17}$$

At this optimum, the value of the objective function $\mathcal{J}$ is given by Bennett (1992),

$$\mathcal{J}(\mathbf{x}^a) = \mathbf{d}^T\mathbf{D}^{-1}\mathbf{d}, \tag{12.18}$$

where $\mathbf{D} = \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}$ denotes the *stabilized representer matrix*, and $\mathbf{d} = \mathbf{y} - \mathbf{H}\mathbf{x}^b$ denotes the *innovation vector*. If the background and observation errors are correctly modeled by **B** and **R**, it may be shown that the minimum value of $\mathcal{J}$ is a chi-squared random variable with m degrees of freedom (Bennett 1992),

$$E\{\mathcal{J}(\mathbf{x}^a)\} = E\{\chi_m\} = m, \tag{12.19}$$

where it is recalled that m is the number of observations, and $E\{\}$ denotes the expected value of its argument. Furthermore, Bennett et al. (2000) notes that the expected values of parts $\mathcal{J}^B$ and $\mathcal{J}^R$ of the objective function $\mathcal{J}$ are

$$E\{\mathcal{J}^B(\mathbf{x}^a)\} = Tr(\mathbf{HBH}^T\mathbf{D}^{-1}), \tag{12.20}$$

and

$$E\{\mathcal{J}^R(\mathbf{x}^a)\} = Tr(\mathbf{RD}^{-1}), \tag{12.21}$$

where $Tr(\mathbf{A})$ denotes the trace of the matrix argument $\mathbf{A}$. These results may be further specialized to compute the expected value of subsets of terms in $\mathcal{J}^B$ and $\mathcal{J}^R$ (Talagrand 1999; Desroziers and Ivanov 2001). Define $\mathbf{\Pi}_l^B$ as a projection operator such that $\mathbf{x}_l = \mathbf{\Pi}_l^B\mathbf{x}$, then the expected value of $\mathcal{J}_l^B$ associated with $\mathbf{x}_l^a$ is given by Desroziers and Ivanov (2001)

$$E\{\mathcal{J}_l^B(\mathbf{x}^a)\} = Tr(\mathbf{\Pi}_l^B\mathbf{HBH}^T\mathbf{D}^{-1}\mathbf{\Pi}_l^{B^T}). \tag{12.22}$$

Likewise, define the projection operator $\mathbf{\Pi}_k^R$ so that $\mathbf{y}_k = \mathbf{\Pi}_k^R\mathbf{y}$, then the expected value for $\mathcal{J}_k^R$ of $\mathcal{J}^R$ is

$$E\{\mathcal{J}_k^R(\mathbf{x}^a)\} = Tr(\mathbf{\Pi}_k^R\mathbf{RD}^{-1}\mathbf{\Pi}_k^{R^T}). \tag{12.23}$$

12.3.2 Validation of Error Variances by Posterior Diagnosis

Desroziers and Ivanov (2001) utilize the above relations (12.22) and (12.23) to validate the error variances in the objective function based on the *posterior diagnosis* of the assimilation system. They demonstrate how to produce realistic error variances for simulated observations in a cost-effective manner. This approach was further evaluated and developed by Chapnik et al. (2004, 2006) and Sadiki and Fischer (2005) for operational data assimilation systems. Following Chapnik et al. (2004), the objective function (12.8) is rewritten as

$$\mathcal{J}(\mathbf{x}) = \sum_{l=1}^{\nu^B} \frac{\mathcal{J}_l^B(\mathbf{x})}{\mathbf{s}_l^B} + \sum_{k=1}^{\nu^R} \frac{\mathcal{J}_k^R(\mathbf{x})}{\mathbf{s}_k^R}, \tag{12.24}$$

where $\mathbf{s}_l^B$ and $\mathbf{s}_k^R$ are scalar tuning parameters for the ν^B and ν^R components of the background and the observations, respectively. The analysis $\mathbf{x}^a(\mathbf{s})$ is now a function of the tuning parameter vector $\mathbf{s} = (\mathbf{s}_l^B, \mathbf{s}_k^R)$ (Chapnik et al. 2004),

$$\mathbf{x}^a(\mathbf{s}) = \mathbf{x}^b + \mathbf{K}(\mathbf{s})(\mathbf{y} - \mathbf{Hx}^b), \tag{12.25}$$

where the tuned *Kalman gain*, $\mathbf{K}(\mathbf{s})$, takes the form

$$\mathbf{K(s)} = \mathbf{B(s)H}^T[\mathbf{HB(s)H}^T + \mathbf{R(s)}]^{-1} = \mathbf{B(s)H}^T\mathbf{D(s)}^{-1}, \tag{12.26}$$

with $\mathbf{B(s)} = \sum_{l=1}^{\nu^B} \mathbf{s}_l^B \mathbf{\Pi}_l^{B^T} \mathbf{B}_l \mathbf{\Pi}_l^B$ and $\mathbf{R(s)} = \sum_{k=1}^{\nu^R} \mathbf{s}_k^R \mathbf{\Pi}_k^{R^T} \mathbf{R}_k \mathbf{\Pi}_k^R$. The reduced values for the sub-parts $\mathcal{J}_l^B$ and $\mathcal{J}_k^R$ of the objective function $\mathcal{J}(\mathbf{s})$ are

$$\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s})) = \mathbf{d}^T\mathbf{D}^{-1}\mathbf{H}\mathbf{\Pi}_l^{B^T}\mathbf{B(s)}\mathbf{\Pi}_l^B\mathbf{H}^T\mathbf{D}^{-1}\mathbf{d}, \tag{12.27}$$

with expected value

$$E\{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}))\} = \mathbf{s}_l^B Tr[\mathbf{\Pi}_l^B\mathbf{HB(s)H}^T\mathbf{D(s)}^{-1}\mathbf{\Pi}_l^{B^T}], \tag{12.28}$$

and

$$\begin{aligned}\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s})) &= [\mathbf{\Pi}_k^R(\mathbf{y} - \mathbf{Hx}^a(\mathbf{s}))]^T\mathbf{R}(s)^{-1}[\mathbf{\Pi}_k^R(\mathbf{y} - \mathbf{Hx}^a(\mathbf{s}))] \\ &= \mathbf{d}^T\mathbf{D(s)}^{-1}\mathbf{\Pi}_k^{R^T}\mathbf{R}(s)\mathbf{\Pi}_k^R\mathbf{D(s)}^{-1}\mathbf{d},\end{aligned} \tag{12.29}$$

with expected value

$$E\{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}))\} = \mathbf{s}_k^R Tr[\mathbf{\Pi}_k^R\mathbf{R(s)D(s)}^{-1}\mathbf{\Pi}_k^{R^T}]. \tag{12.30}$$

The criterion for the tuning parameters is that the relations

$$\mathbf{s}_l^B = \frac{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}))}{Tr[\mathbf{\Pi}_l^B\mathbf{HB(s)H}^T\mathbf{D(s)}^{-1}\mathbf{\Pi}_l^{B^T}]} \tag{12.31}$$

and

$$\mathbf{s}_k^R = \frac{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}))}{Tr[\mathbf{\Pi}_k^R\mathbf{R(s)D(s)}^{-1}\mathbf{\Pi}_k^{R^T}]} \tag{12.32}$$

are exactly satisfied. Desroziers and Ivanov (2001) proposed an iterative approach (*fixed-point algorithm*) to solve (12.31) and (12.32), namely,

$$\mathbf{s}_{l\ i+1}^B = \frac{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}_i))}{Tr[\mathbf{\Pi}_l^B\mathbf{HB}(\mathbf{s}_i)\mathbf{H}^T\mathbf{D}(\mathbf{s}_i)^{-1}\mathbf{\Pi}_l^{B^T}]} \tag{12.33}$$

$$\mathbf{s}_{k\ i+1}^R = \frac{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}_i))}{Tr[\mathbf{\Pi}_k^R\mathbf{R}(\mathbf{s}_i)\mathbf{D}(\mathbf{s}_i)^{-1}\mathbf{\Pi}_k^{R^T}]}, \tag{12.34}$$

observing that the first iteration of the fixed-point algorithm gives a good estimate of the converged results.

12.3.3 Practical Implementation and Application to NAVDAS-AR

Computation of the tuning parameters requires the evaluation of the trace of the large matrices, $Tr[\mathbf{\Pi}_l^B \mathbf{H}\mathbf{B}(\mathbf{s})\mathbf{H}^T \mathbf{D}(\mathbf{s})^{-1} \mathbf{\Pi}_l^{B^T}]$ and $Tr[\mathbf{\Pi}_k^R \mathbf{R}(\mathbf{s})\mathbf{D}(\mathbf{s})^{-1} \mathbf{\Pi}_k^{R^T}]$. Because the matrices $\mathbf{HBH}^T$ and $\mathbf{D}(\mathbf{s})^{-1}$ are not explicitly formed (Chua and Bennett 2001), the trace is computed using the randomized trace estimator (Girard 1989; Hutchinson 1989) which was used by Wahba et al. (1995) for an adaptive tuning of parameters in a numerical weather prediction application.

It is the randomized trace technique which makes feasible the posterior analysis of Desroziers and Ivanov (2001) for large-scale data assimilation, and this approach has been applied to the NAVDAS-AR. The forecast model associated with the NAVDAS-AR system is the United States Navy Operational Global Atmospheric Prediction System (NOGAPS). NOGAPS is a global spectral numerical weather prediction model (Hogan and Rosmond 1991) with 42 vertical levels and T239 spectral horizontal resolution.

The research version of NAVDAS-AR routinely assimilates conventional in situ observations (including radiosondes and pibals, and surface observations from land and sea) and satellite observations (including geostationary rapid-scan and feature-tracked winds; winds from QuikScat, WindSat, ASCAT, ERS-2, AVHRR, MODIS, SSM/I and SSMIS; and total precipitable water from WindSat, SSM/I and SSMIS). NAVDAS-AR also assimilates remotely-sensed microwave and infrared sounder radiances from AMSU-A, SSMIS, AIRS and IASI. The representation of the background error covariance matrix $\mathbf{B}$ (in (12.7)) is based on the NAVDAS 3D-Var analysis system (Daley and Barker 2001), and the observation error covariance matrix $\mathbf{R}$ is diagonal. Because the space-time error covariance $\mathbf{F}$ (in (12.7)) is set to zero, the current system is 4D-Var, rather than the W4D-Var targeted for the future.

Figure 12.4 shows the behavior of the NAVDAS-AR system based on the diagnostics: $\mathcal{J}(\mathbf{x}^a)/m$, $\mathbf{s}^B$ and $\mathbf{s}^R$. The values are computed over a 7 day period from 23 to 29 November 2008, with all available observations assimilated. If the background and observation errors are correctly modeled, one would expect $\mathcal{J}(\mathbf{x}^a)/m = \mathbf{s}^B = \mathbf{s}^R \approx 1$. The figure shows that $\mathcal{J}(\mathbf{x}^a)/m$ varies from 0.4 to 0.6 and is smaller than the expected value of 1. Also, the background errors are underestimated and the observation errors are overestimated, as shown by values of $\mathbf{s}^B$ varying from 1.8 to 2.4, and values of $\mathbf{s}^R$ varying from 0.4 to 0.6, nearly overlapping the values of $\mathcal{J}(\mathbf{x}^a)/m$. The diagnostics also indicate that the analysis system is sensitive to the number of observations (more radiosonde observations at 0 and 12 UTC than at 6 and 18 UTC), with stable values over the observation period.

The observation error tuning coefficient $\mathbf{s}^R$ may be further broken down to diagnose the observation error variances for different types of observations. Table 12.1 shows the components for temperature, wind velocity, wind speed, moisture, total precipitable water, and satellite radiances. The values indicate that the temperature standard errors should be kept unchanged, but the standard error of the zonal and meridional components of wind should be slightly reduced. Likewise, the standard

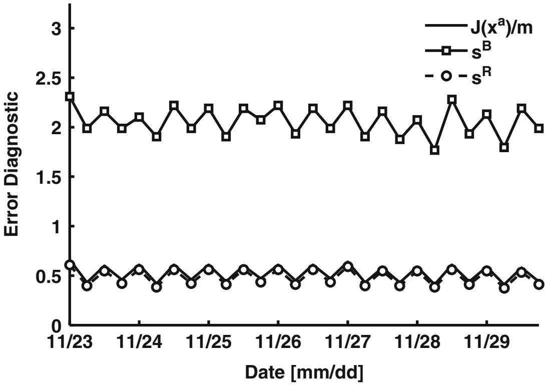


Fig. 12.4 NAVDAS-AR posterior error diagnostics. The reduced value of the objective function divided by the number of observations is consistently smaller than unity ($\mathcal{J}(\mathbf{x}^a)/m < 1$; *solid line*), its expected value if both background and observation errors are correctly scaled (12.19). Analysis of the separate background and observation errors, $\mathbf{s}^B$ (12.31) and $\mathbf{s}^R$ (12.32), respectively, shows that the background error variance is under-estimated ($\mathbf{s}^B > 1$; *solid line*, *square markers*) and the observation error variance is over-estimated ($\mathbf{s}^R < 1$; *dashed-line*, *circle markers*). The sawtooth (*up-down*) pattern in these curves is due to the twice-daily timing of radiosonde observations, resulting in twice-daily changes in the number of observations assimilated.

Table 12.1 Tuning coefficients

Obs-type	TEMP	UWIND	VWIND	WINDSPD	H2O	TPW	RADIANCE
$\mathbf{s}_k^R$	1.15	0.72	0.72	0.23	1.46	0.29	0.28

TEMP tuning coefficients for temperature, *UWIND* zonal wind, *VWIND* meridional wind, *WINDSPD* wind speed, *H2O* moisture, *TPW* total precipitable water, and *RADIANCE* satellite radiances

error for wind-speed, total precipitable water, and radiances should be adjusted downward. In contrast, the standard error for moisture data should be increased.

12.4 Summary

Variational data assimilation systems based on representer-based solution methods are being used to perform analyses and prediction in the ocean and atmosphere. One such weather prediction system, NAVDAS-AR, is currently in operational use (Xu et al. 2005; Rosmond and Xu 2006).

The inner iterative linear solvers at the core of these systems may display non-monotonic convergence in the norm defined by the primal objective function, and this behavior makes problematic the development of practical stopping criteria. One approach to this problem has been described, namely, using an inner solver that permits more diagnostics of the solution progress and objective function to be computed during the minimization. The generalized conjugate residual (GCR) algorithm provides these diagnostics, at the cost of some additional complexity compared with the conjugate gradient algorithm, but it performs reliably when the approximate adjoint of the model is used.

The analysis produced by any data assimilation system is always limited by the quality of the prior covariance models for the background, model forcings, and observations. In Sect. 12.3 it was shown how the posterior error analysis of Desroziers and Ivanov (2001) could be applied to calibrate these covariance models in variational data assimilation systems using representer-based solvers. Application of these methods has been applied to diagnose the observation error in NAVDAS-AR, which utilizes many sources of atmospheric data, each with unique error characteristics.

Acknowledgements Zaron was sponsored by the National Science Foundation (NSF), award OCE-0623540, with additional support from the Naval Research Laboratory, award N00173-08-2-C015. Authors Chua, Xu, Baker, and Rosmond gratefully acknowledge the support of their sponsors, the Naval Research Laboratory, the Office of Naval Research, and the PMW-120, under program elements, 0602435N and 0603207N, respectively. Computational resources for Zaron were provided by the National Center for Atmospheric Research, which is sponsored by NSF.

References

Bennett AF (1985) Array design by inverse methods. Prog Oceanogr 15:129–156

Bennett AF (1992) Inverse methods in physical oceanography, 1st edn. Cambridge University Press, New York, 346p

Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, New York, 234p

Bennett AF, Thorburn MA (1992) The generalized inverse of a nonlinear quasigeostrophic ocean circulation model. J Phys Oceanogr 22:213–230

Bennett AF, Chua BS, Harrison DE, McPhaden MJ (2000) Generalized inversion of tropical atmosphere–ocean (TAO) data and a coupled model of the tropical ocean. Part II: the 1995–96 La Niña and 1997–98 El Niño. J Climate 13:2770–2785

Bennett AF, Chua BS, Pflaum BL, Erwig M, Fu Z, Loft RD, Muccino JC (2008) The inverse ocean modeling system. I: implementation. J Atmos Oceanic Technol 25:1608–1622

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 130:2253–2275

Chapnik B, Desroziers G, Rabier F, Talagrand O (2006) Diagnosis and tuning of observational error in quasi-operational data assimilation setting. Q J R Meteorol Soc 132:543–565

Chua B, Bennett AF (2001) An inverse ocean modeling system. Ocean Model 3:137–165

Cohn SE, Da Silva A, Guo J, Sienkiewicz M, Lamich D (1998) Assessing the effects of data selection with the DAO physical-space statistical analysis system. Mon Weather Rev 126:2913–2926

Courtier P (1997) Dual formulation of four-dimensional assimilation. Q J R Meteorol Soc 123:2449–2461

Courtier P, Thepaut J, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. Q J R Meteorol Soc 120:1367–1387

Daley R, Barker E (2001) NAVDAS: formulation and diagnostics. Mon Weather Rev 129:869–883 de Sturler E (1994) Iterative methods on distributed memory computers. PhD thesis, Delft University of Technology, Delft, the Netherlands

de Sturler E (1996) Nested Krylov methods based on GCR. J Comput Appl Math 67:15–41

Desroziers G, Ivanov S (2001) Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. Q J R Meteorol Soc 127:1433–1452

El Akkraoui A, Gauthier P (2010) Convergence properties of the primal and dual forms of variational data assimilation. Q J R Meteorol Soc 136:107–115

El Akkraoui A, Gauthier P, Pellerin S, Buis S (2008) Intercomparison of the primal and dual formulations of variational data assimilation. Q J R Meteorol Soc 134:1015–1025

Girard D (1989) A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. Numer Math 56:1–23

Giraud L, Ruiz D, Touhami A (2006) A comparative study of iterative solvers exploiting spectral information for spd systems. SIAM J Sci Comput 27:1760–1786

Golub G, Van Loan C (1989) Matrix computations, 2nd edn. Johns Hopkins University Press, Baltimore, 642p

Gratton S, Tshimanga J (2009) An observation-space formulation of variational assimilation using a restricted preconditioned conjugate gradient algorithm. Q J R Meteorol Soc 135:1573–1585

Hogan T, Rosmond T (1991) The description of the Navy Operational Global Atmospheric Prediction System's spectral forecast model. Mon Weather Rev 119:1786–1815

Hutchinson MF (1989) A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Commun Stat Simul Comput 18:1059–1076

Lorenc A (1986) Analysis methods for numerical weather prediction. Q J R Meteorol Soc 112:117–1194

Muccino JC, Arango H, Bennett AB, Chua BS, Cornuelle B, DiLorenzo E, Egbert GD, Hao L, Levin J, Moore AM, Zaron ED (2008) The inverse ocean modeling system. II: applications. J Atmos Oceanic Technol 25:1623–1637

Parks ML, de Sturler E, Mackey G, Johnson DD, Maiti S (2006) Recycling Krylov subspaces for sequences of linear systems. SIAM J Sci Comput 28:1651–1674. doi:10.1137/040607277

Rosmond T, Xu L (2006) Development of NAVDAS-AR: nonlinear formulation and outer loop tests. Tellus 58A:45–58

Sadiki W, Fischer C (2005) A posteriori validation applied to the 3D-var Arpege and Aladin data assimilation systems. Tellus 57A:21–34

Talagrand O (1999) A posterior verification of analysis and assimilation algorithms. In: Proceedings of a workshop on diagnosis of data assimilation systems, ECMWF, Reading, UK

Wahba G, Johnson DR, Gao F, Gong J (1995) Adaptive tuning of numerical weather prediction models: randomized GCV in three- and four-dimensional data assimilation. Mon Weather Rev 123:3358–3369

Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR: formulation and initial tests of the linear problem. Tellus 57:546–559

Zaron ED (2006) A comparison of data assimilation methods using a planetary geostrophic model. Mon Weather Rev 134:1316–1328

Zaron ED, Chavanne C, Egbert GD, Flament P (2009) Baroclinic tidal generation in the Kauai Channel inferred from HF-Radar. Dyn Atmos Oceans 48:93–120. http://dx.doi.org/10.1016/j.dynatmoce.2009.03.002

Chapter 13
Variational Data Assimilation for the Global Ocean

James A. Cummings and Ole Martin Smedstad

Abstract A fully three dimensional, multivariate, variational ocean data assimilation system has been developed that produces simultaneous analyses of temperature, salinity, geopotential and vector velocity. The analysis is run in real-time and is being evaluated as the data assimilation component of the Hybrid Coordinate Ocean Model (HYCOM) forecast system at the U.S. Naval Oceanographic Office. Global prediction of the ocean weather requires that the ocean model is run at very high resolution. Currently, global HYCOM is executed at 1/12 degree resolution (∼7 km mid-latitude grid mesh), with plans to move to a 1/25 degree resolution grid in the near future (∼3 km mid-latitude grid mesh). These high resolution global grids present challenges for the analysis given the huge model state vector and the ever increasing number of satellite and in situ ocean observations available for the assimilation. In this paper the development and evaluation of the new oceanographic three-dimensional variational (3DVAR) data assimilation is described. Special emphasis is placed on documenting the capabilities built into the 3DVAR to make the system efficient for use in global HYCOM.

13.1 Introduction

Eddy-resolving global ocean prediction requires high resolution since the characteristic scale of ocean eddies is on the order of a few tens of kilometers. Only recently have sufficient data and computer power become available to nowcast and forecast the ocean weather at eddy-resolving scales, including processes that control the surface mixed layer, the formation of ocean eddies, meandering ocean

J.A. Cummings (✉)
Oceanography Division, Naval Research Laboratory, Monterey, CA, USA
e-mail: cummings@nrlmry.navy.mil

O.M. Smedstad
QinetiQ North America, Stennis Space Center, MS, USA

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_13,

currents and fronts, and generation and propagation of coastally trapped waves. Hurlburt et al.(2008a) gives a good discussion of the requirements for an ocean model to be eddy-resolving. High resolution global ocean forecast models present challenges for the assimilation component of the forecasting system given the huge model state vector and the ever increasing number of satellite and in situ ocean observations available for the assimilation. Accordingly, the global analysis has to be both computationally efficient and accurate to account for the oceanographic features resolved by the high resolution model. At the same time the analysis must use all of the available observations and create and maintain dynamically adjusted corrections to the model forecast.

The purpose of this chapter is to provide an overview of a new variational ocean data assimilation system that has been developed as an upgrade to an existing multivariate optimum interpolation (MVOI) system (Cummings 2005). Compared to the MVOI the 3DVAR algorithm has several advantages. First, the 3DVAR performs a global solution that does not require data selection. In the MVOI, observations are organized into overlapping analysis volumes and the solution can depend on how the volumes are defined. This is not the case in the 3DVAR, as the global solve allows all observations to influence all grid points, a requirement for an optimum analysis. Second, through the use of observation operators, 3DVAR can incorporate observed variables that are different from the model prognostic variables. Examples of this in the ocean are integral quantities, such as acoustic travel time and altimeter measures of sea surface height, and direct assimilation of satellite radiances of sea surface temperature (SST) through radiative transfer modeling. Finally, 3DVAR permits more powerful and realistic formulations of the background error covariances, which control how information is spread from the observations to the model grid points and model levels. The error covariances also ensure that observations of one model variable produce dynamically consistent corrections in the other model variables.

The 3DVAR referred to in this paper is the Navy Coupled Ocean Data Assimilation (NCODA) system, version 3. NCODA 3DVAR is in operational use at the U.S. Navy oceanographic production centers: Fleet Numerical Meteorology and Oceanography Center (FNMOC) in Monterey, CA, and the Naval Oceanographic Office (NAVOCEANO) at the Stennis Space Center, MS. NCODA is truly a unified and flexible oceanographic analysis system. It is designed to meet all Navy ocean data analysis and assimilation requirements using the same code. In two-dimensional mode, NCODA provides SST and sea ice concentration analyses for lower boundary conditions of the Navy global and regional atmospheric forecast models. In three-dimensional mode, it is executed in a sequential incremental update cycle with the Navy ocean forecast models: the Hybrid Coordinate Ocean Model (HYCOM) on the global scale, and the Navy Coastal Ocean Model (NCOM) on the regional scale. Here, NCODA provides updated initial conditions of ocean temperature, salinity, and currents for the next run of the ocean forecast model. The analysis background fields, or first guess, are generated from a short-term ocean model forecast, and the 3DVAR computes dynamically consistent corrections to the first-guess fields using all of the observations that have become available since the last analysis was made. Further, NCODA 3DVAR is globally relocatable and has been integrated into the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS®[1]), which is used by Navy for rapid environmental assessment. In this mode of operation, the 3DVAR performs multi-scale analyses on nested, successively higher resolution grids. Finally, NCODA provides the data assimilation component for the WAVEWATCH wave model forecasting system at FNMOC (Wittmann and Cummings 2005). In this mode of operation, NCODA computes corrections to the model's two-dimensional wave spectra from assimilation of satellite altimeter and wave buoy observations of significant wave height.

The examples used in the paper are taken from NCODA 3DVAR analyses cycling with global HYCOM. Sections 13.2 and 13.3 of the paper describe the assimilation method and techniques used to specify the error covariances. Section 13.4 lists the ocean observing systems assimilated and outlines the data selection and data pre-processing that is done for the real-time global forecast. Section 13.5 gives an overview of the entire NCODA system, including the diagnostic suite. Section 13.6 presents some verification results from global HYCOM. Section 13.7 describes future capabilities and applications of the NCODA 3DVAR system, while Sect. 13.8 gives a summary.

13.2 Method

The method used in NCODA is an oceanographic implementation of the Navy Variational Atmospheric Data Assimilation System (NAVDAS), a 3DVAR technique developed for Navy numerical weather prediction (NWP) systems (Daley and Barker 2001). The oceanographic 3DVAR analysis variables are temperature, salinity, geopotential (dynamic height), and u, v vector velocity components. All ocean variables are analyzed simultaneously in three dimensions. The horizontal correlations are multivariate in geopotential and velocity, thereby permitting adjustments to the mass fields to be correlated with adjustments to the flow fields. The velocity adjustments (or increments) are in geostrophic balance with the geopotential increments, which, in turn, are in hydrostatic agreement with the temperature and salinity increments. The multivariate aspects of the 3DVAR assimilation are discussed further in Sect. 13.3.3.

The NCODA 3DVAR problem is formulated as:

$$x_a = x_b + P_b H^T (H P_b H^T + R)^{-1} [y - H(x_b)] \tag{13.1}$$

where $\mathbf{x_a}$ is the analysis vector, $\mathbf{x_b}$ is the background vector, $\mathbf{P_b}$ is the positive-definite background error covariance matrix, $\mathbf{H}$ is the forward operator, $\mathbf{R}$ is the observation error covariance matrix, and $\mathbf{y}$ is the observation vector. At the present

[1] COAMPS® is a registered trademark of the Naval Research Laboratory

time, the forward operator in NCODA is spatial interpolation performed in three dimensions by fitting a surface to a 4 × 4 × 4 grid point target and evaluating the surface at the observation location. Thus, $\mathbf{HP_bH^T}$ is approximated directly by the background error covariance between observation locations, and $\mathbf{P_bH^T}$ directly by the error covariance between observation and grid locations. For the purposes of discussion, the quantity $[\mathbf{y}-\mathbf{H}(\mathbf{x_b})]$ is referred to as the innovation vector, $[\mathbf{y}-\mathbf{H}(\mathbf{x_a})]$ is the residual vector, and $\mathbf{x_a}$-$\mathbf{x_b}$ is the increment (or correction) vector.

The observation vector contains all of the synoptic temperature, salinity and velocity observations that are within the geographic and time domains of the forecast model grid and update cycle. Observations can be assimilated at their measurement times within the update cycle time window by comparison against time dependent background fields using the first guess at appropriate time (FGAT) method. An advantage of the FGAT method is that it eliminates a component of the mean analysis error that occurs when comparing observations and forecasts not valid at the same time. As will be described in Sect. 13.6, the use of FGAT in real-time HYCOM allows for assimilation of late receipt observations.

Equation (13.1) is the observation space form of the 3DVAR equation. A dual form of the 3DVAR is the analysis space algorithm, which is defined by the model state vector on some regular grid. Courtier (1997) has shown that the two formulations are equivalent and give the same solution. However, as discussed by Daley and Barker (2000, 2001), there are advantages to the use of an observation space approach in Navy ocean model applications. In the observation space algorithm the matrix to be inverted $(\mathbf{HP_bH^T} + \mathbf{R})^{-1}$ is dimensioned by the number of observations, while in the analysis space algorithm the matrix to be inverted is dimensioned by the number of grid locations. Given the high dimensionality of global ocean forecast model grids, and the relatively sparse ocean observations available for the assimilation, an observation space 3DVAR algorithm will have a clear computational advantage. Further, an observation space system is more flexible and can more easily be coupled to many prediction models. As has been discussed, NCODA must work equally well with multiple atmospheric and oceanographic prediction systems in a wide variety of applications, as well as a wave model prediction system. Finally, due to the local nature of the observation space algorithm, the background error covariances are multivariate and can be formulated and generalized in a straightforward manner. As will be shown, this aspect of the 3DVAR is an important feature of NCODA. On the other hand, analysis space systems typically restrict the background error covariances to be sequences of univariate, one-dimensional digital filters, thereby ignoring the inherent multivariate nature of the background error correlations.

Solution of the observation space 3DVAR problem is done in two steps. First, the equation,

$$(HP_bH^T + R)z = [y - H(x_b)] \tag{13.2}$$

is solved for the vector $\mathbf{z}$. Next, a post-multiplication step is performed by left-multiplying $\mathbf{z}$ using,

$$x_a - x_b = P_bH^Tz \tag{13.3}$$

to obtain the correction field in grid point space. A pre-conditioned conjugate gradient descent algorithm is used to solve (13.2) using block diagonal pre-conditioners. The blocks are defined by decomposing the analysis grid into non-overlapping partitions of a regular quilt laid over the analysis domain in model grid point (i, j) space. The use of i, j blocks rather than latitude-longitude blocks allows the analysis to be completely grid independent. The flexibility of this approach is shown in Fig. 13.1 for the global HYCOM Atlantic basin analysis (see Sect. 13.6 and Fig. 13.9 for a discussion of the HYCOM basins). A total of 1,935, 2,436, and 1,147 blocks are defined for the global HYCOM Atlantic, Indian, and Pacific analysis regions, respectively, which use Mercator grid projections. Observations are sorted into the blocks and the pre-conditioning matrix is formed from a Choleski decomposition of the correlations between observations in the same block. The Choleski decomposed block matrices are calculated once and stored before application of the conjugate gradient descent algorithm. Solution of the pre-conditioned conjugate gradient for the vector **z** n (13.2) typically converges in 6–10 iterations. Determination of convergence is based on the norm of the gradient of the cost function estimated at each iteration step. This gradient is a vector the size of the number of observations and the norm is the square root of the sum of the elements, which are the residuals of the fit of the analysis to the innovations. In practice, convergence is reached when the norm of the gradient is reduced by 2 orders of magnitude. This is considered to be sufficient because an increase in the number of iterations only fits small-scale features. This may appear to be beneficial, but it must be noted that the post-multiplication step is a spatially smoothing operation when the background error correlations are applied. Thus, the extra iterations in the solver required to resolve small-scale features in the observations do not have much effect on the final analyzed increment field because of the smoothing effect of the post-multiplier.

Observation space 3DVAR algorithms converge quickly because very good pre-conditioners can be developed. In fact, the pre-conditioner used in NCODA is perfect. For example, NCODA is configured such that when the data count is less than 2,000 the observation data block is defined as the entire analysis domain. When this global pre-conditioned data block is presented to the conjugate gradient solver the algorithm converges in a single iteration. No further work by the solver is necessary. This sparse data pathway through the code is often encountered when NCODA 3DVAR is executed on nested grids in the relocatable coupled model system where the innermost grid represents a small geographic area.

As noted by Daley and Barker (2001), partitioning of the observations into blocks has no effect on the final solution. The NCODA 3DVAR formulation is guaranteed to include correlations between all observations in all blocks, thereby achieving a global solution. After the vector **z** is obtained it is post-multiplied by $\mathbf{P_b H^T}$ to create the analysis correction fields for each analysis variable. This step is performed using blocks in grid space that are defined differently from the observation blocks used to compute the solution vector **z**. To accommodate high resolution ocean model forecast grids and minimize computer memory resource requirements for the analysis, the grid space blocks are defined smaller by simply

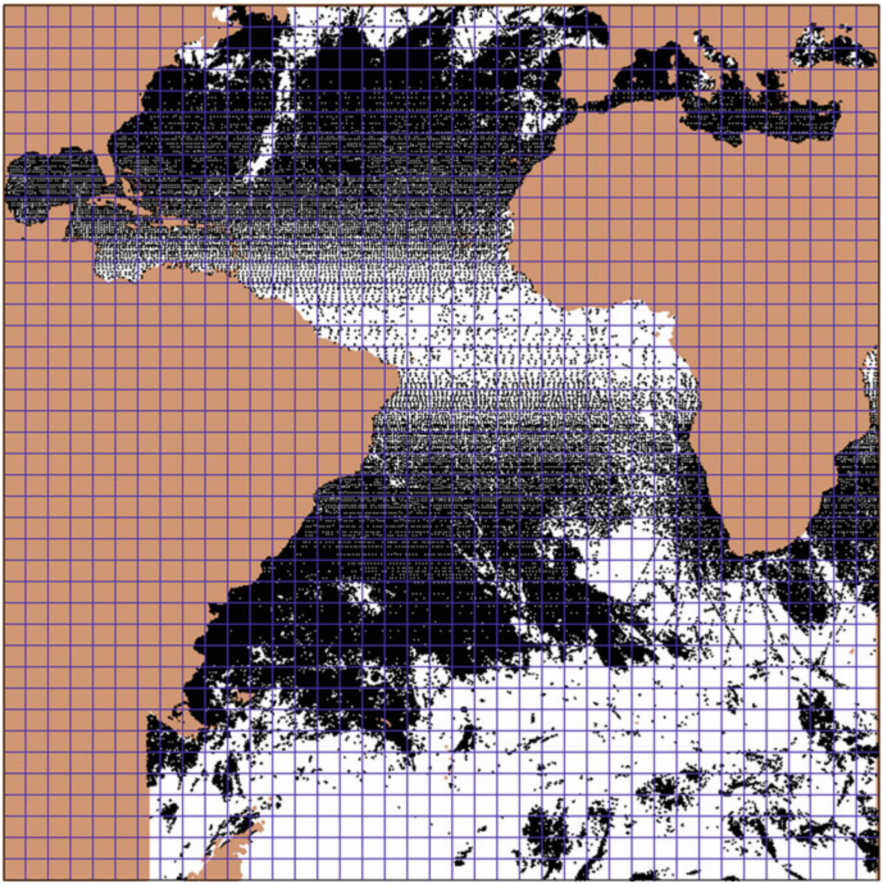


Fig. 13.1 Observation data blocks for HYCOM Atlantic basin grid. *Blue lines* give observation block edges; observation locations are indicated by *black dots*. A total of 1,935 data blocks are defined (43 in the X direction, 45 in the Y direction)

sub-setting the previously defined observation blocks. Again, it must be emphasized that partitioning the grid domain into blocks in the post multiplication does not affect the final results. The correction fields are guaranteed to contain the correlations between all observations and all grid points, thereby creating a seamless and continuous analysis.

Parallelization of the 3DVAR algorithm is achieved in three ways. The first parallelization is done over the observation-defined blocks in the pre-conditioner, the second parallelization is done over observation-defined blocks in the conjugate gradient solver, and the third parallelization is done over grid point-defined blocks in the post-multiplication step (mapping from observation space to grid space). The number of processors used to execute the 3DVAR can be as few as one or as many as the maximum number of observation or grid node blocks. A load balancing algorithm is used to spread the work related to the block-dependent calculations out evenly across the processors. In the conjugate gradient descent step, the work load for an observation block is calculated as the sum of the observation-observation interactions. In the post-multiplication step, the work estimate is based on the sum of the observation-grid point interactions. Observation and grid point blocks are determined to be close enough to contribute to the solution if the block centers are within 8 correlation length scales. Thus, for a given block size, the number of observation-observation and observation-grid point block interactions varies with the horizontal correlation length scales and will be more numerous where length scales are long. Further efficiency is achieved by keeping communication among the processors minimal. To do this matrix elements are calculated, stored, and used on each processor, they are never passed between processors. Only elements of the solution and correction vectors scattered across the processors have to be communicated and reassembled and, in the case of the solution vector, broadcast for the next iteration. Note that memory utilization for the conjugate gradient solver in the 3DVAR is reduced as the number of processors is increased. This feature allows the 3DVAR to scale very well across many processors on large machines, and run equally well on small platforms with limited memory.

13.3 Error Covariances

Specification of the background and observation error covariances in the assimilation is very important. As previously noted, the background error covariances control how information is spread from the observations to the model grid points and model levels, but they also ensure that observations of one model variable produce dynamically consistent corrections in the other model variables. The background error covariances in the NCODA 3DVAR are similar to the error covariances defined for the MVOI, but with some notable exceptions. As in the MVOI, the error covariances in the 3DVAR are separated into a background error variance and a correlation. The correlation is further separated into a horizontal ($\mathrm{C_h}$) and a vertical ($\mathrm{C_v}$) component. Correlations are modeled as either second order auto-regressive (SOAR) functions of the form,

$$\begin{aligned} C_h &= (1 + s_h)\exp(-s_h) \\ C_v &= (1 + s_v)\exp(-s_v) \end{aligned} \tag{13.4}$$

or Gaussian functions of the form,

$$\begin{aligned} C_h &= \exp(-s_h^2) \\ C_v &= \exp(-s_v^2) \end{aligned} \tag{13.5}$$

where $\mathrm{s_h}$ and $\mathrm{s_v}$ are the horizontal and vertical distances between observations or observations and grid points, normalized by the arithmetic mean of the horizontal or the vertical correlation length scales at the two locations. The horizontal correlation length scales vary with location and the vertical correlation length scales vary with depth and location in the analysis. As described in the subsequent sections, both correlation components evolve with time in accordance with information obtained from the model forecast background valid at the update cycle interval.

13.3.1 Horizontal Correlations

The horizontal correlation length scales are set proportional to the first baroclinic Rossby radius of deformation using estimates computed from the historical profile archive by Chelton et al. (1998). Rossby length scales qualitatively characterize scales of ocean variability and vary from 10 km at the poles to greater than 200 km in the tropics. The Rossby length scales increase rapidly near the equator which allows for stretching of the zonal scales in the equatorial wave guide. Flow-dependence is introduced in the analysis by modifying the horizontal correlations with a tensor computed from forecast model sea surface height (SSH) gradients. The flow-dependent tensor spreads innovations along rather than across the SSH contours, which are used as a proxy for the circulation field. Flow dependence is a desirable outcome in the analysis, since error correlations across an ocean front are expected to be characteristically shorter than error correlations along the front. Note that other gradient fields can be used as a flow-dependent tensor in the analysis, such as SST or potential vorticity (Martin et al. 2007). The flow dependent correlation tensor (C_f) is computed using either a SOAR or Gaussian model defined in (13.4) and (13.5), where the SSH difference between two locations is normalized by a scalar that defines the strength of the flow dependence. Because the flow dependent correlations are computed directly from the forecast SSH fields they depend strongly on the accuracy of the model forecast. This dependence may prove not to be very useful in practice if the forecast model fields are inaccurate. Accordingly, the normalization scalar can be set to a relatively large value in order to reduce the strength of the flow dependence in the analysis and prevent a model with systematic errors from adversely affecting the analysis. Alternatively, the flow dependence can be switched completely off. Figure 13.2 shows a zoom of the analysis increments off South Africa from a global high resolution SST analysis executed using a 6-h update cycle. The analysis has 12-km resolution at the equator, 9-km mid-latitude, and is a FNMOC contribution to the Group for High Resolution SST (GHRSST). Background SST gradients are used as the flow dependent tensor, with the result that the SST analysis increments are constrained by the meanders and eddies associated with the Agulhas retroflection current. The increments are both positive and negative along the front and eddy locations, indicating that application of the flow dependent tensor is a relatively weak constraint and the strength and position of features can change from one update cycle to the next in the analysis.

To account for the discontinuous and non-homogeneous influence of coastlines in the analysis a second tensor is introduced (C_l) that rotates and stretches horizontal

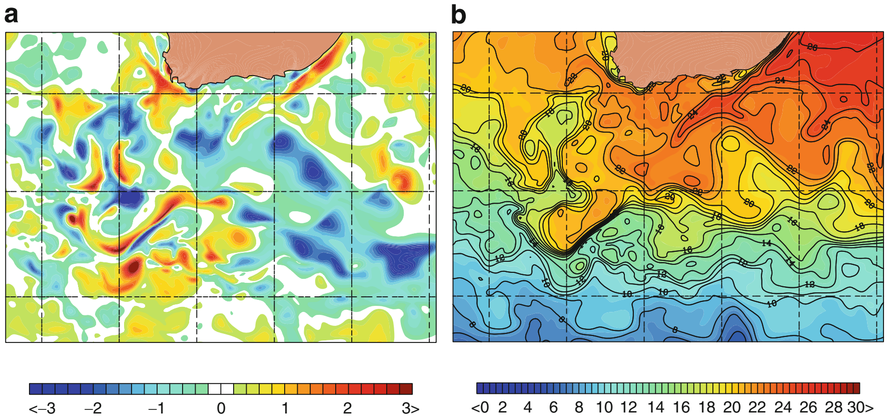


Fig. 13.2 Analysis example of flow dependent tensor based on SST gradients in Agulhas Current region; scalar value defining gradient strength of flow dependence set to 0.5°C. (**a**) analyzed increments; (**b**) analyzed SST field

correlations along the coast while minimizing or removing correlations into the coast. First, all observations and model grid points are assigned an orthogonal distance to land value based on a 1-km global coastline database. Land distances greater than some minimal value (say, 20 km) are set to the minimal value. This operation results in land distance gradients greater than zero along coastlines and zero elsewhere. Similar to the flow dependence tensor, the coastline tenor is then calculated using the difference in land distance between two locations normalized by a scalar that specifies the strength of the coastline dependence. Away from the coast (>20 km) this difference is zero resulting in no modification of the horizontal correlations. However, in the vicinity of the coast (<20 km) land distance differences are non-zero, resulting in $\mathrm{C_l} < 1$ and a modification of the horizontal correlations. Background error correlations close to the coast are expected to be anisotropic because horizontal advection from coastal currents will elongate the corrections and spread the information along the coast. Figure 13.3 illustrates the coastline tensor applied to an observation ~5 km from the coast in Monterey Bay. In this example, the horizontal correlations are specified as homogenous with a length scale of 30 km. The effect of the coastline tensor is clearly seen as the correlations adjust to prominent coastal features like the Monterey peninsula to the south and the rotation of the coastline to an east–west orientation north of the observation location.

The total horizontal background error correlation ($\mathrm{C_b}$) is then computed as the product of the two correlation components and the two correlation tensors according to,

$$C_b = C_h C_v C_f C_l \tag{13.6}$$

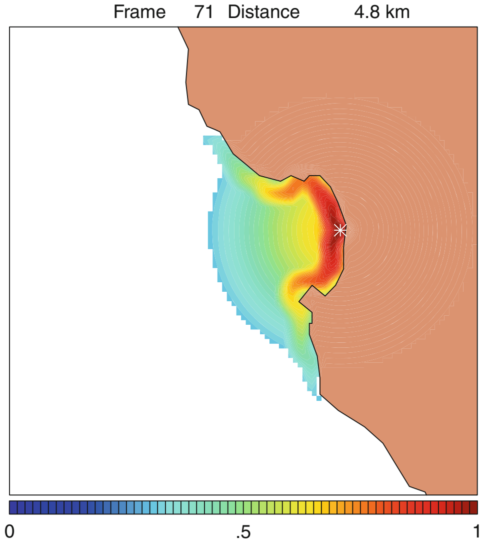


Fig. 13.3 Example of land distance correlation tensor for point 4.8 km from coast in Monterey Bay, California, USA. Observation point is given by white X mark. Horizontal length scales are assumed homogenous at 30 km. The land distance tensor spreads the correlations from the observation point along the contours of the Monterey Bay coastline

13.3.2 Vertical Correlations

Vertical correlation length scales vary with location and depth and evolve from one analysis cycle to the next in the 3DVAR. They are defined on the basis of either: (1) background density vertical gradients in pressure space, or (2) background density differences in isopycnal space. In the vertical density gradient option, a change in density stability criterion is used to define a well-mixed layer. The change in density criterion is then scaled by the background vertical density gradient at each grid location and grid level according to,

$$h_v = \rho_s / (\partial\rho / \partial z) \tag{13.7}$$

where h_v is the vertical correlation length scale, ρ_s is the change in density criterion ($\sim$0.15 kg m^{-3}), and $\partial\rho/\partial z$ is the vertical density gradient. Surface mixed layer depths, calculated at each grid point using the same change in density criteria (Karra et al. 2000), are spliced onto the three-dimensional vertical length scale field computed using (13.7). With this modification, surface-only observations decorrelate at the base of the spatially varying mixed layer. The vertical density gradient correlations are computed each update cycle from the background density fields, thereby allowing the vertical scales to evolve with time and capture changes in mixed layer, thermocline depths, and the formation of mode waters. Overall, the method produces vertical correlation length scales that vary with depth and location, and are long when the water column stratification is weak and short when the water column is strongly stratified.

In the isopycnal option, observation or grid point differences in density are scaled by ρ_s to form a correlation. This procedure essentially derives the vertical correlations relative to a density vertical coordinate. Observations are more correlated along an isopycnal than across an isopycnal, which introduces considerable flow dependence into the correlations. The procedure is cost free and does not require a transformation of the model background to isopycnal coordinates. All that is needed is knowledge of the density for any point of interest, which can be obtained from the observation itself or the model forecast. Use of the isopycnal vertical correlation option is ideally suited for HYCOM, since each coordinate surface in the model is assigned a reference isopycnal. Vertical correlation defined along isentropic surfaces is well known in atmospheric data assimilation (e.g., Riishøjgaard 1998). Note that vertical correlations in the analysis are calculated either via a SOAR, (13.4) or Gaussian, (13.5) function using lengths scales derived from either the vertical density gradient or isopycnal formulations.

Figure 13.4 gives cross sections through the vertical correlation length scale field and the model density field for the HYCOM Pacific domain (Sect. 13.6). The length scales were computed using the vertical density gradient option with $\rho_s = 0.15$. The cross sections extend from the coast of Japan at 42°N, 140°E along a great circle path to the equator at 0°N, 160°E. Figure 13.4a shows vertical correlation length scales shorter near the surface and longer at depth in agreement with the density stratification (Fig. 13.4b). The influence of the Kuroshio front is clearly seen, with longer length scales at increasingly shallower depths as the permanent thermocline shoals towards the equator. Relatively longer length scales are also seen in the 17–19°C mode-water layer immediately south of the Kuroshio, which has relatively uniform density at depths of 200–400 m.

13.3.3 Multivariate Correlations

The horizontal and vertical correlation functions described above are used in the analysis of temperature, salinity, and geopotential. Temperature and salinity are analyzed as uncorrelated scalars, while the analysis of geopotential is multivariate with velocity. Geopotential is computed in the analysis from vertical profiles of temperature and salinity by integrating the specific volume anomaly (Fofonoff and Millard 1983) from a level of no motion (2,000 m depth) to the surface. The multivariate correlations require specification of a parameter γ, which measures the divergence permitted in the velocity correlations, and a parameter φ, which specifies the strength of the geostrophic coupling of the velocity/geopotential correlations. Typically, γ is set to a small, constant value ($\gamma = 0.05$) that produces weakly divergent velocity increments and assumes that the divergence is not correlated with changes in the mass field. The geostrophic coupling parameter φ varies with location from 0 to 1. It is scaled to zero within 1° of latitude from the equator, where geostrophy is not defined, and in shallow water (<50 m deep), where friction rather than pressure gradient forces control ocean flow. The multivariate correlations

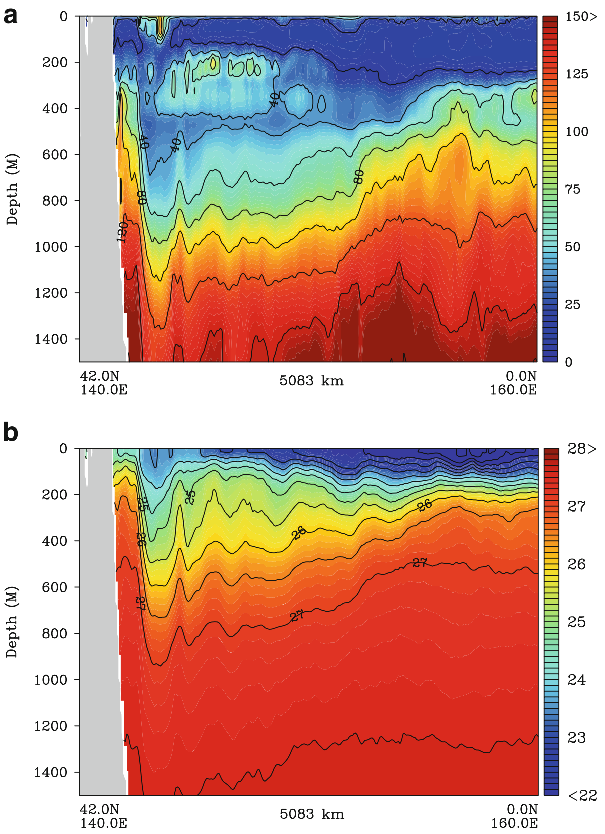


Fig. 13.4 Cross sections of vertical correlation length scales and density from Pacific basin run of global HYCOM. (**a**) Vertical length scales (m); (**b**) Density (kg/m^3)

also include auto- and cross-correlations of the u, v vector velocity components. However, at the present time, there are no operational sources of ocean current observations available for the assimilation, although the capability to assimilate velocity data is built into the 3DVAR system. A full derivation of the multivariate horizontal correlations is provided in Daley (1991). The multivariate correlations are derived from the first and second derivatives of the SOAR (or Gaussian) model function and require precise calculation of the angles between any two locations in order to guarantee a symmetric correlation matrix.

13.3.4 Background Error Variances

Background error variances are poorly known in the ocean and are likely to be strongly dependent on model resolution and other factors, such as atmospheric model forcing errors and ocean model parameterization errors. In the analysis, the background error variances ($\mathbf{e_b^2}$) vary with location, depth, and analysis variable. The variances are computed prior to an analysis from a weighted time history of differences in forecast fields valid at the update cycle interval and issued from a series of analyzed states according to,

$$e_b^2 = \sum_{k=1}^{n} w_k (x_k - x_{k-1})^2 \tag{13.8}$$

where $\mathbf{x_k} - \mathbf{x_{k-1}}$ are the differences in model forecasts (indices indicating grid location and depth are omitted for clarity), k is the update cycle index, n is the number of update cycles into the past to use in the summation, and $\mathrm{w_k}$ is a weight vector computed using a geometric series, $\mathrm{w_k} = (1 - \phi)^{\mathrm{k}-1}$, where ϕ is typically set to 0.1. The background error variances computed according to (13.8) are normalized such that the weighted averages are unbiased. In practice, the background error variances tend to evolve to a quasi-steady state over time. The model forecast difference fields include the influence of observations from the assimilation, so in well observed areas the background errors are consistent with the innovations (model-data errors at the update cycle interval). However, in the case of poorly observed or strong flow areas the background error variances are more likely dominated by model variability arising from atmospheric forcing and baroclinic and barotropic instabilities. Figure 13.5 shows background temperature error standard deviation computed using Eq. (13.8) for different vertical levels in the global HYCOM analysis domains (see Sect. 13.6). Figure 13.6 shows the background salinity error standard deviation and Fig. 13.7 the background velocity error standard deviation at the surface. Relatively high background errors are evident at all depths in boundary current areas: Gulf Stream, Kuroshio, Agulhas, Brazil-Malvinas, East Australia. Surface salinity error levels are also large near some river outflow areas, in tropical regions, and in the marginal ice zone around Antarctica during the Austral summer. Surface velocity error standard deviations tend to be large in western boundary currents and in the inter-tropical convergence zone (ITCZ) due to the variable wind and solar forcing in that area.

The adaptive scheme implemented here is designed to provide background errors that: (1) are appropriate for the time interval at which data are inserted into the model; (2) are coherent with the variance of the innovation time series; (3) reflect the variable skill of the different ocean forecast models that are used with the analysis system; and (4) adjust quickly to new ocean areas when the analysis is re-located in a rapid environmental assessment mode of operation. One difficulty with this approach is that differences in model fields contain a mixture of forecast and analysis error. Forecast errors result from initial condition, model, and atmospheric forcing deficiencies, while analysis errors result from the fact that the statistical parameters used in the analysis represent expected values and are unlikely to be correct at all places and at all times.

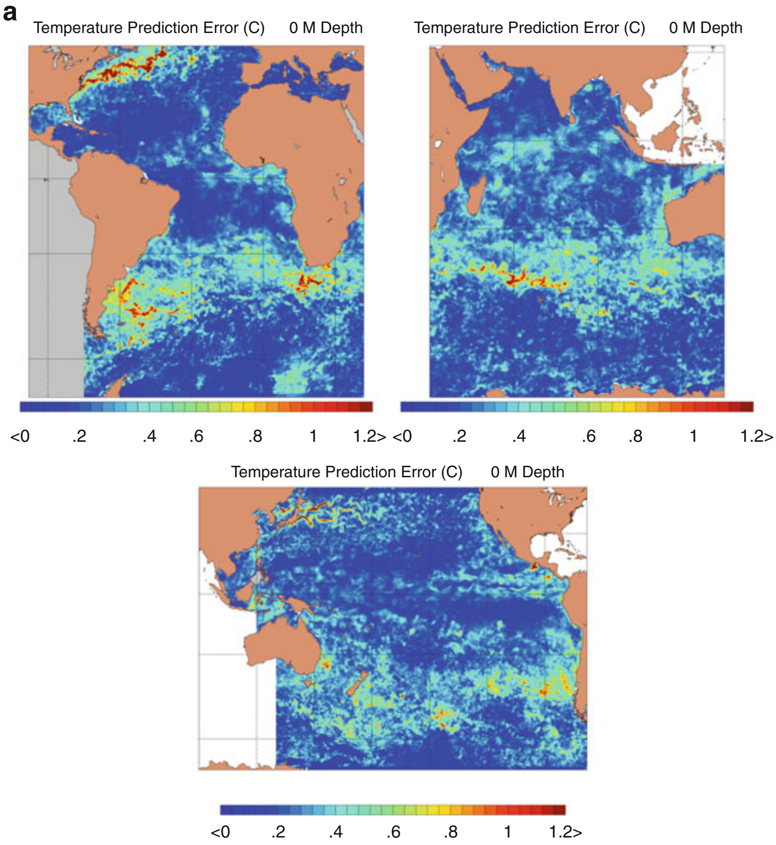

a
Temperature Prediction Error (C) 0 M Depth
Temperature Prediction Error (C) 0 M Depth
<0 .2 .4 .6 .8 1 1.2>
<0 .2 .4 .6 .8 1 1.2>
Temperature Prediction Error (C) 0 M Depth
<0 .2 .4 .6 .8 1 1.2>

Fig. 13.5 Temperature (°C) background error standard deviations valid 20 January 2012 in global HYCOM analysis domains: Atlantic, Indian, and Pacific. (**a**) 0 M depth, (**b**) 150 M depth, (**c**) 300 M depth

13.3.5 Observation Error Variances

The observation errors and the background errors are assumed to be uncorrelated, and errors associated with observations made at different locations and at different times are also assumed to be uncorrelated. As a result of these assumptions, the observation error covariance matrix **R** is set equal to $1 + E_o^2$ along the diagonal and zero elsewhere. Note that E_o^2 represents observation error variances (e_o^2) normalized by the background error variances interpolated to the observation location ($E_o^2 = e_o^2/e_b^2$). Observation errors are computed as the sum of a measurement error and a representation error. Measurement errors reflect the accuracy of the instruments and the ambient conditions in which the instruments operate. These errors are fairly well known for many ocean observing systems, although the errors can change in time due to calibration drift of the instruments and other

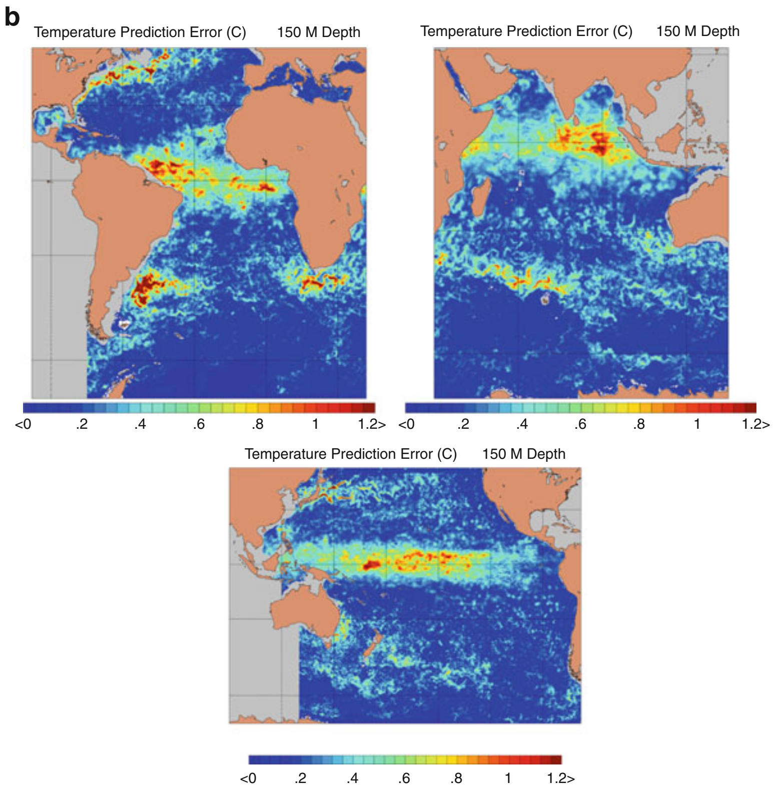

b
Temperature Prediction Error (C) 150 M Depth
<0 .2 .4 .6 .8 1 1.2>
Temperature Prediction Error (C) 150 M Depth
<0 .2 .4 .6 .8 1 1.2>
Temperature Prediction Error (C) 150 M Depth
<0 .2 .4 .6 .8 1 1.2>

Fig. 13.5 (continued)

factors. Representation errors, however, are a function of the resolution of the model and the resolution of the observing network. For satellite retrievals with known measurement footprints, representation errors are set equal to the gradient of the background field at the observation location when the retrieval footprint exceeds the model grid resolution. Representation error of profile observations consists of two additive components. The first component is set proportional to the observed profile vertical gradients of temperature and salinity as a proxy for uncertainty associated with internal waves. The second component is estimated from the variability of multiple observed profile level data averaged into layers defined by the model vertical grid (see Sect. 13.4.2).

13.4 Ocean Observations

The analysis makes full use of all sources of the operational ocean observations. Ocean observing systems assimilated by the 3DVAR are listed in Table 13.1, along

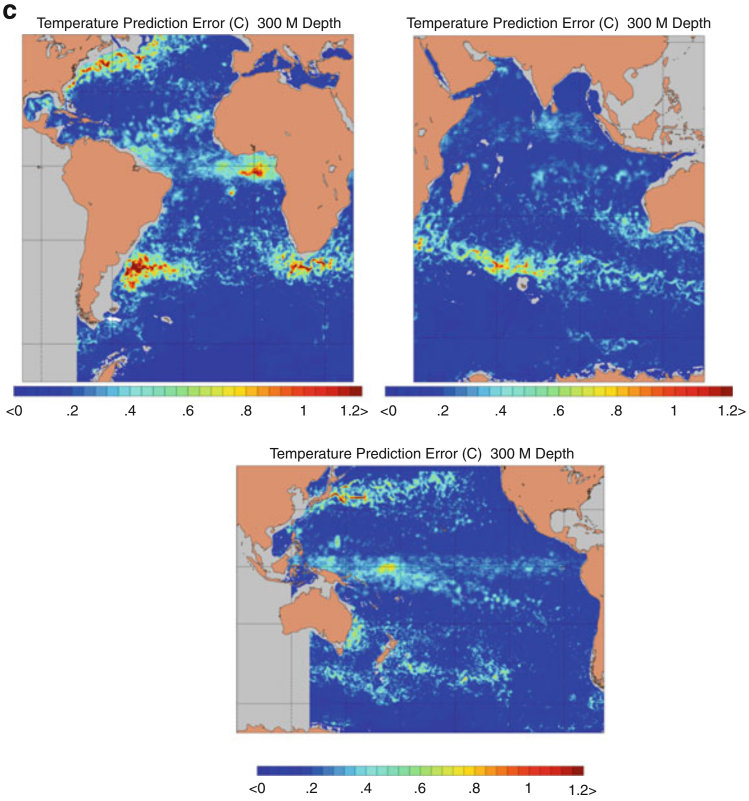

c
Temperature Prediction Error (C) 300 M Depth
<0
.2
.4
.6
.8
1
1.2>
Temperature Prediction Error (C) 300 M Depth
<0
.2
.4
.6
.8
1
1.2>
Temperature Prediction Error (C) 300 M Depth
<0
.2
.4
.6
.8
1
1.2>

Fig. 13.5 (continued)

with typical global data counts per day. All ocean observations are subject to data quality control (QC) procedures prior to assimilation. The need for quality control is fundamental to a data assimilation system. Accepting erroneous data can cause an incorrect analysis, while rejecting extreme, but valid, data can miss important events. The NCODA 3DVAR analysis was co-developed and is tightly coupled to an ocean data QC system. Cummings (2011) provides an overview of the NCODA ocean data quality control procedures.

13.4.1 Surface Observations

Table 13.1 indicates that there are many high volume sources of satellite and in situ SST, SSH, and sea ice observations. It is not uncommon to assimilate ∼40 million satellite SST retrievals, ∼2 million sea ice concentration retrievals, and ∼500, 000

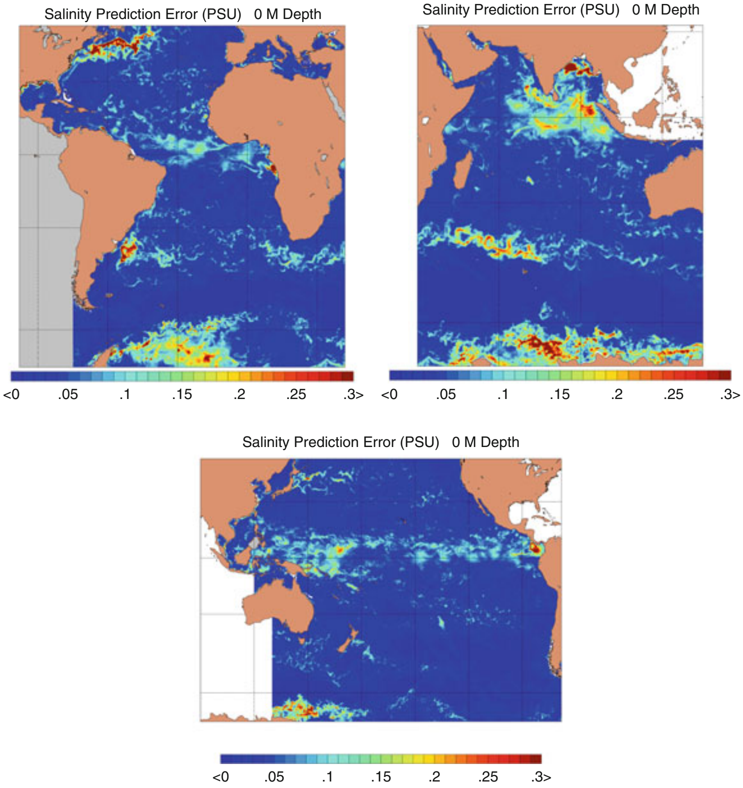

Salinity Prediction Error (PSU) 0 M Depth
<0
.05
.1
.15
.2
.25
.3>
Salinity Prediction Error (PSU) 0 M Depth
<0
.05
.1
.15
.2
.25
.3>
Salinity Prediction Error (PSU) 0 M Depth
<0
.05
.1
.15
.2
.25
.3>

Fig. 13.6 Surface salinity (PSU) background error standard deviations valid 20 January 2012 in global HYCOM analysis domains: Atlantic, Indian, and Pacific

altimeter SSH observations in a single day. These high-density, surface-only, data types must be thinned prior to the analysis to remove redundancies in the data and minimize horizontal correlations among the observations. The data thinning is achieved by averaging innovations into bins with spatially varying sizes defined using the ratio of horizontal correlation length scales and horizontal grid resolution. Innovations are inversely weighted based on observation error in the data thinning process, and in the case of SST observations the water mass of origin is maintained (see Cummings 2005 for a discussion of the Bayesian water mass classification scheme). The length scale to grid mesh ratio bin sizes automatically adjust to changes in the spatially varying horizontal correlation length scales, but are never smaller than the underlying model grid mesh. As a result, fewer data are thinned as the grid resolution decreases or as the correlation length scales shorten. This adaptive feature of the data thinning process can be used to decrease (increase) the amount of data thinning by artificially shortening (lengthening) the horizontal correlation length scales given a fixed model grid. Note that simply increasing data

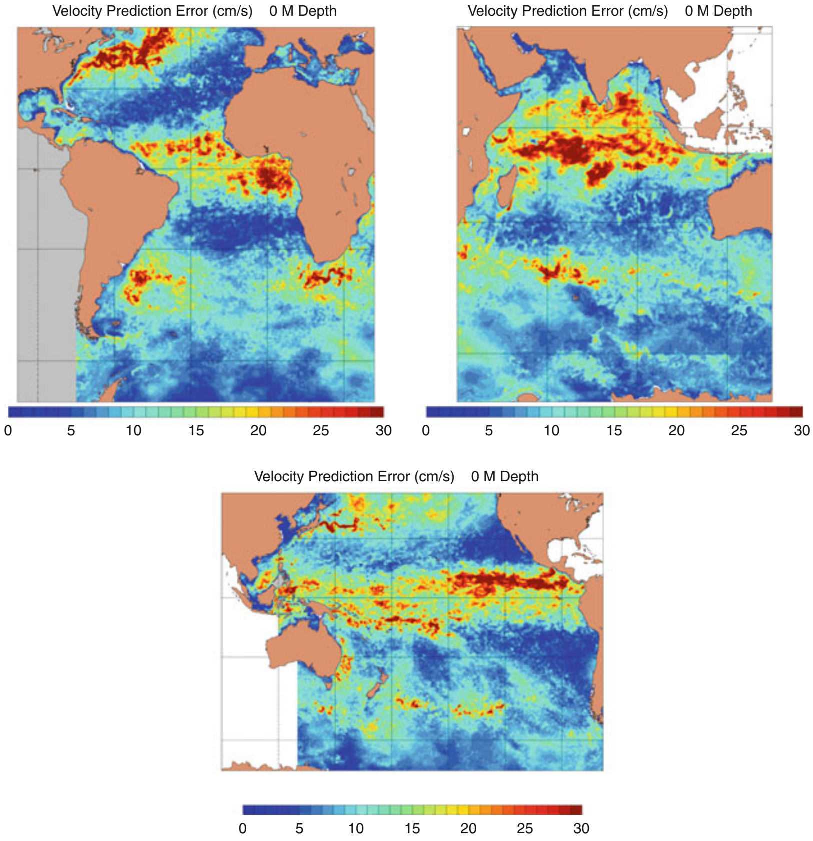

Velocity Prediction Error (cm/s) 0 M Depth
0 5 10 15 20 25 30
Velocity Prediction Error (cm/s) 0 M Depth
0 5 10 15 20 25 30
Velocity Prediction Error (cm/s) 0 M Depth
0 5 10 15 20 25 30

Fig. 13.7 Surface velocity (cm/s) background error standard deviations valid 20 January 2012 in global HYCOM analysis domains: Atlantic, Indian, and Pacific

density does not necessarily improve the analysis. More data will require more conjugate gradient iterations while, more importantly, it may not noticeably alter the results given the smoothing operation of the post-multiplication step (see discussion in Sect. 13.2). Figure 13.8 shows an example of data thinning results for 6 h of satellite SST observations in the FNMOC GHRSST analysis. Even with just 6 h of SST data the various satellite missions and in situ sources show a high degree of spatial overlap. The data thinning removes this data redundancy and creates a sampling pattern consistent with the horizontal correlation length scales defined for the analysis. In this case, length scales are based on Rossby radius of deformation, which varies significantly across the grid. As a result, there is increased data thinning near the equator where length scales are ∼200 km. Elsewhere, especially at high latitude, the data thinning is much less, and satellite retrievals with footprint resolutions of 2 km and 8 km are directly assimilated without any spatial averaging.

Table 13.1 Data types assimilated in NCODA 3DVAR with typical daily data counts. Note that the profile data counts are for the entire profile. Profiles typically contain hundreds of levels that are assimilated as unique latitude, longitude, level observations

Data type	Data source	Specifications	Number daily obs
Satellite SST	NOAA-18 NOAA-19	Infrared 2-km day, night retrievals	4,800,000
	NOAA-18 NOAA-19	Infrared 8-km day, night, relaxed day retrievals	800,000
	AMSR-E	Microwave 25-km day, night retrievals	3,600,000
	METOP-A	Infrared 2-km day, night retrievals	15,000,000
	METOP-A	Infrared 8-km day, night, relaxed day retrievals	450,000
	GOES E/W	Infrared 12-km day, night retrievals	2,000,000
	MeteoSat-2	Infrared 8-km day, night retrievals	220,000
	AATSR	Infrared 1-km day, night retrievals	12,000,000
In Situ SST	Ships	Engine room intake	6,500
		Hull contact sensor	1,000
		Bucket temperature	100
		CMAN Station	100
	Drifting Buoy		34,000
	Fixed Buoy		7,000
Satellite altimeter	Jason 1, 2 Envisat	SSHA	150,000
		SWH	180,000
Sea ice concentration	DMSP F13, F14, F15	SSM/I 25-km retrievals	900,000
	DMSP F16, F17, F18	SSMIS 25-km retrievals	1,200,000
Profiles	Drifting buoy	Temperature	50
	Fixed buoy		1,200
	Argo		600
	XBT		100
	TESAC (CTD)		3,500
	Drifting buoy	Salinity	50
	Fixed buoy		800
	Argo		600
	TESAC (CTD)		3,000

13.4.2 Profile Observations

Preparation of profile observations for the assimilation consists of several steps. First, observed profiles are extended to the bottom using the model forecast. The

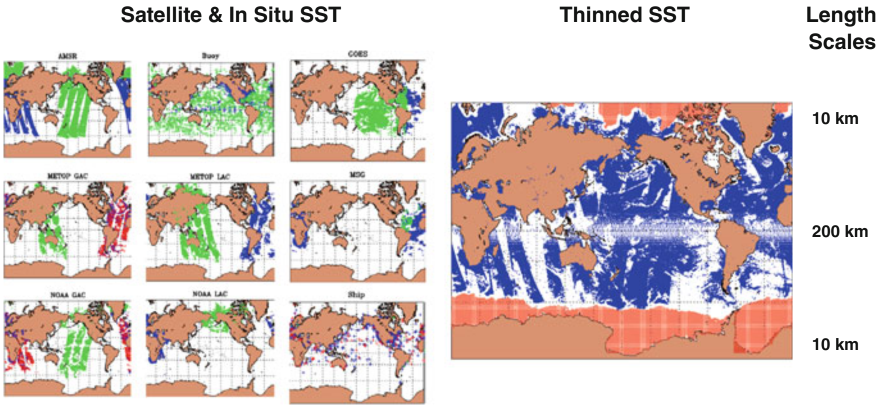


Fig. 13.8 Data thinning of global SST data. Satellite and in situ sources SST show in *left panel* (blue daytime, green nighttime, red relaxed day satellite retrieval types). The SST data sources are (in order from top to bottom): AMSR-E, Drifting and Fixed Buoy, GOES E/W, METOP-A GAC, METOP LAC, MeteoSat-2, NOAA 18,19 GAC, NOAA 18,19 LAC, Surface Ship (engine room intake, bucket, hull contact sensor). Thinned data for assimilation is show in *middle panel* (*blue*—SST observation; *red*—freezing sea water under ice covered seas). Schematic of how correlation lengths vary as a function of latitude shown on *right*

observed profile is merged to the forecast profile by selecting the depth at which the merge is complete based on the shape of the extracted forecast model profile. This target depth is set to be the second zero crossing of the forecast profile curvature. Note that the merge can fail if a suitable target depth is not found or if the difference between the observed and model profile at the merge depth is too large (>3°C for temperature; >0.1 PSU for salinity). Second, similar to the high density surface-only data, profile observations are thinned in the vertical to remove redundant data. The profile thinning is done by averaging temperature and salinity observations at observed levels within vertical layers defined by the mid-points of the model vertical grid. Since the ocean circulation models interfaced with the 3DVAR have very different vertical coordinates (NCOM uses a sigma/z vertical grid; HYCOM uses a z/isopycnal/sigma hybrid vertical grid), model vertical levels at the grid point closest to the profile location are used to define layer thicknesses. Third, in cases where profile vertical sampling is inadequate to resolve the local vertical correlation length scales, the profile is expanded in the vertical by linearly interpolating data to interleaving levels in order to form a more vertically dense profile. This scheme ensures vertically smooth analysis increments at all model levels even when vertical correlations are short due to strong density stratification. This situation routinely occurs in the tropics with the sparse vertical sampling in profiles received from the Tropical Atmosphere Ocean (TAO), Triangle Trans-Ocean Buoy Network (TRITON), and Prediction and Research Moored Array in the Atlantic (PIRATA) buoys. It is clear that the vertical sampling of the tropical mooring arrays needs to be improved.

13.4.3 Altimeter Sea Surface Height

Table 13.1 shows that most ocean observations are remotely sensed and measure ocean variables only at the surface. The lack of synoptic real-time data at depth places severe limitations on the ability of the data assimilation system to resolve and maintain an adequate representation of the ocean mesoscale. Subsurface properties in the ocean, therefore, must be inferred from surface-only observations. The most important observing system for this purpose is satellite altimetry, which measures the time varying change in SSH. Changes in sea level are strongly correlated with changes in the depth of the thermocline in the ocean, and the ocean dynamics generating sea level change are for the most part the mesoscale eddies and meandering ocean fronts. The SSH data are provided as anomalies relative to a time-mean field. The time mean removes the unknown geoid, but it also removes the mean dynamic topography (MDT), which needs to be added back in order to allow the data to be compared with model fields. The 3DVAR determines the satellite altimeter SSH sampling locations in two alternative ways: (1) direct assimilation of the along-track data at the observed locations, or (2) by first performing a 2D horizontal analysis of SSH and then generate a sampling pattern of synthetic profile locations within contours of sea level change that exceed a prescribed noise level threshold (see Cummings 2005 for details). Once the altimeter sampling locations are known there are two alternative methods available in the 3DVAR to project the SSH data to depth in the form of synthetic temperature and salinity profiles. One method is the Modular Ocean Data Assimilation System (MODAS) database, which models the time averaged co-variability of dynamic height vs. temperature at depth and temperature vs. salinity at a fixed location from an analysis of historical profile data (Fox et al. 2002). The MDT used in the MODAS method is derived from historical hydrographic data. Note that an upgrade to the MODAS synthetic profile capability, the Improved Synthetic Ocean Profile (ISOP) system (Helber et al. 2012), is currently being evaluated. The second “direct” method adjusts the model forecast density field to be in agreement with the difference found between the model forecast sea surface height field and the SSH measured by the altimeter (Cooper and Haines 1996). The MDT used in the direct method is the mean SSH from the model derived from a hindcast run. Output of the direct method is in the form of innovations of temperature and salinity from the forecast model background field, which are directly input into the assimilation. An advantage of the direct method is that it relies on model dynamics for its prior information rather than statistics fixed at the start of the assimilation. However, a disadvantage is that it cannot explicitly correct for forecast model errors in stratification due to model drift in the absence of any real data constraints. MODAS does not suffer from these limitations, although MODAS may have marginal skill due to: (1) sampling limitations of the historical profile data, (2) non-steric signals in the altimeter data, or (3) weak correlations between steric height and temperature at depth due to other factors, such as the influence of salinity on steric height at high latitudes. Needless to say, neither of the methods available for assimilating altimeter SSH data is ideal. A new method under development assimilates altimeter SSH by conversion of the along-track SSH slopes to geostrophic velocity profiles. This method is described briefly in Sect. 13.7.

While having the potential of adding important information in data-sparse areas, the number of altimeter-derived synthetic observations computed can greatly exceed and overwhelm the in situ observations in the analysis. Accordingly, the synthetic observations are thinned prior to the analysis in four ways. First, it is assumed that directly observed temperature and salinity profiles are a more reliable source of subsurface information wherever such observations exist. Altimeter-derived synthetic profiles, therefore, are not generated in the area surrounding an in situ profile observation. Second, the observed SSH from the along-track data or the analyzed incremental change in sea level must exceed a threshold value, defined as the noise level of the satellite altimeters, to trigger the generation of a synthetic observation. This value is typically set to 4 cm. Third, projection of the SSH signal onto the model subsurface density field can produce unrealistic results when the vertical stratification is weak. In the absence of specific knowledge about how to partition SSH anomaly into baroclinic and barotropic structures in these weakly stratified regions, synthetic profiles are rejected for assimilation when either of the following occurs: (1) the top-to-bottom temperature difference of the MODAS synthetic profile is less than 5°C; or (2) the maximum value of the Brunt-Väisälä frequency calculated from the model density profile in the direct method is less than 1.4. Fourth, there are problems with the SSH data in shallow water due to contamination of the altimeter signal by tides. Accordingly, SSH data are not assimilated in water depths less than 400 m.

13.5 NCODA System

NCODA is a comprehensive ocean data assimilation system. In addition to the 3DVAR it contains other components that perform functions useful for many applications. These component capabilities are briefly summarized in this section.

13.5.1 Analysis Error Covariance

The analysis error covariance $\mathbf{P_a}$ is estimated from the equation,

$$P_a = P_b - P_b H^T (H P_b H^T + R)^{-1} H P_b \tag{13.9}$$

where $\mathbf{P_b}$ and $\mathbf{R}$ are the background and observation error covariances previously defined for (13.1). Unlike (13.1), which involves matrix–vector operations, (13.9) requires the use of matrix-matrix operations and is computationally expensive to perform. The NCODA 3DVAR provides an estimate of the analysis error variance (the diagonal of the second right-hand term) in the form of a normalized reduction of the forecast error ranging from 1 (0 % reduction) to 0 (100 % reduction) for each analysis variable at all model grid points. The analysis error solution is a local approximation performed within the grid decomposition blocks that is improved upon though the use of halo regions to bring in the influence of additional observations. The analysis error estimation uses the same data inputs as the 3DVAR other than the innovations. In this way the analysis error calculation can be done at the same time as the analysis, albeit on a different set of processors, to improve throughput of the entire data assimilation system. The primary application of the analysis error covariance program is as a constraint in the Ensemble Transform technique (Sect. 13.5.3).

13.5.2 Adjoint

Adjoint-based observation sensitivity provides a feasible all at once approach to estimating observation impact. Observation impact is calculated in a two-step process that involves the adjoint of the forecast model and the adjoint of the assimilation system. First, a cost function (**J**) is defined that is a scalar measure of some aspect of the forecast error. The forecast model adjoint is used to calculate the gradient of the cost function with respect to the forecast initial conditions $(\partial \mathbf{J}/\partial \mathbf{x_a})$. The second step is to extend the initial condition sensitivity gradient from model space to observation space using the adjoint of the assimilation procedure $(\partial \mathbf{J}/\partial \mathbf{y} = \mathbf{K}^T \partial \mathbf{J}/\partial \mathbf{x_a})$, where $\mathbf{K} = \mathbf{P_b H^T [H P_b H^T + R]^{-1}}$ is the Kalman gain matrix of (13.1) and the adjoint of **K** is given by $\mathbf{K^T} = \mathbf{[H P_b H^T + R]^{-1} H P_b}$. The only difference between the forward and adjoint of the analysis system is in the post-multiplication of going from the solution in observation space to grid space. The pre-conditioned, conjugate gradient solver $\mathbf{[H P_b H^T + R]}$ is symmetric or self-adjoint and operates the same way in the forward and adjoint directions. The NCODA 3DVAR adjoint was coded directly from the forward 3DVAR by transposition of the post-multiplier to a pre-multiplier that is invoked first to convert adjoint sensitivities from grid space to observation space prior to execution of the solver for calculation of observation sensitivities and data impacts.

13.5.3 Ensemble Transformation

The ensemble transform (ET) ensemble generation technique (Bishop and Toth 1999) transforms an ensemble of forecast perturbations into an ensemble of analysis perturbations. The method ensures that the analysis perturbations are consistent with the analysis error covariance matrix $(\mathbf{P_a})$, computed using (13.9). To compute the required transform matrix an eigenvector decomposition is performed,

$$(X_f^T P_a^{-1} X_f)/n = C \lambda C^T \tag{13.10}$$

where $\mathbf{X_f}$ is the matrix of ensemble forecast perturbations about the ensemble forecast mean, $\mathbf{P_a}$ is the analysis error covariance matrix, n is the number of model variables (state vector), and C are the eigenvectors and λ the eigenvalues of the left hand side of (13.10). Superscript T indicates matrix transpose. Given the eigenvector decomposition the transformation matrix $\mathbf{T}$ is given by $\mathbf{T} = \mathbf{C}\lambda^{-1/2}\mathbf{C^T}$, which is used to transform a matrix of forecast perturbations to a matrix of analysis perturbations according to $\mathbf{X_a} = \mathbf{X^fT}$. If the ensemble size is large enough it can be shown that the covariance of the analysis perturbations equals the prescribed analysis error covariance $\mathbf{P_a}$ (McLay et al. 2008). Thus the analysis error covariance is an effective constraint in the ET, ensuring that the ensemble generation system is consistent with the data assimilation system.

The NCODA ET is multivariable and computes the transformation matrix for temperature, salinity, and velocity simultaneously. As a result the NCODA ET perturbations are balanced and flow dependent. In an ET ensemble generation scheme the control run is the only ensemble member that executes the 3DVAR. This results in a considerable savings in computational time as compared to a perturbed observation approach where the analysis must be executed by all of the ensemble members. Given a 3DVAR control run analysis and its corresponding analysis error covariance estimate, the system calculates the ET analysis perturbations and adds the perturbations to the control run to form new initial conditions for each ensemble member. The forecast model is then integrated creating a new set of ensemble forecasts for the next cycle of the ET. The NCODA ET and 3DVAR have been successfully implemented in a coupled ocean atmosphere mesoscale ensemble prediction system (Holt et al. 2011).

13.5.4 Residual Vector

The residual vector $[\mathbf{y} - \mathbf{H}(\mathbf{x_a})]$ is very useful in assessing the fit of the analysis to specific observations or observing systems. It is usually calculated at the end of the analysis after the post-multiplication step by horizontally and vertically interpolating the analysis vector $(\mathbf{x_a})$ to the observation locations and application of the nonlinear forward operators $\mathbf{H}$ to obtain $\mathbf{H}(\mathbf{x_a})$ in observation space. This result is then subtracted from the observations to form the residual vector. The problem here is that horizontal and vertical interpolations of the analysis grid to the observation locations and subsequent application of the $\mathbf{H}$ operator introduces error into the residual vector, which may change interpretation of the quality of the fit of the analysis to an observing system. A better approach is to estimate the analysis result, and the residual vector, while still in observation space, that is, before application of the post-multiplication (13.3). Daley and Barker (2000) show that a good approximation of the true residuals while in observation space can be obtained from $\mathbf{y_a} = \mathbf{y} - \mathbf{Rz}$, where $\mathbf{y}$ is observation vector, $\mathbf{y_a}$ the residual vector, $\mathbf{R}$ is the observation error covariance matrix, and $\mathbf{z}$ is defined in (13.2). Using this formulation to calculate residuals gives a better indication of the performance of the 3DVAR assimilation algorithm and how best to tune the background and observation error statistics to improve the analysis. The NCODA 3DVAR system routinely computes residual vectors while still in observation space and saves the residual and innovation vectors for each update cycle in a diagnostics file. As noted, a time history of the innovations and the residuals is the basic information needed to compute *a posteriori* refinements to the 3DVAR statistical parameters. Analysis of the innovations is the most common, and the most accurate, technique for estimating observation and forecast error covariances and the method has been successfully applied in practice (e.g. Hollingsworth and Lonnberg 1986). Similarly, a spatial autocorrelation analysis of the residuals is used to determine if the analysis has extracted all of the information in the observing system. Any spatial correlation remaining in the residuals at spatial lags greater than zero represents information that has not been extracted by the analysis and indicates an inefficient analysis (Hollingsworth and Lonnberg 1989).

13.5.5 Internal Data Checks

Internal data checks are those quality control procedures performed by the analysis system itself. These data consistency checks are best done within the assimilation algorithm, since it requires detailed knowledge of the background and observation error covariances, which are available only when the assimilation is being performed. The first step is to scale the innovations $(\mathbf{y} - \mathbf{H}(\mathbf{x_b}))$ by the diagonal of $(\mathbf{HP_bH^T} + \mathbf{R})^{1/2}$, the symmetric positive-definite covariance matrix of (13.1). The elements of this scaled innovation vector $(\mathbf{d}\hat{})$ should have a standard deviation equal to 1 if the background and observation error covariances have been specified correctly. Assuming this to be the case, set a tolerance limit $(\mathrm{T_L})$ to detect and reject any observation that exceeds it. Since $\mathbf{P_b}$ and $\mathbf{R}$ are never perfectly known, it is best to use a relatively high tolerance limit $(\mathrm{T_L} = 4.0)$ to identify marginally acceptable observations.

The second part of the internal data check is a consistency check. It compares the marginally acceptable observations with all of the observations. The procedure is a logical extension of the tolerance limit check described above. In the data consistency test, the innovations are scaled by the full covariance matrix (not just the diagonal). The elements of this scaled innovation vector $(\mathbf{d}^*)$ are also dimensionless quantities normally distributed. However, because the scaling in $\mathbf{d}^*$ involves the full covariance matrix, it includes correlations between all of the observations. By comparing the vectors $\mathbf{d}\hat{}$ and $\mathbf{d}^*$ it can be shown (Daley and Barker 2000) which marginally acceptable observations are inconsistent with other observations and can therefore be rejected. The $\mathbf{d}^*$ metric should increase (decrease) with respect to $\mathbf{d}\hat{}$ when that observation is inconsistent (consistent) with other observations, as specified by the background and observation error statistics.

13.6 Global HYCOM

As mentioned in the introduction, the NCODA 3DVAR analysis is currently cycling with global HYCOM in real-time at NAVOCEANO. The 3DVAR is expected to replace the MVOI as the data assimilation component in the operational HYCOM, which is referred to as the Global Ocean Forecast System (GOFS) version 3.

As configured within GOFS v3, HYCOM has a horizontal equatorial resolution of .08° or $\sim 1/12°$($\sim$7 km mid latitude) resolution. This makes HYCOM eddy resolving. Eddy-resolving models can more accurately simulate western boundary currents and the associated mesoscale variability and they better maintain more accurate and sharper ocean fronts. In particular, an eddy resolving ocean model allows upper ocean topographic coupling via flow instabilities, while an eddy-permitting model does not because fine resolution of the flow instabilities is required to obtain sufficient coupling (Hurlburt et al. 2008b). The coupling occurs when flow instabilities drive abyssal currents that in turn steer the pathways of upper ocean currents (Hurlburt et al. 1996 in the Kuroshio; Hogan and Hurlburt 2000 in the Japan/East Sea; and Hurlburt and Hogan 2008 in the Gulf Stream). In ocean prediction this coupling is important for ocean model dynamical interpolation skill in data assimilation/nowcasting and in ocean forecasting, which is feasible on time scales up to about a month (Hurlburt et al. 2008a).

The global HYCOM grid is on a Mercator projection from 78.64°S to 47°N and north of this it employs an Arctic dipole patch where the poles are shifted over land to avoid a singularity at the North Pole. This gives a mid-latitude (polar) horizontal resolution of approximately 7 km (3.5 km). This version employs 32 hybrid vertical coordinate surfaces with potential density referenced to 2,000 m and it includes the effects of thermobaricity (Chassignet et al. 2003). Vertical coordinates can be isopycnals (density tracking), often best in the deep stratified ocean, levels of equal pressure (nearly fixed depths), best used in the mixed layer and unstratified ocean, and sigma-levels (terrain-following), often the best choice in shallow water. HYCOM combines all three approaches by choosing the optimal distribution at every time step. The model makes a dynamically smooth transition between coordinate types by using the layered continuity equation. The hybrid coordinate extends the geographic range of applicability of traditional isopycnic coordinate circulation models toward shallow coastal seas and unstratified parts of the world ocean. It maintains the significant advantages of an isopycnal model in stratified regions while allowing more vertical resolution near the surface and in shallow coastal areas, hence providing a better representation of the upper ocean physics. HYCOM is configured with options for a variety of mixed layer sub-models (Halliwell 2004) and this version uses the K-Profile Parameterization (KPP) of Large et al. (1994). A more complete description of HYCOM physics can be found in Bleck (2002). The ocean model uses 3-hourly Navy Operational Global Atmospheric Prediction System (NOGAPS) forcing from FNMOC that includes: air temperature at 2 m, surface specific humidity, net surface short-wave and long-wave radiation, total (large scale plus convective) precipitation, ground/sea temperature, zonal and meridional wind velocities at 10 m, mean sea level pressure and dew-point temperature at 2 m. The first six fields are input directly into the ocean model or used in calculating components of the heat and buoyancy fluxes while the last four are used to compute surface wind stress with temperature and humidity based stability dependence. Currently the system uses the 0.5° degree resolution application grid NOGAPS products (i.e. already interpolated by FNMOC to a constant 0.5° latitude/longitude grid); however HYCOM can also (and preferably) use the NOGAPS T319 computational grid (i.e. a Gaussian grid—constant in longitude, nearly constant in latitude) products. Typically atmospheric forcing forecast fields extend out to 120 h (i.e. the length of the HYCOM/NCODA forecast). On those instances when atmospheric forecasts are shorter than 120 h, an extension is created based on climatological products. The last available NOGAPS forecast field is then gradually blended toward climatology to provide forcing over the entire forecast period. The current version of the global HYCOM forecast system includes a built-in energy loan, thermodynamic ice model. In this non-rheological system, ice grows or melts as a function of SST and heat fluxes. For an extensive validation of the global forecast system see Metzger et al. (2008, 2010a,b).

The NCODA 3DVAR analysis system consists of three separate programs that are executed in sequence. The first program does the analysis and data preparation, including computation of the innovation vector. The second program performs the 3DVAR, where it reads the innovation vector and outputs the analysis increment correction fields. The third program performs several post-processing tasks, such as updating the background error fields and computing some diagnostic and verification statistics. The global HYCOM 3DVAR analysis is split into seven overlapping regions covering the global ocean (Fig. 13.9). The Atlantic, Indian and Pacific Ocean regions cover the Mercator part of the model grid. The remaining four regions cover the irregular part of the model domain, one region in the Antarctic, one each in the northern part of the Atlantic and Pacific and the last region covering the Arctic Ocean. The boundary between the different regions follows the natural boundary of the continents. The regions overlap to ensure that the analyses will be smooth across the boundaries that fall over the ocean. At present the forecast system is running on 624 Cray XT5 processors. The processors are split among the sub-regions so that each regional analysis can run in parallel and finish at about the same time. Note that performing the 3DVAR in sub-regions is a holdover from the old MVOI system. There are no limitations in the 3DVAR that prevent the analysis from being executed on the full global HYCOM grid. However, at the present time, memory limitations in the data prep program do not allow the system to be executed globally. This problem is being addressed.

Two assimilative runs of the 3DVAR cycling with global HYCOM on a daily basis (24-h update cycle) are reported here. Both runs were initialized from a non-assimilative spin-up of the model. The run initialized on 1 May 2010 was executed in hindcast mode and has the advantage of assimilating synoptic ocean observations. The run initialized on 29 November 2011 is a real-time run and must deal with data latency issues associated with some of the ocean observing systems. Satellite altimeter and profile observations have the longest time delays before the data are

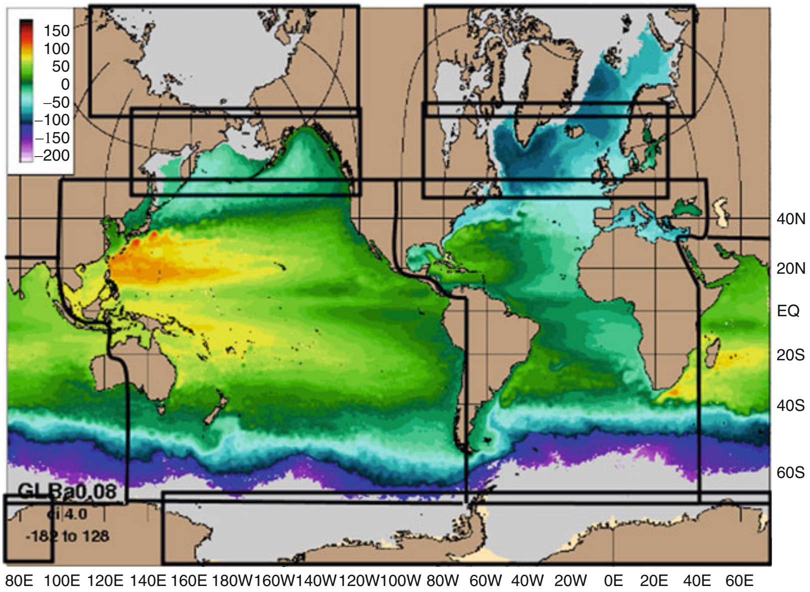

150
100
50
0
–50
–100
–150
–200
GLBa0.08
ci 4.0
-182 to 128
40N
20N
EQ
20S
40S
60S
80E 100E 120E 140E 160E 180W 160W 140W 120W 100W 80W 60W 40W 20W 0E 20E 40E 60E

Fig. 13.9 NCODA 3DVAR analysis regions for global HYCOM. The three regions in the Atlantic, Indian and Pacific Ocean cover the Mercator projection part of the global model grid. The three regions in the Arctic Cap cover the irregular bi-polar part of the global grid: northern part of the Atlantic, northern part of the Pacific, and a region covering the Arctic Ocean. A spherical grid projection is used in the vicinity of Antarctica

available for assimilation in real-time. The delays in the altimeter data are at least 7296 h due to orbit corrections that have to be applied to improve the accuracy of the measurements. Profile data can be delayed up to ∼72 h. Since ocean data are so sparse it is important to use all of the data in the assimilation. Accordingly, in real-time applications the 3DVAR has the capability to select data for the assimilation based on receipt time (the time the observation is received at the center) instead of observation time. In this way all data *received* since the previous analysis are used in the next real-time run of the 3DVAR. However, data selected this way will necessarily contain non-synoptic measurement times. This source of error in the analysis is reduced by comparing observations against time dependent background fields using FGAT. Hourly forecast fields are used in the FGAT for assimilation of SST observations in order to maintain a diurnal cycle in the model. Daily averaged forecast fields are used in FGAT for profile data types (both synthetic and real). SSH data are assimilated in global HYCOM using the MODAS synthetic profile approach. The 3D temperature, salinity, and u, v velocity analysis increments are incrementally inserted into the model over a 6 h time period using the incremental analysis update procedure (Bloom et al. 1996). A separate 2D ice concentration analysis is used to update the ice concentration in the thermodynamic ice model.

Figures 13.10, 13.11, and 13.12 give time series of innovation and residual error statistics in the Pacific domain of the hindcast run. The statistics are computed in

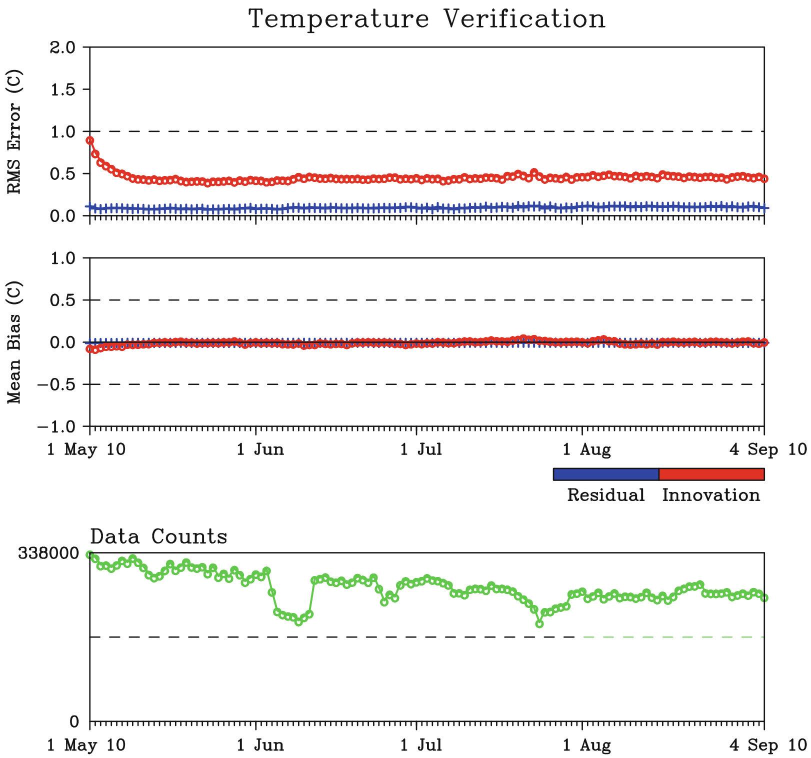


Fig. 13.10 Time of RMS and mean bias error statistics for temperature observations in HYCOM Pacific basin. *Upper panel* reports RMSE, middle panel reports mean bias, and bottom gives temperature data counts. Tick marks along time axis indicate 24-h update cycle periods

observation space and represent averages across all data assimilated for a particular analysis variable. Innovation RMS errors for temperature (Fig. 13.10) and salinity (Fig. 13.11) show increased errors for the first few update cycles while the free running model adjusts to the data. After this initial adjustment time, RMS errors are very stable, with temperature errors ∼0.4°C and salinity errors ∼0.1 PSU. The model innovations are remarkably unbiased in both temperature and salinity. The 3DVAR analysis produces a reduction in error from the innovations to the residuals of about 60 %, which is clearly seen in both temperature and salinity. However, the time series of the layer pressure error statistics (Fig. 13.12) are the most interesting. When cycling with HYCOM, the 3DVAR includes a sixth analysis variable, layer pressure. Layer pressure innovations are computed as differences in the depths of density layers in the observations and the model forecast. The layer pressure correction fields are then used to correct isopycnal layer depths in the model. Unlike the fairly rapid response of the free-running model to the assimilation of temperature and salinity observations, bias in the layer structure of the model spin-up takes about a month to adjust to the data. Layer pressure RMS errors remain

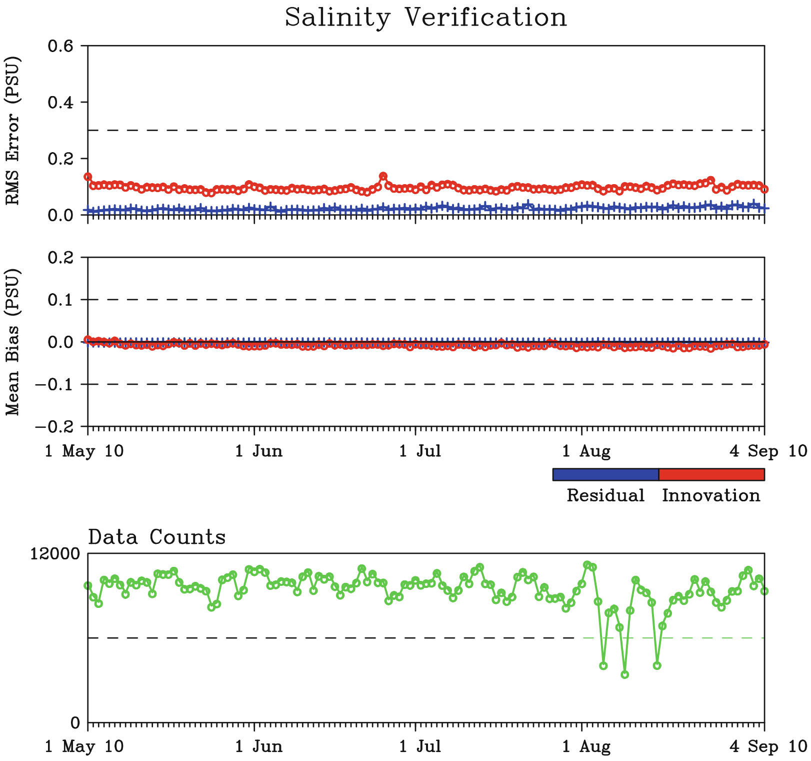


Fig. 13.11 Same as Fig. 13.10, except for salinity observations

high (∼100 db) after the adjustment time period due to the assimilation of MODAS synthetic profiles at high latitudes. MODAS synthetics were not thinned based on stratification (Sect. 13.4.3) in these model runs. Layer pressure RMS errors are reduced more than 50 % when weakly stratified MODAS synthetics are rejected (not shown).

Figure 13.13 shows a verification result from the real-time run for sea surface height in the Kuroshio region on 12 January 2012. The assimilation of SSH anomalies is crucial to accurately map the circulation in these highly chaotic regions dominated by flow instabilities. The white (black) line overlain is an independent analysis of available infrared observations of the north edge of the current system performed at the Naval Oceanographic Office. The frontal analysis clearly indicates that the forecast system is able to accurately map the mesoscale features in the western boundary current.

Table 13.2 gives run times for the 3DVAR conjugant gradient solver and post-multiplication steps. The run times are listed for a typical day (28 January 2012) in six of the global HYCOM analysis subdomains. A total of 2.2 million observations were assimilated into the HYCOM grid that contained more than 520 million grid

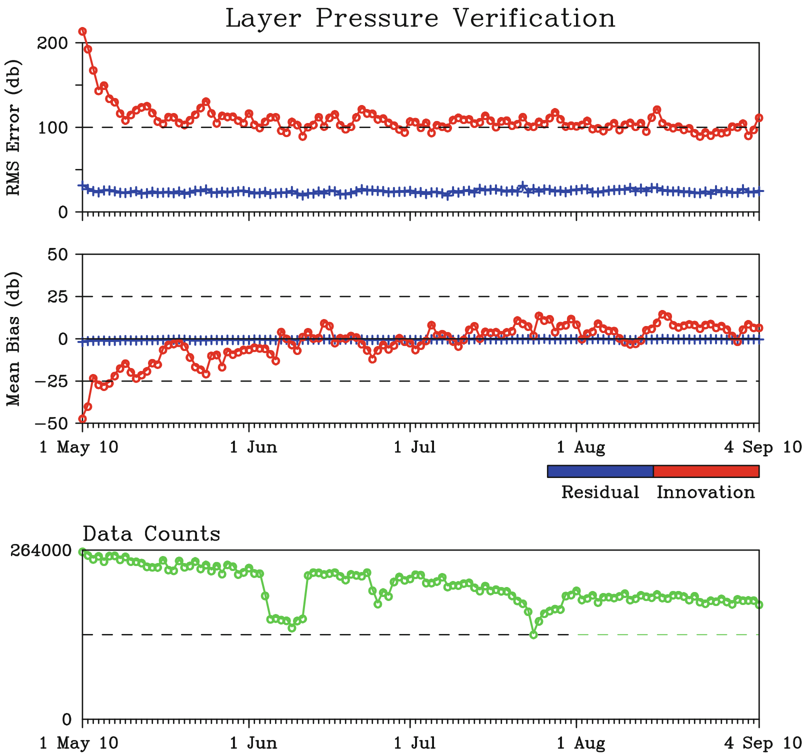


Fig. 13.12 Same as Fig. 13.10, except for layer pressure observations. Layer pressure is computed from density using temperature and salinity profiles (see text for details)

points. The total time of the 3DVAR step in the NCODA analysis system is the maximum time needed to complete any of the subdomains—in this case 14.2 min to complete the Indian Ocean analysis. Efficiency of the 3DVAR is clear, especially in the large Pacific basin where >1 million observations were assimilation into 195.2 million grid points in ∼9.8 min wall clock time. Table 13.2 also shows how well the analysis scales using different numbers of processors. Reduction of the Indian Ocean run time, and thus speed-up of the 3DVAR analysis step in global HYCOM analysis/forecast system, can easily be achieved by simply increasing the number of processors. In general, the post-multiplication step of the analysis is more computationally expensive than the observation space solver. Accordingly, the analysis contains an option to perform the post-multiplication step on a reduced resolution grid. The innovations are always formed from the full resolution model grid, and the solution vector is calculated using all of the observations, but now the solution is mapped to every other (or any multiple) horizontal grid point. This option results in a considerable saving in computational time with no loss of information when analysis correlation length scales generally exceed the model grid resolution.

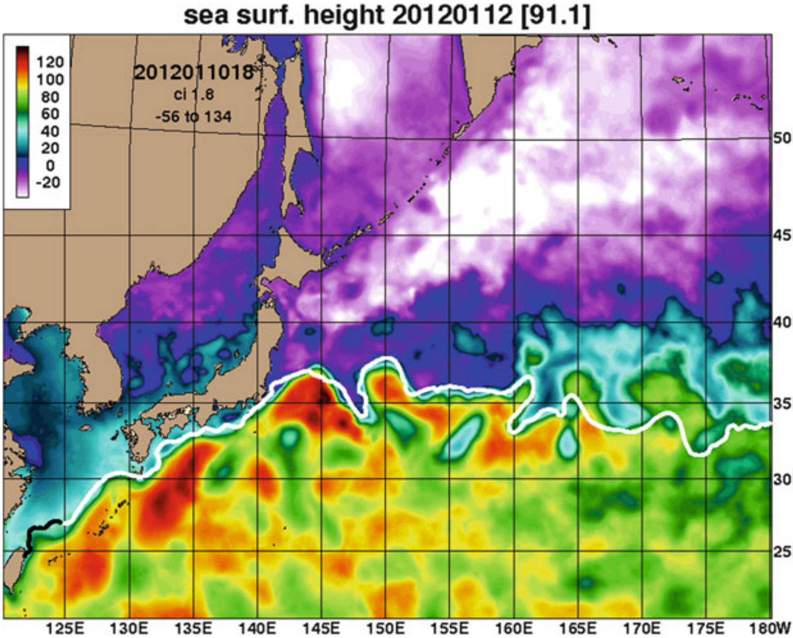


Fig. 13.13 Sea surface height in the Kuroshio region from the 1/12° global HYCOM/NCODA forecast system on January 12, 2012. An independent infrared (IR) analysis of the north edge of the current system performed by the Naval Oceanographic Office is overlain. A *white* (*black*) line means the IR analysis is based on data less (more) than four days old

Table 13.2 3DVAR run times for six of the seven global HYCOM analysis domains on 28 January 2012

Domain	Grid size	Number procs	Number obs	Solver (min)	Post proc (min)	Total (min)
Atlantic	1,751 × 1,841 × 42	88	613,525	4.8	5.6	10.7
Indian	1,313 × 1,569 × 42	64	468,828	6.6	7.3	14.2
Pacific	2,525 × 1,841 × 42	416	1,028,369	6.7	2.6	9.8
Arctic Ocean	1,630 × 551 × 42	16	11,879	0.1	0.2	1.7
Arctic Atlantic	1,490 × 551 × 42	16	82,137	0.1	0.6	2.3
Arctic Pacific	1,335 × 551 × 42	16	17,630	0.4	0.2	1.6
Totals	520,250,556[a]	616	2,222,368			

[a] Total for grid size is the total number of grid points

Full resolution correction fields for the model update are produced for each analysis variable in the NCODA 3DVAR post-processing step by interpolation. This reduced resolution grid option is used in global HYCOM where the solution vector is mapped to every other model grid point.

13.7 Future Capabilities

The NCODA 3DVAR and Navy global ocean forecasting systems continue to be developed and improved. These new developments and capabilities are summarized in this section.

13.7.1 HYCOM GOFS

The present $1/12°$ global HYCOM/NCODA system is the first step towards a $1/25°$ global forecast system. The first phase of the upgrade will continue to use the $1/12°$ model. In this phase the simple thermodynamic ice model will be replaced by the Los Alamos Community Ice CodE (CICE). CICE is the result of an effort to develop a computationally efficient sea ice component for a fully coupled forecast system. CICE has several interacting components: a thermodynamic model that computes local growth rates of snow and ice due to vertical conductive, radiative and turbulent fluxes, along with snowfall; a model of ice dynamics, which predicts the velocity field of the ice pack based on a model of the material strength of the ice; a transport model that describes advection of the areal concentration, ice volumes and other state variables; and a ridging parameterization that transfer ice among thickness categories based on energetic balances and rates of strains. HYCOM and CICE will be fully coupled via the Earth System Modeling Framework (ESMF: Hill et al. 2004). An interim, fully coupled, real time Arctic Cap HYCOM/CICE/NCODA-3DVAR forecast system has been set up until CICE is implemented in the global model (Posey et al. 2010). The second phase of the upgrade includes the implementation of a fully coupled $1/25°$ HYCOM/CICE model that includes tidal forcing and uses NCODA 3DVAR as the data assimilation component for both HYCOM and CICE. Preliminary experiments with the assimilative $1/25°$ model are under way. This model will have $\sim$3 km mid latitude resolution.

13.7.2 Satellite SST Radiance Assimilation

At the present time, SST retrievals are empirically derived using stored regressions between cloud cleared satellite SST radiances and drifting buoy SSTs. The regressions are global, calculated once, and held constant. The coefficients represent a very broad range of atmospheric conditions with the result that subtle systematic errors are introduced into the empirical SST when the method is uniformly applied to new radiance data. In the 3DVAR, work is underway to develop an observation operator for direct assimilation of satellite SST radiances. This new physical SST algorithm uses an incremental approach. It takes as input prior estimates of SST and short-term predictions of air temperature and water vapor profiles from NWP. The algorithm is forced by differences between observed and predicted top-of-the-atmosphere (TOA) brightness temperatures (BTs) for the different satellite SST channel wavelengths. Calculation of the TOA-BTs requires use of a fast radiative transfer model. For this purpose the Community Radiative Transfer Model (CRTM; Han et al. 2006) is being integrated into the 3DVAR. In addition to the TOA forward model, CRTM provides the tangent linear radiance sensitivities (Jacobians) with respect to the prior SST, water vapor, and atmospheric temperature predictor variables as a function of the infrared satellite 3.5, 11 and 12 μm wavelengths. The physical SST inverse model for a given channel is given by,

$$\begin{bmatrix} \delta BT \cdot J_{sst} \\ \delta BT \cdot J_t \\ \delta BT \cdot J_q \end{bmatrix} = \begin{bmatrix} \varepsilon_{sst}^{-1} \cdot J_{sst} \cdot J_{sst} & J_{sst} \cdot J_t & J_{sst} \cdot J_q \\ J_t \cdot J_{sst} & \varepsilon_t^{-1} \cdot J_t \cdot J_t & J_t \cdot J_q \\ J_q \cdot J_{sst} & J_q \cdot J_t & \varepsilon_q^{-1} \cdot J_q \cdot J_q \end{bmatrix} \begin{bmatrix} \delta T_{sst} \\ \delta T_a \\ \delta Q_a \end{bmatrix} \tag{13.11}$$

where δBT are the TOA-BT innovations, J_{sst}, J_t, and J_q are the radiative transfer model Jacobians for SST, atmospheric temperature, and water vapor, respectively, ε_{sst}, ε_t, and ε_q are the errors of the priors, and δT_{sst}, δT_{atm}, and δQ_{atm} are the corrections output for each of the priors that take into account the variable SST and temperature and water vapor content of the atmosphere at the time and location of the radiance measurement. The prior corrections are calculated and summed over the SST channels (3 channels at night, 2 channels during the day). With this approach, coefficients that relate radiances to SST in the observation operator are dynamically defined for each atmospheric situation observed. The method removes atmospheric signals in the radiance data and extracts more information on the SST, which improves the time consistency of the SST estimate, especially in the tropics where water vapor variations create unrealistic sub-daily variations in the empirically derived SST. However, the physical SST method requires careful consideration of biases and error statistics of the NWP fields. Biases are expected since the NWP information may represent areas that are both cloudy and clear, while the satellite radiance data, by definition, are only available in clear-sky, cloud free conditions. Accordingly, a bias correction step is under development following the ideas developed by Merchant et al. (2008). Proper specification of the error statistics of the priors is also required to correctly partition the observed TOA-BT differences into the various sources of variability (atmospheric temperature, water vapor, or sea surface temperature). Sensitivity experiments are underway to evaluate situation dependent error statistics for the atmospheric temperature and water vapor priors using the 96-member global NWP ensemble operational at FNMOC.

Implementation of the physical SST method via an observation operator will have many advantages in the 3DVAR. First, in a coupled model forecast, the prior SST will come from the coupled ocean model forecast and differences between observed and predicted TOA-BTs will be computed using the coupled model atmospheric state. This is a true example of coupled data assimilation: an observation in one fluid (atmospheric radiances) creates an innovation in a different fluid (ocean SST). Second, the method can easily be extended to incorporate the effects of aerosols; the presence of which tends to introduce a cold bias in infrared estimates of SST. To do this prior information on the microphysical properties of dust and its amount and vertical distribution is obtained from the Navy Aerosol Analysis Prediction System (NAAPS; http://www.nrlmry.navy.mil/aerosol/). The contribution of NAAPS aerosol information to the TOA-BTs is determined using CRTM, which contains aerosol Jacobians defined for 91 wavelengths and 6 aerosol species. Equation (13.11) is then expanded to a 4 × 4 matrix to further partition differences between observed and simulated TOA BTs into an additional aerosol source of variability. Third, the method can be applied to radiances from ice covered seas to determine ice surface temperature (IST). Knowledge of IST is important since it controls snow metamorphosis and melt, the rate of sea ice growth, and modification of air–sea heat exchange. IST has been added as an analysis variable in the 3DVAR and is analyzed simultaneously with SST to form a seamless depiction of surface temperature from the open ocean to ice covered seas. This capability will be used in the coupled HYCOM/CICE system (Posey et al. 2010).

13.7.3 SSH Velocity Assimilation

An alternative to assimilating SSH information referenced to the along-track mean is to assimilate the dynamically important along-track SSH slope. Altimeter SSH slopes provide the cross-track component of the vertically averaged geostrophic current. As noted in Sect. 13.4.3, current methods for assimilating altimeter SSH data via synthetic temperature and salinity profiles have known deficiencies. One major difficulty is the need to specify a reference MDT matching that contained in the altimeter data; a non-trivial problem. The mean height of the ocean includes the Geoid (a fixed gravity equipotential surface) as well as the MDT, which is not known accurately enough relative to the centimeter scales of variability contained in the dynamic topography. The use of SSH slopes obviates the need for a MDT.

To derive geostrophic currents from SSH slopes appropriate for the ocean mesoscale, noise in the along-track altimeter data must be suppressed. For this purpose a quadratic LOESS smoother (LOcally wEighted Scatterplot Smoother: Cleveland and Devlin 1988; Schlax and Chelton 1992) with varying cutoff wave lengths is applied. The wave lengths are adjusted in accordance with the Rossby radius of deformation to account for the varying eddy length scales. The advantage of this method is that noise in the data, the SSH slope derivative, and the u, v vector velocity components are all computed in a single operation. Figure 13.14 shows the LOESS smoothing of the altimeter SSH data along two tracks; track 109 across the Gulf Stream (Fig 13.14a) and track 106 across the Kuroshio (Fig. 13.14b). The quality of the LOESS filter is clearly seen when the altimeter data exhibit considerable noise (distance points 1,000–3,000, track 109; distance points 1,200–2,440, track 106), and when the altimeter data show strong signals from crossing the Gulf Stream and Kuroshio fronts (distance points 3000–3800, track 109; distance points 400–1,000, track 106). Figure 13.15 shows the Atlantic and Pacific basin

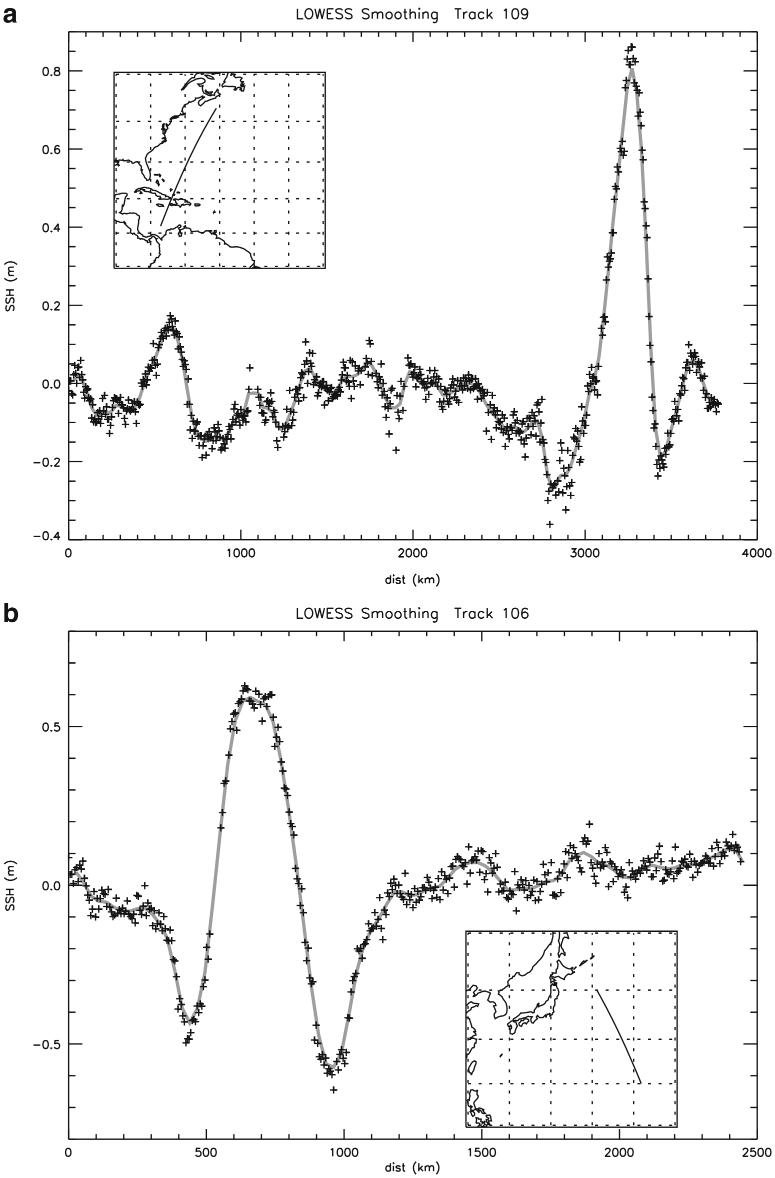


Fig. 13.14 Smoothed along-track SSH computed using LOESS filter. All data from 10 January 2012. (**a**) LOESS filter fit to altimeter SSH data along track 109 in the Gulf Stream area; (**b**) LOESS filter fit to altimeter SSH data along track 106 in the Kuroshio area. Plus marks give raw altimeter SSH data values, solid line gives LOESS fit

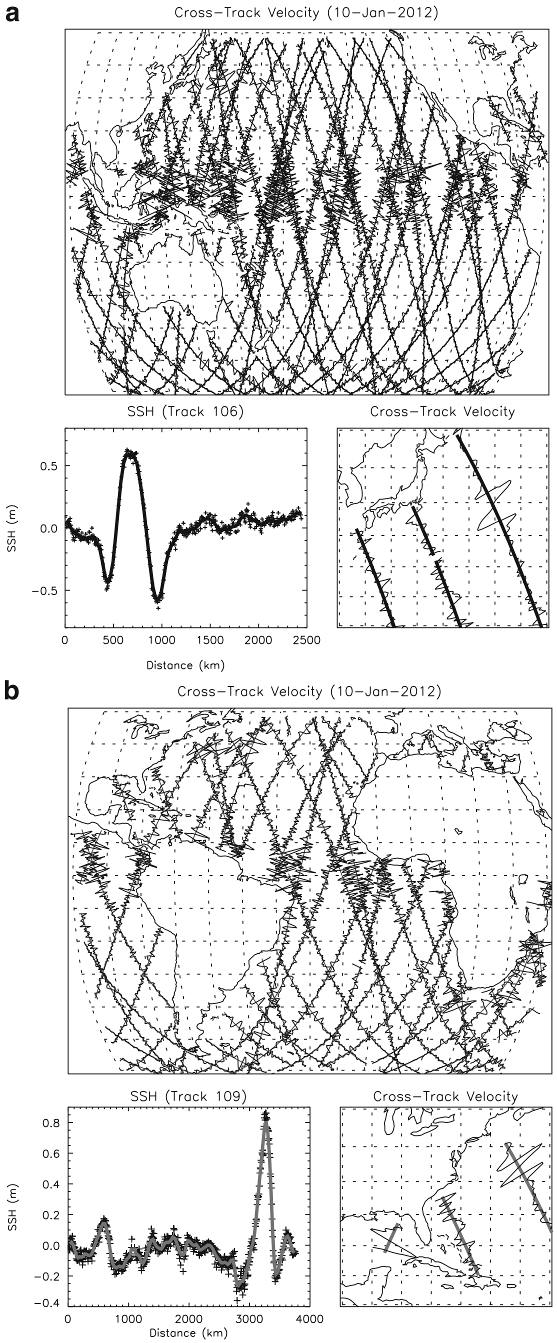


Fig. 13.15 Basin scale geostrophic velocity data calculated from smoothed along-track altimeter SSH data using LOESS filter. All data from 10 January 2012. Top part of each Figure gives basin scale results, lower left gives LOESS filter results, lower right gives zoom on geostrophic velocity along tracks intersecting the Kuroshio and Gulf Stream fronts. (**a**) HYCOM Pacific basin; (**b**) HYCOM Atlantic basin

cross-track geostrophic velocities computed using the LOESS filter for one day of along-track altimeter data (10 January 2012). It is readily apparent that a tremendous amount of mesoscale oceanographic information is contained in the geostrophic velocities derived from the along-track altimeter data.

Once the altimeter SSH along-track geostrophic currents are calculated the model equivalents are determined. Cross-track geostrophic velocity relative to a deep level of no motion (2,000 m) is computed from the model using dynamic height differences at points adjacent to the along-track estimate of the SSH slope. The difference between the vertically averaged model and altimeter cross-track geostrophic velocities is used to correct the relative geostrophic shear from the model and form the velocity profile $u_a(z)$ for the assimilation according to:

$$u_a(z) = u_g(z) - \overline{u_g} + c \tag{13.12}$$

where $u_g(z)$ is the model relative geostrophic shear profile, $\bar{u}_g$ is its vertical average, and c is the integral cross track velocity component calculated from the altimeter slope. Assimilation of the u, v velocity vectors formed this way via the multivariate correlations in the 3DVAR provide balanced geopotential increments, which in turn are decomposed into balanced temperature and salinity increments using a linearized equation of state. The velocity profiles in this scheme are very sensitive to the reference level of no motion. One option here is to use Argo trajectory data to infer a time dependent geopotential field at the float parking depth (cf. Davis 2005). A dynamic geopotential field would go a long way in solving a long-standing problem of hydrography: properly referencing geostrophic shear.

13.7.4 Hybrid Ensemble Four Dimensional Data Assimilation

A four-dimensional (4D) ensemble-enhanced data assimilation scheme for global HYCOM is being developed to better deal with the late receipt, temporally distributed observations than the current 3DVAR methodology. As previously noted, a crucial aspect of all ocean data assimilation schemes is the way in which the background error covariances are specified. The data assimilation process is optimal if the background error covariances are perfectly known, which is never the case. A major challenge then is to find ways to estimate accurate and comprehensive background error covariances. Ensemble methods provide a method for doing this, including the ability to provide a flow-dependent estimate of the background error covariances.

When ensemble covariances are used in a variational data assimilation framework to augment the existing background-error covariance, analyses are further improved. This method is called a hybrid ensemble variational method. In comparison with conventional ensemble-based data assimilation, a hybrid scheme is attractive for the following reasons. First, the hybrid schemes build upon existing variational systems enabling the ensemble information to be incorporated relatively easily. Existing variational ocean data assimilation technology and capabilities are not lost. Second, when ensemble variances are imperfect the optimal error variance estimate is a linear combination of a climatological covariance and an ensemble covariance. The superiority of hybrids over conventional ensemble assimilation schemes is particularly marked when the ensemble size is small or the model error is large.

A static 4D ensemble covariance data base will be computed from an ensemble of mesoscale anomalies using the long term integration of global HYCOM in the 1993–2009 reanalysis product, which includes NCODA 3DVAR assimilation. Covariances calculated in this way have clear physical meanings and represent 4D model climate flow dependence and model variable interactions. Existing 3DVAR initial covariances will be extended to 4D by assuming that the error covariances between variables are a separable function of space and time. The computational overhead of imparting this 4D aspect to the 3DVAR covariances is expected to be very small. The 4D extension of the NCODA covariances will then be linearly combined with the 4D localized HYCOM static ensemble covariances forming a fully 4D hybrid data assimilation scheme. Optimum values for weighting the ensemble and extended 3DVAR covariances in the hybrid are determined from model statistics.

13.8 Summary

This paper describes the development, implementation, and validation of a new oceanographic 3DVAR assimilation system. The system is unified and flexible and a key component of many Navy ocean and atmosphere applications. It is run globally or regionally, where it can be applied to nested, successfully higher-resolution grids, providing analyses on a range of scales. NCODA 3DVAR provides the assimilation component for both ocean and wave model prediction systems as well as multiple atmospheric prediction systems, where it is used to provide sea ice and SST lower boundary conditions. It assimilates a wide range of ocean data types and it contains numerous diagnostic features for assessing and tuning the statistics needed for the assimilation as well as quality control. The background error covariance formulation permits considerable anisotropy with adaptive horizontal and vertical length scales and error variances that vary with location and evolve with time. It is shown to be efficient for very large scale, high resolution global ocean model grids, assimilating millions of observations a day. The intelligent, adaptive data thinning algorithm permits all sources of the high density surface data types to be assimilated with minimal loss of information. The parallel implementation has minimal communication overhead, with granularity of the code (important for load balancing) easily controlled by the number and size of the observation data blocks. The NCODA 3DVAR system is operational at the Navy oceanographic production centers and is in the final phase of pre-operational testing as the data assimilation component for the global HYCOM forecasting system.

Acknowledgements The first author gratefully acknowledges the work of Roger Daley and Ed Barker in developing the NAVDAS solution algorithm which forms the basis of the NCODA 3DVAR. This work was funded in part by NRL base projects "Variational Data Assimilation for Ocean Prediction" and "Observation Impact using a Variational Adjoint System". Funding was also received from the National Ocean Partnership Program (NOPP) through the project "US GODAE: Global-Ocean Prediction with the HYbrid Coordinate Ocean Model (HYCOM)" and the Office of Naval Research (ONR) under program element number 61153 N. The Department of Defense High Performance Computing Modernization Program provided grants of computer time at Major Shared Resource Centers operated by the Naval Oceanographic Office, Stennis Space Center, MS, USA. This is NRL contribution NRL/BC/7320-12-1125 and has been approved for public release. Distribution is unlimited.

References

Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. J Atmos Sci 56:1748–1765

Bleck R (2002) An oceanic general circulation model framed in hybrid isopycnic- Cartesian coordinates. Ocean Model 4:55–88

Bloom SC, Takacs LL, Da Silva AM, Ledvina D (1996) Data assimilation using incremental analysis updates. Mon Weather Rev 124:1256–1271

Chassignet EP, Smith LT, Halliwell GR, Bleck R (2003) North Atlantic simulations with the HYbrid coordinate ocean model (HYCOM): impact of the vertical coordinate choice, reference pressure, and thermobaricity. J Phys Oceanogr 33(12):2504–2526

Chelton DB, DeSzoeke RA, Schlax MG, Naggar KE, Siwertz N (1998) Geographical variability of the first baroclinic Rossby radius of deformation. J Phys Oceanogr 28:433–460

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83:596–610

Cooper M, Haines KA (1996) Altimetric assimilation with water property conservation. J Geophys Res 24:1059–1077

Cummings JA (2005) Operational multivariate ocean data assimilation. Q J R Met Soc 131: 3583–3604

Cummings JA (2011) Ocean data qaulity control. In: Schiler A, Brassington GB (eds) Operational oceanography in the 21st century. Springer, Dordrecht, pp 91–121

Courtier P (1997) Dual formulation of four-dimensional variational assimilation. Q J R Meteorol Soc 123:2449–2461

Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge, p 457

Daley R, Barker E (2000) The NAVDAS source book. Naval Research Laboratory NRL/PU/7530-00-418, Monterey, 151pp

Daley R, Barker E (2001) NAVDAS formulation and diagnostics. Mon Weather Rev 129:869–883

Davis R (2005) Intermediate-depth circulation of the Indian and South Pacific Oceans measured by autonomous floats. J Phys Oceanog 35:683–707

Fofonoff NP, Millard RC (1983) Algorithms for computation of fundamental properties of seawater. Tech Pap Mar Sci UNESCO 44:53

Fox DN, Teague WJ, Barron CN, Carnes MR, Lee CM (2002) The modular ocean data assimilation system. J Atmos Ocean Technol 19:240–252

Halliwell GR (2004) Evaluation of vertical coordinate and vertical mixing algorithms in the HYbrid Coordinate Ocean Model (HYCOM). Ocean Model 7(3–4):285–322

Han Y, van Delst P, Liu Q, Weng F, Yan B, Treadon R, Derber J (2006) JCSDA Community Radiative Transfer Model (CRTM)—version 1, NOAATechnical Report. NESDIS 122:40

Helber RW, Carnes MR, Townsend TL, Barron CN, Dastugue JM (2012) Validation test 1 report for the Improved Synthetic Ocean Profile (ISOP) system, Part I: Stand-alone capability (In preparation)

Hill C, DeLuca C, Balaji V, Suarez M, da Silva A (2004) The architecture of the earth system modeling framework. Comp Sci Eng 6:18–28

Hogan PJ, Hurlburt HE (2000) Impact of upper ocean—topographic coupling and isopycnal outcropping in Japan/East Sea models with $1/8^\circ$ to $1/64^\circ$ resolution. J Phys Oceanogr 30:2535–2561

Hollingsworth A, Lonnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. Tellus 38A:111–136

Hollingsworth A, Lonnberg P (1989) The verification of objective analyses: diagnostics of analysis system performance. Meteor Atmos Phys 40:3–27

Holt TR, Cummings JA, Bishop CH, Doyle JD, Hong X, Chen S, Jin Y (2011) Development and testing of a coupled ocean–atmosphere mesoscale ensemble prediction system. Ocean Dynam 61:1937–1954

Hurlburt HE, Hogan PJ (2008) The Gulf Stream pathway and the impacts of the eddy-driven abyssal circulation and the deep western boundary current. Dynam Atmos Oceans 45:71–101

Hurlburt HE, Wallcraft AJ, Schmitz WJ Jr., Hogan PJ, Metzger EJ (1996) Dynamics of the Kuroshio/Oyashio current system using eddy-resolving models of the North Pacific Ocean. J Geophys Res 101(C1):941–976

Hurlburt HE, Chassignet EP, Cummings JA, Kara AB, Metzger EJ, Shriver JF, Smedstad OM, Wallcraft AJ, Barron CN (2008a) Eddy-resolving global ocean prediction. In: Hecht M, Hasumi H (eds) Ocean modeling in an eddying regime. Geophysical monograph, vol 177. American Geophysical Union, Washington, DC, pp 353–381

Hurlburt HE, Metzger EJ, Hogan PJ, Tilburg CE, Shriver JF (2008b) Steering of upper ocean currents and fronts by the topographically constrained abyssal circulation. Dynam Atmos Oceans 45:102–134. doi:10.1016/j.dynatmoce.2008.06.003

Large WG, Mc Williams JC, Doney SC (1994) Oceanic vertical mixing: a review and a model with a nonlocal boundary layer parameterization. Rev Geophys 32:363–403

Karra B, Rochford PA, Hurlbut H (2000) An optinal definition for mixed layer depth. J Geophys Res 105:16803–16821

Martin MJ, Hines A, Bell MJ (2007) Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. Q J R Meteorol Soc 133:981–995

McLay J, Bishop CH, Reynolds CA (2008) Evaluation of the ensemble transform analysis perturbation scheme at NRL. Mon Weather Rev 136:1093–1108

Merchant CJ, Le Borgne P, Marsouin A, Roquet H (2008) Optimal estimation of sea surface temperature from split-window observations. Rem Sens Environ 112:2469–2484

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 Memo. Report NRL/MR/7320–08- 9148

Metzger EJ, Smedstad OM, Thoppil PG, Hurlburt HE, Franklin DS, Peggion G, Shriver JF, Townsend TL, Wallcraft AJ (2010a) Validation Test Report for the Global Ocean Forecast System V3.0 $-$ $1/12^\circ$ HYCOM/NCODA: Phase II. NRL Memo. Report NRL/MR/7320–10-9236

Metzger EJ, Thoppil PG, Smedstad OM, Franklin DS (2010b) Global Ocean Forecast System V3.0 Validation Test Report Addendum: addition of the Diurnal Cycle. NRL Memo. Report NRL/MR/7320–10-9305

Posey PG, Metzger EJ, Wallcraft AJ, Smedstad OM, Phelps MW (2010) Validation of the $1/12^\circ$ Arctic Cap Nowcast/Forecast System (ACNFS). NRL Memo. Report NRL/MR/7320–10-9287

Riishøjgaard LP (1998) A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. Tellus 50A:42–57

Schlax MG, Chelton DB (1992) Frequency domain diagnostics for linear smoothers. J Am Stat Assoc 87:1070–1081

Wittmann P, Cummings J (2005) Assimilation of altimeter wave measurements into WAVE-WATCH III. In: Proceedings of the 8th international workshop on wave hindcasting and forecasting, North Shore Oahu, Hawaii 16 pp

Chapter 14
A 4D-Var Analysis System for the California Current: A Prototype for an Operational Regional Ocean Data Assimilation System

Andrew M. Moore, Christopher A. Edwards, Jerome Fiechter, Patrick Drake, Emilie Neveu, Hernan G. Arango, Selime Gürol, and Anthony T. Weaver

Abstract In this chapter we will describe a comprehensive 4-dimensional variational ocean data assimilation system that is currently being used in the Regional Ocean Model System for the production of both near real-time and historical ocean analyses of the California Current circulation. The main focus of this article is on the practical aspects of the data assimilation system as applied to an energetic coastal mesoscale circulation environment.

14.1 Introduction

For many years, ocean data assimilation has lagged behind meteorological data assimilation primarily for two reasons. First, developments in data assimilation in meteorology have been driven primarily by the demands of numerical weather prediction (NWP), and second, until relatively recently, observations of the oceans have been relatively scarce compared to the data-rich atmosphere. However, with the revolution in new ocean observing systems such as Argo drifting floats, ocean gliders, and autonomous underwater vehicles, and the push to develop ocean observing and forecasting systems, ocean data assimilation has rapidly reached an advanced level of maturity.

A.M. Moore (✉) · C.A. Edwards · J. Fiechter · P. Drake · E. Neveu
Department of Ocean Sciences, University of California at Santa Cruz, Santa Cruz, CA, USA
e-mail: ammoore@ucsc.edu

H.G. Arango
Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA

S. Gürol
ECMWF, Shinfield Park, Reading, UK

A.T. Weaver
CERFACS, Toulouse, France

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_14,

Global ocean data assimilation efforts have been propelled in part by the coordinated efforts of the Global Ocean Data Assimilation Experiment (GODAE), the growing need for seasonal forecasts, as well as growing concerns over climate change. An increasing number of groups produce global ocean analyses of the circulation of the past (see for example http://www.godae.org/Ocean-products.html). In addition, some operational centers routinely generate analyses for present conditions. Global ocean data assimilation, however, presents a considerable challenge because of the size of the inverse problem involved, so the resolution (horizontal and vertical) of global assimilation products is often limited. The highest resolution products currently available are performed on grids with horizontal grid-spacing typically $\sim 1/4$–$1/6$ degree, so the effective resolutions are probably 2–3 times lower than this considering that anything less than 3 or 4 grid lengths is poorly resolved. This is marginal for resolving much of the important mesoscale variability in the open ocean, and certainly inadequate for capturing important circulation features in coastal regions. For this reason, there has also been a push to develop regional ocean data assimilation systems which utilize higher resolution grids. Two approaches to regional ocean data assimilation are typically used: either the regional model is nested within a global data assimilating model, or the regional model is run stand-alone and boundary condition information is provided by a global assimilating model. While there are clear advantages and disadvantages to both approaches, the stand-alone approach offers greater flexibility since the analyses can be produced using a variety of global circulation estimates as boundary conditions, thus providing a range of uncertainty estimates. Some current regional ocean data assimilation efforts can also be found at http://www.godae.org/Ocean-products.html.

In this chapter we will review a state-of-the-art regional ocean data assimilation system that is run routinely for the U.S. west coast to provide both near-real time and historical analyses for the California Current System (CCS). This is one of several such systems currently in operation in support of the U.S. Integrated Ocean Observing System (IOOS) which comprises seven regional centers, three of which are focused on different parts of the CCS. The CCS is one of 65 Large Marine Ecosystems (LMEs) that have been identified by NOAA and the United Nations Environment Program (UNEP) which collectively account for $\sim 95\,\%$ of global fisheries biomass (see http://www.lme.noaa.gov). The CCS is particularly noteworthy because it is one of five LMEs that are subject to seasonal variations in coastal upwelling in which cold, nutrient rich water is brought to the surface, creating conditions that are favorable for high levels of primary productivity. The CCS is therefore a region of considerable environmental and socio-economic importance.

In Sect. 14.2 we describe the Regional Ocean Modeling System (ROMS) and a detailed summary of the important features of the ROMS 4-dimensional variational (4D-Var) data assimilation algorithms. The specific configuration of ROMS and 4D-Var for the CCS is introduced in Sect. 14.3, while in Sects. 14.4 and 14.5 we describe two ongoing applications: an historical analysis of the CCS circulation, and a near real-time analysis system. We end with a summary in Sect. 14.6.

14.2 ROMS 4D-Var

The ocean analysis system described here is based on the Regional Ocean Modeling System (Shchepetkin and McWilliams 2005). ROMS is a hydrostatic, primitive equations model that uses terrain following coordinates in the vertical and orthogonal curvilinear coordinates in the horizontal to resolve the complex bathymetry and land geometry that characterize many coastal regions. While ROMS is primarily designed with coastal and regional applications in mind, it is also run routinely at basin scales (e.g. Haidvogel et al. 2008). One of the great advantages of ROMS over other ocean models is the high degree of flexibility that it affords the user in terms of available numerical schemes, physical parameterizations, and open boundary conditions. A detailed description of ROMS is beyond the scope of this chapter, but more details about the model can be found at http://www.myroms.org.

14.2.1 Primal Versus Dual Formulation

The ROMS data assimilation system is based on an incremental 4-dimensional variational approach (4D-Var). Two different approaches to 4D-Var are available as part of ROMS: one based on the primal formulation, and the other based on the dual formulation (Courtier 1997). The primal version of ROMS 4D-Var is very similar to that used at several operational NWP centers, and follows closely that of Weaver et al. (2003) for the ocean. The dual version of ROMS 4D-Var comes in two flavors: one approach is based on the Physical-space Statistical Analysis System approach (PSAS) of Cohn et al. (1998), while the other uses the indirect representer method of Egbert et al. (1994). The details and differences between the three ROMS 4D-Var algorithms are not important for what we describe here, and the interested reader is referred to Moore et al. (2011a) for a complete description of each algorithm. In later sections, however, it will be necessary to refer to some specific features of the primal and dual formulations of ROMS 4D-Var, so we begin with a brief overview of the fundamental ideas that underpin each system.

The ocean state vector for ROMS is composed of all ocean grid point values of temperature (T), salinity (S), sea surface height (ζ), and the two components of velocity (u,v). Following the standard notation of Ide (1997) and later extended for the ocean by Daget et al. (2009) we will denote the ROMS state vector as $\mathbf{x} = \left(\mathbf{T}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}}, \boldsymbol{\zeta}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}}\right)^{\mathrm{T}}$, where the vector elements of $\mathbf{x}$ denote column vectors of the grid point values of the state variables. Using this notation, ROMS can then be represented symbolically as:

$$\mathbf{x}(t_i) = M(t_i, t_{i-1})(\mathbf{x}(t_{i-1}), \mathbf{f}(t_i), \mathbf{b}(t_i)) \tag{14.1}$$

where $M(t_i, t_{i-1})$ denotes the operators of the non-linear ROMS that advance the state vector forward in time over the interval $[t_{i-1}, t_i]$, while $\mathbf{f}(t_i)$ denotes the ocean surface forcing (i.e. surface fluxes of momentum, heat and freshwater), and $\mathbf{b}(t_i)$ denotes the lateral open boundary conditions, both over the same interval $[t_{i-1}, t_i]$.

Data assimilation requires the specification of *prior* or background estimates of all the control variables for the model, which in the case of ROMS includes the model initial conditions, $\mathbf{x_b}(t_0)$, the surface forcing, $\mathbf{f_b}(t_i)$, and open boundary conditions, $\mathbf{b_b}(t_i)$. For each *prior*, there will also be an associated *prior* or background error covariance matrix, namely $\mathbf{B_x}$, $\mathbf{B_f}$, and $\mathbf{B_b}$ which embody all of the hypotheses about errors and uncertainties in the *prior* fields. For ease of notation, we will denote the vector that comprises all control variables as $\mathbf{z} = \left(\mathbf{x}^{\mathrm{T}}(t_0), \mathbf{f}^{\mathrm{T}}, \mathbf{b}^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\mathbf{f}$ and $\mathbf{b}$ in the absence of a time argument denote the concatenation in time of the vectors of surface forcing and open boundary conditions over the entire time interval of interest, $[t_0, t_N]$. Similarly we will denote the combined *prior* error covariance matrix of all control variables by the block diagonal matrix $\mathbf{D} = \mathrm{diag}(\mathbf{B_x}, \mathbf{B_f}, \mathbf{B_b})$. In the systems described in later sections it is assumed that the model is free of errors, so the inclusion of control vector elements to account for model errors is not considered here. Furthermore, the *prior* or background error covariance matrix $\mathbf{D}$ is assumed to be time invariant.

In addition to the *prior* estimate of the circulation $\mathbf{x_b}(t)$ from ROMS, there will also be available observations during the same interval $[t_0, t_N]$. The vector of observations is traditionally denoted as $\mathbf{y}^{\mathrm{o}}$ with an associated observation error covariance matrix $\mathbf{R}$. According to Bayes' theorem, the optimal choice of $\mathbf{z}$ that yields the most likely *posterior* circulation estimate is that which minimizes the cost function:

$$J_{NL} = (\mathbf{z} - \mathbf{z_b})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{z} - \mathbf{z_b}) + (\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}^{\mathrm{o}} - H(\mathbf{x})) \tag{14.2}$$

where H denotes the observation operator, and $H(\mathbf{x})$ denotes the circulation estimate $\mathbf{x}(t)$ evaluated at the appropriate observation times and locations (Lorenc 1986; Wikle and Berliner 2007). Since $\mathbf{x}(t)$ is the solution of the nonlinear model (14.1), the cost function J_{NL} in (14.2) is a non linear function of the state vector, which in practical terms means that it may not possess a unique global minimum value. Even in the event that a global minimum does exist for (14.2), locating the global minimum in what may be a complicated topology may be very challenging. Therefore instead of minimizing (14.2) directly, Courtier et al. (1994) proposed the incremental approach in which the desired *posterior* circulation estimate $\mathbf{x_a}(t)$ can be considered as a small departure from the *prior*, namely $\mathbf{x_a}(t) = \mathbf{x_b}(t) + \delta\mathbf{x_a}(t)$. The increment $\delta\mathbf{x_a}(t)$ is a solution of the tangent linearization of (14.1) subject to the surface forcing increments $\delta\mathbf{f_a}(t)$ and open boundary increments $\delta\mathbf{b_a}(t)$ which are also assumed to be small compared to the *prior* estimates $\mathbf{f_b}(t_i)$ and $\mathbf{b_b}(t_i)$ respectively. Specifically, we will denote the tangent linearization of (14.1), the so called tangent linear ROMS (hereafter TLROMS), as:

$$\delta\mathbf{x}(t_i) = \mathbf{M_b}(t_i, t_0)\delta\mathbf{x}(t_0) \tag{14.3}$$

where $\mathbf{M_b}(t_i, t_0)$ denotes the linear operator (also sometimes referred to as the resolvent or propagator matrix) that advances the initial increment $\delta\mathbf{x}(t_0)$ forward in time, and the subscript $\mathbf{b}$ indicates that the linearization is about the *prior* circulation $\mathbf{x_b}(t)$ subject to the *prior* forcing and *prior* open boundary conditions $\mathbf{f_b}$ and $\mathbf{b_b}$. Since $\delta\mathbf{x}(t_i)$ also depends on $\delta\mathbf{f}(t)$ and $\delta\mathbf{b}(t)$, it is sometimes more convenient to express TLROMS as:

$$\delta\mathbf{x}(t_i) = \mathcal{M}_\mathbf{b}(t_i, t_0)\delta\mathbf{z} \tag{14.4}$$

where $\delta\mathbf{z}$ is the vector of control increments composed of $\delta\mathbf{x}(t_0)$, and $\delta\mathbf{f}$ and $\delta\mathbf{b}$ at all times in the interval $[t_0, t_N]$. Therefore, $\mathcal{M}_\mathbf{b}(t_i, t_0)$ is the linear operator that maps a control vector increment into a state vector increment.

Using the incremental approximation, the cost function can be re-expressed as:

$$J = \delta\mathbf{z}^\mathrm{T}\mathbf{D}^{-1}\delta\mathbf{z} + (\mathbf{d} - \mathbf{G}\delta\mathbf{z})^\mathrm{T}\mathbf{R}^{-1}(\mathbf{d} - \mathbf{G}\delta\mathbf{z}) \tag{14.5}$$

where $\mathbf{d} = \mathbf{y}^\mathrm{o} - H(\mathbf{x_b})$ is referred to as the innovation vector, and $\mathbf{G}$ represents the convolution in time of $\mathcal{M}_\mathbf{b}$ with $\mathbf{H}$, where $\mathbf{H}$ is the tangent linearization of the observation operator H. Since the constraints in (14.5) are linear in $\delta\mathbf{z}$, a unique global minimum value of J exists. The vector of control increments that yields the optimal circulation estimate will be denoted as $\delta\mathbf{z_a}$ and the vector of the total control vector as $\mathbf{z_a} = \mathbf{z_b} + \delta\mathbf{z_a}$. At the minimum of (14.5) the gradient is $\partial J/\delta\mathbf{z} = 0$, and the optimal control vector increment is given by $\delta\mathbf{z_a} = \mathbf{Kd}$ where $\mathbf{K}$ is referred to as the Kalman gain matrix. The Kalman gain matrix can be expressed in two equivalent forms as:

$$\mathbf{K} = (\mathbf{D}^{-1} + \mathbf{G}^\mathrm{T}\mathbf{R}^{-1}\mathbf{G})^{-1}\mathbf{G}^\mathrm{T}\mathbf{R}^{-1} \tag{14.6}$$

$$\mathbf{K} = \mathbf{D}\mathbf{G}^\mathrm{T}(\mathbf{G}\mathbf{D}\mathbf{G}^\mathrm{T} + \mathbf{R})^{-1}. \tag{14.7}$$

Equation (14.6) is referred to as the primal form, and corresponds to the case where the minimum of J in (14.5) is found by searching for $\delta\mathbf{z}$ directly in control space. Conversely, (14.7) is referred to as the dual form, and corresponds to the case where the minimum of J is found by searching for $\delta\mathbf{z}$ in observation space. Both approaches yield the same optimal circulation estimate, as demonstrated in ROMS by Moore et al. (2011b). The main advantage of the dual formulation over the primal formulation is that the control vector can be expanded in the former to include corrections for model error without any increase in the dimension of the matrix inverse in (14.7). Until recently the dual formulation was known to suffer from poor convergence properties (El Akkraoui and Gauthier 2010; Moore et al. 2011b) making it difficult to use for large problems. However, Gratton and Tshimanga (2009) have shown that the same rate of convergence of the primal and dual formulations toward the minimum of J can be guaranteed by using a restricted preconditioned conjugate gradient method (RPCG), as confirmed recently by Gürol et al. (2013) in two complex ocean general circulation models, including ROMS.

It is important to note that while (14.6) and (14.7) are written in matrix notation, the matrix inverse in each case is never directly evaluated, but is instead computed by solving an equivalent system of linear simultaneous equations. The latter is achieved iteratively by the direct minimization of J using a conjugate gradient descent algorithm. Similarly, none of the implied matrix multiplications in (14.6) and (14.7) are ever performed explicitly, but instead involve the direct integration of a model. Specifically, $\mathbf{G}$ represents an integration of TLROMS sampled at the appropriate space-time observation points, while $\mathbf{G}^{\mathrm{T}}$ is the adjoint of TLROMS, hereafter referred to as ADROMS. The operator $\mathbf{G}$ is a linear map from the space of the control vector to the space of the observations, while $\mathbf{G}^{\mathrm{T}}$ is a linear map from the dual of the observation space to the dual of the control vector space. The *prior* covariance matrix $\mathbf{D}$ is also described by a model using the diffusion operator approach introduced by Derber and Rosati (1989). Only the observation error covariance $\mathbf{R}$ is explicitly treated as a matrix since for the applications considered here it is assumed to have a simple diagonal structure (i.e. spatially and temporally uncorrelated errors).

ROMS 4D-Var also supports weak constraint data assimilation in which errors in the model formulation can be admitted in the calculation of the ocean circulation estimate. However, in the applications presented here, no explicit account is taken of model errors (i.e. the so-called strong constraint problem), so important considerations as they relate to the treatment of model errors will not be discussed further. Full details of weak constraint 4D-Var in ROMS can be found in Moore et al. (2011a, b).

14.2.2 Inner- and Outer-Loops

In general, we are interested in identifying the minimum of J_{NL} in (14.2), and in practice this proceeds via a sequence of linear minimizations of J in (14.5). Each minimization of (14.5) proceeds iteratively where each iteration is referred to as an inner-loop. During the first set of inner-loop iterations, $\mathbf{G}$ and $\mathbf{G}^{\mathrm{T}}$ are linearized about the time evolving *prior* circulation estimate $\mathbf{x_b}(t)$ resulting from the *prior* control vector $\mathbf{z_b}$. The kth sequence of inner-loops will be represented in sequel by the superscript k. For the first sequence of inner-loops $k = 1$, and when the increment $\delta\mathbf{z}^1$ has been identified that minimizes J, a new circulation estimate $\mathbf{x}^1(t)$is computed using the updated control vector $\mathbf{z}^1 = \mathbf{z_b} + \delta\mathbf{z}^1$, and a new sequence of inner-loops performed during which $\mathbf{G}$ and $\mathbf{G}^{\mathrm{T}}$ are linearized about $\mathbf{x}^1(t)$. The repeated application of this procedure is equivalent to minimizing (14.2) using a Gauss-Newton method (Lawless et al. 2005), and the updates of the circulation $\mathbf{x}^k(t)$ about which TLROMS and ADROMS are linearized are referred to as outer-loops.

14.2.3 Conjugate Gradient Descent and Preconditioning

Following the customary approach adopted in NWP, the minimization of (14.5) in ROMS is preconditioned by a change of variable, namely $\delta\mathbf{v} = \mathbf{D}^{-1/2}\delta\mathbf{z}$.

Such an approach is essential for any practical application of 4D-Var, since it greatly improves the convergence of the conjugate descent algorithm toward the minimum of J. In the primal algorithm of ROMS 4D-Var, a Lanczos formulation of the conjugate gradient approach is used following Fisher and Courtier (1995). In the dual algorithm of ROMS 4D-Var, a Lanczos formulation of the RPCG algorithm of Gratton et al. (2009) is used following Gürol et al. (2013), and is hereafter referred to as RLanczos. RLanczos is a specific version of the Range Space Full Orthogonalization Method (RSFOM) developed by Gratton et al. (2009). The use of the Lanczos formulation in both the primal and dual formulations introduces considerable utility to the ROMS 4D-Var system. For example, due to limited computer resources it is neither possible (or even desirable) to iterate 4D-Var to complete convergence, in which case $\delta\mathbf{z_a} = \tilde{\mathbf{K}}\mathbf{d}$ where $\tilde{\mathbf{K}}$ is the gain matrix that is used in practice to compute the optimal control vector increment, and represents a reduced rank approximation of the true gain matrix. It is straightforward to express $\tilde{\mathbf{K}}$ (hereafter referred to as the "practical gain matrix") in terms of the Lanczos vectors of either the primal or dual form, which in turn can be used to compute the impact of observations on the resulting analysis, as well as other useful diagnostics such as model error (see Moore et al. 2011c, 2012 for details).

14.2.4 Covariance Models and Balance Operators

As noted above, each block diagonal component of the *prior* error covariance matrix $\mathbf{D} = \mathrm{diag}(\mathbf{B_x}, \mathbf{B_f}, \mathbf{B_b})$ is also expressed as a model in ROMS. Specifically, we follow the approach of Weaver et al. (2005) in which first the initial condition increment $\delta\mathbf{x}(t_0)$ is expressed as the sum of the balanced and unbalanced components of the circulation. The unbalanced components of $\delta\mathbf{x}(t_0)$ are assumed to be mutually uncorrelated with covariance matrix $\varSigma\mathbf{C}\varSigma^{\mathrm{T}}$, where $\mathbf{C}$ is a univariate correlation matrix, and $\varSigma$ is a diagonal matrix of standard deviations. The initial condition *prior* error covariance matrix $\mathbf{B_x}$ can then be factorized as:

$$\mathbf{B_x} = \mathbf{K_b}\varSigma\mathbf{C}\varSigma^{\mathrm{T}}\mathbf{K_b^{\mathrm{T}}} \tag{14.8}$$

where $\mathbf{K_b}$ is a multivariate balance operator, and describes the covariances between errors in the balanced components of $\delta\mathbf{x}(t_0)$. As in Weaver et al. (2005) and Ricci et al. (2005), $\mathbf{K_b}$ is based on the T-S characteristics and dominant dynamical balances in the ocean, namely hydrostatic balance and geostrophic balance.

Following Weaver and Courtier (2001), the univariate correlation matrix $\mathbf{C}$ is assumed to be separable in the horizontal and the vertical. The horizontal (vertical) correlation function is then modeled as the solution of a 2-dimensional (1-dimentional) pseudo-diffusion equation. The product of the pseudo time interval and diffusion coefficient determines the desired correlation length, and can be varied spatially. At the present time, the balance operator is applied only to the initial condition *prior* error covariance matrix $\mathbf{B_x}$.

14.2.5 Background Quality Control Checks

A new feature of the ROMS 4D-Var systems is the recent introduction of a background quality control of the observations to reject those data that are subject to gross errors (Hollingsworth et al. 1986; Lorenc and Hammon 1988). The approach used is based on that described by Järvinen and Undén (1997) and Andersson and Järvinen (1999) and used in NWP. Specifically, the elements of the innovation vector **d** are compared with their expected error (assuming observation and background errors are uncorrelated errors) according to:

$$\left(y_i^o - y_i^b\right)^2/\sigma_b^2 < \alpha\left(1 + \sigma_o^2/\sigma_b^2\right) \tag{14.9}$$

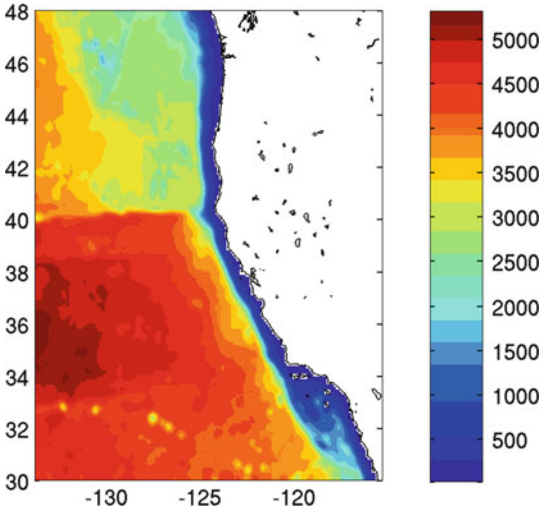
where y_i^o is the ith observation, y_i^b is the ith element of the vector $H(\mathbf{x_b})$, the background evaluated at the observation locations, and σ_o and σ_b are the standard deviations of the observation and background errors at the observation points. The threshold parameter α generally depends on each observation type, and appropriate values can be determined from historical analyses as described by Andersson and Järvinen (1999). Observations that do not satisfy (14.9) are rejected prior to the analysis. This has the effect of eliminating from the analysis observations that are subject to large gross errors. In addition, observations that are inconsistent with the model due, for example, to model deficiencies, are also eliminated from the analysis. The introduction of the background quality control in ROMS based on (14.9) has been found to yield substantial improvement in the behavior and convergence properties of the dual 4D-Var algorithm in particular.

14.3 Configuration of ROMS CCS and 4D-Var

The California Current System (CCS) is an eastern boundary current characterized by a pronounced seasonal cycle of upwelling and by energetic mesoscale circulations (Hickey 1998; Checkley and Barth 2009), and provides a challenge for linear data assimilation methods such as 4D-Var. The ROMS CCS domain and circulation is described in detail by Veneziani et al. (2009) and Broquet et al. (2009a, b), and spans the region 134°W to 116°W and 31°N to 48°N, with 1/10th degree resolution in the horizontal and 42σ-levels in the vertical. The model domain and bathymetry are shown in Fig. 14.1.

The model forcing is derived from either 6 hourly or daily averaged atmospheric boundary layer fields from different sources depending on the application. In the case of the 13 year reanalysis project described in Sect. 14.4.1 and the near real-time analysis system described in Sect. 14.5, daily averaged atmospheric variables at standard heights were taken from the Naval Research Laboratory's Coupled Ocean-Atmosphere Mesoscale Prediction System (COAMPS) (Doyle

Fig. 14.1 The ROMS CCS domain and bathymetry



et al. 2009). The ocean surface fluxes were derived using the bulk formulations of Liu et al. (1979) and Fairall et al. (1996a, b). However, historical COAMPS analyses are not available prior to 1999, so in the case of the 31 year reanalysis described in Sect. 14.4.2, a combination of 6 hourly fields from the ECMWF ERA40 and ERA Interim projects were used, along with the cross-calibrated, multiplatform (CCMP) ocean wind product of Atlas et al. (2011). In either case, the surface forcing fields obtained represent the background surface forcing, $\mathbf{f_b}$, for 4D-Var introduced in Sect. 14.2.1.

The model domain has open boundaries at the northern, southern, and western edges, and at these boundaries the tracer and velocity fields are prescribed, while the free surface and vertically integrated flow are subject to Chapman (1985) and Flather (1976) boundary conditions respectively. The prescribed open boundary solution was taken from the Simple Ocean Data Assimilation product (SODA) of Carton and Giese (2008) in the case of the reanalyses of Sect. 14.4, and from the World Ocean Atlas 2005 (WOA05) in the case of the near real-time system of Sect. 14.5. In either case, these fields represent the background open boundary conditions, $\mathbf{b_b}$, for 4D-Var introduced in Sect. 14.2.1.

A sponge layer was also used adjacent to each open boundary where viscosity increased linearly from 4 $\mathrm{m^2s^{-1}}$ in the interior to 400 $\mathrm{m^2s^{-1}}$ at the boundary over a distance of 100 km.

The observations assimilated into the model were collected by various platforms, and will be described in more detail in Sects. 14.4 and 14.5. To reduce data redundancy, all observations of the same state variable within each model grid cell, over a 6 h time window, were combined to form "super observations," and the standard deviation of the observations that contribute to the super observation in each grid cell was used as an estimate of the error of representativeness.

Observation errors were assumed to be uncorrelated in space and time, resulting in a diagonal observation error covariance matrix, **R**. The variances along the main diagonal of **R** were assigned as the sum of measurement error and the error of representativeness. Measurement errors were chosen with the following standard deviations: 0.02–0.04 m for dynamic topography depending on the instrument; 0.3–1°C for SST depending on the satellite platform; 0.1–0.5°C for in situ T; and 0.01–0.1 for in situ S.

The background error standard deviations for the initial condition components of the control vector (i.e. the elements of the diagonal matrix Σ) were estimated based on the variance of a long run of the model subject only to surface forcing and boundary conditions (i.e. no data assimilation). The surface forcing and boundary condition fields used depend on the application as described above. However, in the case of salinity, past experience has revealed that the background errors computed using this method are too large, so the standard deviations for S were capped at 0.1. The temporal variability of the surface forcing fields for the appropriate period was used as the variance for the background surface forcing error, and the open boundary condition background error variances were chosen to be the variances of the appropriate data (as described above) at the boundaries.

As noted in Sect. 14.2.4, each block diagonal component of the background error covariance matrix **D** was modeled using the diffusion operator approach of Weaver and Courtier (2001). The capability to have spatially varying correlation lengths is a fairly recent addition to the ROMS 4D-Var code, so in the calculations and applications presented in Sects. 14.4 and 14.5, the horizontal and vertical correlation lengths were held constant over the model domain. The decorrelation length scales used to model the *prior* errors of all initial condition control variable components of $\mathbf{B_x}$ were 50 km in the horizontal and 30 m in the vertical. Horizontal correlation scales chosen for the background surface forcing error components of $\mathbf{B_f}$ were 300 km for wind stress and 100 km for heat and freshwater fluxes. The correlation lengths for the background open boundary condition error components of $\mathbf{B_b}$ were chosen to be 100 km in the horizontal and 30 m in the vertical. No explicit account is taken of temporal correlations in any of the background errors in the current version of ROMS 4D-Var, although this capability is currently under development. However, the surface forcing, and boundary condition increments were only computed daily and interpolated to each intervening model time step, a procedure which effectively introduces some temporal correlation of the errors. The correlation lengths chosen for the *prior* errors are typically estimated using semi-variogram techniques that have traditionally been applied to observational data (e.g. Banerjee et al. 2004; Milliff et al. 2003; Matthews et al. 2011). However, some level of subjective tuning of the correlation lengths is also typically required to optimize the performance of the 4D-Var algorithm. A discussion of the choice of the aforementioned background error covariance parameters for the CCS can be found in Broquet et al. (2009a, 2009b, 2011) and Moore et al. (2011b). In all of the calculations presented here, the multivariate balance operator was not used.

14.4 CCS Historical Analyses

ROMS CCS is currently being used in conjunction with 4D-Var to construct two sequences of historical analyses for the circulation along the west coast of North America. The first of these analyses, referred to as WCRA13, is a 13 year sequence that spans the period Jan. 1999–Dec. 2011, while the second sequence, referred to as WCRA31, covers the 31 year period Jan. 1980–Dec. 2011. The two analyses use identical configurations for ROMS, but differ in the *prior* surface forcing used. During the overlapping period 1999–2011 of the two analyses, the observations assimilated into the model are identical.

14.4.1 WCRA13

In WCRA13, the *prior* surface forcing is derived from the NRL COAMPS model introduced in Sect. 14.3. While the COAMPS fields for the CCS are not available before Jan. 1999, they do span the full period of WCRA13. The standard height atmospheric variables are actually derived from four different nests of COAMPS with horizontal resolution ranging from 3 to 81 km from the inner to the outer nest (Doyle et al. 2009). Only data from the three inner-most nested grids are used in ROMS and yield surface fields with a resolution of 3–9 km near the coast. This is the highest resolution atmospheric forcing data set currently available for the CCS region, and COAMPS verifies well against independent observations, indicating that it is a high quality product. High horizontal resolution is important for the surface forcing because many of the important regions of coastal upwelling along the U.S. west coast are due to topographically enhanced regions of wind stress curl.

14.4.2 WCRA31

In WCRA31, the *prior* surface forcing is derived from a combination of atmospheric analysis products. The surface winds are taken from the cross-calibrated multiplatform product (CCMP) of Atlas et al. (2011) which is a 2D-Var analysis of all available surface wind observations, using the ECMWF ERA40 reanalysis as the *prior* estimate for the period 1987–1999, and the operational ECMWF analysis after 1999. Consequently, we use 6 hourly sea level pressure, radiation fluxes, precipitation, and standard height temperature, humidity from the ERA40 analysis so that the ROMS derived heat and fresh water fluxes are consistent with the *prior* used for the winds. The resolution of the CCMP wind fields is 25 km, while that of the ERA-40 reanalysis fields is 2.5°. Prior to 1987, the ERA-40 reanalysis fields are used, while after 1987 the ERA-40 reanalysis fields are used in conjunction with CCMP winds since ERA40 is the *prior* for the CCMP analyses.

After 2001 ERA-Interim reanalysis fields, which have a resolution of 0.7°, are used in conjunction with CCMP winds.

A comparison of the WCRA13 and WCRA31 analyses during the overlapping period 1999–2011 will reveal the impact of the resolution of the *prior* surface forcing fields on the circulation estimates.

14.4.3 WCRA Observations

The observations assimilated into the model during each WCRA analysis were collected by various platforms, and are summarized in Table 14.1, along with the combined measurement error and error of representativeness that is assumed for the diagonal entries of **R**.

All of the in situ hydrographic profiles of T and S were taken from the quality controlled EN3 data archive maintained by the UK Met Office as part of the European Union ENSEMBLES project (Ingleby and Huddleston 2007). The version of EN3 used here is version 2a which includes the XBT and MBT temperature error corrections of Levitus et al. (2009).

The in situ observations from the EN3 archive are available from a variety of different observing platforms including: expendable bathythermographs (XBTs), mechanical bathythermographs (MBTs), conductivity temperature depth devices (CTDs), free drifting Argo profiling floats, and autonomous pinniped bathythermographs (APBs) in the form of tagged marine mammals. Figure 14.2 shows a time series of $\log_{10}$ of the total number of super observations from EN3 and each satellite platform that fall within the ROMS CCS model domain during each month of the year spanning the full period of WCRA31.

The SSH observations assimilated into the model are in the form of 1 day gridded composites of the mean dynamic topography from Aviso. Before assimilation, the mean dynamic topography of the Aviso data averaged over the ROMS CCS domain was corrected to match that of the model. This data is used rather than the raw along-track data because at the present time there is no temporal correlation included in the *prior* error covariance matrices **D** or **R**. The result is that information from individual along track observations is lost quite quickly and becomes ineffective for constraining the model unless some additional effort is made to persist the along-track observations over time. This problem is alleviated by using the gridded products, although we appreciate that this is not an ideal solution because of the limitations of the objective mapping technique used to map the altimeter observations onto a regular grid. In addition, satellite SSH observations near the coast are known to be unreliable (Saraceno et al. 2008) so only observations that are more than 50 km from the coast are assimilated into ROMS.

As indicated in Table 14.1, SST observations are available from several different platforms. The along track data from each platform are used, and when multiple platforms are concurrently available, the data from each platform are combined to form super observations. Only the number of super observations for each individual platform are shown in Fig. 14.2.

Table 14.1 A summary of the observation types, observing platforms, data sources, the nominal combined measurement and representation errors, and the period covered. The combined error is replaced by the standard deviation of the observations about the super observation value if larger than the assumed nominal error

Observation type	Observing platform	Source	Combined error	Period covered
SSH	Altimeter	Aviso, 1 day average	0.04 m	1993–2010
SST	AVHRR/ Pathfinder	NOAA Coast Watch	0.6°C	1981–2011
SST	AMSR-E	NOAA Coast Watch	0.7°C	2002–2010
SST	GOES	NOAA Coast Watch	1°C	2001–2010[a]
SST	MODIS-Terra	NASA JPL	0.5°C	2000–2011
Hydrographic data	Various	UK Meteorological Office	0.5°C for T 0.1 for S	1950–2011

[a]The GOES SST are seriously biased during the period 2001–2002, so they are not used in ROMS 4D-Var until 2003

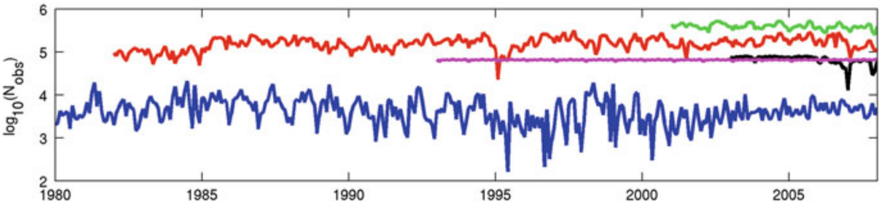


Fig. 14.2 A time series of $\log_{10}$ the total number of super observations available each month from EN3 and each satellite observing platform within the ROMS CCS model domain during each month of the year during the period spanned by WCRA31. *Blue*: In situ observations from EN3; *Red*: SST from AVHRR/PathFinder; *Black*: SST from AMSR-E; *Green*: SST from GOES; *Magenta*: SSH from Aviso

14.4.4 WCRA 4D-Var Configuration

Both WCRA13 and WCRA31 take the form of sequences of overlapping 8 day analyses each separated by 4 days as illustrated in Fig. 14.3. During each analysis cycle, all of the available observations during the cycle time interval are assimilated into ROMS using the dual formulation of 4D-Var in the form of PSAS. Each analysis cycle utilizes 1 outer-loop and 15 inner-loops which has been demonstrated by Broquet et al. (2009a), Moore et al. (2011b) and Gürol et al. (2013) to yield adequate convergence toward the cost function minimum in the CCS region. The *prior* estimate for the initial conditions for each assimilation cycle is the *posterior*

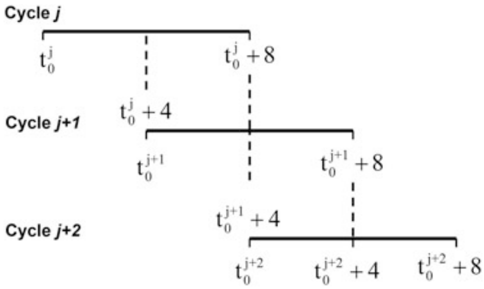


Fig. 14.3 A schematic illustrating the overlapping 8 day data assimilation cycles used in WCRA13 and WCRA31. The starting time for cycle j is denoted as t_0^j and the mid-point and ending times as $t_0^j + 4$ and $t_0^j + 8$ respectively. As indicated, the ending time of 4D-Var analysis cycle j corresponds to the mid-point of cycle $j + 1$ and the starting time of cycle $j + 2$. The *prior* circulation initial condition for cycle $j + 1$ is taken as the *posterior* circulation estimate at the mid-point of cycle j

circulation estimate at the mid-point of the previous analysis cycle. The advantages of overlapping cycles are two-fold. First, it is well known that the 4D-Var analysis cycle is equivalent to a Kalman Smoother, in which case the uncertainty in the analysis will be at a minimum at the mid-point of the cycle, hence each analysis cycle will start from the best possible *prior* initial condition. Second, at each initial analysis time, an ensemble of three circulation estimates will be available allowing for the possibility of ensemble averaging to further minimize the uncertainty of the *posterior* circulation estimate. However, with overlapping analysis cycles, the observations collected during the first half of each cycle will be correlated with the background circulation during the same period. Since these correlations are not accounted for in the current 4D-Var analysis system, this may lead to overweighting of the analysis to some of the observations.

The dual formulation of ROMS 4D-Var was chosen for the historical analyses because of the added utility that is available in the form of diagnostic post-processing tools. As part of the ROMS 4D-Var suite, drivers are available for computing the *a posteriori* impact of each observation on different scalar measures of the circulation via $\tilde{\mathbf{K}}^T$, the transpose of the practical gain matrix. In addition, the adjoint of the entire dual 4D-Var algorithm is available, and the sensitivity of the same scalar circulation indices to uncertainties in the observations can be quantified, as well as the expected errors in each index (Moore et al. 2011c, 2012).

14.4.5 Background Quality Control of Observations for WCRA

Andersson and Järvinen (1999) describe a procedure by which suitable values of the threshold parameter α in (14.9) can be estimated from the frequency

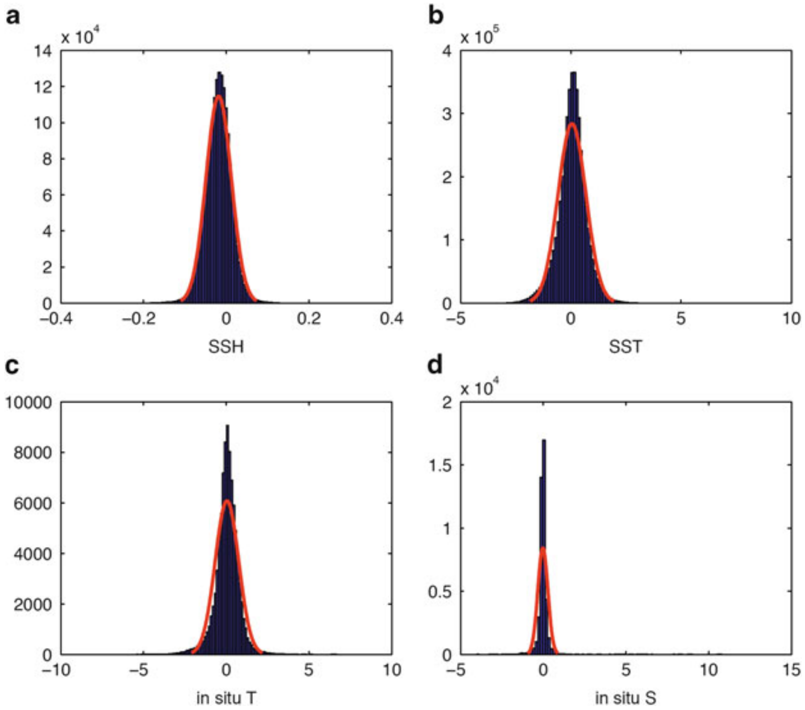


Fig. 14.4 Frequency distributions f of the elements of the innovation vector **d** for (**a**) SSH observations (m), (**b**) SST observations (K), (**c**) in situ temperature observations (K), and (**d**) in situ observations of salinity. The distributions are computed from 1 year of 7 day 4D-Var cycles during 1999. The *red curves* show the best fit Gaussian distribution in each case

distribution of the elements of the innovation vector **d** computed from historical analyses. In our case, no sequence of historical analyses is available, so instead we examined the innovations from a randomly chosen year (1999) during which all observations were assimilated into the model. The frequency distributions, f, of the innovation elements for each observation platform, and the transformed distribution $\hat{f} = \sqrt{-2\ln[f/\max(f)]}$ were computed for the random year following Andersson and Järvinen (1999), where f is the number of data in each bin of the histogram. The resulting histogram distributions of f and $\hat{f}$ for satellite SST (AVHRR/Pathfinder), SSH (Aviso), in situ temperature and in situ salinity (both from EN3) are shown in Figs. 14.4 and 14.5 respectively. Also shown in Fig. 14.4 is the best fit Gaussian distribution for each histogram. The transformed distribution $\hat{f}$ highlights the tails of the distribution and is therefore more convenient for viewing the outlier innovations. The slopes of the lines superimposed on the transformed distributions in Fig. 14.5 represent the standard deviation for the best fit Gaussian distributions in Fig. 14.4, and are estimates of $(\sigma_b^2 + \sigma_o^2)^{1/2}$. Figure 14.5a, b indicate that for satellite SST and SSH, there are relatively few outliers in the elements of **d**. Conversely,

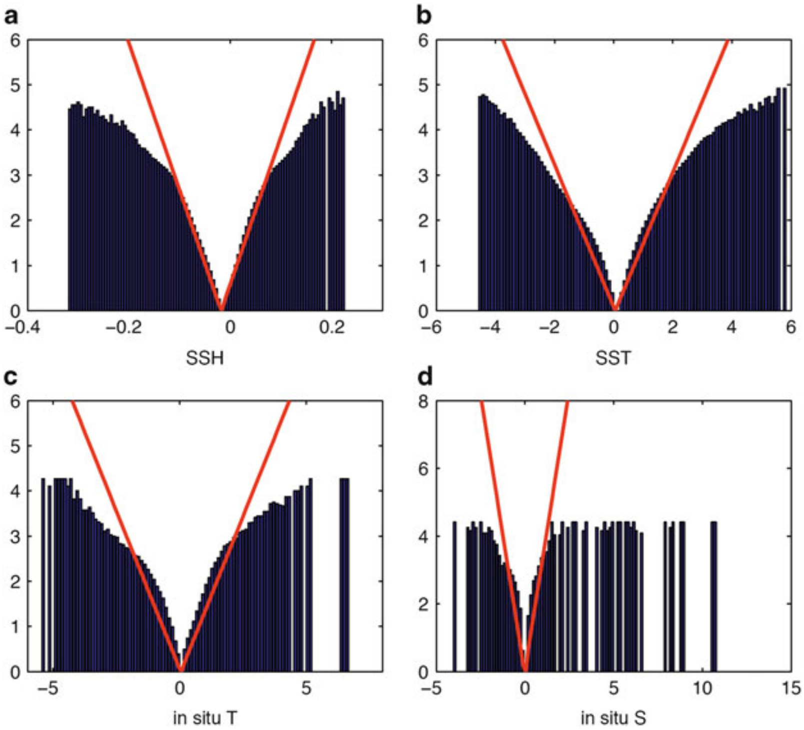


Fig. 14.5 The transformed frequency distributions $\hat{f}$ corresponding to the same data groupings shown in Fig. 14.4. The *red curves* in each case represent the lines $y = \text{abs}\left(\text{x}/\sigma_f\right)$ where σ_f is the estimate of $(\sigma_o^2 + \sigma_b^2)^{1/2}$ corresponding to the best fit Gaussian distributions shown in Fig. 14.4

for the in situ observations, there are typically a significant number of outliers, particularly in the case of salinity. Following Andersson and Järvinen (1999), the threshold parameter α was chosen to reflect a significant departure of the straight lines in Fig. 14.5 from the distribution. Based on Fig. 14.5, we apply (14.9) only to the in situ observations with $\alpha = 16$, which corresponds to rejecting observations that are greater than $\sim 4(\sigma_b^2 + \sigma_o^2)^{1/2}$ from the mean of the distribution, which corresponds to ∼2.8°C and ∼1.2 for in situ temperature and salinity respectively. When applying (14.9), the prescribed σ_b and σ_o evaluated at the observation locations are used during each 4D-Var cycle. For this choice of α it was found that on average less than 1 % of the in situ observations are rejected during 4D-Var.

14.4.6 Preliminary Results

At the time of writing, WCRA31 is underway with ∼5 years completed so far and will require approximately another 3 months to complete. In this section we are

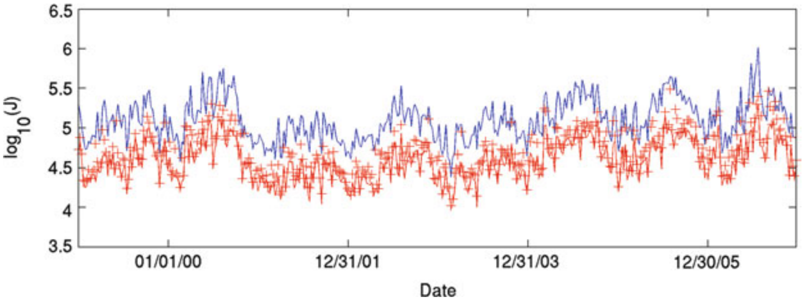


Fig. 14.6 Time series of the initial (*blue*) and final (*red*) values of the cost function $\log_{10}(J)$ in (14.5) for each 4D-Var cycle of the preliminary WCRA13 analysis sequence. Also shown are the values of J_{NL} from (14.2) computed at the end of the single outer-loop (*red pluses*)

therefore only able to show some preliminary results from a sequence of benchmark integrations for WCRA13 which were performed to assess storage requirements, execution time, efficacy of the assimilation set-up, etc. The final analyses for WCRA31 and WCRA13 are now available to the community via a dedicated Opendap web server at http://oceanmodeling.pmc.ucsc.edu.

To illustrate the performance of 4D-Var in ROMS CCS, Fig. 14.6 shows a time series of the initial and final values of the cost function J from (14.5) for each data assimilation cycle of the preliminary WCRA13 sequence. In general, the cost function is reduced by $\sim$ 50–70 % during most cycles. Also shown in Fig. 14.6 is the value of J_{NL} from (14.2) which is consistent with J during most cycles. Clearly the 4D-Var procedure moves the model circulation estimates closer to the observations.

14.5 The CCS Near Real-Time Analysis and Forecast System

In addition to the historical analyses described in Sect. 14.4, ROMS CCS is also being used in conjunction with 4D-Var to compute analyses for the ocean in near real-time. The system described here builds on previous experience using ROMS 4D-Var in real-time analysis and ensemble prediction mode in the Caribbean Sea and the Gulf of Mexico by Powell et al. (2009). The near real-time aspect of this system places strict constraints on the *prior* information and observations that can be used in the assimilation system. The *prior* surface forcing fields are computed by ROMS based on standard height atmospheric variables, surface radiation fluxes and precipitation from COAMPS that is run at NRL Monterey in a near real-time mode. There is no near real-time product available for constraining the model at the open boundaries, so climatological open boundary conditions derived from WOA05 are used as the prior in this case.

The CCS near real-time analysis system is run every Monday morning using all of the available ocean observations from the previous week. The primal 4D-Var

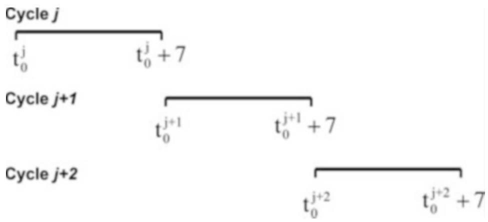


Fig. 14.7 A schematic to illustrate the 7 day data assimilation cycles used in the near real-time system. The starting and ending times of cycle j are denoted t_0^j and $t_0^j + 7$ respectively, and the starting time of cycle $j + 1$ corresponds to the ending of time of cycle j. The *prior* circulation initial condition for cycle $j + 1$ is taken as the *posterior* circulation estimate at the end of cycle j

Table 14.2 A summary of the observations and platforms currently used in the UCSC near real-time analysis system for the CCS. CalCOFI Line 67 is a repeat glider line that runs offshore from the California coast just south of Monterey Bay out to 316 km offshore. This CalCOFI line is maintained by the Monterey Bay Aquarium Research Institute (MBARI). *OSTIA*: operational sea surface temperature and sea ice analysis, and is described by Stark et al. (2007), *CaLCOFI*: the California cooperative fisheries investigation

Observation type	Observation platform	Source	Combined error
SSH	Altimeter	Aviso	0.02 m
SST	Various	OSTIA UK Met Office	0.4°C
Hydrographic data	Glider, CalCOFI Line 67	MBARI	0.1°C for T 0.01 for S

algorithm is used in this case, for historical reasons, in conjunction with a 7 day data assimilation window in which the *prior* initial condition for each analysis cycle is the *posterior* circulation estimate at the end of the previous cycle. In this system, the 4D-Var control vector is composed of the initial conditions only. The procedure used is illustrated schematically in Fig. 14.7.

Because of the near real-time aspect of this system, the number and type of ocean observations that are available for assimilation is limited. Information about the observations that are currently used is given in Table 14.2.

An example analysis is shown in Fig. 14.8 which shows the *prior* and increments for SST, sea surface salinity (SSS), sea surface height (SSH), for the analysis cycle starting on 2 Feb., 2012. The *prior* fields for SST, SSS and SSH all reveal the complex nature of the meso-scale circulation environment associated with the CCS. The increments reveal that most of the corrections that are being made to the *prior* during this cycle are generally at the mesoscale also, although some larger scale corrections are present as well, as for example in SSH (Fig. 14.8e) in the northern part of the domain.

The near real-time analyses, like that of Fig. 14.8, are produced in support of the Central and Northern California Ocean Observing System (CeNCOOS) and are freely available at http://oceanmodeling.pmc.ucsc.edu/ccsnrt. Typical users include fisherman, marine planners, and search and rescue organizations.

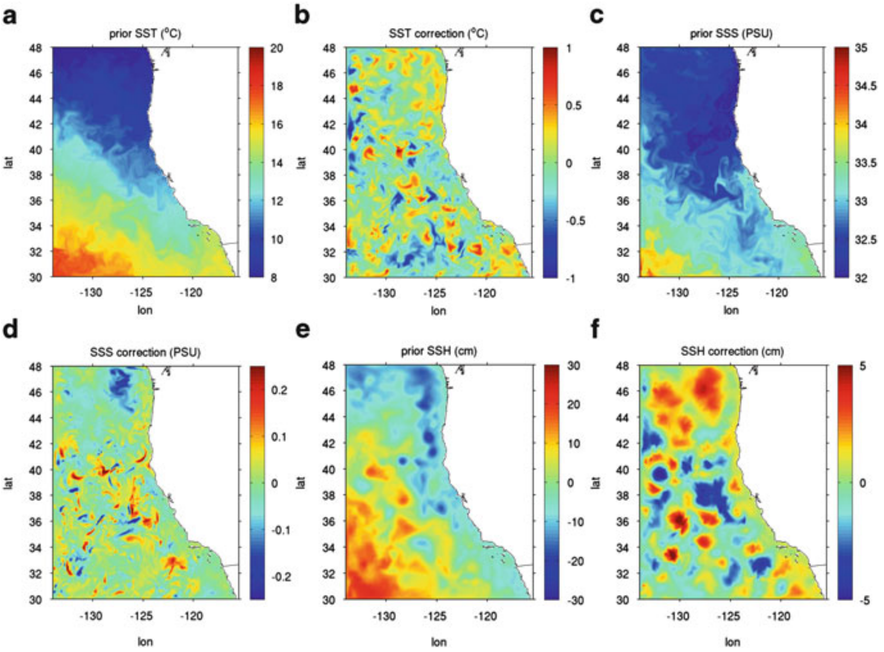


Fig. 14.8 An example of the surface analysis for the 4D-Var cycle starting on 2 Feb., 2012 as part of the near real-time system. (**a**) *prior* SST, (**b**) SST increment, (**c**) *prior* sea surface salinity (SSS), (**d**) SSS increment, (**e**) *prior* sea surface height (SSH) and (**f**) SSH and increments

14.6 Summary

Our aim in this chapter is to demonstrate that ocean data assimilation has reached a maturity whereby sophisticated, state-of-the-art systems have been developed for a widely used community ocean model. The ROMS 4D-Var system has features that are comparable to those used in some operational NWP systems, and this article demonstrates that it is now possible to compute regional ocean analyses at eddy resolving resolutions to generate not only historical circulation estimates but also analyses in near real-time. The ROMS 4D-Var system used here is freely available to the ocean modeling community at large (http://myroms.org), and the experience gained from the activities described here will provide valuable guidance for future efforts in the California Current system and elsewhere.

Acknowledgements The authors gratefully acknowledge the past and continuing support of the U.S. National Science Foundation, the U.S. Office of Naval Research (N00014-10-1-0476), and the U.S. National Oceanic and Atmospheric Administration (CeNCOOS Award NA08NOS4730282). This material is based on work supported by the National Science Foundation under Grant No. 1061434. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Andersson E, Järvinen H (1999) Variational quality control. Q J Roy Meteorol Soc 125:679–722

Atlas R, Hoffman RN, Ardizzone J, Leidner SM, Jusem JC, Smith DK, Gombos D (2011) A cross-calibrated multiplatform ocean surface wind velocity product for meteorological and oceanographic applications. Bull Am Meteorol Soc. doi:10.1175/2010BAMS2946.1

Broquet G, Edwards CA, Moore AM, Powell BS, Veneziani M, Doyle JD (2009a) Application of 4D-variational data assimilation to the California Current System. Dyn Atmos Oceans 48:69–91

Broquet G, Moore AM, Arango HG, Edwards CA, Powell BS (2009b) Ocean state and surface forcing correction using the ROMS-IS4DVAR data assimilation system. Mercator Ocean Q Newsl 34:5–13

Broquet G, Moore AM, Arango HG, Edwards CA (2011) Corrections to ocean surface forcing in the California Current System using 4D-variational data assimilation. Ocean Model 36:116–132

Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Monographs on statistics and applied probability, vol 101. Chapman and Hall/CRC, Boca Raton

Carton JA, Giese BS (2008) A reanalysis of ocean climate using simple ocean data assimilation (SODA). Mon Weather Rev 136:2999–3017

Chapman DC (1985) Numerical treatment of cross-shelf open boundaries in a barotropic coastal ocean model. J Phys Oceanogr 15:1060–1075

Checkley DM, Barth JA (2009) Patterns and process in the California current system. Prog Oceanogr 83:49–64

Cohn SE, Da Silva A, Guo J, Sienkiewicz M, Lamich D (1998) Assessing the effects of data selection with the DAO physical-space statistical analysis system. MonWeather Rev 126:2913–2926

Courtier P (1997) Dual formulation of four-dimensional variational assimilation. Q J Roy Meteorol Soc 123:2449–2461

Courtier P, Thépaut J-N, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var using an incremental approach. Q J Roy Meteorol Soc 120:1367–1388

Daget N, Weaver AT, Balmaseda MA (2009) Ensemble estimation of background-error variances in a three-dimensional variational data assimilation system for the global ocean. Q J Roy Meteorol Soc 135:1071–1094

Derber J, Rosati A (1989) A global oceanic data assimilation system. J Phys Oceanogr 19:1333–1347

Doyle JD, Jiang Q, Chao Y, Farrara J (2009) High-resolution atmospheric modeling over the Monterey Bay during AOSN II. Deep Sea Res II 56:87–99

Egbert GD, Bennett AF, Foreman MCG (1994) TOPEX/POSEIDON tides estimated using a global inverse method. J Geophys Res 99:24,821–24,852

El Akkraoui A, Gauthier P (2010) Convergence properties of the primal and dual forms of variational data assimilation. Q J Roy Meteorol Soc 136:107–115

Fairall CW, Bradley EF, Godfrey JS, Wick GA, Ebson JB, Young GS (1996a) Cool-skin and warm layer effects on the sea surface temperature. J Geophys Res 101:1295–1308

Fairall CW, Bradley EF, Rogers DP, Ebson JB, Young GS (1996b) Bulk parameterization of air-sea fluxes for tropical ocean global atmosphere coupled-ocean atmosphere response experiment. J Geophys Res 101:3747–3764

Fisher M, Courtier P (1995) Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF Tech Memo 220:1–26

Flather RA (1976) A tidal model of the northwest European continental shelf. Memoires de la Societe Royale des Sciences de Liege 6(10):141–164

Gratton S, Tshimanga J (2009) An observation-space formulation of variational assimilation using a restricted preconditioned conjugate gradient algorithm. Q J Roy Meteorol Soc 135:1573–1585

Gratton S, Toint PL, Tshimanga J (2009) Inexact range-space Krylov solvers for linear systems arising from inverse problems. FUNDP Tech. Report, 09/20

Gürol, S, Weaver AT, Moore AM, Piacentini A, Arango HG, Gratton S (2013) B-Preconditioned minimization algorithms for variational data assimilation with the dual formulation. Q J Roy Meteorol Soc (In press)

Haidvogel DB, Arango HG, Budgell WP, Cornuelle BD, Curchitser E, Di Lorenzo E, Fennel K, Geyer WR, Hermann AJ, Lanerolle L, Levin J, McWilliams JC, Miller AJ, Moore AM, Powell TM, Shchepetkin AF, Sherwood CR, Signell RP, Warner JC, Wilkin J (2008) Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the regional ocean modeling system. J Comput Phys 227:3595–3624

Hickey BM (1998) Coastal oceanography of western North America from the tip of Baja, California to Vancouver Island. Sea 11:345–393

Hollingsworth A, Shaw DB, Lönnberg P, Illari L, Arpe K, Simmons A (1986) Monitoring of observation and analysis quality by data assimilation systems. Mon Weather Rev 114:861–879

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

Ingleby B Huddleston M (2007) Quality control of ocean temperature and salinity profiles–historical and real-time data. J Mar Syst 65:158–175

Järvinen H, Undén P (1997) Observation screening and first guess quality control in the ECMWF 3D-Var data assimilation system. ECMWF Tech Memo 236:33, EMCWF, Shinfield Park, Reading

Lawless AS, Gratton S, Nichols NK (2005) Approximate iterative methods for variational data assimilation. Int J Numer Meth Fluid 1:1–6

Levitus S, Antonov JI, Boyer TP, Locarnini RA, Garcia HE, Mishonov AV (2009) Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. Geophys Res Letts 36:L07608. doi:10.1029/2008GL037155

Liu WT, Katsaros KB, Businger JA (1979) Bulk parameterization of the air-sea exchange of heat and water vapor including the molecular constraints at the interface. J Atmos Sci 36:1722–1735

Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE (2006) World Ocean Atlas 2005, Volume 1: Temperature. In: Levitus S (ed) NOAA Atlas NESDIS 61. U.S. Government Printing Office, Washington, DC, 182 pp

Lorenc AC (1986) Analysis methods for numerical weather prediction. Q J Roy Meteorol Soc 112:1177–1194

Lorenc AC, Hammon P (1988) Objective quality control of observations using Bayesian methods. Theory, and a practical implementation. Q J Roy Meteorol Soc 114:515–543

Matthews D, Powell B, Milliff RF (2011) Dominant spatial variability scales from observations around the Hawaiian Island. Deep Sea Res I 58:979–987

Milliff RF, Niiler P, Morzel J, Sybrandy A, Nychka D, Large W (2003) Mesoscale correlation length scales from NSCAT and Minimet surface wind retrievals in the Labrador Sea. J Atmos Ocean Technol 20:513–533

Moore AM, Arango HG, Broquet G, Powell BS, Zavala-Garay J, Weaver AT (2011a) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. Part I: system overview and formulation. Prog Oceanogr 91:34–49

Moore AM, Arango HG, Broquet G, Edwards CA, Veneziani M, Powell BS, Foley D, Doyle J, Costa D, Robinson P (2011b) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. Part II: performance and application to the California Current System. Prog Oceanogr 91:50–73

Moore AM, Arango HG, Broquet G, Edwards CA, Veneziani M, Powell BS, Foley D, Doyle J, Costa D, Robinson P (2011c) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. Part III: observation impact and observation sensitivity in the California Current System. Prog Oceanogr 91:74–94

Moore AM, Arango HG, Broquet G (2012) Analysis and forecast error estimates derived from the adjoint of 4D-Var. Mon Weather Rev 140:3183–3203

Powell BS, Moore AM, Arango HG, Di Lorenzo E, Milliff RF, Leben RR (2009) Near real-time assimilation and prediction in the Intra-Americas Sea with the Regional Ocean Modeling System (ROMS). Dyn Atmos Oceans 48:46–68

Ricci S, Weaver AT, Vialard J, Rogel P (2005) Incorporating state-dependent temperature-salinity constraints in the background-error covariance of variational ocean data assimilation. Mon Weather Rev 133:317–338

Saraceno M, Strub PT, Kosro PM (2008) Estimates of sea surface height and near surface alongshore coastal currents from combinations of altimeters and tide gauges. J Geophys Res 113:C11013. doi:10.1029/2008JC004756

Shchepetkin AF, McWilliams JC (2005) The regional oceanic modeling system (ROMS): a split explicit, free-surface, topography-following-coordinate oceanic model. Ocean Model 9:347–404

Stark JD, Donlon CJ, Martin MJ, McCulloch ME (2007) OSTIA: an operational, high resolution, real time, global sea surface temperature analysis system. In: Oceans '07 IEEE Aberdeen, conference proceedings. Maritime challenges: coastline to deep sea, Aberdeen

Veneziani M, Edwards CA, Doyle JD, Foley D (2009) A central California coastal ocean modeling study: 1. Forward model and the influence of realistic versus climatological forcing. J Geophys Res. doi:10.1029/2008JC004774

Weaver AT, Courtier P (2001) Correlation modelling on the sphere using a generalized diffusion equation. Q J Roy Meteorol Soc 127:1815–1846

Weaver AT, Vialard J, Anderson DLT (2003) Three- and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part I: formulation, internal diagnostics and consistency checks. Mon Weather Rev 131:1360–1378

Weaver AT, Deltel C, Machu E, Ricci S, Daget N (2005) A multivariate balance operator for variational ocean data assimilation. Q J Roy Meteorol Soc 131:3605–3625

Wikle CK, Berliner LM (2007) A Bayesian tutorial for data assimilation. Physica D 230:1–16

Chapter 15
A Weak Constraint 4D-Var Assimilation System for the Navy Coastal Ocean Model Using the Representer Method

Hans Ngodock and Matthew Carrier

Abstract A 4D-Variational system was recently developed for assimilating ocean observations with the Navy Coastal Ocean Model. It is described here, along with initial assimilation experiments in the Monterey Bay using a combination of real and synthetic ocean observations. For testing a new assimilation system it is advantageous to use this combination of real and synthetic data over simplified cases of climatology and twin data. Assimilation experiments are carried out in a weak constraint formulation, with the model's external forcing assumed to be erroneous in addition to initial conditions. The system's ability to fit assimilated and non assimilated observations is assessed, as well as the consistency and relevance of the retrieved model forcing. Experiment results show that the assimilation system fits the data with relatively high prior errors in the initial conditions and surface forcing fluxes. However, the retrieved model forcing errors are well within the range of acceptable corrections according to an independent study.

15.1 Introduction

This paper presents the development of a weak constraint 4D-Var data assimilation system based on the representer method (Bennett 1992, 2002) for the Navy Coastal Ocean Model (NCOM). NCOM is an operational ocean model that has been validated (Martin 2000; Barron et al. 2006). A major effort to implement state-of-the-art assimilation schemes was undertaken a few years ago, with the development of a 3DVAR, and a 4D-Var system based on the NCOM numerical code. The 3DVAR system is used for assimilation in global to regional scales, while the 4D-Var is to be used in limited area models with in-situ observations, provided initial and

H. Ngodock (✉) · M. Carrier
Naval Research Lab, code 7321, Stennis Space Center Hancock County, MS 39529, USA
e-mail: hans.ngodock@nrlssc.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_15,

boundary conditions from a global or regional model assimilating with 3DVAR. Both the adjoint and linear perturbation (also called the forward representer) model codes were derived for the most part with the help of the Parametric Fortran compiler (PFC), Erwig et al. 2007.

Some general circulation models of the complexity of NCOM have seen similar efforts undertaken in the past decade: a 4D-Var assimilation system was developed for the Ocean Parallelisé (OPA) model (Weaver et al. 2003), for the MIT general circulation model (MITgcm, Marotzke et al. 1999) also used in the ECCO consortium assimilation experiments (Stammer et al. 2002), and a similar system was built for the regional ocean model system (ROMS), Moore et al. 2004. Unlike the other models using fixed z-levels (OPA and MITgcm) or s-coordinates (ROMS) NCOM uses a combination of both sigma layers, z-levels and a generalized vertical coordinate.

It is a common practice to test a recently developed assimilation system with climatological data or identical twin experiments in which the observations are simulated by the numerical model. There is hardly a case of failure in twin experiments, yet a successful assimilation with twin experiments never guarantees success with real data. On the other hand, climatological data are overly smooth in both space and time (due mostly to linear interpolation) and lack the variability associated with real observations. To avoid these simplified cases, the newly developed NCOM 4D-Var system is tested with real and synthetic observations generated by the modular ocean data assimilation system (MODAS) Fox et al. 2002, as well as with real observations collected from satellites and a fleet of gliders during the second autonomous ocean sampling network (AOSN II) in the Monterey Bay.

There are no specific applications of 4D-Var in the Monterey Bay, let alone its weak constraint formulation. Strong constraint variational assimilation (Broquet et al. 2009) has been applied to the California current system (CCS), including an application to estimate surface forcing correction (Broquet et al. 2011), using the inverse Regional Ocean Modeling System (IROMS, Di Lorenzo et al. 2007) with horizontal resolutions of 10 and 30 km. The CCS is a large area that includes the Monterey Bay, although these applications did not specifically target the Monterey Bay, given their rather coarse resolutions. Most of the assimilation experiments that have been carried for the Monterey Bay were based on sequential methods such as 3DVAR and ensemble-based Kalman filters: Chao et al. (2009), Haley et al. (2009), and Shulman et al. (2009). This study presents an application of the weak constraint 4D-Var in the Monterey Bay in a proof-of-concept context, using synthetic and real observations. The first objective is to demonstrate the system's ability to reduce large discrepancies between the model and the observations, when the latter are assigned very low errors. Therefore, this paper is more focused on the technical development of the weak constraint 4D-Var system.

A brief description of the numerical model is presented in the next section, followed by the 4D-Var system derivation and implementation in Sect. 15.3. Section 15.4 deals with the experiments setup and results, and concluding remarks follow in Sect. 15.5.

15.2 The Model

NCOM is described in the literature (Martin 2000; Barron et al. 2006). The description of the model equations given in the appendix is only repeated in order to exhibit the nonlinear terms in the model equations, as they directly affect the development of the linearized and adjoint models associated with NCOM. NCOM is a free surface model based on the primitive equations and employs the hydrostatic, Boussinesq and incompressible approximations. The model is discretized using finite differences on an Arakawa C-grid in the spatial dimensions. The equations are solved in three dimensions for momentum (both zonal and meridional components of velocity), temperature and salinity, and two dimensions for the free-surface mode: surface elevation and barotropic velocities.

The leapfrog scheme is used for time stepping in conjunction with an Asselin filter to avoid time splitting. All terms are treated explicitly in time except for the solution for the free surface and vertical diffusion. In the solution for the free surface, the surface pressure gradient terms in the depth-averaged momentum equations and the divergence terms in the depth-averaged continuity equation are evenly split between the old and new time levels to minimize the damping of surface waves. The model equations discretized with finite differences in flux-conservative form are given in the appendix.

The model domain used for this experiment contains the Monterey Bay, California region. This location is favorable for ocean modeling due to its strong land/sea breeze circulation patterns, complex coastline with steep topography, and the existence of frequent local upwelling and relaxation events (Shulman et al. 2002). The domain covers latitudes 35.6°–37.49° North and longitudes 121.38°–123.2° West with a horizontal resolution of 2 km and 41 layers in the vertical. The model was initialized on 01 August, 2003 and ran for one month to 01 September, 2003. The initial conditions were obtained from downscaling the operational $1/8°$ resolution global NCOM to an intermediate model with horizontal resolution of 6 km, and then via a 3-to-1 nesting ratio to the 2 km model. Horizontal viscosities and diffusivities are computed using either the grid-cell Reynolds number (Re) or the Smagorinsky schemes, both of which tend to decrease as resolution is increased. The grid-cell Re scheme sets the mixing coefficient K to maintain a grid cell Re number below a specified value, e.g. if $\mathrm{Re} = u^*\mathrm{dx}/K = 30$, then $K = \mathrm{u}^*\mathrm{dx}/30$. Hence, as dx decreases, K decreases proportionally. A similar computation is performed for the Smagorinsky scheme.

Surface boundary conditions (e.g. wind stress, IR radiation flux, etc.) are provided by the atmospheric mesoscale model COAMPS (Hodur 1997), which is run at the same horizontal resolution as the ocean model, with forcings archived every 12 h at the synoptic times of 0000 and 1200 UTC. Open boundary conditions use a combination of radiative models and prescribed values provided by the $1/8°$ Global NCOM (GNCOM). Different radiative options are used at the open boundaries depending on the model state variables: a modified Orlanski radiative model is used for the tracer fields (temperature and salinity), an advective model for the zonal velocity (u), a zero gradient condition for the meridional velocity (v) as well as the barotropic velocities, and the Flather boundary condition for elevation.

15.3 The 4D-Var System

15.3.1 Linearization

Nonlinear terms in the model consist of all the advection terms in the momentum and tracer equations, the horizontal mixing with the Smagorinsky formula, the curvature correction, the vertical mixing with coefficients computed using the Mellor-Yamada 2.5 turbulence closure. Additional nonlinearities stem from the discretization in flux conservative form where vertical increments Δz in the sigma layers depend on the free surface elevation. As a consequence, even the time discretization is nonlinear. Nonlinearities also appear in the free surface, or barotropic mode, with the multiplication by the depth variables D^u and D^v in (15.23) and (15.24). However, the barotropic transports $D^u\bar{u}$ and $D^v\bar{v}$ are computed explicitly first, then the barotropic velocities ($\bar{u}$ and $\bar{v}$) are derived by dividing the barotropic transports by the depth variable, which is a nonlinear operation. The baroclinic pressure gradient is computed from the density field obtained from the state equation as a nonlinear function of temperature and salinity. Other nonlinearities appear in the various radiative conditions at the open boundaries of the model domain mentioned above.

With the exception of the Mellor-Yamada turbulence closure, all of these nonlinear terms are linearized according to the first-order Taylor's approximation for the derivation of the tangent linear model.

For the sake of clarity, let's rewrite the leap-frog time discretization of (15.14), see the appendix, in the form

$$\frac{\Delta x^u \Delta y^u}{2\Delta t}\left(\left(\Delta z^u\right)^{n+1} u^{n+1} - \left(\Delta z^u\right)^{n-1} u^{n-1}\right) = G^n, \tag{15.1}$$

where G^n represents the terms in the right hand side of (15.14) evaluated at time level n, and the depth increment $(\Delta z^u)^{n+1}$ is available from a previously computed elevation. The numerical model is updated by

$$u^{n+1} = \frac{1}{\left(\Delta z^u\right)^{n+1}}\left[\left(\Delta z^u\right)^{n-1} u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u} G^n\right] \tag{15.2}$$

The linearization of (15.2) is

$$\begin{aligned}\delta u^{n+1} = {} & \frac{1}{\left(\Delta z^u\right)^{n+1}}\left[\left(\Delta z^u\right)^{n-1} \delta u^{n-1} + \left(\delta\Delta z^u\right)^{n-1} u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u}\delta G^n\right] \\ & - \frac{\left(\delta\Delta z^u\right)^{n+1}}{\left[\left(\Delta z^u\right)^{n+1}\right]^2}\left[\left(\Delta z^u\right)^{n-1} u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u} G^n\right]\end{aligned} \tag{15.3}$$

where u is the background solution, i.e. the solution around which the model is linearized, G and Δz are computed using the background solution, and δu, δG and $\delta\Delta z$ are the linear perturbations of u, G and Δz respectively. In both (15.2)

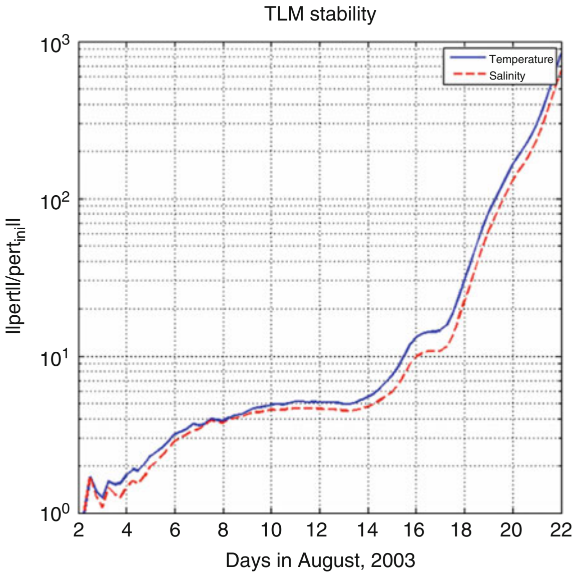


Fig. 15.1 Time evolution of the magnitude of the perturbation to the temperature and salinity fields normalized by the magnitude of their respective initial perturbations

(and hence (15.3)) a small positive number is usually added to the denominator to prevent it from vanishing. As mentioned above the depth increments in the vertical discretization in NCOM depend on the time varying elevation only in the sigma layers. In the z-level portion of the vertical grid, (15.2) and (15.3) take the form

$$u^{n+1} = u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u \Delta z^u} G^n \tag{15.4}$$

and

$$\delta u^{n+1} = \delta u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u \Delta z^u} \delta G^n . \tag{15.5}$$

As for the vertical mixing coefficients from the Mellor-Yamada turbulence closure scheme, they are provided by the nonlinear model trajectory around which the model is linearized.

The stability of the linearized model is assessed by the time evolution of small perturbations: the tangent linear model is initialized by random three dimensional perturbations of the temperature and salinity fields and integrated over time. At each time step the norms of the perturbed temperature and salinity fields are computed and divided by the norms of their respective initial perturbations. Results plotted in Fig. 15.1 show that the linear perturbations are stable and bounded for about 12–15 days before they start to grow exponentially. Initial perturbations here are generated by the adjoint integration forced by Dirac impulses at randomly selected grid points. This process produces three-dimensional initial fields with dynamically coherent structures compared to purely random fields. However, the TLM test with purely random fields did not yield different results (not shown).

15.3.2 Adjoint Derivation

Once the linear perturbation model was obtained, the adjoint model was derived by transposition of the perturbation model as follows for both sigma layers and z-levels:

$$
\begin{aligned}
\lambda^* &= \lambda_u^{n+1}\\
\lambda_u^{n+1} &= 0\\
\lambda_{\Delta z^u}^{n+1} &= \lambda_{\Delta z^u}^{n+1} - \frac{\lambda^*}{\left[(\Delta z^u)^{n+1}\right]^2}\left[(\Delta z^u)^{n-1}u^{n-1} + \frac{2\Delta t}{\Delta x^u \Delta y^u}G^n\right]\\
\lambda_G^n &= \lambda_G^n + \frac{1}{(\Delta z^u)^{n+1}}\frac{2\Delta t}{\Delta x^u \Delta y^u}\lambda^*\\
\lambda_{\Delta z^u}^{n-1} &= \lambda_{\Delta z^u}^{n-1} + \frac{u^{n-1}}{(\Delta z^u)^{n+1}}\lambda^*\\
\lambda_u^{n-1} &= \lambda_u^{n-1} + \frac{(\Delta z^u)^{n-1}}{(\Delta z^u)^{n+1}}\lambda^*
\end{aligned}
\tag{15.6}
$$

and

$$
\begin{aligned}
\lambda^* &= \lambda_u^{n+1}\\
\lambda_u^{n+1} &= 0\\
\lambda_G^n &= \lambda_G^n + \frac{2\Delta t}{\Delta x^u \Delta y^u \Delta z^u}\lambda^*\\
\lambda_u^{n-1} &= \lambda_u^{n-1} + \lambda^*
\end{aligned}
\tag{15.7}
$$

where λ_a^i denotes the adjoint variable associated with the state variable a at the time level i, and λ^* is a temporary variable. In (15.6) and (15.7) it is assumed that the adjoint variables have been initialized at a prior time level. In practice, the model is usually computer programmed by subroutines, with individual terms of the model equations computed in separate subroutines. Similarly, the linearization and the adjoint derivation were carried out one subroutine at a time, and care was taken to ensure that symmetry between the linearized subroutine and its adjoint was preserved. The entire linearized model was obtained once every subroutine was linearized, and the entire adjoint was obtained with individual adjoint subroutines appearing in reverse order as compared to the linearized model.

In practice, both the linearized and adjoint models were obtained with the help of the Parametric Fortran compiler (PFC). Parametric Fortran is an extension of Fortran that supports defining Fortran program templates by allowing the parameterization of arbitrary Fortran constructs. A Fortran program template can be translated into a regular Fortran program guided by values for the parameters. The Parametric Fortran compiler is written in Haskell (Peyton Jones 2003), and the parameter values are represented as Haskell values so they can be used by the Parametric Fortran compiler directly. Parametric Fortran is particularly useful in scientific computing. The applications include defining generic functions, removing duplicated code, and automatic differentiation. Parametric Fortran thus has broader and more general uses than previous tools in the likes of TAMC (Giering and Kaminski 1998), TAPENADE (Hascoet and Pascual 2004) or ADIFOR (Bischof et al. 1992), developed just for the purpose of automatic differentiation. The differentiation is based on the chain rule, with special treatment for non-differentiable functions.

15.3.3 How PFC Works for TL and Adjoint Generation

The Parametric Fortran compiler is publicly available from http://web.engr.oregonstate.edu/~erwig/pf/. It is a command line program in which the differentiation operation has been parameterized by "Diff". Assuming it has been installed on a user's computer, it can be used to generate tangent linear and adjoint of Fortran subroutines or programs in the following manner:

1. The user creates a parameter text file, say "param_file", in the format:

   ```
   Diff = TL [var1, var2, var3 . . .]
   ```

   where var1, var2, var3 . . . , form a list of all active variables and all variables that depend or operate on active variables (including temporary variables), "TL" will indicate to the compiler that the tangent linear model is being created, and "Diff" is the differentiation parameter for Parametric Fortran.
2. For a subroutine "test.f" to be differentiated the user also creates a file "test.pf" that contains the subroutine in the form

   ```
   {Diff:
   Subroutine test(var1,var2. . . )
   Body of subroutine
   end
   }
   ```
3. Finally, the compiler is invoked by typing the following from the command line: pfc -p param_file test.pf test_TL.f

   The output of the compiler will be the tangent linearized subroutine "test_TL.f".
4. The procedure for generating the adjoint is the same except that in steps 1 and 3 "TL" is replaced with "AD".

For generic state variables x and y and a subroutine computing a quantity Ax, the symmetry between the linearized subroutine and its adjoint is evaluated by

$$\langle \mathrm{Ax}, \mathrm{y} \rangle = \left\langle \mathrm{x}, \mathrm{A}^{\mathrm{T}}\mathrm{y} \right\rangle \tag{15.8}$$

where $\langle .\,,. \rangle$ denotes an inner product. This equality should hold to machine precision (regardless of computer architecture) not only for individual subroutines, but also for the entire linearized model and its adjoint. For randomly generated x and y as initial and final conditions for the linearized and adjoint models respectively, equality (15.8) was tested for integration periods of 1 and 5 days with an absolute difference in the order of 10^{-14} between the left and right hand side of (15.8), the computations being done in double precision.

Alternatively, this symmetry is also assessed by the symmetry of the representer matrix (Bennett 1992, 2002). For a given number M of observation locations (in the space-time domain), regardless of which model variable is observed, representer functions are computed, one per observation location. A representer function associated with a given observation location is obtained by integrating the adjoint model forced by a Dirac delta function centered at the chosen observation location, then using the adjoint solution (in space and time) as forcing for the perturbation model. A column of the representer matrix is computed by evaluating a representer function at all observation locations. If the adjoint model is consistently derived from the perturbation model, the representer matrix should be symmetric to the machine precision which is the case for our model and its adjoint.

15.3.4 The Cost Function

For sake of clarity, the model equations are rewritten in a simpler form

$$\begin{cases} \frac{\partial X}{\partial t} = F\left(X\right) + f, 0 \le t \le T \\ X\left(t = 0\right) = I\left(x\right) + i(x) \end{cases} \tag{15.9}$$

where X stands for all the dependent model state variables: two dimensional sea surface height and barotropic velocities, and three dimensional temperature, salinity and baroclinic velocities, F is the model tendency terms in the right hand side of (15.14, 15.15, 15.16, 15.17 and 15.18) and (15.23, 15.24 and 15.25), f is the model error, a function of the independent variables (x and t) of the space-time domain Ω with covariance C_f, $I(x)$ is the prior initial condition, $i(x)$ is the initial condition error with covariance C_i. Given a vector Y of M observations of the model state in the space-time domain, with the associated vector of observation errors ε (with covariance C_ε),

$$y_m = H_m X + \varepsilon_m, \quad 1 \le m \le M \tag{15.10}$$

where H_m is the observation operator associated with the mth observation, one can define a weighted cost function

$$J = \int_0^T \int_\Omega \int_0^T \int_\Omega f(x,t) W_f(x,t,x',t') f(x',t') dx' dt' dx dt + \int_\Omega \int_\Omega i(x) W_i(x,x') i(x') dx' dx + \varepsilon^T \mathrm{W}_\varepsilon \varepsilon \qquad (15.11)$$

where Ω denotes the model domain, the weights W_f and W_i are defined as inverses of C_f and C_i in a convolution sense, and W_ε is the matrix inverse of C_ε. The latter is usually considered a diagonal matrix, from the assumption that observation errors are uncorrelated. Boundary condition errors are omitted from (15.9) to (15.11) only for the sake of clarity. The model error covariance is assumed to take the form

$$C_f\left(x,t,x',t'\right) = \mathbf{v}(x)^{1/2}\,\mathbf{v}(x')^{1/2} \exp\left(-\frac{|x-x'|^2}{2L^2}\right) \exp\left(-\frac{|t-t'|}{\tau}\right) \qquad (15.12)$$

where $\mathbf{v}(x)$ is the error variance and L and τ are the length and time scales respectively. The initial error covariance C_i assumes the form of (15.12) with the exception of the time correlation term and different (higher) variance. Horizontal correlations in (15.12) are obtained by solving a diffusion equation (Derber and Rosati 1989; Egbert et al. 1994; Weaver and Courtier 2001), while the time correlation is obtained by solving a pair of coupled Langevin equations (Chua and Bennett 2001; Bennett 2002; Ngodock 2005). Correlations in (15.12) are univariate and are implemented layer by layer for each model state variable. The cross correlations are provided by the model dynamics through the integration of the adjoint and the tangent linear models. Note that although the cost function is written with the inverse of the covariance functions, the actual inverses are not needed in practice, when the solution of the Euler-Lagrange equations associated with the minimization of (15.11) is sought through the representer method (Bennett 1992, 2002).

15.3.5 Error Standard Deviations: $v(x)^{1/2}$

Assigning model errors and prescribing their covariances is the most difficult task in data assimilation, as acknowledged by most assimilation experts: Daley (1992), Talagrand (1999), Bennett (2002), Wunsch (2006). Not only are there many error sources (external forcing, initial and boundary conditions, bad parameterization, empirical formulation, unresolved processes), but also the errors cannot be measured. Therefore one can only make assumptions about them. Since NCOM includes all resolvable processes and sub-gridscale parameterization, errors are attributed to the initial conditions and external forcing for all the dynamical equations, and the derivation of their estimates is given below. Note that there is no external forcing applied to the continuity equation, and thus it is not assigned a model error either, as in Jacobs and Ngodock (2003).

Consider the momentum equation (15.14) in its non-discretized form

$$\frac{\partial u}{\partial t} + \ldots = \ldots + \rho^{-1} F \tag{15.13}$$

where F represents the wind stress atmospheric forcing (in Nm^{-2}), the volume flux source and the tidal potential, and ρ is the water density. The model error at the surface consists of errors in the wind stress. For the subsurface, errors are assumed to arise from the volume flux and the tidal potential terms. We consider errors to be high in magnitude at the surface and decreasing with depth. Although the wind stress varies in space and time, its associated error is assumed uniform in the horizontal directions. The error magnitude is considered to be 50 % of the actual wind stress at the surface and decreasing with depth in order to mimic the decreasing impact of wind stress with depth. Two terms contribute to the forcing for the temperature equation: the net longwave, latent and sensible heat flux on one hand, and the solar radiation on the other hand. Both are assumed to be 30 % in error and the sum of their errors constitutes the forcing error in the temperature equation, with a spatial distribution similar to the one used for the errors in the momentum equation. A similar approach is taken for the errors in the salinity equation, where the forcing consists of the river inflow and evaporation minus precipitation. Forcing terms here are also considered to be 30 % in error. Finally the standard deviations for the initial condition errors are 1 m for the surface elevation, $0.5\,\mathrm{ms}^{-1}$ for both components of the velocity field, 2 K for temperature and 0.5PSU for salinity. These rather high errors indicate the lack of confidence in the forcing fields and initial conditions. Spatial and temporal correlation scales in (15.12) are set to 10 km and 30 h. The errors and scales above are obviously arguable, and it is not our intention to defend their choice. Rather, they are selected in this preliminary assimilation setup to demonstrate the functionality of the NCOM 4D-Var system. Smaller errors will be adopted when the system is used with real observations.

15.3.6 The Minimization

The solution of the assimilation problem is found by solving the Euler-Lagrange (EL) system of equations associated with the minimization of the cost function (15.11). The EL system is a linear yet coupled system between the adjoint and state variables. The representer methods uncouples the system by expanding the solution as the sum of a first guess and a finite linear combination of representer functions, with the representer coefficients computed by solving a linear system in data space involving the representer matrix, the data error covariance matrix and the innovation vector. The entire representer matrix need not be computed since the linear system can be solved using an iterative algorithm (e.g. the conjugate gradient), by taking advantage of the symmetry of each matrix involved. The representer coefficients constitute the right hand side of the adjoint equation in the EL system. Once the representer coefficients are computed, they are substituted in the adjoint equation which is then solved and substituted in the forward linear equation for the final solution. A background solution around which the model is linearized is needed. Usually it is the solution of the nonlinear model. For the first guess solution, one may consider either the background or the tangent linear solution around the background. Also, the new optimal solution may replace the background for another minimization process (i.e. outer loops) until formal convergence (Bennett et al. 1996, 1998, 2002; Ngodock et al. 2000, 2007, 2009).

15.4 Experiment Setup and Results

Assimilation experiments are carried out with two different data sets, and the results shown below are primarily aimed at evaluating the 4D-Var system's ability to fit both the assimilated and the non-assimilated observations.

15.4.1 MODAS Data

MODAS generates synthetic vertical profiles of temperature and salinity in the two following steps: first, a subsurface temperature is computed at a given depth using a regression from sea surface temperature and the steric component of the sea surface height anomaly. Once the subsurface temperature is computed, a corresponding subsurface salinity is computed using a climatology-based temperature/salinity relationship, Fox et al. (2002). MODAS data are thus a combination of real sea surface data (SSH and SST) and simulated sub-surface data derived from the real surface data using regression and historical relationships.

MODAS synthetics are saved and utilized in the 4D-Var analysis at intervals of 6 h. There are approximately fifty-six uniformly distributed profiles of temperature and salinity across the model domain. Each profile is represented on a vertical grid of 46 layers that do not coincide with the model's vertical grid of 41 layers, but the observation operator H in (15.10) handles the projection from the model grid to the data grid. Temperature (salinity) observation errors are set to 0.2°C (0.1 psu), and held constant through the entire assimilation window. These observation errors are purposefully set low, not because MODAS data are very accurate, but to test the assimilation's ability to reduce large discrepancies with the model, i.e. to drive the model with large errors to fit observations with small errors.

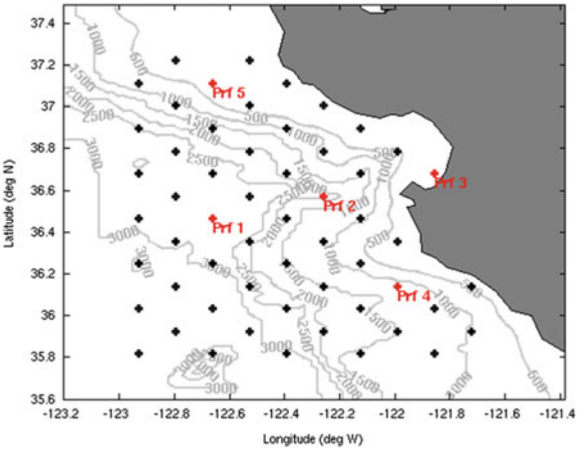


Fig. 15.2 The model domain with bathymetry contours and the profile locations, including the numbered profiles (in *red*) where the assimilated solution is evaluated

15.4.2 Results with MODAS Data

Starting from an initial condition on August 02, the model was integrated and the assimilation performed for 5 days at a time, with the analysis at the end of the 5 days becoming the initial condition for the following 5-day assimilation, the overall assimilation experiment interval being 30 days.

In order to assess how well the assimilation fits the observations, the analysis is examined at 5 locations in the model domain shown on Fig. 15.2. These locations are selected according to their geographic position with respect to the bay: offshore (location 1), slightly outside of the bay mouth (location 2), inside the bay (location 3), and south and north of the bay (locations 4 and 5). Results at locations 2 and 4 are similar to those at location 5, and therefore are not shown.

Examining the solution in the top 500 m at the offshore location 1, it can be seen that the assimilation is able to correct large and small discrepancies between the first guess and the observations for both the temperature and salinity fields, as seen in Fig. 15.3. In the first 5 days temperature discrepancies range between 2 K in the upper 50 m, and about 1 K from 100 m and below. Likewise salinity discrepancies range from 0.15psu in the upper 200 m to 0.05psu below. These discrepancies are gradually corrected in the analysis (bottom panels of Fig. 15.3) and by the end of the first 5-day assimilation window, they have vanished. For the subsequent 5-day assimilation windows, the model temperature and salinity appear to be well constrained below 100 m with minimal to no discrepancies between the first guess and the data. Discrepancies are confined to the upper 100 m. They are small at the beginning of each 5-day window and grow with time. This is to be expected since the first guess is initialized with the previous 5-day analysis at the final time, and because the NOGAPS forcing fields are not necessarily compatible with MODAS data. That the discrepancies are confined to the upper ocean also suggests that the model error is driven by erroneous surface fluxes, although the simulation of the

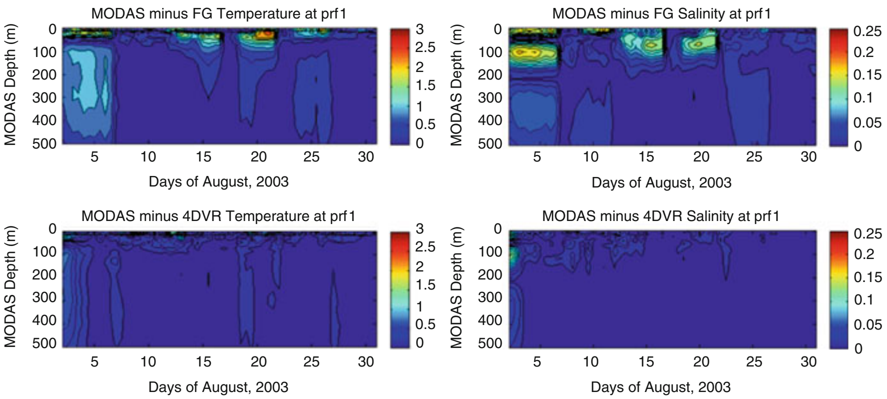


Fig. 15.3 Time evolution of the absolute value of the innovation (*top*) and the analysis error (*bottom*) at profile location 1, for temperature (*left*) and salinity (*right*)

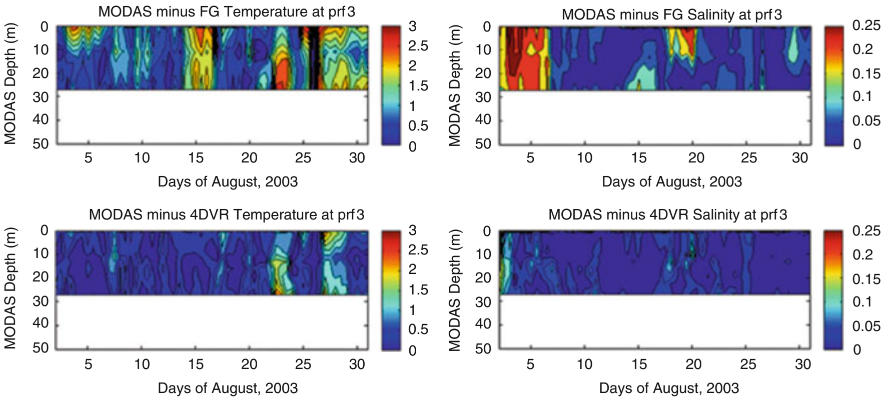


Fig. 15.4 Same as Fig. 15.3, except for location 3

mixed layer could also be incompatible with MODAS data. Yet both the data and the forcing fields are purposefully chosen in order to test the assimilation's ability to efficiently reduce these discrepancies while estimating a reasonable (magnitude-wise) correction to the surface fluxes. The assimilation effectively reduces all the discrepancies to within the data standard deviation for both temperature and salinity.

The maximum depth at location 3 inside the bay is 28 m. Results at this location, shown in Fig. 15.4, indicate that high salinity discrepancies sometimes exceeding .25psu are distributed through the water column during the first 5-day assimilation period. Some large salinity discrepancies also appear between days 18–20. Temperature discrepancies on the other hand are more prevalent, distributed over space and time. It appears that the initialization of the model using the previous 5-day analysis has less influence on the current 5-day first-guess. This may be due to the fact that in this shallow location, temporal variability of the solution is mostly governed by the

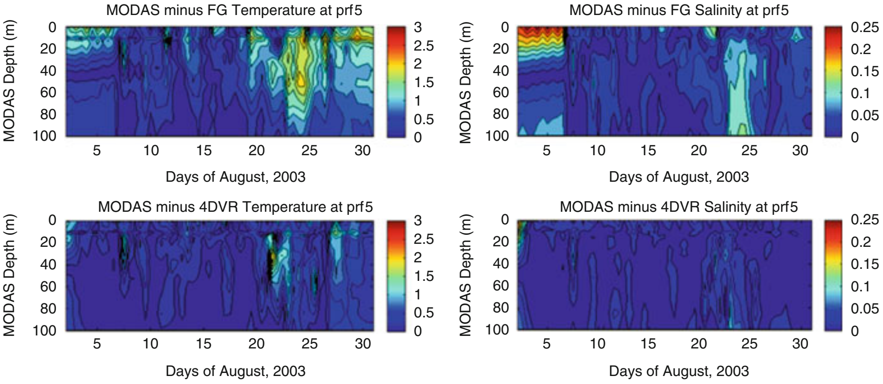


Fig. 15.5 Same as Fig. 15.3, except for location 5

local external/surface forcing coupled with strong mixing, and not by the short-lived initial conditions. Nevertheless, the assimilation still significantly reduces these discrepancies (bottom panels of Fig. 15.4) through the depth-time domain except for some isolated places. Assimilation results at location 4 (south of the mouth of the bay) are very similar to those at location 2, and therefore are not shown here.

At location 5 (north of the mouth of the bay) the maximum depth is 100 m. The largest salinity discrepancies are in the upper 20 m during the first 5-day, as seen in Fig. 15.5. There are also some moderate discrepancies in the lower layers around day 24. Temperature discrepancies are initially moderate (less than 1.5 K during the first 5-day period) and remain low until day 20, after which they start growing again, reaching 2 K. For most of the assimilation period these discrepancies are significantly reduced below 0.5 K, except for some isolated locations, e.g. around 40 m depth at days 21 and 22.

15.4.3 AOSN II Data

The dataset comprises SST from satellite and aircraft, a few SSH from satellite altimetry (due to the limited area of the model domain), vertical profiles of temperature and salinity from Slocum and Spray gliders and two moorings (M1 and M2) and AXBTs. All the vertical profiles are projected on a static grid of 42 levels.

Slocum glider tracks covered a portion of the bay, the mouth of the bay and the area to the northwest of the bay, i.e. the upwelling center around Año Nuevo. Spray glider tracks originated from the nearshore and went offshore in transec-like trajectories as seen in Fig. 15.6. To avoid redundancy some of the glider data are withheld from the assimilation and used for validation of the analyses. Withholding the data takes into account the model grid resolution and the prescribed horizontal decorrelation scale of the model error. The observations are assigned a constant error of 0.5 K and 0.3psu in temperature and salinity respectively.

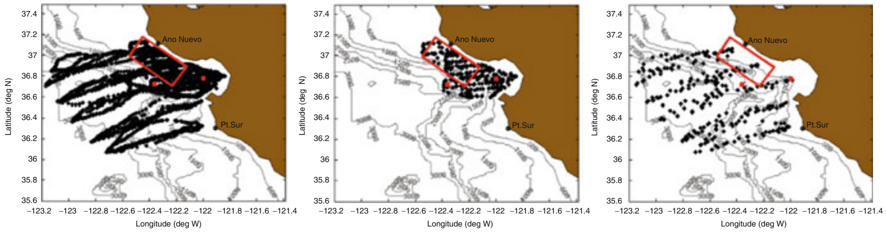


Fig. 15.6 All glider and the two mooring positions (*left*), and assimilated Slocum (*center*) and Spray (*right*) glider tracks during August, 2003. The *red dots* represent the location of the moored buoys M1 (*right*) and M2 (*left*). The *red box* indicates the upwelling center near Año Nuevo

15.4.4 Results with AOSN II Data

The assimilation covers the time window of August 2 to August 27, 2003, and is carried out in cycles of 5 days, with the analysis at the end of a cycle becoming the initial condition for the following cycle. Although the observations are processed and stored in 6-h intervals, the 4D-Var system assimilates all observations within the 5-day cycle simultaneously. The performance of the assimilation system is examined by computing the difference between the observations and three model solutions: (1) the free running (non assimilative) model that is integrated from the given initial conditions and forcing fields, (2) the first guess (also non assimilative) for which the initial condition is updated from the assimilation in the previous cycle, with the exception of the first cycle where both the first guess and the free running model are equal, and (3) the analysis. The first guess is also the background trajectory for the tangent linear model and the adjoint, i.e. the trajectory around which the model is linearized. It is stored in intervals of 6 h. It is anticipated that due to the re-initialization from assimilating in a previous cycle, the first guess should have smaller discrepancies with the observations than the free running model, and the analysis should have smaller discrepancies with the observations than the first guess. This should be the case for discrepancies computed with the assimilated and non-assimilated observations. It is expected of every assimilation system to fit the assimilated observations within one observation standard deviation. Unassimilated observations consist of withheld observations within the current assimilation window and future observations, those in the next cycle before the assimilation. The assimilation is expected to fit the former as a measure of the system's ability to propagate the information from the assimilated observations sites through the model spate-time domain within the assimilation window. However, there is no expectation to fit future observations, i.e. the innovations in the next cycle are not expected to be smaller than the observation standard deviation. One only hopes that having initialized the model from the previous cycle's assimilation, the model forecast will remain sufficiently accurate to maintain small innovations. However, integrating the model from the initial conditions with uncorrected forcing fields is prone to drive the model away from the observations.

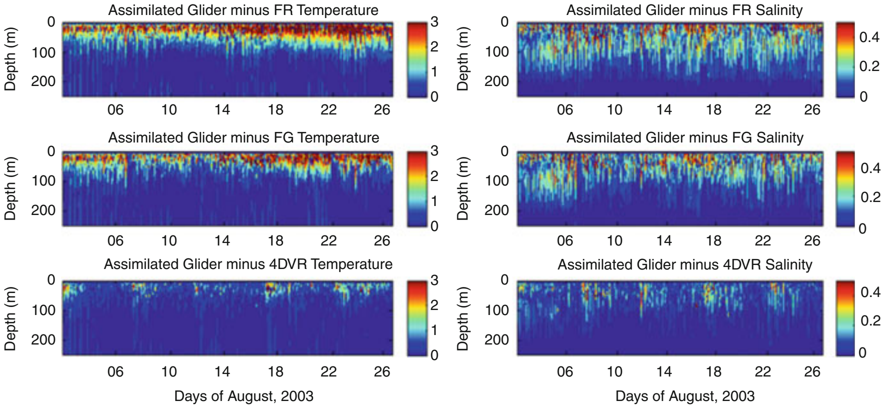


Fig. 15.7 Absolute model temperature (*left*) and salinity (*right*) discrepancies to assimilated observations for the free run (*top*), first guess (*middle*) and analysis (*bottom*)

The difference between the observations and the model is computed for all assimilated profiles of temperature and salinity and plotted in chronological order in Fig. 15.7. It can be seen that the temperature differences are confined in the upper 100 m of the water column, with magnitudes sometimes reaching 3 K for both the free run and the first guess. Salinity differences extend deeper in the water column, to about 200 m, although the largest differences are confined to the upper 100 m. A slight improvement can be noticed from the free run to the forecast solutions in the temperature field, but not as much in the salinity field. However, the assimilation is able to significantly reduce the forecast discrepancies in both the temperature and salinity fields, with the exception of a few profiles at the beginning of each cycle. The assimilation is able to reduce discrepancies as high as 3 K and 0.4psu to less than 0.5 K and 0.1psu in temperature and salinity respectively.

The forecast solution is expected to have smaller discrepancies to the observations than the free run, because it is initialized with the analysis at the end of a previous cycle. So, having only a marginal improvement from the free run to the first guess is an indication that the gains from the assimilation are short-lived in the forecast run as a consequence of inadequate forcing fields driving the model away from future observations

15.4.5 Independent Observations

For verification and evaluation purposes, discrepancies are computed between the withheld glider observations and the three model solutions: the free run, the first guess and the analysis. Results in Fig. 15.8 show that all three solutions have similar error levels with respect to the un-assimilated as to the assimilated observations.

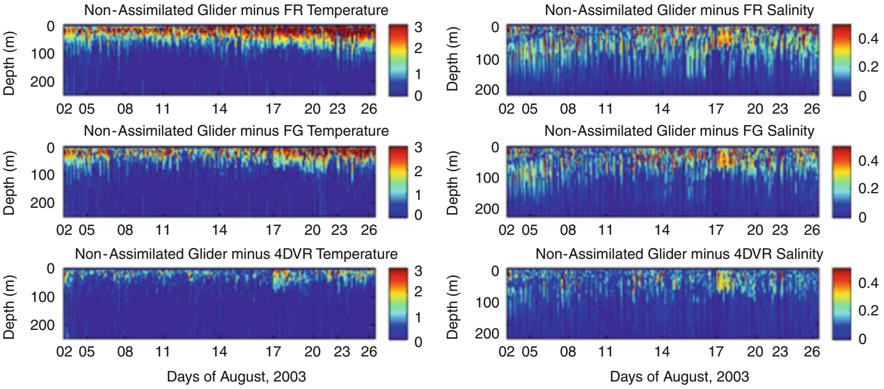


Fig. 15.8 Same as Fig. 15.7, except for non-assimilated glider observations

This result was expected because in most cases the withheld observations were located in the vicinity of assimilated observations. There are still some large temperature and salinity discrepancies in the analysis, usually around the beginning of the assimilation cycle.

15.4.6 Qualitative Fitting of the Data

The assimilation system's ability to fit the observations is further examined by comparing the differences between the observations and the free running model, the first guess and the analysis for all the observations and at all times, for both MODAS and AOSN II data. The free running model is integrated from the initial conditions and is never re-initialized, while the first guess for an assimilation cycle is initialized by the analysis at the end of the previous cycle. Elements of these difference vectors are binned by comparing their magnitude to the observations standard deviation. For example, all elements that are smaller than a standard deviation in absolute value are binned together, and so are all elements whose absolute value is between one and two standard deviations, and so on. The number of elements in each bin is then converted into a percentage of the number of assimilated observations. The results plotted as a cumulative bar chart on Fig. 15.9 show that the assimilated solution with MODAS data fits 80 % and 90 % of the observations to within one and two standard deviations respectively, while the corresponding numbers for the first guess are 60 % and 75 %, and 45 % and 63 % for the free running model. Some posterior misfits, although only a small percentage, are larger than 7 observations standard deviations, which obviously violate the Gaussian assumption on the errors in general. Similarly, for the AOSN II data, assimilated solution fits 86 % and 95 % of the observations to within one and two standard deviations respectively, while

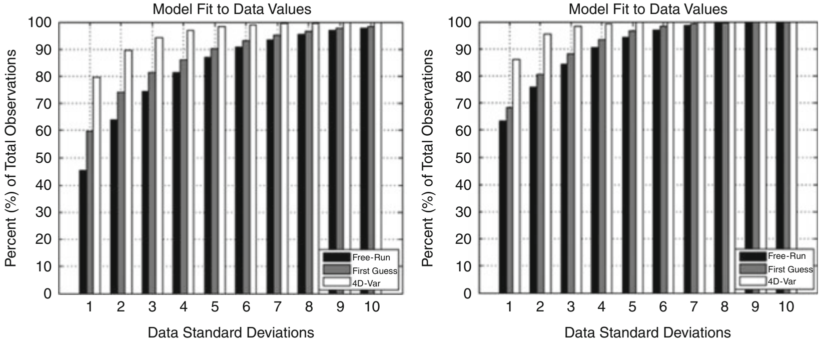


Fig. 15.9 Cumulative bar chart showing the percentage of the number of observations that are matched by the free running model (*black*), the first guess (*grey*) and the analysis (*white*) as a function of the number of observation standard deviations. MODAS experiment is shown on the *left* and AOSN II experiment on the *right*

the corresponding numbers for the first guess are 68 % and 80 %, and 64 % and 76 % for the free running model.

The large posterior misfits happened for some temperature observations with prior misfits sometimes higher than 5°, and the assimilation reduces these misfits to about 1.5°. They are larger than 7 standard deviations primarily because of very low data errors and possibly high model errors. It is assumed that a better fit would be achieved with larger observation errors and lower model errors. Such experiments (not shown here) are carried in the context of real observations and are the subject of another study.

15.5 Conclusion

A 4D-Var assimilation system for NCOM has been developed based on the indirect representer method. The system produces analysis increments for all prognostic variables (3D temperature, salinity, u- and v- components of velocity, and sea surface elevation) from a time-window of observations in a weak-constraint environment. The adjoint model has been checked against the linearized model using well established methods, verifying that the system is symmetric to within machine precision. Assimilation experiments were carried out with two different data sets.

The first experiment involved MODAS synthetic data (T, S, SSH) that were sampled every 6 h and assimilated in a sequence of 5-day time windows. Starting from an initial condition on August 02, the model was integrated and the assimilation performed for 5 days at a time, with the analysis at the end of the 5 days becoming the initial condition for the following 5-day assimilation. The results indicate that the assimilation system is performing correctly, with the model-data misfit is reduced substantially as examined at individual profiles.

The second experiment used the observations collected during AOSN II. Starting from a free run solution that completely misrepresented both the data and the dynamics of the region during the selected time period, the assimilation was able to accurately fit the assimilated data. Also, contrary to the free run and the first guess, the upwelling and relaxation events that dominate the dynamics of the regions were accurately described by the analysis which benefited from a good observation coverage of the domain and a robust assimilation system.

To avoid redundancy, some glider profiles were withheld from the assimilation and used for evaluation. The analysis fitted the withheld observations with the same accuracy as the assimilated observations. This was due in part to the proximity of the withheld observations with those that were assimilated.

The assimilated solution with MODAS data fits 80 % and 90 % of the observations to within one and two standard deviations respectively, while the corresponding numbers for the first guess are 60 % and 75 %, and 45 % and 63 % for the free running model. Some posterior misfits, although only a small percentage, are larger than 7 observations standard deviations, which obviously violate the Gaussian assumption on the errors in general. Similarly, for the AOSN II data, the assimilated solution fits 86 % and 95 % of the observations to within one and two standard deviations respectively, while the corresponding numbers for the first guess are 68 % and 80 %, and 64 % and 76 % for the free running model.

The largest discrepancies between the first guess and the observations were mostly confined to the upper ocean. After the first 5-day assimilation the first guess discrepancies grew quickly from their small initial values, confirming that the model is being forced by surface fluxes that are not compatible with the observations. This was purposefully set up in order to test the assimilation's ability to efficiently reduce these discrepancies while estimating what appears to be magnitude-wise a reasonable correction to the surface fluxes.

Acknowledgements This work was sponsored by the office of Naval Research program element number 0601153N as part of the projects "Exploring Covariances for Ocean Variational Data Assimilation" and "Variational Data Assimilation for Ocean Predicition." This paper is the Naval Research Laboratory paper contribution number JA/7320-11-646. The authors are thankful to the anonymous reviewers for their constructive remarks that helped improve the quality of this manuscript.

Appendix

The discretization of NCOM uses second-order interpolation and differentiation as defined with the notations:

$$\begin{aligned} \overline{\phi}^{x} &= 0.5\left(\phi_{x+\Delta x/2} + \phi_{x-\Delta x/2}\right), \\ \left.\frac{\partial \phi}{\partial x}\right|_{x} &= \frac{1}{\Delta x}\delta_x \phi = \frac{1}{\Delta x}\left(\phi_{x+\Delta x/2} - \phi_{x-\Delta x/2}\right), \end{aligned}$$

and

$$\delta_{2t}\phi = (\phi_{t+\Delta \mathrm{t}} - \phi_{t-\Delta \mathrm{t}})$$

The NCOM equations are then discretized in flux conservative form as follows

$$\begin{aligned}
\frac{\Delta x^u \Delta y^u}{2\Delta t}\delta_{2t}\left(\Delta z^u u\right) = {} & \overline{\Delta x \Delta y \Delta z\left(f + C_{curv}\right)\bar{v}^{y}}^{x} - \Delta y^u \Delta z^u g \delta_x(\zeta^* + \zeta_{atm} - \zeta_{tp}) \\
& - \Delta x^u \Delta z^u \frac{1}{\rho_0}\delta_x(p_i) \\
& - \delta_x\left(\overline{\Delta y^u \Delta z^u u^a}^{x}\bar{u}^{x}\right) - \delta_y\left(\overline{\Delta x^v \Delta z^v v^a}^{x}\bar{u}^{y}\right) \\
& - \delta_z\left(\overline{\Delta x \Delta y w}^{x}\bar{u}^{z}\right) \\
& + \overline{\Delta x \Delta y \Delta z Q}^{x} u_{sor} + F_u^* \\
& + \Delta x^u \Delta y^u \delta_z\left(\frac{\overline{K_M}^{x}}{\left(\overline{\Delta z^w}^{x}\right)^{n+1}}\delta_z u^{n+1}\right)
\end{aligned} \qquad (15.14)$$

$$\begin{aligned}
\frac{\Delta x^v \Delta y^v}{2\Delta t}\delta_{2t}\left(\Delta z^v v\right) = {} & \overline{\Delta x \Delta y \Delta z\left(f + C_{curv}\right)\bar{u}^{x}}^{y} - \Delta x^v \Delta z^v g \delta_y\left(\zeta^* + \zeta_{atm} - \zeta_{tp}\right) \\
& - \Delta x^v \Delta z^v \frac{1}{\rho_o}\delta_y\left(p_i\right) \\
& - \delta_x\left(\overline{\Delta y^u \Delta z^u u^a}^{y}\overline{v}^{x}\right) - \delta_y\left(\overline{\Delta x^v \Delta z^v v^a}^{x}\overline{v}^{y}\right) \\
& - \delta_z\left(\overline{\Delta x \Delta y w}^{y}\overline{v}^{z}\right) \\
& + \overline{\Delta x \Delta y \Delta z Q}^{y} v_{sor} + F_v^* \\
& + \Delta x^v \Delta y^v \delta_z\left(\frac{\overline{K_M}^{y}}{\left(\overline{\Delta z^w}^{y}\right)^{n+1}}\delta_z v^{n+1}\right)
\end{aligned} \qquad (15.15)$$

$$\frac{\Delta x \Delta y}{2\Delta t}\delta_{2t}\left(\Delta z\right) = -\delta_x\left(\Delta y^u \Delta z^u u^a\right) - \delta_y\left(\Delta x^v \Delta z^v v^a\right) - \delta_z\left(\Delta x \Delta y w\right) \qquad (15.16)$$

$$\begin{aligned}
\frac{\Delta x \Delta y}{2\Delta t}\delta_{2t}\left(\Delta z T\right) = {} & -\delta_x\left(\Delta y^u \Delta z^u u^a \overline{T}^{x}\right) - \delta_y\left(\Delta x^v \Delta z^v v^a \overline{T}^{y}\right) \\
& - \delta_z\left(\Delta x \Delta y w \overline{T}^{z}\right) \\
& + \Delta x \Delta y \Delta z Q T_{sor} + \delta_x\left(\frac{\Delta y^u \Delta z^u A_H^u}{\Delta x^u}\delta_x T^{n-1}\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \delta_y \left(\frac{\Delta y^v \Delta z^v A_H^v}{\Delta y^v} \delta_y T^{n-1} \right) \\
&+ \Delta x \Delta y \delta_z \left(\frac{K_H}{(\Delta z^w)^{n+1}} \delta_z T^{n+1}) \right) + \Delta x \Delta y Q_r \delta_z \gamma \qquad (15.17)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\Delta x \Delta y}{2\Delta t} \delta_{2t}(\Delta z S) = &-\delta_x(\Delta y^u \Delta z^u u^a \overline{S}^x) - \delta_y(\Delta x^v \Delta z^v v^a \overline{S}^y) - \delta_z(\Delta x \Delta y w \overline{S}^z) \\
&+ \Delta x \Delta y \Delta z Q S_{sor} + \delta_x \left(\frac{\Delta y^u \Delta z^u A_H^u}{\Delta x^u} \delta_x S^{n-1} \right) \\
&+ \delta_y \left(\frac{\Delta y^v \Delta z^v A_H^v}{\Delta y^v} \delta_y S^{n-1} \right) \\
&+ \Delta x \Delta y \delta_z \left(\frac{K_H}{(\Delta z^w)^{n+1}} \delta_z S^{n+1} \right) \qquad (15.18)
\end{aligned}
$$

In (15.14, 15.15, 15.16, 15.17 and 15.18), F_u and F_v are the horizontal mixing terms, ζ_{atm} and ζ_{tp} are the atmospheric surface pressure and tidal potential respectively, and ζ^* is the surface elevation term that can be distributed among any of the three time levels, $\zeta^* = \alpha_1 \zeta^{n+1} + \alpha_2 \zeta^n + \alpha_3 \zeta^{n-1}$, according to the temporal weighting terms α_1, α_2, or α_3, which are specified by the user. A_M and A_H are the horizontal mixing coefficients for the velocity and scalar fields (temperature and salinity) respectively, likewise K_M and K_M for the vertical mixing, Q is a volume flux source term (with T_{sor}, S_{sor}, u_{sor}, and v_{sor} as the term source values), Q_r is the solar radiation, γ is a function describing the solar extinction, Δx, Δy and Δz denote the grid-cell dimensions defined at the center of the grid cells, and the superscripts u, v and w indicate the grid-cell dimensions computed at those velocity locations on the staggered Arakawa C-grid. f is the Coriolis term, ρ_0 and p_{i} are the reference density of seawater and the internal pressure, respectively, and the horizontal advection velocity terms are given by u^{a} and v^{a}. The term C_{curv} is used to correct the horizontal advection of momentum for the horizontal curvature of the grid. It is calculated as

$$
C_{curv} = \bar{v}^y \frac{\delta_{2x}(\Delta y)}{2\Delta x \Delta y} - \bar{u}^x \frac{\delta_{2y}(\Delta x)}{2\Delta x \Delta y} \qquad (15.19)
$$

The horizontal mixing terms for the momentum equations are given by

$$
\begin{aligned}
F_u^* = \delta_x &\left(2 \overline{\left(\frac{\Delta y^u \Delta z^u A_M^u}{\Delta x^u} \right)}^x \delta_x u^{n-1} \right) \\
&+ \delta_y \left(\overline{\left(\frac{\Delta x^v \Delta z^v A_M^v}{\Delta y^v} \right)}^x \delta_y u^{n-1} + \overline{\left(\frac{\Delta x^v \Delta z^v A_M^v}{\Delta x^v} \right)}^x \delta_x v^{n-1} \right) \qquad (15.20)
\end{aligned}
$$

$$F_v^* = \delta_x \left(\overline{\left(\frac{\Delta y^u \Delta z^u A_M^u}{\Delta y^u} \right)}^y \delta_y u^{n-1} + \overline{\left(\frac{\Delta y^u \Delta z^u A_M^u}{\Delta x^u} \right)}^y \delta_x v^{n-1} \right)$$

$$+ \delta_y \left(2 \overline{\left(\frac{\Delta x^v \Delta z^v A_M^v}{\Delta y^v} \right)}^y \delta_y v^{n-1} \right) \tag{15.21}$$

where the mixing coefficient is modeled according the Smagorinsky formula

$$A_M = C_{Smag} \Delta x \Delta y \left[\left(\frac{1}{\Delta x} \delta_x u^n \right)^2 + \frac{1}{2} \left(\frac{1}{2\Delta y} \delta_{2y} \overline{u^n}^x + \frac{1}{2\Delta x} \delta_{2x} \overline{v^n}^y \right)^2 \right.$$

$$\left. + \left(\frac{1}{\Delta y} \delta_y v^n \right)^2 \right]^{\frac{1}{2}} \tag{15.22}$$

with the magnitude of the eddy coefficient being scaled by the constant C_{smag}. The vertical mixing coefficients are computed using the turbulence closure by Mellor and Yamada in either 2 or 2.5 version.

The computation for the free-surface mode is governed by the equations:

$$\frac{\Delta x^u \Delta y^u}{2\Delta t} \delta_{2t} \left(D^u \bar{u} \right) = -\Delta y^u D^u g \delta_x \left(\alpha_1 \zeta^{n+1} + \alpha_2 \zeta^n + \alpha_3 \zeta^{n-1} \right) + D^u \overline{G_u} \tag{15.23}$$

$$\frac{\Delta x^v \Delta y^v}{2\Delta t} \delta_{2t} \left(D^v \bar{v} \right) = -\Delta x^v D^v g \delta_y \left(\alpha_1 \zeta^{n+1} + \alpha_2 \zeta^n + \alpha_3 \zeta^{n-1} \right) + D^v \overline{G_v} \tag{15.24}$$

$$\frac{\Delta x \Delta y}{2\Delta t} \delta_{2t} \zeta = -\delta_x \left(\Delta y^u \left(\beta_1 \left(\overline{D}^u u \right)^{n+1} + \beta_2 \left(\overline{D}^u u \right)^n + \beta_3 \left(\overline{D}^u u \right)^{n-1} \right) \right)$$

$$- \delta_y \left(\Delta x^v \left(\beta_1 \left(\overline{D}^v v \right)^{n+1} + \beta_2 \left(\overline{D}^v v \right)^n + \beta_3 \left(\overline{D}^v v \right)^{n-1} \right) \right)$$

$$+ \Delta x \Delta y D \overline{Q}, \tag{15.25}$$

where β_1, β_2 and β_3 are positive constants define by the user with $\beta_1 + \beta_2 + \beta_3 = 1$, $D^u \overline{G_u}$ and $D^v \overline{G_v}$ are the vertical integrals of all the terms in the right hand side of (15.14) and (15.15) respectively, with the exception of the surface elevation gradient terms and the vertical mixing, and $D^u = \bar{D}^x$ and $D^v = \bar{D}^y$. The free-surface mode (15.25) is solved by first substituting $(D^u \bar{u})^{n+1}$ and $(D^v \bar{v})^{n+1}$ from the time discretized (15.23) and (15.24) into (15.25), resulting in an elliptic equation that is solved for the surface elevation at time level n + 1, which is then substituted back in (15.23) and (15.24) for computing the barotropic transports $D^u \bar{u}$ and $D^v \bar{v}$ from which the barotropic velocities are obtained.

The vertical discretization uses a combination of sigma layers and z-levels in a three-tiered distribution with (1) free sigma layers near the surface that expand and contract with the free surface elevation, (2) fixed sigma layers that do not vary with the free surface, and (3) fixed z levels that allow for partial bottom cells for a better match of the bottom topography.

References

Amodei L (1995) Solution approchée pour un problème d'assimilation de données avec prise en compte de l'erreur du modèle. Comptes Rendus de l'Académie des Sciences 321:Série IIa, 1087–1094

Barron CN, Kara AB, Martin PJ, Rhodes RC, Smedstad LF (2006) Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). Ocean Model 11:347–375

Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press, New York, 347 pp

Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, Cambridge, UK/NY

Bennett AF, Chua BS, Leslie LM (1996) Generalized inversion of a global numerical weather prediction model. Meteorol Atmos Phys 60:165–178

Bennett AF, Chua BS, Harrison ED, McPhaden MJ (1998) Generalized inversion of Tropical Atmosphere-Ocean (TAO) data and a coupled model of the tropical Pacific. J Climate 11:1768–1792

Bischof C, Corliss G, Green L, Griewank A, Haigler K, Newman P (1992) Automatic differentiation of advanced CFD codes for multidisciplinary design. Comput Syst Eng 3(6):625–637

Broquet G, Edwards CA, Moore AM, Powell BS, Veneziani M, Doyle JD (2009) Application of 4D-variational data assimilation to the California Current System. Dyn Atmos Oceans 48:69–92

Broquet G, Moore AM, Arango HG, Edwards CA (2011) Corrections to ocean surface forcing in the California Current System using 4D variational data assimilation. Ocean Model 36:116–132

Chao Y, Li ZJ, Farrara J, McWilliams JC, Bellingham J, Capet X, Chavez F, Choi JK, Davis R, Doyle J, Fratantoni DM, Li P, Marchesiello P, Moline MA, Paduan J, Ramp S (2009) Development, implementation and evaluation of a data-assimilative ocean forecasting system off the central California coast. Deep-Sea Res Part II-Top Stud Oceanogr 56(3–5):100–126

Chua BS, Bennett AF (2001) An inverse ocean modeling system. Ocean Model 3:137–165

Daley R (1992) Atmospheric data analysis. Cambridge University Press, Cambridge, NY, 472 pp

Daley R, Barker E (2001) NAVDAS formulation and diagnostics. Mon Weather Rev 129:869–883

Derber J, Rosati A (1989) A global oceanic data assimilation system. J Phys Oceanogr 19:1333–1347

Di Lorenzo E, Moore A, Arango H, Chua, B, Cornuelle BD, Miller AJ, Powell B, Bennett A (2007) Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): development and application for a baroclinic coastal upwelling system. Ocean Model 16(3–4):160–187

Egbert GD, Bennett AF, Foreman MGG (1994) TOPEX/POSEIDON tides estimated using a global inverse method. J Geophys Res 99:24821–24852

Erwig M, Fu Z, Pflaum B (2007) Parametric fortran: program generation in scientific computing. J Software Mainten Evol 19(3):155–182

Fox DN, Teague WJ, Barron CN, Carnes MR, Lee CM (2002) The modular ocean data assimilation system (MODAS). J Atmos Ocean Technol 19:240–252

Giering R, Kaminski T (1998) Recipes for adjoint code construction. ACM Trans Math Software 24(4):437–474

Goerss JS, Phoebus PA (1992) The Navy's operational atmospheric analysis. Weather Forecast 7:232–249

Haley PJ, Lermusiaux PFJ, Robinson AR, Leslie WG, Logoutov O, Cossarini G, Liang XS, Moreno P, Ramp SR, Doyle JD, Bellingham J, Chavez F, Johnston S (2009) Forecasting and reanalysis in the Monterey Bay/California current region for the autonomous ocean sampling network-II experiment. Deep-Sea Res Part II-Top Stud Oceanogr 56(3–5):127–148

Hascoet L, Pascual V (2004) Tapenade 2.1 user's guide. Rapport INRIA n 300, 2004; 78. Available from http://www.inria.fr/rrrt/rt-0300.html

Hodur RichardM (1997) The naval research laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). Mon Weather Rev 125:1414–1430

Jacobs GA, Ngodock HE (2003) The maintenance of conservative physical laws within data assimilation systems. Mon Weather Rev 131:2595–2607

Marotzke J, Giering R, Zhang KQ, Stammer D, Hill C, Lee T (1999) Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. J Geophys Res 104(C12):29, 529–29, 547

Martin P (2000) Description of the navy coastal ocean model version 1.0. NRL report NRL/FR/7322—00–9961

Moore AM, Arango HG, Di Lorenzo E, Cornuelle BD, Miller AJ, Neilson DJ (2004) A comprehensive ocean prediction and analysis system based on the tangent linear and adjoint of a regional ocean model. Ocean Model 7:227–258

Ngodock HE (2005) Efficient implementation of covariance multiplication for data assimilation with the representer method. Ocean Model 8(3):237–251

Ngodock HE, Chua BS, Bennett AF (2000) Generalized inversion of a reduced gravity primitive equation ocean model and tropical atmosphere ocean data. Mon Weather Rev 128:1757–1777

Ngodock HE, Jacobs GA, Chen M (2006) The representer method, the ensemble Kalman filter and the ensemble Kalman smoother: a comparison study using a nonlinear reduced gravity ocean model. Ocean Model 12:378–400

Ngodock HE, Smith SR, Jacobs GA (2007) Cycling the representer algorithm for variational data assimilation with the Lorenz attractor. Mon Weather Rev 135:373–386

Ngodock HE, Smith SR, Jacobs GA (2009) Cycling the representer method with nonlinear models. In: Data assimilation for atmospheric, oceanic & hydrologic applications. Springer, Berlin

Peyton Jones SL (2003) Haskell 98 language and libraries: the revised report. Cambridge University Press, Cambridge, UK, 270 pp

Rosenfeld LK, Schwing FB, Garfield N, Tracy DE (1994) Bifurcated flow from an upwelling center: a cold water source for Monterey Bay. Continent Shelf Res 14:931–964

Rosmond TE (1992) The design and testing of the Navy Operational Global Atmospheric Prediction System. Weather Forecast 7:262–262

Shulman I, Wu CR, Lewis JK, Paduan JD, Rosenfeld LK, Kindle JC, Ramp SR, Collins CA (2002) High resolution modeling and data assimilation in The Monterey Bay. Continent Shelf Res 22(8):1129–1151

Shulman I, Rowley C, Anderson S, DeRada S, Kindle J, Martin P, Doyle J, Cummings J, Ramp S, Chavez F, Frantoni D, Davis R (2009) Impact of glider data assimilation on the Monterey Bay Model. Deep Sea Res 56:188–198

Smith SR, Ngodock HE (2008) Cycling the representer algorithm for 4D-variational data assimilation with the Navy Coastal Ocean Model. Ocean Model 24:92–107

Stammer D, Wunsch C, Giering R, Eckert C, Heimbach P, Marotzke J, Adcroft A, Hill CN, Marshall J (2002) The global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model. J Geophys Res 107:C9 3118

Talagrand O (1999) A posteriori evaluation and verification of analysis and assimilation algorithms. In: Proceedings of workshop on diagnosis of data assimilation system held at ECMWF, Shinfield Park, pp 17–28

Weaver A, Courtier P (2001) Correlation modeling on the sphere using a generalized diffusion equation. Q J Roy Meteorol Soc 127:1815–1846

Weaver AT, Vialard J, Anderson DLT (2003) Three- and four-dimensional variational assimilation with a general circulation model of the tropical Pacific Ocean. Part I: formulation, internal diagnostics, and consistency checks. Mon Weather Rev 131(7):1360–1378

Wunsch C (2006) Discrete inverse and state estimation problems. With geophysical fluid applications. Cambridge University Press, Cambridge, 371 pp

Chapter 16
Ocean Ensemble Forecasting and Adaptive Sampling

Xiaodong Hong and Craig Bishop

Abstract An ocean adaptive sampling algorithm, derived from the Ensemble Transform Kalman Filter (ETKF) technique, is illustrated in this Chapter using the glider observations collected during the Autonomous Ocean Sampling Network (AOSN) II field campaign. This algorithm can rapidly obtain the prediction error covariance matrix associated with a particular deployment of the observation and quickly assess the ability of a large number of future feasible sequences of observations to reduce the forecast error variance. The uncertainty in atmospheric forcing is represented by using a time-shift technique to generate a forcing ensemble from a single deterministic atmospheric forecast. The uncertainty in the ocean initial condition is provided by using the Ensemble Transform (ET) technique, which ensures that the ocean ensemble is consistent with estimates of the analysis error variance. The ocean ensemble forecast is set up for a 72 h forecast with a 24 h update cycle for the ocean data assimilation. Results from the atmospheric forcing perturbation and ET ocean ensemble mean are examined and discussed. Measurements of the ability of the ETKF to predict 24–48 h ocean forecast error variance reductions over the Monterey Bay due to the additional glider observations are displayed and discussed using the signal variance, signal variance summary map, and signal variance summary bar charts, respectively.

16.1 Introduction

The impact of supplemental observations on the forecast error reduction depends on: (a) the size of the forecast error at the location where the observation is taken, (b) the assumptions used in the data assimilation scheme about the strength of the correlation between errors in forecasts of the observed variable and errors in all other

X. Hong (✉) · C. Bishop
Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA
e-mail: xd.hong@nrlmry.navy.mil; craig.bishop@nrlmry.navy.mil

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_16,

variables defining the model state, (c) the actual correlation between errors of the observed variable and the model state variables, and (d) the growth and movement of the change in the estimated state imparted by the supplemental observations. In many applications, there is a special region called a *verification region* and a special time called a *verification time*. One often wishes to collect and use supplemental observations at an earlier *observation time* to minimize the forecast error variance within the verification region at the verification time. The problem of identifying the best location for deploying mobile observation platforms is often called the adaptive sampling or targeting problem. The importance of this problem has been heightened in oceanic applications by the advent of Autonomous Underwater Vehicles (AUVs) and underwater gliders. These observing platforms need to be told where to go and when. Since one must decide where to take the supplemental observations well before the targeting time, it is critical to solve the adaptive sampling problem in an accurate and timely manner. The ETKF based technique is used to provide the guidance of the ocean adaptive sampling for the supplemental ocean observations.

The ETKF uses an ensemble forecast initialized at an *initialization time* to quickly obtain the prediction error covariance matrix associated with a particular deployment of observation by solving a low rank Kalman filter equation. The technique can quickly assess the ability of a large number of future feasible sequences of observations to reduce the forecast error variance. The ETKF was developed by Bishop et al. (2001) and first used to provide the optimal flight tracks, where Global Positioning System (GPS) dropwindsondes were released during the Winter Storm Reconnaissance (WSR) program (Szunyogh et al. 2000), for improving the 24–72 h forecasts over the continental United States (Majumdar et al. 2002). It was also used for the medium range forecasts through a single model ensemble (Buizza et al. 2003; Sellwood et al. 2008), and a multi-model ensemble (Majumdar et al. 2010), as well as for tropical cyclone predictions (Majumdar et al. 2011). While the ETKF technique is increasingly used in the area of atmospheric adaptive sampling, there are relatively few applications in the area of ocean adaptive sampling.

In this study, the ETKF ocean adaptive sampling technique is applied to the glider data collected during the AOSN II field campaign that took place in the Monterey Bay in August 2003. The goal for the month-long field experiment was to build a fundamental understanding for upwelling and relaxation processes as well as their impact on the other biological (ecosystem productivity) and chemical (nutrient fertilization) counterparts in the Monterey Bay. To achieve the goal, it was important to develop strategies to command sophisticated robotic vehicles to the locations where the observations collected by them could be the most useful ones (AOSN 2003). Multiple AUVs and underwater gliders were deployed during the field campaign to collect data so that the data could be integrated into ocean forecast models for improving the model performance.

The ocean ensemble and adaptive sampling technique presented here is a continued effort of the verification of ocean modeling project (Hong et al. 2009a). The deterministic run in Hong et al. (2009a) is used as the control run of the ensemble simulation in this study. Consequently, the model, model configuration,

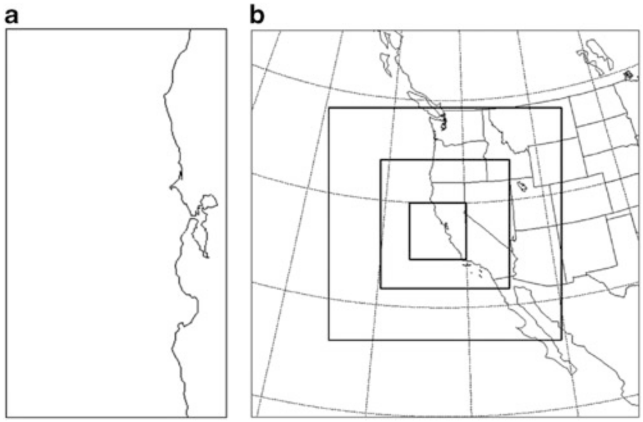


Fig. 16.1 The NCOM and NCODA domain; (**b**) The COAMPS nested domain

and domain setting are exactly the same in both studies. The ocean model is the Navy Coastal Ocean Model (NCOM, Martin 2000) with the multivariate analysis of Navy Coupled Ocean Data Assimilation (NCODA, Cummings 2005). The atmospheric forcing is obtained from a deterministic operational forecast using Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS, Doyle et al. 2008). Figure 16.1 shows the domain setting for the atmospheric and oceanic components of COAMPS as well as NCODA, respectively. The domain for the ocean components is within the innermost nested domain of the atmospheric component of COAMPS.

The rest of the Chapter is organized as follows. In Sect. 16.2, the description of the ETKF adaptive sampling is provided. Section 16.3 contains the discussion of the atmospheric forcing ensemble generation. Section 16.4 presents the results from the ocean ensemble forecast. Section 16.5 illustrates the application of the ocean adaptive sampling for the AOSN II glider observations. Summary and discussion are presented in Sect. 16.6.

16.2 Ocean Adaptive Sampling Technique

In ETKF adaptive sampling, the observations are divided into: (1) non-adaptive or routine observations such as satellite and buoy observed SST, satellite observed altimeter, mooring observed ocean profiles and high frequency radar observed surface current, and (2) adaptive observations such as aircraft observed SST and observations collected by autonomous underwater gliders. The first step is to estimate routine analysis error covariance matrix valid for the ocean routine observations. The second step is to estimate the reduction in forecast error variance due to the supplemental ocean adaptive observations.

16.2.1 Analysis Error Covariance for the Routine Observations with the ET Technique

To be consistent with the ET technique of ensemble generation, we need to utilize a guess of the analysis error covariance matrix $\mathbf{P}_g^a$ associated with the routine observational network. Let the columns of the nxK matrices $\mathbf{X}^o$ and $\mathbf{X}^v$ list the raw ensemble perturbations at the observation and verification times, respectively, of the ensemble forecast initialized at the initialization time.

The forecast perturbations $\mathbf{X}^o$ can be transformed into a set of perturbations $\mathbf{X}^r$ that are consistent with $\mathbf{P}_g^a$ using

$$\mathbf{X}^r = \mathbf{X}^o\mathbf{T} \tag{16.1}$$

where

$$\mathbf{T} = \mathbf{B}\cancel{\Lambda}^{-1/2}\mathbf{B}^T \tag{16.2}$$

and where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K]$ is a $\mathbf{K} \times \mathbf{K}$ orthogonal matrix containing the eigenvectors of the symmetric matrix $\left(\mathbf{X}^{oT}\mathbf{P}_g^{a-1}\mathbf{X}^0/N\right)$. In other words,

$$\frac{\mathbf{X}^{oT}\mathbf{P}_g^{a-1}\mathbf{X}^o}{N} = \mathbf{B}\Lambda_{K\,\times\,K}\mathbf{B}^T. \tag{16.3}$$

where $\Lambda = diag\,(\lambda_{11}, \lambda_{22}, \ldots, \lambda_{KK})$ is a $\mathbf{K} \times \mathbf{K}$ diagonal matrix listing the eigenvalues of $\left(\mathbf{X}^{oT}\mathbf{P}_g^{a-1}\mathbf{X}^0/N\right)$. Since the sum of the forecast perturbations is equal to zero, one of these eigenvalues will be equal to zero. Consequently, provided each ensemble contains K−1 linearly independent perturbations, Λ can be written in the form,

$$\Lambda_{K\,\times\,K} = \begin{bmatrix} \Lambda_{(K-1)\,\times\,(K-1)} & 0 \\ 0 & 0 \end{bmatrix} \tag{16.4}$$

where $\Lambda_{(K-1)\,\times\,(K-1)}$ is a (K−1)×(K−1) diagonal matrix whose diagonal elements are all greater than zero. The $\cancel{\Lambda}$ used in (16.4) is obtained from Λ by setting its zero eigenvalue equal to 1, in other words,

$$\cancel{\Lambda}_{K\,\times\,K} = \begin{bmatrix} \Lambda_{(K-1)\,\times\,(K-1)} & 0 \\ 0 & 1 \end{bmatrix} \tag{16.5}$$

Note that while $\cancel{\Lambda}$ has an inverse, the inverse of Λ does not exist. This adjustment of the eigenvalue matrix is permissible because it does not affect the sample covariance matrix of initial perturbations implied by (16.3). To see this, first note that pre and post multiplying (16.5) by the eigenvector $\mathbf{b}_K$ corresponding to the zero eigenvalue $\lambda_K = 0$ shows that

$$\frac{\mathbf{b}_K^T\mathbf{X}^{oT}\mathbf{P}_g^{a-1}\mathbf{X}^0\mathbf{b}_K}{N} = 0, \text{ and consequently } |\mathbf{X}^o\mathbf{b}_K| = 0. \tag{16.6}$$

Second, note that if λ_{ii} and $\cancel{\lambda}_{ii}$ denote the diagonal elements of Λ and $\cancel{\Lambda}$, respectively, we may deduce that the perturbation ensemble sample covariance matrix $\mathbf{P}_e^r$ associated with the transformed ensemble perturbations is given by

$$\begin{aligned}\mathbf{P}_e^r &= \frac{\mathbf{X}^r\mathbf{X}^{rT}}{K-1} = \frac{\mathbf{X}^o\mathbf{T}\mathbf{T}^T\mathbf{X}^{oT}}{K-1} = \frac{\mathbf{X}^o\mathbf{B}\cancel{\Lambda}^{-1}\mathbf{B}^T\mathbf{X}^{oT}}{K-1} \\ &= \frac{1}{K-1}\sum_{i=1}^{K}\frac{\mathbf{x}_i^o\mathbf{b}_i\mathbf{b}_i^T\mathbf{x}_i^{oT}}{\cancel{\lambda}_{ii}^{1/2}} = \frac{1}{K-1}\sum_{i=1}^{K-1}\frac{\mathbf{x}_i^o\mathbf{b}_i\mathbf{b}_i^T\mathbf{x}_i^{oT}}{\lambda_{ii}^{1/2}}\end{aligned} \tag{16.7}$$

where $\mathbf{b}_i$ is the ith column of $\mathbf{B}$. Equation 16.7 shows that because $|\mathbf{X}^o\mathbf{b}_K| = 0$, $\mathbf{P}_e^r$ is entirely independent of the value assigned to Kth eigenvalue. Throughout this discussion we will assume that every ensemble contains K−1 linearly independent ensemble perturbations.

16.2.2 Signal Variance and Forecast Error Variance Reduction for Adaptive Observation with the ETKF Technique

If the true analysis error covariance at the observation time after assimilating all routine observations was given by $\mathbf{P}_e^r = \frac{\mathbf{X}^r\mathbf{X}^{rT}}{K-1}$ then the posterior analysis error covariance $\mathbf{P}_i^a$ after assimilating the ith feasible deployment of adaptive observations $\mathbf{y}_i^a$ in addition to the routine observations is given by

$$\mathbf{P}_i^a = \mathbf{P}_r^e - \mathbf{P}_r^e\tilde{\mathbf{H}}_i^{aT}\left(\tilde{\mathbf{H}}_i^a\mathbf{P}_r^e\tilde{\mathbf{H}}_i^{aT} + \mathbf{I}\right)^{-1}\tilde{\mathbf{H}}_i^a\mathbf{P}_r^e \tag{16.8}$$

where $\tilde{\mathbf{H}}_i^a$ describes the mapping from the model state vector to the observation vector normalized by the inverse square root of the observation error covariance $\mathbf{R}_i^{-1/2}$ associated with the ith feasible deployment; in other words,

$$\tilde{\mathbf{H}}_i^a\mathbf{x}^t = \mathbf{R}_i^{-1/2}\mathbf{y}_i^t \tag{16.9}$$

where $\mathbf{x}^t$ denotes the true model state and $\mathbf{y}_i^t$ denoted the true value of the observed variable. As shown in Bishop et al. (2001), if

$$\mathbf{P}_i^a = \frac{\mathbf{X}_i^a\mathbf{X}_i^{aT}}{K-1} \tag{16.10}$$

where $\mathbf{X}_i^a$ is a n×K matrix then

$$\mathbf{X}_i^a = \mathbf{X}^r \mathbf{C}_i \left(\Gamma_i + \mathbf{I}\right)^{-1/2} \mathbf{C}_i^T \tag{16.11}$$

where the K×K orthonormal matrix $\mathbf{C}_i$ and the K × K diagonal matrix Γ_i is given by the eigenvector decomposition

$$\frac{\mathbf{X}_r^{aT}\mathbf{H}_i^{aT}\mathbf{R}_i^{-1}\mathbf{H}_i^a\mathbf{X}_r^a}{K-1} = \mathbf{C}_i \Gamma_i \mathbf{C}_i^T. \tag{16.12}$$

The columns of $\mathbf{X}_i^a$ may be interpreted as transformed ensemble perturbations whose covariance gives the analysis error covariance at the observation time assuming that the ith deployment of adaptive observations had been assimilated. To see the impact of the adaptive observations at the verification time, one needs to be able to propagate each of the columns of $\mathbf{X}_i^a$ through time in a manner consistent with the governing dynamical equations. A computationally expensive way of doing this would be to define a tangent linear model $\mathbf{M}$ such that

$$M\left(\mathbf{x}_c^o + \mathbf{x}_{ji}^a\right) - M\left(\mathbf{x}_c^o\right) \approx \mathbf{M}\mathbf{x}_{ji}^a \tag{16.13}$$

where M is the non-linear dynamics propagator that maps state vectors from the observation time to the verification time, $\mathbf{x}_c^o$ is the control forecast at the observation time and $\mathbf{x}_{ji}^a$ is the jth column of $\mathbf{X}_i^a$. If one had this operator in hand, then the forecast error covariance matrix given the ith deployment of observations $\mathbf{P}_i^v$ would be given by

$$\mathbf{P}_i^v = \frac{\mathbf{M}\mathbf{X}_i^a\left(\mathbf{M}\mathbf{X}_i^a\right)^T}{K-1} \tag{16.14}$$

However, using (16.11) and (16.1) in (16.14) gives

$$\mathbf{M}\mathbf{X}_i^a = \left(\mathbf{M}\mathbf{X}^o\right)\mathbf{T}\mathbf{C}\left(\Gamma + \mathbf{I}\right)^{-1/2}\mathbf{C}^T. \tag{16.15}$$

Now $\mathbf{M}\mathbf{X}^o$ represents a tangent linear approximation to the propagation of the raw untransformed ensemble perturbations at the observation time to the verification time. Of course, the non-linear equations map the observation time raw perturbations $\mathbf{X}^o$ to the verification time perturbations $\mathbf{X}^v$. These are directly available from the raw ensemble without any additional computational expense. Hence, a computationally inexpensive way of computing $\mathbf{P}_i^v$ that is more accurate than that given by (16.14) is

$$\mathbf{P}_i^v = \frac{\mathbf{X}_i^v\mathbf{X}_i^{vT}}{K-1}, \text{where} \quad \mathbf{X}_i^v = \mathbf{X}^v\mathbf{T}\mathbf{C}_i\left(\Gamma_i + \mathbf{I}\right)^{-1/2}\mathbf{C}_i^T. \tag{16.16}$$

Equation 16.16 gives the forecast error covariance of the model variables given the ith deployment of adaptive observations. Often the controller of adaptive observational resources will want to use them to minimize the error variance of some q-vector function $\mathbf{f}^v$ of some subset(s) of the forecasted variables.

A perfect raw ensemble would provide K draws from the distribution of verifying functions given the forecast. In particular, the jth ensemble member gives

$$\begin{aligned} \mathbf{f}_j^v &= H^v\left(\mathbf{x}_j^v\right) = H^v\left[\bar{\mathbf{x}}^v + \left(\mathbf{x}_j^v - \bar{\mathbf{x}}^v\right)\right] \\ &\simeq H^v\left(\bar{\mathbf{x}}^v\right) + \mathbf{H}^v\left(\mathbf{x}_j^v - \bar{\mathbf{x}}^v\right) \end{aligned} \tag{16.17}$$

where $\bar{\mathbf{x}}^v$ is the mean of the ensemble forecast and where H^v is the non-linear function of interest and $\mathbf{H}^v$ is the derivative of the non-linear function with respect to the model variables about the mean of the ensemble forecast $\bar{\mathbf{x}}^v$. Thus, the estimate of the qxq forecast error covariance matrix of the vector function $\mathbf{f}$ associated with the forecast upon which targeting decisions is made is given by

$$\begin{aligned} \left\langle\left(\mathbf{f}-\mathbf{f}^t\right)\left(\mathbf{f}-\mathbf{f}^t\right)^T\right\rangle &\simeq \frac{1}{K-1}\sum_{j=1}^{K}\left[H^v\left(\mathbf{x}_j^v\right) - \overline{H^v\left(\mathbf{x}_j^v\right)}\right]\left[H^v\left(\mathbf{x}_j^v\right) - \overline{H^v\left(\mathbf{x}_j^v\right)}\right]^T \\ &\simeq \frac{1}{K-1}\sum_{j=1}^{K}\mathbf{H}^v\left(\mathbf{x}_j^v - \bar{\mathbf{x}}^v\right)\left(\mathbf{x}_j^v - \bar{\mathbf{x}}^v\right)^T\mathbf{H}^{vT} \\ &= \frac{\mathbf{H}^v\mathbf{X}^v\left(\mathbf{H}^v\mathbf{X}^v\right)^T}{K-1}. \end{aligned} \tag{16.18}$$

where $\overline{H^v\left(\mathbf{x}_j^v\right)}$ denotes the mean of the ensemble of vector functions. Using (16.18) and (16.16) leads to the following estimate of forecast error covariance matrix $\left\langle\left(\mathbf{f}-\mathbf{f}^t\right)\left(\mathbf{f}-\mathbf{f}^t\right)^T\right\rangle_i$ for the vector function $\mathbf{f}$ given routine observations and the ith deployment of adaptive observations.

$$\begin{aligned} \left\langle\left(\mathbf{f}-\mathbf{f}^t\right)\left(\mathbf{f}-\mathbf{f}^t\right)^T\right\rangle_i &\approx \frac{\mathbf{H}^v\mathbf{X}_i^v\mathbf{X}_i^{vT}\mathbf{H}^{vT}}{K-1} \\ &= \frac{\mathbf{H}^v\mathbf{X}^v\mathbf{T}\mathbf{C}_i\left(\Gamma_i+\mathbf{I}\right)^{-1}\mathbf{C}_i^T\mathbf{T}^T\mathbf{X}^{vT}\mathbf{H}^{vT}}{K-1} \\ &\approx \frac{\left[H^v\left(\mathbf{X}^v\right)\right]\mathbf{T}\mathbf{C}_i\left(\Gamma_i+\mathbf{I}\right)^{-1}\mathbf{C}_i^T\mathbf{T}^T\left[H^v\left(\mathbf{X}^v\right)\right]^T}{K-1} \end{aligned} \tag{16.19}$$

where the qxK matrix $\left[H^v\left(\mathbf{X}^v\right)\right]$ is given by

$$\begin{aligned} &\left[H^v\left(\mathbf{X}^v\right)\right] \\ &\quad= \left[\left(H^v\left(\mathbf{x}_1^v\right) - \overline{H^v\left(\mathbf{x}^v\right)}\right), \left(H^v\left(\mathbf{x}_2^v\right) - \overline{H^v\left(\mathbf{x}^v\right)}\right), \ldots, \left(H^v\left(\mathbf{x}_K^v\right) - \overline{H^v\left(\mathbf{x}^v\right)}\right)\right]. \end{aligned} \tag{16.20}$$

Thus, the ETKF allows non-linear cost functions without the need for the first derivative (Jacobian) of the non-linear verification time functions of interest.

Equation 16.19 gives the forecast error covariance of the user specified functions of interest for the ith deployment of adaptive observations. Often, users will reduce the information in this matrix to a single *cost* function by, for example, evaluating the trace of the matrix. To find which of all feasible deployments of adaptive observations minimizes the user specified cost function, one simply evaluates (16.19) for all feasible deployments of adaptive observations and chooses the deployment which minimizes the cost. Since the transformation matrix (16.2) associated with the routine observational network and the $[H^v(\mathbf{X}^v)]$ matrix only need to be evaluated once, the main computational expense associated with each deployment is the K × K eigenvector decomposition (16.12). For ensemble sizes smaller than 100, this is a trivial expense on today's CPUs and thousands of networks can be evaluated in a matter of minutes on moderate computing resources.

To highlight and predict the impact of the targeted observations, it is also of interest to predict the covariance of the distribution of changes to the forecast that would be imparted by the ith observational network given an infinite sampling of the distributions of observation and forecast. As shown in Bishop et al. (2001), at the observation time this covariance is given by

$$\begin{aligned}\left\langle\left(\mathbf{x}_i^o-\mathbf{x}_r^o\right)\left(\mathbf{x}_i^o-\mathbf{x}_r^o\right)^T\right\rangle &= \mathbf{P}_r^e\tilde{\mathbf{H}}_i^{aT}\left(\tilde{\mathbf{H}}_i^a\mathbf{P}_r^e\tilde{\mathbf{H}}_i^{aT}+\mathbf{I}\right)^{-1}\tilde{\mathbf{H}}_i^a\mathbf{P}_r^e\\ &= \frac{\mathbf{X}^o\mathbf{T}\mathbf{C}_i\Gamma_i\left(\Gamma_i+\mathbf{I}\right)^{-1}\mathbf{C}_i^T\mathbf{T}^T\mathbf{X}^{oT}}{K-1}\end{aligned} \tag{16.21}$$

where $\mathbf{x}_r^o$ represents the minimum error variance state estimates at the observation time given routine observations while $\mathbf{x}_i^o$ represents the minimum error variance state estimates at the observation time given routine observations *and* the ith deployment of adaptive observational resources. Thus, it represents the covariance of changes to the state estimate due to adaptive observations. The changes due to the adaptive observations are called *signals* and the covariance of these changes is called the *signal covariance.* The expression for the signal covariance at the verification time is

$$\left\langle\left(\mathbf{x}_i^v-\mathbf{x}_r^v\right)\left(\mathbf{x}_i^v-\mathbf{x}_r^v\right)^T\right\rangle = \frac{\mathbf{X}^v\mathbf{T}\mathbf{C}_i\Gamma_i\left(\Gamma_i+\mathbf{I}\right)^{-1}\mathbf{C}_i^T\mathbf{T}^T\mathbf{X}^{vT}}{K-1} \tag{16.22}$$

As can be seen by comparing (16.21) with (16.8) and as was discussed in Bishop et al. (2001), for an optimal data assimilation scheme, the signal variance is precisely equal to the reduction in forecast error variance due to the observations that created the signals. Comparison of geographical plots of the diagonal elements of (16.21) and (16.22) with actual changes in forecasts due to targeted observations can give a good indication of whether the ETKF signal variance predictions are reasonable or not.

16.3 Atmospheric Forcing Ensemble Generation

Based on the theory that model forecast errors are often well described in terms of shifting and timing errors (Hoffman et al. 1995), the uncertainty of atmospheric forcing can be represented by adding perturbations to surface fields from a single deterministic atmospheric forecast through spatial and temporal deformation. The amplitude of the perturbations is chosen to be small enough to ensure that the perturbed field lies within the error bounds of the forecast. To control the amplitude and horizontal correlation length scale of the random perturbations, the covariance matrix of the shift-vector $\delta\mathbf{t}$ of shifts at a certain time is given by:

$$\left\langle \delta\mathbf{t}\delta\mathbf{t}^T \right\rangle = \mathbf{DE}\,\Lambda\,\mathbf{E}^T\mathbf{D} \tag{16.23}$$

where $\mathbf{D}$ is a diagonal matrix of the variances we wish to assign to the random process at each grid point and $\mathbf{E}\,\Lambda\,\mathbf{E}^T$ defines a correlation matrix whose diagonal values are all equal to 1. For simplicity, we chose the columns of $\mathbf{E}$ to be the two-dimensional sinusoids and cosinusoids that define a basis for the two-dimensional domain upon which the ocean state is defined. Let $\mathbf{a}$ be a random normal vector with zero mean and covariance $\left\langle \mathbf{aa}^T \right\rangle = \Lambda$. Now consider random vectors $\mathbf{y}$ obtained using $\mathbf{y} = \mathbf{Ea}$. Note that since the columns of $\mathbf{E}$ are the sinusoidal basis used in inverse Fourier transform, the operation $\mathbf{Ea}$ is simply an inverse Fourier transform. To ensure that the random perturbations satisfy (16.1), we generate each perturbation using

$$\delta\mathbf{t} = \mathbf{DEa}, \text{ where } \langle \mathbf{a} \rangle = 0 \text{ and } \left\langle \mathbf{aa}^{\mathrm{T}} \right\rangle = \Lambda \tag{16.24}$$

In other words, a random perturbation is created by

1. Creating a vector $\mathbf{b}$ of n normally independently identically distributed numbers each of which has a mean of zero and a variance of 1.
2. Letting $\mathbf{a} = \Lambda^{1/2}\mathbf{b}$.
3. Performing the inverse Fourier transform implied by $\mathbf{Ea}$.
4. Performing the operation $\delta\mathbf{t} = \mathbf{DEa}$.

To see that this process creates random perturbations that satisfy (16.1) note that

$$\begin{aligned}\left\langle \delta\mathbf{t}\delta\mathbf{t}^T \right\rangle &= \left\langle \mathbf{DEaa}^T\mathbf{E}^T\mathbf{D} \right\rangle \\ &= \mathbf{DE}\left\langle \mathbf{aa}^T \right\rangle \mathbf{E}^T\mathbf{D} \text{ because } \mathbf{E} \text{ and } \mathbf{D} \text{ are constant} \\ &= \mathbf{DE}\,\Lambda\,\mathbf{E}^T\mathbf{D}, \text{ because } \left\langle \mathbf{aa}^T \right\rangle = \Lambda \end{aligned} \tag{16.25}$$

The scales and magnitudes of the random perturbations are thus determined by the user's specification of $\mathbf{D}$ and Λ. Here, we chose $\mathbf{D} = \alpha\mathbf{I}$ so that the constant α gives the variance at each point and let the diagonal elements λ_{ii} of Λ be given by the Gaussian function of the total wavenumber to which they pertain that is given by

$$\lambda_{ii}(k,l) = C \times \exp\left(\frac{-(k^2+l^2)}{L^2}\right) \tag{16.26}$$

where k and l are non-dimensional wave numbers (associated with the indexing of grid points in the FFT routine), L (a non-dimensional length scale) controls the horizontal correlation length scale in spectral space. Decreasing L increases the spatial scale of the random fields by (16.24). The scale C is an amplitude factor that is used to ensure that the diagonal elements of $\mathbf{E}\,\Lambda\,\mathbf{E}^T$ are equal to unity and hence that $\mathbf{E}\,\Lambda\,\mathbf{E}^T$ is a valid correlation matrix. The values of C, L and α used in our experiments are 0.5, 10, and 0.5 h, respectively. With these parameters, (16.24) produces a spatially correlated field of time shifts with a standard deviation of $\alpha = 0.5\,\text{h}$.[1]

To create a time shift vector $\delta\mathbf{t}(t)$ that varies in time as well as space, we used (16.24) to create two entirely independent time-shift vector shifts $\delta\mathbf{t}(t_i)$ and $\delta\mathbf{t}(t_{i+1})$ corresponding to the discrete times t_i and t_{i+1}. These two times might be 24 or 72 h apart depending on the perceived decorrelation time of atmospheric forcing errors. (In our study independent fields were generated every 24 h). To ensure that the time shift vector varied smoothly between these two times, we set

$$\delta\mathbf{t}(t) = \delta\mathbf{t}(t_i)\cos\left[\frac{\pi}{2}\left(\frac{t-t_i}{t_{i+1}-t_i}\right)\right] + \delta\mathbf{t}(t_{i+1})\sin\left[\frac{\pi}{2}\left(\frac{t-t_i}{t_{i+1}-t_i}\right)\right] \tag{16.27}$$

Equation 16.27 implies that the evolution of the covariance of time shifts is given by

$$\begin{aligned}\left\langle \delta\mathbf{t}\,(t)\,\delta\mathbf{t}\,(t)^T \right\rangle = {} & \mathbf{D}_i\mathbf{E}\Lambda_i\mathbf{E}^T\mathbf{D}_i\cos^2\left[\frac{\pi}{2}\left(\frac{t-t_i}{t_{i+1}-t_i}\right)\right] \\ & + \mathbf{D}_{i+1}\mathbf{E}\Lambda_{i+1}\mathbf{E}^T\mathbf{D}_{i+1}\sin^2\left[\frac{\pi}{2}\left(\frac{t-t_i}{t_{i+1}-t_i}\right)\right]\end{aligned} \tag{16.28}$$

This formulation allows both the scale and magnitude of the deformations to be a function of time. Note also that in the special case that $\mathbf{D}_{i+1}\mathbf{E}\Lambda_{i+1}\mathbf{E}^T\mathbf{D}_{i+1} = \mathbf{D}_i\mathbf{E}\Lambda_i\mathbf{E}^T\mathbf{D}_i$, the trigonometric rule $\cos^2\theta + \sin^2\theta = 1$ ensures that the covariance of the time shifts given by (16.27) and (16.28) is constant even though each individual time shift is smoothly evolving through time.

For the experiments reported in this Chapter, the eigenvector matrix $\mathbf{E}$ was comprised by the set of sinusoidal basis functions spanning a two dimensional plane. By making the domain on which the time shifts $\delta\mathbf{t}$ were generated larger than that of the regional ocean model, it was possible to produce aperiodic time-shifts.

The temporally shifted fields include surface wind, air temperature, relative humidity, precipitation, sea-level pressure, and short- and long-wave radiation. Each

[1] This technique has been used to perturb an initial best-guess unperturbed state of sea surface temperature (SST) to provide an ensemble of ocean-surface lower boundary conditions for atmospheric ensemble forecast (Hong et al. 2011).

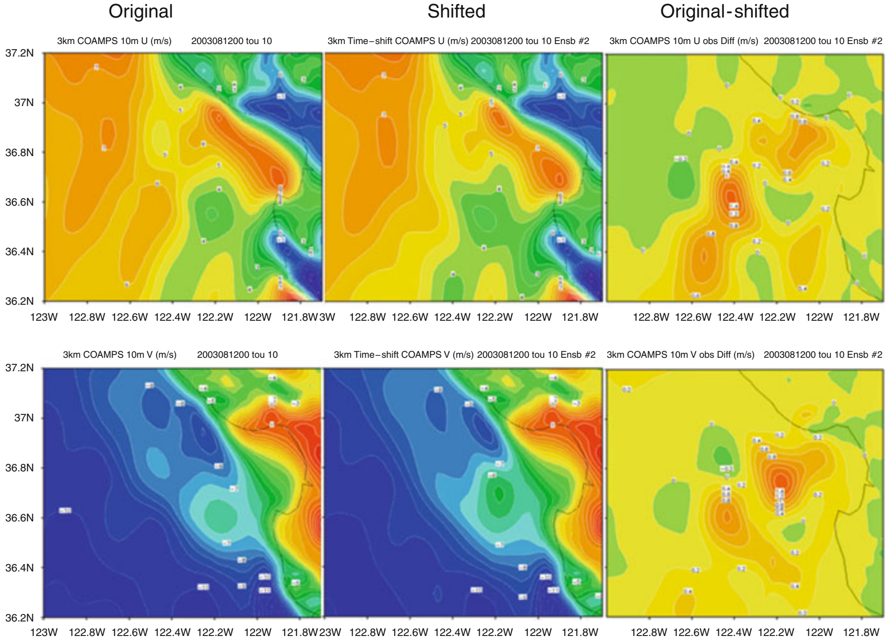

Original
Shifted
Original-shifted
3km COAMPS 10m U (m/s) 2003081200 tou 10
3km Time-shift COAMPS U (m/s) 2003081200 tou 10 Ensb #2
3km COAMPS 10m U obs Diff (m/s) 2003081200 tou 10 Ensb #2
3km COAMPS 10m V (m/s) 2003081200 tou 10
3km Time-shift COAMPS V (m/s) 2003081200 tou 10 Ensb #2
3km COAMPS 10m V obs Diff (m/s) 2003081200 tou 10 Ensb #2
37.2N
37N
36.8N
36.6N
36.4N
36.2N
123W
122.8W
122.6W
122.4W
122.2W
122W
121.8W

Fig. 16.2 Original (*first column*), shifted (*second column*) and difference between original and shifted u-component (*upper panel*) and v-component of surface 10-m wind speed from COAMPS forecast for AOSN II domain (Monterey Bay)

randomly shifted field is used to compute the surface wind stress and heat fluxes for each ensemble member. The NCOM-predicted SST is interactively feedback to the surface latent and sensible heat fluxes using the drag coefficient from the standard bulk formulas of Kondo (1975) (Martin and Hodur 2003; Hong et al. 2007, 2009b). The surface salt flux for NCOM is calculated from the computed latent heat flux and the COAMPS precipitation.

Figure 16.2 shows u- and v- components of surface 10-m wind from a single COAMPS deterministic forecast, a time shifted field and the difference between the original and shifted fields. The high-resolution COAMPS atmospheric forecast presents a strong northwesterly, which is favorable for the ocean coastal upwelling for the Monterey Bay during the AOSN II field campaign (Doyle et al. 2008). The northwesterly lasts from 7 to 19 August and induces an upwelling period. The perturbed atmospheric forcing fields for a particular ensemble member and forecast lead time present smooth features over the entire domain. The northwesterly wind is preserved in the perturbed fields so that the upwelling will be induced in each ocean ensemble forecast with the inclusion of atmospheric forcing uncertainty. The difference between the original and shifted fields displays various locations of maximum perturbation, which explains the feature of random distribution from space and time shifting.

16.4 Ocean Ensemble Forecast

Ocean ensemble generation is based on the ET technique, which has been used for atmospheric ensemble generation (Bishop et al. 2009) and for coupled atmosphere/ocean ensemble generation (Holt et al. 2011). The ET technique provides initial perturbations that (1) have an initial variance consistent with the best available estimates of initial condition error variance, (2) are dynamically conditioned by a process similar to that used in the breeding technique (Toth and Kalnay 1993, 1997), (3) add to zero at the initial time, (4) are quasi-orthogonal and equally likely, and (5) partially respect mesoscale balance constraints by ensuring that each initial perturbation is a linear sum of forecast perturbations from the preceding forecast. The analysis error variance is used to constrain the magnitude of initial perturbations that represent transformations or linear combinations of ensemble forecast perturbations, so called ET perturbations (Bishop and Toth 1999; Bishop et al. 2009). The analysis error variance used in this study is scaled from the NCODA ocean analysis to adjust large untruthful values from the sparse ocean observations. A complete description of the ET technique and the detailed steps to creating an ET ensemble can be found in Bishop et al. (2009).

The ocean ensemble with 20 ensemble members is initialized from a set of perturbations derived from a control deterministic NCOM run for one month from August 1–31, 2003. The NCOM monthly run is performed in a sequential incremental update cycle with an update interval of 24 h and produces 72 h forecast at each analysis update time (Hong et al. 2009a). The differences between every 12 h forecast (up to 24 h) and monthly mean generate 62 perturbations, which provide a database for random selection of initial ensemble perturbations.

From August 7–19, the winds are upwelling favorable with north/northwesterly (Doyle et al. 2008) and induce strong upwellings with two upwelling centers developed off Point Ano Nuevo and Point Sur (Hong et al. 2009a). Ensemble means display stronger upwellings from the two upwelling centers than in the control run and provide features more comparable with the observation (Fig. 16.3). Stronger horizontal SST gradients occur between the upwelled cold water and the offshore warm water. The seaward advection is more consistent with the observation from the ensemble mean on August 12 (upper panel in Fig. 16.3). Later in the upwelling period, a cold tongue of upwelled water off Point Ano Nuevo is advected southward across the mouth of the Monterey Bay and joins with the upwelled cold water from Point Sur, resulting in a large, cold-water region located just off the coast both in ensemble mean and the observation. These results indicate that the ensemble means are more accurate to the observation MCSST than the control run.

The ensemble spread increases with the forecast lead time as shown for SST forecast in Fig. 16.4. Large ensemble spread transports southward with time, reflecting the upwelled cold water movement. It indicates that the forecasted transport of upwelled cold water across the mouth of the Monterey Bay during the upwelling period has high uncertainty.

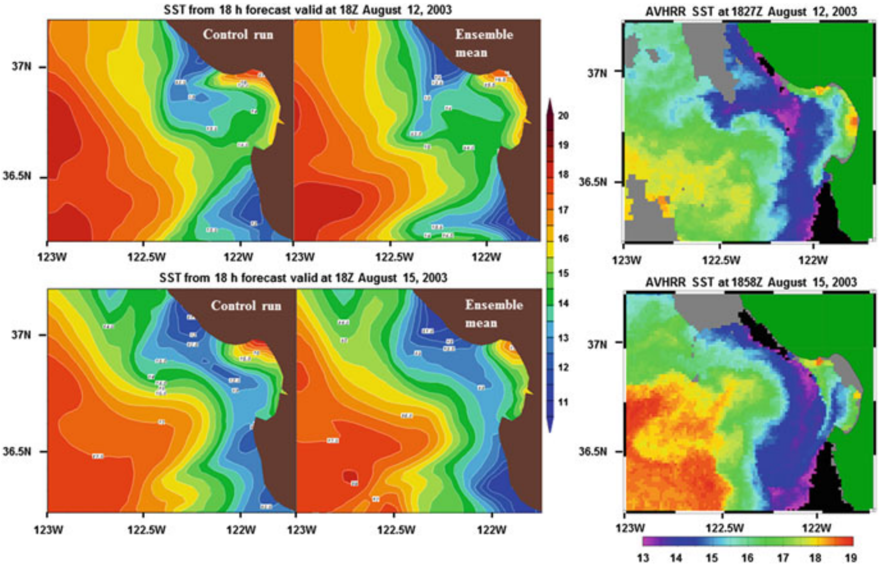

SST from 18 h forecast valid at 18Z August 12, 2003
Control run
Ensemble mean
SST from 18 h forecast valid at 18Z August 15, 2003
AVHRR SST at 1827Z August 12, 2003
AVHRR SST at 1858Z August 15, 2003
37N
36.5N
123W
122.5W
122W
20
19
18
17
16
15
14
13
12
11
13 14 15 16 17 18 19

Fig. 16.3 SST from NCOM control run, ensemble mean and NOAA POES AVHRR HRPT (Courtesy NWS and NOAA Coastwatch). The model outputs are from 18 h forecast valid at 18Z August 12, 2003 for the *upper panel* and 18Z August 15, 2003 for the *lower panel*

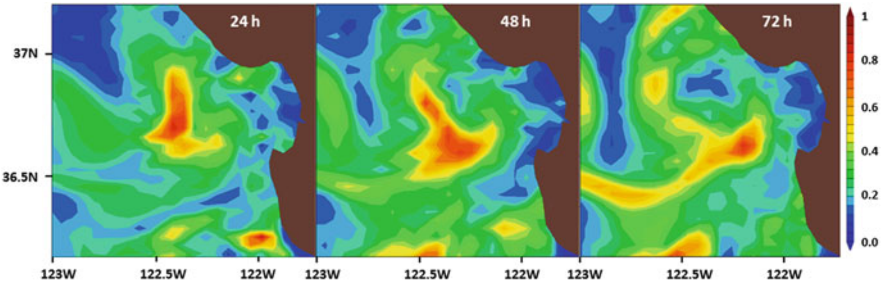


Fig. 16.4 Ensemble spread for 24, 48 and 72 h forecast initiated from August 12, 2003

16.5 Adaptive Sampling for the AOSN II Glider Observation

The underwater vehicle network features a fleet (up to 15 gliders) of autonomous underwater gliders during the AOSN-II field campaign. Underwater gliders are small, relatively simple and inexpensive, winged, buoyancy-driven submersibles. They are ideal platforms to collect scientific data for the ocean adaptive sampling. The deployment of the gliders are efficient and effective by allowing them to change plans on-line in response to the state and environmental measurement needs with

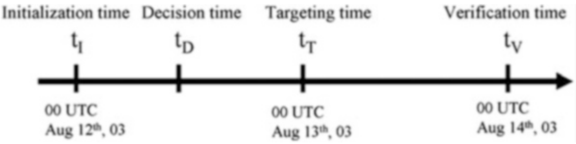


Fig. 16.5 Illustration of ensemble initialization time, decision time, targeting time and verification time for adaptive sampling used in this study

daily time scale and faster time scale (on the order of every two hours) (Leonard and Robinson 2003). With the ability to frequently update the glider plan, the time for decision-making for optimal glider deployment can be shorter than other type of platform deployment, such as aircraft equipped with GPS dropwindsondes for upstream observation of significant weather event (Majumdar et al. 2002).

Key times involved in the decision-making process for the adaptive sampling application of AOSN II glider observation are illustrated in Fig. 16.5. The goal of the adaptive sampling is to use an available ensemble forecast to identify the future glider path that would maximally reduce the forecast error variance in the verification region at the verification time. As an example, consider the ensemble forecast initialized at the initialization time of 00 UTC Aug 12th. A new forecast will be initialized at the targeting time of 00 UTC Aug 13th using targeted observations. The decision time is the time when one must decide the location to which the glider should be sent in order to minimize the error norm of the forecast to be initialized on 00 UTC Aug. 13th. The verification time selected here is 00 UTC Aug 14th to verify the forecast error reduction for the upwelled cold water transport across the mouth of the Monterey Bay.

For a group of adaptive observations, the signal variance, which would be equal to the reduction in forecast error variance in an optimal system, is used to identify the best location for the deployment. The verification region is placed in a location within which the ensemble variance is large at the verification time. This choice of verification region increases the chances that the targeted observations will result in a significant reduction in forecast error (Bishop et al. 2006). Figure 16.4b illustrates the fact that for a verification time 48 h from the ensemble initialized, at 00 UTC Aug. 12th, there is a large ensemble spread across the mouth of the Monterey Bay due to the uncertainty of the southward transportation of upwelled cold water from Point Ano Nuevo. The verification region selected to enclose some of this high spread region is shown by the ellipse on Fig. 16.4b. The possible location for optimal adaptive deployment can be tested in the two areas where the ensemble spread is significant at the targeting time. As shown in Fig. 16.4a, there are two possible locations with one off the mouth of the Monterey Bay (location #1) and another one in the south off Point Sur coast (location #2).

Nine adjacent “test” observations of surface temperature are placed for these two locations centered at 36.7°N, 122.5°W and 36.2°N, 122.1°W, respectively and used to calculate signal variance at the targeting and verification times (Fig. 16.6). There are high signal variances for both locations of the adaptive observation at

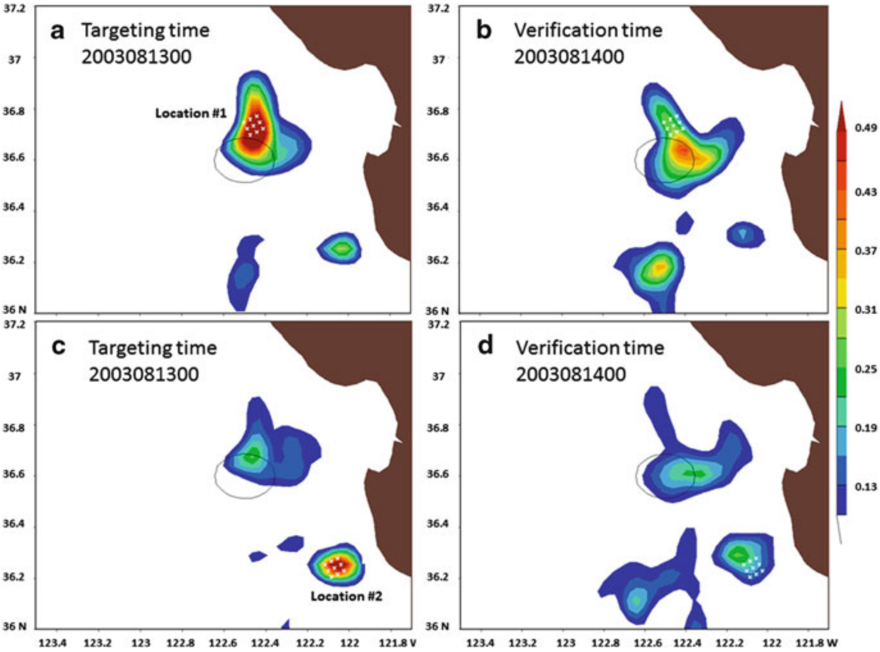


Fig. 16.6 Signal variance for nine observations centered at 36.7°N, 122.5° W for (**a**) targeting time 00 UTC 13 Aug 2003 and (**b**) verification time 00 UTC 14 Aug 2003. Signal variance for nine observations centered at 36.2° N, 122.1° W for (**c**) targeting time same as (**a**) and verification time same as (**b**). The black ellipse contour indicates verification region

the targeting time. It shows larger signal variances at the location #1 (Fig. 16.6a) compared to the location #2 (Fig. 16.6c) due to larger ensemble spread at the targeting time. The signal variance at the verification time has larger values within the verification region from the location #1 (Fig. 16.6b) compared to the location #2 (Fig. 16.6d). This suggests the first location for the deployment is more likely to improve the forecast than the second location.

Figure 16.7a depicts the predicted reduction in forecast error variance at the verification time due to a surface temperature observation at the targeting time at the location indicated by the white cross. By integrating this field across the verification region we obtain a prediction of the reduction in forecast error variance due to an observation at the white cross. Figure 16.7b plots the mean reduction in forecast error variance as a function of the location of the test observation. We refer to maps like Fig. 16.7b as a “summary map”.

If gliders are available for adaptive sampling, summary bar charts can be used to choose among several feasible glider paths. At a particular location, a glider needs to be directed which direction it will be towards to. To demonstrate how signal variance summary bar chart can be used, assuming that for a particular location, a glider can have eight possible tracks (red lines in Fig. 16.7b). The predicted

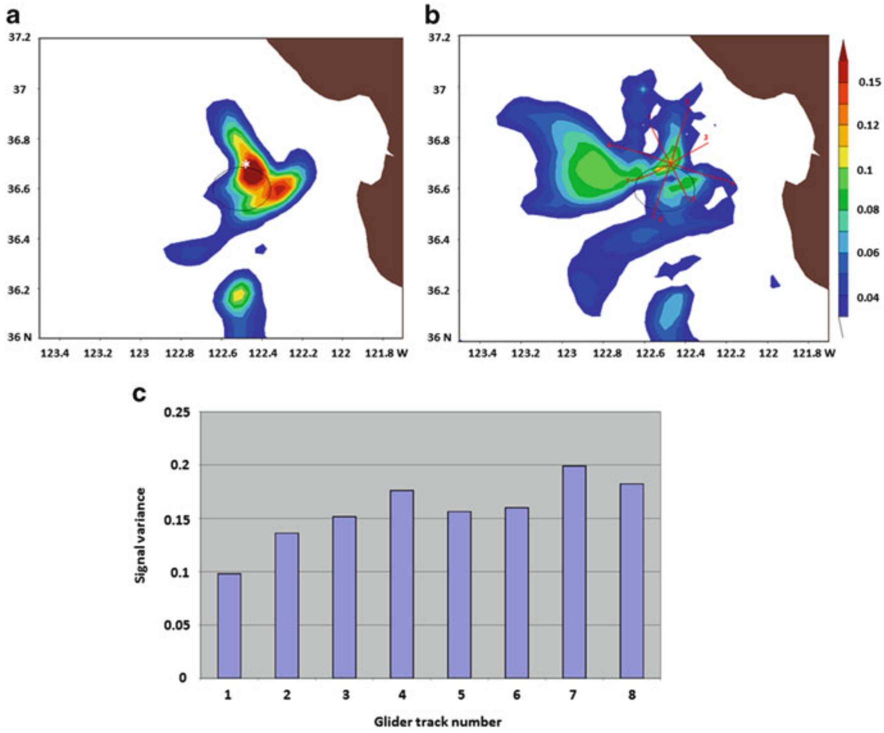


Fig. 16.7 (**a**) Signal variance at the verification time for the feasible deployment of adaptive observations indicated by the white crosses, (**b**) Summary map of average signal variance over the verification area at the verification time as a function of a single temperature observation, (**c**) Bar chart of signal variance for eight glider tracks displayed in (**b**)

reduction in forecast error variance within the verification region at the verification time as a function of each of the eight possible glider paths is plotted as a bar chart (Fig. 16.7c). Each bar gives the ETKF prediction of the reduction of forecast error variance within the verification region to be associated with a particular glide track. Given knowledge about where a glider is at the beginning of the targeting time, these bar charts can be used to direct the glider along the path predicted to have the maximum impact on the forecast error reduction. Thus, the signal variance given on the bar chart suggests that track seven is the best of these eight glider deployments.

During the AOSN II field campaign, up to 15 different gliders are crisscrossing the Monterey Bay at any given time. For example, thirteen gliders are deployed on Aug 13, 2003 during a 24 h observation time window and each takes the path indicated in Fig. 16.8a. As a test of a target technique, it is of a great interest to see which of these 13 glider paths would have been the best choice if one were only going to assimilate observations from just one of the 13 gliders. Figure 16.8b gives the ETKF predicted reduction in forecast error variance in the verification region at

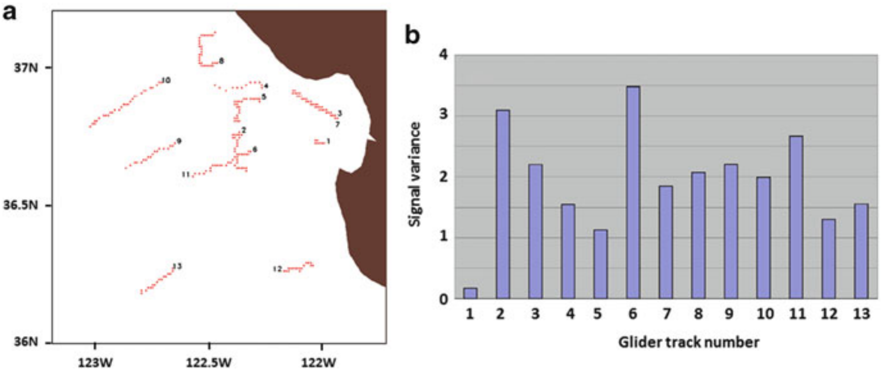


Fig. 16.8 (**a**) Glider tracks on Aug 13, 2003. (**b**) Signal variance summary bar chart for 13 glider tracks shown in (**a**)

the verification time as a function of the glider track. It shows that, according to our implementation of the ETKF, path 6 would have been the best followed by path 2 and 11.

16.6 Summary

The purpose of this Chapter is to illustrate the development and present preliminary results of the ETKF ocean adaptive sampling system that incorporates three distinctive techniques: (1) a time-shifting technique that enables an ensemble of very high resolution atmospheric forecasts to be generated from a single high resolution ensemble member, (2) an ET ensemble generation technique for the generation of ocean ensemble, and (3) an ETKF technique for ocean adaptive sampling. The system is applied to the Monterey Bay area during the AOSN II field campaign in the month of August 2003.

The atmospheric forcing from COAMPS AOSN II forecast is shifted smoothly in time to transfer a single deterministic forecast to an ensemble for ocean ensemble forecast. The shifted atmospheric forcing fields are able to preserve the important aspects of the atmospheric features so that each ocean ensemble member is forced with an approximation to a realization of the true atmospheric state given previous observations.

The NCOM ensemble mean is found to be able to give a better representation of the upwelling features than the single deterministic run during the upwelling period. Two upwelling centers are found. One is near the coast of Point Ano Nuevo and the other near Point Sur. The ensemble mean is also found to be closer to the features in the satellite observations than the ones in the control forecast. Furthermore, the ensemble mean is closer to the observed cold water seaward movement and transport across the mouth of the Monterey Bay during an earlier and later time of the upwelling period, respectively. The ensemble spread is found to be maximized near the upwelled cold water transport across the mouth of the Monterey Bay.

An ocean adaptive sampling system derived from the ETKF technique is illustrated using the data collected during the AOSN II field campaign. For a large number of possible adaptive observations, a signal variance summary map provides an overview of the predicted reduction in forecast error variance within the verification region as a function of the location of a plausible future observation. The predicted reduction in forecast error variance for a large number of possible glider tracks is summarized and displayed in a bar chart for each feasible deployment. The real glider tracks from the AOSN II field campaign are used to derive a signal variance bar chart with 13 possible glider deployments. The ETKF adaptive sampling distinguishes one path with a large summarized signal variance near the verification area. The use of this path, in our view, would have been most likely to reduce the forecast error within the verification region.

As discussed in Majumdar et al. (2002), the quantitative assessments of the accuracy of ETKF signal variance predictions require a large number of events. Unfortunately, the limited events during the AOSN II do not provide enough cases for such quantitative assessments to be made. Nevertheless, the aforementioned experiment indicates that the adaptive sampling locations selected using the technique presented here are, at the very least, consistent with the group velocity of wave packets of ocean forecast errors that are unlikely to propagate very far over a 24 h period in the ocean. For the future work, we hope to use a large number of cases to quantitatively measure the accuracy of the ETKF prediction of forecast error variance reduction in the ocean prediction.

Acknowledgements The support of the sponsors, the Office of Naval Research, Ocean Modeling and Prediction Program, through program element N0001405WX20669 is gratefully acknowledged. Computations were performed on zornig, which is a SGI ORIGIN 3800 with IRIX 6.5 OS and 512 R12000 400 MHz PEs and is located at the U.S. Army Research Laboratory (ARL) DoD Supercomputing Resource Center (DSRC), Aberdeen Proving ground, MD.

References

AOSN: Autonomous ocean sampling network (2003) http://www.mbari.org/aosn/Accessed 30 May 2012

Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. J Atmos Sci 56:1748–1765

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter part I: Theoretical aspects. Mon Wea Rev 129:420–435

Bishop CH, Etherton BJ, Majumdar SJ (2006) Verification region selection in adaptive sampling. Quart J Roy Met Soc 132:915–933

Bishop CH, Holt T, Nachamkin J, Chen S, McLay JG, Doyle JD, Thompson WT (2009) Regional ensemble forecasts using the ensemble transform technique. Mon Wea Rev 137:288–298

Buizza R, Richardson DS, Palmer TN (2003) Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man’s ensembles. Quart J Roy Meteor Soc 129:1269–1288

Cummings JA (2005) Operational multivariate ocean assimilation. Quart J Roy Meteorol Soc 131:3583–3604

Doyle JD, Jiang Q, Chao Y, Farrara J (2008) High resolution atmospheric modeling over the Monterey Bay during AOSN II. Deep Sea Res. doi:10.1016/j.dsr2.2008.08.009

Hoffman RN, Liu Z, Louis J-F, Grassotti C (1995) Distortion representation of forecast errors. Mon Wea Rev 123:2758–2770

Holt RT, Cummings JA, Bishop CH, Doyle JD, Hong X, Chen S, Jin Y (2011) Development and testing of a coupled ocean-atmosphere mesoscale ensemble prediction system. Ocean Dyn 61:1937–1954. doi:10.1007/s10236-011-0449-9

Hong X, Hodur RM, Martin PJ (2007) Numerical simulation of deep-water convection in the Gulf of Lion. Pure Appl Geophys 164:2101–2116

Hong X, Cummings JA, Martin PJ, Doyle JD (2009a) Ocean data assimilation: a coastal application. In: Parks S, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications. Springer, Berlin/Heidelberg, pp 269–292. doi:10.1007/978-3-540-71056-1_14

Hong X, Martin PJ, Wang S, Rowley C (2009b) High SST variability south of Martha's Vineyard: observation and modeling study. J. Mar Syst 78:59–76

Hong X, Bishop CH, Holt T, O'Neill L (2011) Impacts of sea surface temperature uncertainty on the western north Pacific subtropical high (WNPSH) and rainfall. Weather Forecast 26:371–387

Kondo J (1975) Air-sea bulk transfer coefficients in diabatic conditions. Boundary-Layer Met 9:91–112

Leonard N, Robinson A (2003) Adaptive sampling and forecasting plan. http://www.princeton.edu/∼dcsl/aosn/. Accessed 25 May 2012

Majumdar SJ, Bishop CH, Etherton BJ, Toth Z (2002) Adaptive sampling with the ensemble transform Kalman filter part II: Field program implementation. Mon Wea Rev 130:1356–1369

Majumdar SJ, Sellwood KJ, Hodyss D, Toth Z, Song Y (2010) Characteristics of target areas selected by the ensemble transform Kalman filter for medium-range forecasts of high-impact winter weather. Mon Wea Rev 138:2803–2824

Majumdar SJ, Chen S-G, Wu C-C (2011) Characteristics of ensemble transform Kalman filter adaptive sampling guidance for tropical cyclones. Quart J Roy Meteorol Soc 137:503–520

Martin PJ (2000) Description of the navy coastal ocean model version 1.0. Naval Research Laboratory, NRL/FR/7322—00-9962, pp 1–42

Martin PJ, Hodur RM (2003) Mean COAMPS air-sea fluxes over the mediterranean during 1999 report. Naval Research Laboratory, Stennis Space Center, Mississippi

Sellwood KJ, Majumdar SJ, Mapes BE, Szunyogh I (2008) Predicting the influence of observations on mediumrange forecasts of atmospheric flow. Quart J Roy Meteor Soc 134:2011–2027

Szunyogh I, Toth Z, Morss RE, Majumdar S, Etherton BJ, Bishop CH (2000) The effect of targeted dropsonde observations during the 1999 winter storm reconnaissance program. Mon Wea Rev 128:3520–3537

Toth Z, Kalnay E (1993) Ensemble forecasting at NMC: the generation of perturbations. Bull Am Meteor Soc 74:2317–2330

Toth Z, Kalnay E (1997) Ensemble forecasting at NCEP and the breeding method. Mon Wea Rev 125:3297–3319

Chapter 17
Climate Change and Its Impacts on Streamflow: WRF and SCE-Optimized SWAT Models

Shie-Yui Liong, Srivatsan V. Raghavan, and Minh Tue Vu

Abstract It has been noted that global warming is likely to increase both the frequency and severity of weather events such as heat waves and heavy rainfall. These could lead to large scale effects such as melting of large ice sheets with major impacts on low-lying regions throughout the world (Intergovernmental Panel on Climate Change, IPCC 2007a). Since these projected climate changes will impact water resources, agriculture, bio-diversity and health, one of the key challenges of climate research is the application of climate models to quantify both future climate change and its impacts on the physical and biological environment. One of the widely studied impacts is on hydrology, right from large scale river basins, river deltas through to small scale urban reservoirs. In this context, this chapter discusses some hydrological impact studies and presents results of a study done over the Sesan catchment in Lower Mekong Basin (in Southeast Asia). Sensitivity analysis and an optimization calibration scheme, SCE-UA algorithm, are applied to the SWAT model.

17.1 Introduction

17.1.1 General Introduction

The Intergovernmental Panel on Climate Change (IPCC) has mentioned in its Fourth Assessment Report (AR4) that "physical and biological systems on all continents and in most oceans are already being affected by recent climate changes and climatic effects on human systems, although more difficult to discern due to adaptation and

S.-Y. Liong (✉) · S.V. Raghavan · M.T. Vu
Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore, 119227, Singapore
e-mail: tmslsy@nus.edu.sg

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_17,

non-climatic drivers, are emerging" (IPCC 2007a). Since these climate changes are likely to alter global surface temperatures, precipitation patterns, sea levels, extreme events and other aspects of climate on which the natural environment and human systems depend, substantial impacts are expected on, for example, water resources, bio-diversity, agriculture and human health. Hence, adaptation to climate change has become globally very important and especially is of great concern to the developing countries. Though the impacts of climate change are expected to affect all natural systems, water resources are likely to be impacted deeply. Current vulnerabilities to climate are strongly correlated with climate variability, in particular precipitation variability. Increased changes in precipitation intensity and variability are projected to increase the risks of flooding and drought in many areas. While temperatures are expected to increase everywhere over land and during all seasons of the year, although at different increments, precipitation is expected to increase globally and in many river basins, but to decrease in many others (IPCC 2007a). However, quantitative projections of changes in precipitation, river flows and water levels at the river-basin scale remain uncertain (IPCC 2001).

The IPCC also notes that semi-arid and arid areas are particularly exposed to the impacts of climate change on freshwater and many of these areas such as the Mediterranean basin, western USA, southern Africa and north-eastern Brazil are likely to suffer a decrease in water resources due to climate change (IPCC 2007b). The IPCC also reports that current water management practices are insufficient to counter the possible negative impacts of climate change on water supply, flood risk, health, energy and aquatic ecosystems. For the development of adaptation policies, realistic projections of climate change and its impacts on water resources are needed. Global Climate Models, also known as General Circulation Models or GCMs are primary tools for prediction of global climate. Precipitation, a principal input signal to water systems, is not reliably simulated in these global climate models due to their coarse resolutions (IPCC 2007b). GCM projections are subject to substantial uncertainties in the modelling processes so that climate projections are not easy to be incorporated into hydrological impact studies (Mearns et al. 2001; Allen and Ingram 2002; Forest et al. 2002). It has been noticed that such uncertainties have produced biases in the simulation of river flows when using direct GCM outputs for hydrological impact studies. Some studies have found that uncertainties in climate change impacts on water resources are primarily due to the uncertainty in precipitation inputs and less due to the uncertainties in greenhouse gas emissions, in climate sensitivities or in hydrological models themselves (IPCC 2007b). Most climate change impact studies consider only changes in precipitation and temperature, based on changes in the averages of long-term monthly values. A major problem in the use of GCM outputs for impact studies is the mismatch of spatial grid scales between GCMs (typically a few hundred kilometers) and the hydrological processes. Water is managed at the catchment scale and adaptation is local, while GCMs work on large spatial grids. Generally, precipitation projections are less consistent than those of temperature due to its high variability in space and time, with large inter-model ranges for seasonal mean rainfall responses. These inconsistencies are explained partly by the inability of GCMs to reproduce the mechanisms responsible

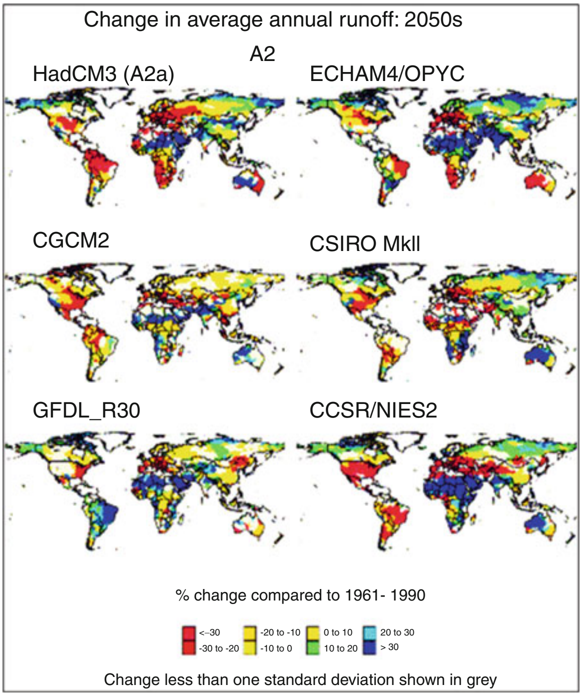

Change in average annual runoff: 2050s
A2
HadCM3 (A2a)
ECHAM4/OPYC
CGCM2
CSIRO MkII
GFDL_R30
CCSR/NIES2
% change compared to 1961- 1990
<-30
-30 to -20
-20 to -10
-10 to 0
0 to 10
10 to 20
20 to 30
> 30
Change less than one standard deviation shown in grey

Fig. 17.1 Changes in average annual runoff for 2050 using A2 IPCC Emission scenario shown by different GCMs. Percentage change compared to 1961–1990. (GCMs HadCM3, ECHAM4, CGCM2, CSIRO, GFDL and CCSR/NIES) (Adopted from Arnell (2004))

for precipitation such as the convection processes and the hydrological cycle or to account for orography (IPCC 2007b). With uncertainties in such climate projections, impacts studies are difficult.

17.1.2 Hydrological Studies Based on Global Climate Projections

Arnell (2004) conducted a study of the future climate change impact on water resources by applying GCM outputs for estimating river flows under both present and future climates. The results of this study are shown in Fig. 17.1 which provides an indication of the effects of future climate change on long-term average annual river runoff by the 2050s across the world, under the IPCC A2 emission scenario, estimated by different climate models. It was reported that climate change is likely to increase water resources stresses in some parts of the world where runoff decreases, including around the Mediterranean, in parts of Europe, central and southern America, and southern Africa. In other water-stressed parts of the world, particularly in southern and eastern Asia, climate change was likely to increase runoff. It was also reported by the author that there were differences in the magnitude and direction of climate change over some parts of the world, including Asia. It was seen that even for large river basins, climate change scenarios from different climate models resulted in very different projections of future runoff change, such as in Australia, South America and Southern Africa. The study recognized the uncertainties that exist amongst the climate projections of various GCMs. Although this study highlighted uncertainties on a global scale, impact studies over regions like Africa are even more difficult due to lack of sufficient technological resources and under-developed scientific research compared to many other parts of the world (Washington et al. 2006).

Some hydrological research studies over the Okavango River basin and Okavango Delta have been made. The most comprehensive study was conducted within the EU WERRD project (Water and Ecosystem Resources in Regional Development). The general objective of this project was to increase understanding of livelihoods, the environment and policies relating to international river basins. In this case, the project refers to the Okavango River Basin and was being designed by many researchers from Botswana, UK, Namibia, South Africa and Sweden, funded by the European Union. As a part of this project, Andersson et al. (2006) applied scenario modelling as a tool for integrated water resource management over the Okavango River basin. The Pitman hydrological model (Pitman 1973) was used to assess the impact of various climate change scenarios on downstream river flow.

Pitman model of the river basin was applied to both present day historical conditions and future climate change scenarios to assess the impact on river flows. Four GCMs (HadCM3, CCSR/NIES, CCCma and GFDL) with present day conditions and future A2 IPCC emission scenario were applied in the study. Their results showed that there was considerable uncertainty about the magnitude and direction of any future discharge response associated with both the GCM and the IPCC emission scenarios. Results of the study showed that the modeled experiments indicated a reduction in future flow after about 2050 for both the A2 and B2 GHG scenarios that increases over time. This is seen in Fig. 17.2 which shows the mean monthly flow at a particular station (Mukwe) in the Okavango River basin that was simulated by the hydrological model used in this study. The key conclusion from the study was that different GCMs predicted future conditions in the Okavango Basin ranging from drier than present to wetter than present and there are differences in both degree of change and direction of change between the Okavango river catchment area and the Okavango Delta.

In a related hydrological modelling study of the Okavango Delta, Murray-Hudson et al. (2006) applied a mathematical model to assess the impacts of changing hydrological inputs on the flooding in the delta. The assessment of effects of possible future changes (2020–2050) on the hydrological characteristics of the Okavango Delta was done by running a hydrological model of the Okavango Delta with discharge inputs from the Pitman model of the river basin. Three different GCMs (HadCM3, CCC and GFDL) with the future A2 IPCC emission scenario were used to drive the hydrological model. The GCMs produced a wide spectrum of possible future conditions in the Delta as shown in Fig. 17.3. The authors concluded that there was a large uncertainty about future climatic conditions and the modeled effects of climatic variation on the hydrology of the Delta. Figure 17.3 shows the different flood plain classes categorized according to Permanent Flooded (PF),

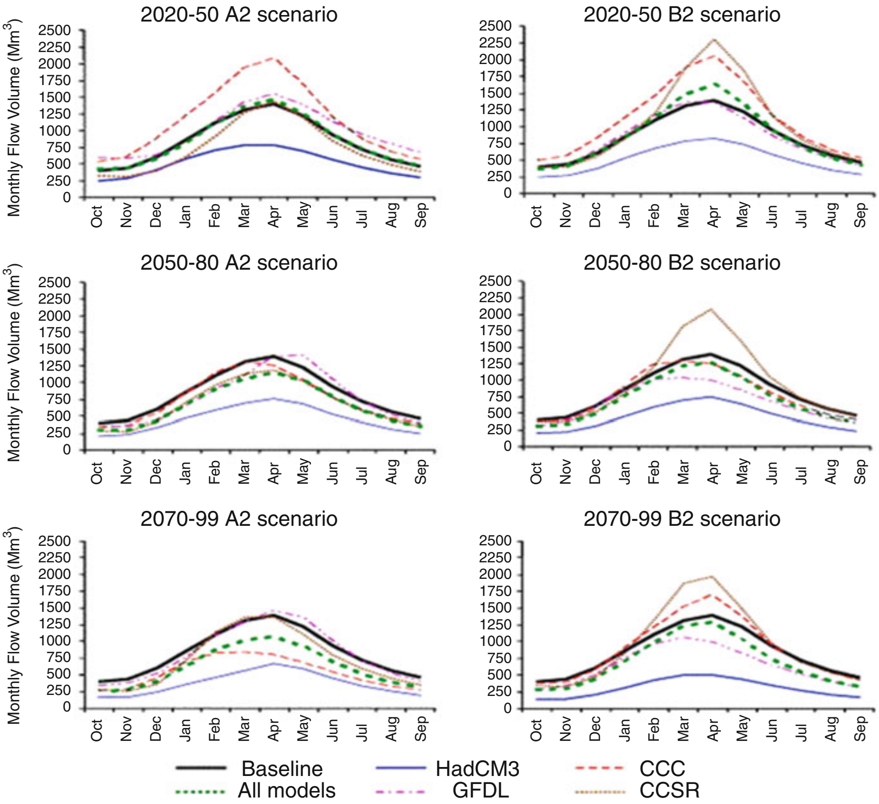


Fig. 17.2 Mean monthly flow at Mukwe with baseline simulations and with assessment of changes of precipitation and evaporation derived from various GCMs, driven by the A2 and B2 greenhouse gas emission scenarios (Adopted from Andersson et al. (2006))

Regularly Flooded (RF), Occasionally Flooded (OF), High Flooded Only (HFO) and Dry Land (DL). The PF floodplain was further classified into Proper and Fringe and hence denoted as PF1 and PF2 respectively and so is the case with RF—Annual (RF1) and Biennial (RF2).

The future change simulated by the hydrological model using these GCMs is shown as either ‘wetter’ or ‘drier’, which makes adaptation policy difficult. These studies have also recognized the uncertainties in current GCM climate change projections and the need to reduce these uncertainties followed by the need to develop appropriate adaptation strategies for the use of water resources over the Okavango River basin region.

It has been noted that changes in future precipitation may be more adequately specified on the sub-basin scale by downscaling the coarse GCM data using Regional Climate Models (RCMs) allowing for more detailed assessments of spatial heterogeneities in climate change impacts on water resources since these are limited area models run at a higher resolution compared to GCMs (Andersson et al. 2006).

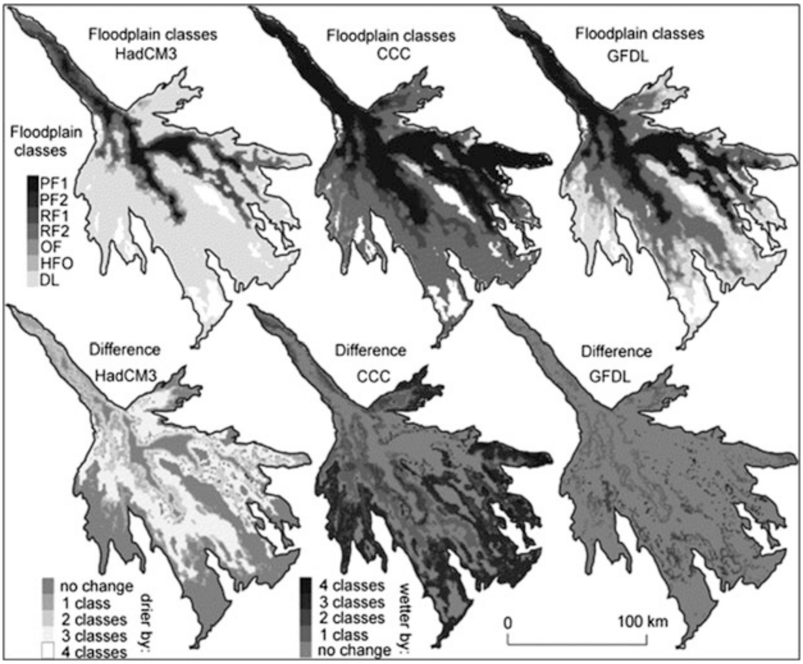


Fig. 17.3 Effects of change in hydrological inputs on the Okavango Delta as obtained from various climate models (HadCM3, CCC and GFDL) under A2 greenhouse gases scenario for 2020–2050 period (Adopted from Murray-Hudson et al. (2006))

The IPCC reports that during recent years several studies have focused on diverse applications of RCMs for impact studies which include downscaling from the climate model scale to the catchment scale, using regional climate models to create scenarios to drive hydrological models and quantifying the effect of hydrological model uncertainty on estimated impacts of climate change (IPCC 2007b).

17.1.3 Dynamically Downscaled Climate Model Input for Hydrological Studies

Although the GCMs provide reasonable simulation accuracy of climate in a global, hemispheric or a continental scale, at a regional/sub-regional scale representation, the simulation accuracies are poor due to their coarse spatial resolution (Giorgi 1990, 1996a, b).

Regional climate is often affected by forcings and circulations that occur at a sub-grid scale of the GCM. Some of the regional and local scale climate forcings due to land-use characteristics, complex topography, land-ocean contrasts, aerosols, radiatively active gases, snow, sea ice and ocean currents are not resolved well by

GCMs. It is therefore obvious that the GCMs cannot explicitly capture the fine scale structure that characterizes climate variables in many regions of the world that is required to run impact models. This becomes particularly problematic for important climate variables like precipitation that have high variability in space and time. Due to their coarse spatial resolutions and their inability to include mesoscale atmospheric features in their large scale circulation, the GCMs do not simulate the precipitation fields with adequate fine scale details to be applied to impact models such as hydrological models. Hence, before the GCM output information of certain key variables like rainfall can be used to drive the impact models at a regional or a local scale, there is an intermediate step which requires the 'downscaling' of this large scale GCM information to regional scale information. One of the techniques that is employed to this end is what is called as regional climate modelling that uses a high resolution limited area climate model, also called Regional Climate Model (RCM), for climate simulations. Later, the output from this RCM, usually precipitation and temperature, are used as input for hydrological models.

This chapter describes a study that assesses hydrological responses over a couple of river basins in one of the most climate vulnerable regions, the Mekong River basin in Southeast Asia. This study uses the output of a regional climate model as the input to the popular hydrological model Soil Water Assessment Tool (SWAT). Sensitivity analysis has been carried out to sort out the most affected parameters to modelling scheme and observed discharge. Shuffled Complex Evolution (SCE) algorithm is employed to calibrate SWAT model's most sensitive parameters. Minimum discrepancy between the observed and the simulated runoff results in an optimal set of values for the calibration parameters.

The following section describes the study region, the Sesan catchment, which is a small part of the Lower Mekong Basin. The hydrological model and the regional climate model that were used for this study and the modelling approaches are also described.

17.2 Study Catchment and Hydrological Model

17.2.1 Sesan Catchment

The catchment considered in this study is the Sesan catchment, lying over the central highland region of Vietnam between 107°19′E and 108°38′E and 13°33′N and 15°15′N. Together with SrePok and SeKong rivers, the catchment forms the main river system of Sub-Area 7 (SA7) of Lower Mekong Basin. Detailed information about the SA7 region is available from the Mekong River Commission website at: http://www.laoiwrm.com/BDPatlas/BDPatlas_6-10(finalune2006)/SA7/SA7_Main.htm

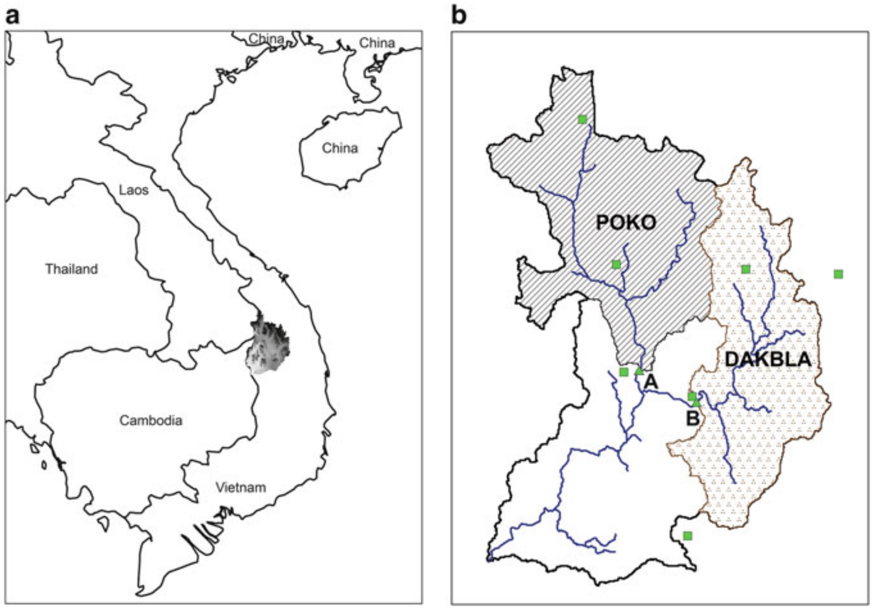


Fig. 17.4 Sesan catchment in Vietnam

The Sesan river basin has a catchment area of 9,030 km^2 which consists of two sub-catchments, called Poko and Dakbla. Each of these sub-catchments has a discharge station located closed to the catchment outlet (Fig. 17.4). There are six rainfall stations inside and outside the study catchment with three stations for each sub-basin. The stations marked ‘A’ and ‘B’ as small triangles indicate the discharge stations and the rest of the six square markers indicate the six rainfall stations. The shaded region in the diagram in the left is magnified for clarity in the right which shows the two sub-basins considered in this study.

17.2.2 SWAT Model

The SWAT model is a river basin scale model, developed at the United States Department of Agriculture (USDA)—Agriculture Research Service (ARS) in the early 1990s (Arnold et al. 1998). SWAT is a physically based model which is designated to work for large river basins and catchments and uses readily available inputs. It is widely used to study the impacts of land management practices on water, sediment and agriculture chemical yields with varying soil, land use and water management conditions. A suite of information about the model and its varied applications are available at the SWAT website: http://swatmodel.tamu.edu/.

17.2.2.1 Model Setup

In this study, the main objective was to quantify the impacts of climate change to hydrological stream flow over a long period of time. SWAT was chosen as the hydrological model to which the precipitation output of a regional climate model was applied as input to simulate stream flow. The input for ArcSWAT included a spatial reference map such as the DEM (Digital Elevation Model) with a resolution of 250 m * 250 m, a land use map and a soil map (converted to raster format with the same resolution) and meteorological data (precipitation and temperature time-series of all stations in daily format). The DEM was obtained from Department of Survey and Mapping (DSM) of Vietnam. The land use map was obtained from the Forest Investigation and Planning Institute (FIPI), of Vietnam for the year 2000. The soil map was obtained through the Ministry of Agriculture and Rural Development (MARD) of Vietnam categorized by the FAO (Food and Agriculture Organization).

17.2.2.2 Model Sensitivity Analysis

The sensitivity analysis is a method that analyzes the sensitivity of model parameters to the model performance. This method entails to filter the model parameters that either have or have not any significant influence on the model results. On the other hand, it also aims to reduce the number of parameters required in fitting to a model input-output. Traditional methods of sensitivity analysis have been classified by Saltelli et al. 2000. They are (1) Local method (Melching and Yoon 1996) (2) Integration of local to global method using Random One-Factor-At-a-Time (OAT) proposed by Morris (1991) and (3) Global methods like Monte Carlo and Latin-Hypercube (LH) simulation (McKay et al. 1979; McKay 1988). By studying the advantages and disadvantages of each of the above methods, van Griensven and Meixner (2006) developed the LH-OAT method which performs LH sampling followed by OAT sampling. This method samples the full range of all parameters using LH design along with the precision of OAT sampling to ensure that the changes in each model output could be attributed to the changed parameter. The LH-OAT design has been coupled to the SWAT model for sensitivity analysis module. Model parameters are analysed based on the performance of its output compared against observed data and the model itself. In the SWAT model, there are 26 parameters sensitive to water flow, 6 parameters sensitive to sediment transport and other 9 parameters sensitive to water quality. In this study, since the stream flow is the main focus, 10 most sensitive parameters out of the available 26 options are analysed (Table 17.1).

17.2.2.3 Auto-Calibration by ParaSol Method (Parameter Solution) Using SCE-UA Algorithm

SWAT model has the options to choose either manual or auto-calibration. Calibration is applied to those most sensitive parameters to yield the optimal set of values

Table 17.1 Sensitivity analysis ranking of 10 most sensitive parameters in SWAT model to stream flow

Sensitivity analysis order	Parameter	Description	Parameter range
1	Cn2	Moisture condition II curve no	35 ∼ 98
2	Alpha_Bf	Baseflow recession constant	0 ∼ 1
3	Ch_K2	Effective hydraulic conductivity in main channel	−0.01 ∼ 500
4	Surlag	Surface runoff lag coefficient	1 ∼ 24
5	Ch_N2	Manning n value for the main channel	−0.01 ∼ 0.3
6	Blai	Maximum potential leaf area index for land cover	0 ∼ 8
7	Sol_Awc	Available water capacity	0 ∼ 1
8	Esco	Soil evaporation compensation factor	0 ∼ 1
9	Canmx	Maximum canopy storage	0 ∼ 100
10	Gwqmn	Threshold water level in shallow aquifer for base flow	0 ∼ 5,000

for the model parameters which results in the minimum discrepancy between the observed and the simulated discharge data. While manual calibration can be used by trained, experienced users who are familiar with the model and the catchment under consideration, auto-calibration is recommended especially for the new user. Parameter Solution method (ParaSol) is a built-in auto-calibration model since the SWAT 2005 version was implemented (van Griensven and Meixner 2004). ParaSol operates by a parameter search method for model parameter optimization followed by a statistical method that was performed during the optimization to provide parameter uncertainty bounds and the corresponding uncertainty bounds on the model outputs. The ParaSol method aggregates objective functions (OF) into a global optimization criterion (GOC), minimizes these OF's or a GOC using the Shuffled Complex Evolution Method (SCE-UA) algorithm and performs uncertainty analysis with a choice between two statistical concepts. The SCE-UA (Duan et al. 1992) method is based on a synthesis of all the best functions from many other existing methods consisting of the Genetic Algorithm (GA), simplex method (Nelder and Mead 1965), controlled random search (Price 1987), competitive evolution (Holland 1975) and the newly developed concept of complex shuffling. SCE-UA conducts a global minimization of a single function for up to 16 parameters. This method is also capable for non-linear optimization problems.

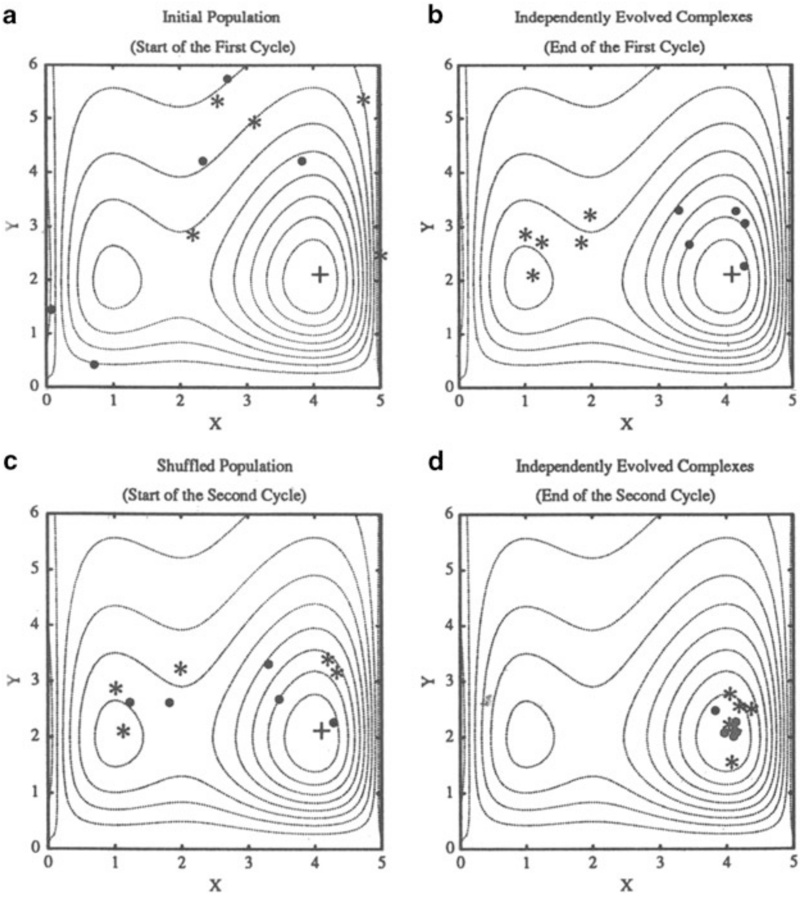


Fig. 17.5 Illustration of the SCE-UA method (Adopted from Duan et al. 1994)

In SCE-UA, the initial set of parameters (first step) is chosen randomly throughout the feasible parameters space for p parameters to be optimized. Then the set is partitioned to several "complexes" that have 2p + 1 points in which each complex evolves independently using the simplex algorithm. The complexes are then shuffled to form new complexes in order to share information between the complexes. SCE-UA method is briefly illustrated in Fig. 17.5. The processes of competitive evolution and complex shuffling introduced in SCE-UA ensure that the information contained in the sample is efficiently and thoroughly exploited. Overall, SCE-UA appears to be capable of efficiently and effectively identifying the optimal values for the model parameters (Duan et al. 1992). SCE-UA has been used widely in watershed model calibration and other areas like soil erosion, subsurface hydrology, land surface modelling (Duan et al. 2003).

The model is based on a synthesis of all the best functions from many other existing methods consisting of the Genetic Algorithm (GA), simplex method and controlled random search of Nelder and Mead (1965). This method is also capable for non-linear optimization problems. This method has been found to be robust, effective and efficient (Duan et al. 2003). There are two objective functions which can be used in the model calibration using SCE-UA. They are (1) the sum of the squares of the residuals (SSQ) and (2) the sum of the squares of the difference of the measured and simulated values after ranking (SQQR). In this study the SSQ objective function is used. The SSQ, used to target at matching the simulated with the observed data, is expressed as follows:

$$SSQ = \sum_{i=1,n} \left[TF\left(x_{i,obs}\right) - TF\left(x_{i,sim}\right)\right]^2 \tag{17.1}$$

where, n is the number of pairs of observed and simulated variable and ‘TF’ is a user defined transformation function. Detailed description of ParaSol method can be found in van Griensven and Meixner (2004).

17.2.2.4 Model Results

Using the above methodology, the SWAT model performance is established. A period of five years from 2000 to 2005 was used to calibrate the SWAT model for stream flow with the first year 2000 as a warm up period. The validation for the model was done for another period 1996–2000 to check if the calibrated model holds good for a different period. Station discharge data from the station Kon Tum (on the Dakbla river) and the station Trung Nghia (on the Poko river) were used to calibrate the model separately. The precipitation data on a daily scale were used from 1996–2005 from the stations inside and outside catchment (Fig. 17.4) for the calibration and validation processes of the SWAT model. The coefficient of determination (R^2) and the Nash-Sutcliffe Index (NE) were used as statistical indices to assess the goodness of fit of the model.

The NE index is defined by:

$$NE = 1 - \frac{\sum_{i=1}^{N} (O_i - S_i)^2}{\sum_{i=1}^{N} \left(O_i - \overline{O}\right)^2} \tag{17.2}$$

and the coefficient of determination R^2 is defined by

Table 17.2 Statistical Indices of model calibration and validation: R^2 and NE

	Calibration				Validation			
River	Daily		Monthly		Daily		Monthly	
Basin	R^2	NE	R^2	NE	R^2	NE	R^2	NE
Dakbla	0.71	0.68	0.94	0.86	0.47	0.46	0.59	0.58
Poko	0.83	0.82	0.90	0.89	0.56	0.55	0.68	0.67

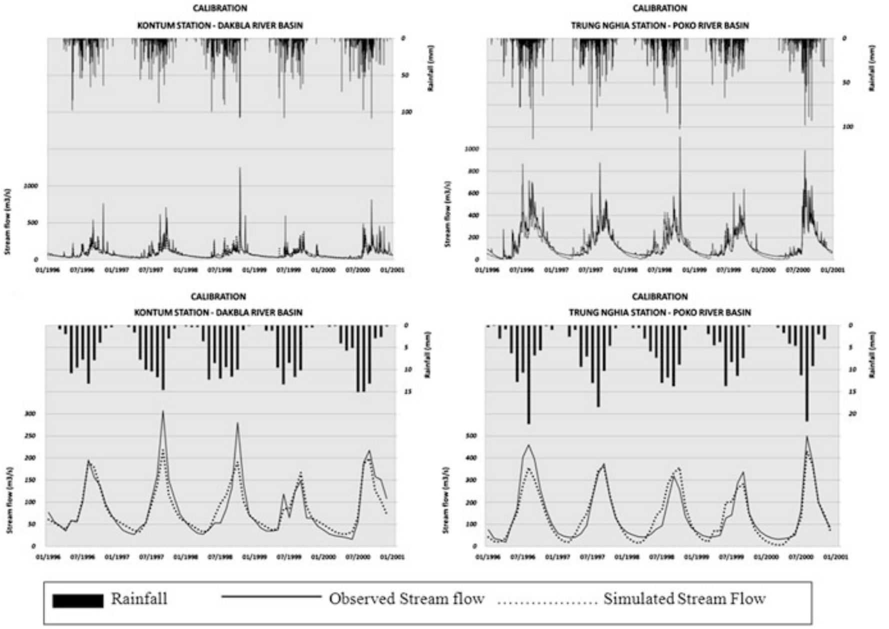


Fig. 17.6 Calibration of the model: Dakbla (*left*) and Poko (*right*) river basins on the daily (*top*) and monthly (*bottom*) scales

$$R^2 = \left\{ \frac{N \sum_{i=1}^{N} S_i O_i - \sum_{i=1}^{N} S_i \sum_{i=1}^{N} O_i}{\sqrt{N \sum_{i=1}^{N} S_i^2 - \left(\sum_{i=1}^{N} S_i \right)^2} \sqrt{N \sum_{i=1}^{N} O_i^2 - \left(\sum_{i=1}^{N} O_i \right)^2}} \right\}^2 \tag{17.3}$$

where, O and S are observed and simulated discharge values, respectively.

The statistical indices in Table 17.2 show that the values of validation indices were, as expected, higher on a monthly scale than a daily scale because daily variability is higher than monthly variability. For climate change impacts study, longer temporal information is of the main concern. The monthly statistical indices, for the validation period, showed a value of about 0.6 for both R^2 and NE; this suggested that the model was reasonably well calibrated. These results suggest that the SWAT model showed good performance and hence suitable for climate change

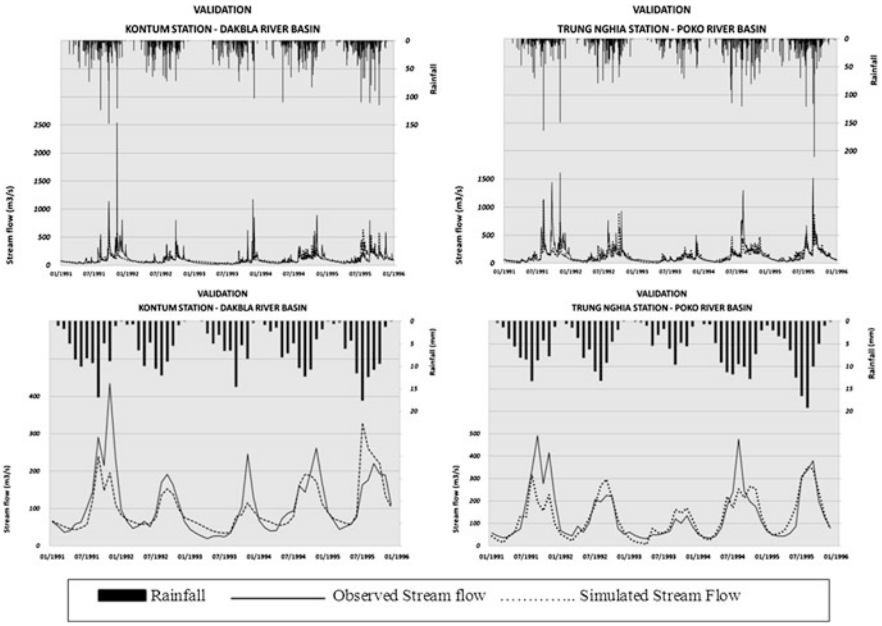


Fig. 17.7 Validation of the model: Dakbla (*left*) and Poko (*right*) river basins on the daily (*top*) and monthly (*bottom*) scales

applications. The calibration and validation results for Dakbla and Poko rivers are shown in Figs. 17.6 and 17.7, respectively.

17.3 RCM Output for Climate Applications

In this study, a regional climate model, Weather Research and Forecasting (WRF), was used to simulate climate over a part of the Lower Mekong Region at a horizontal resolution of 30 km centered over the Sesan catchment area. The RCM WRF model was run for the period 1991–2000 using the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis to assess its performance of the present-day climate. Later, the WRF model was also run driven by the global climate model ECHAM5 T63 over the same period to assess the model's performance on the 10 year climatology of the present-day climate. The rainfall output from the WRF model obtained from the two above mentioned simulations were then used as input to the calibrated SWAT model to simulate the stream flow. The simulated stream flow is shown in Fig. 17.8. Results also indicate the precipitation output derived from the WRF model driven by the GCM ECHAM5 were better than the NCEP/NCAR reanalysis as the stream flow simulated using WRF ECHAM5 output agreed better against the station data than the stream flow

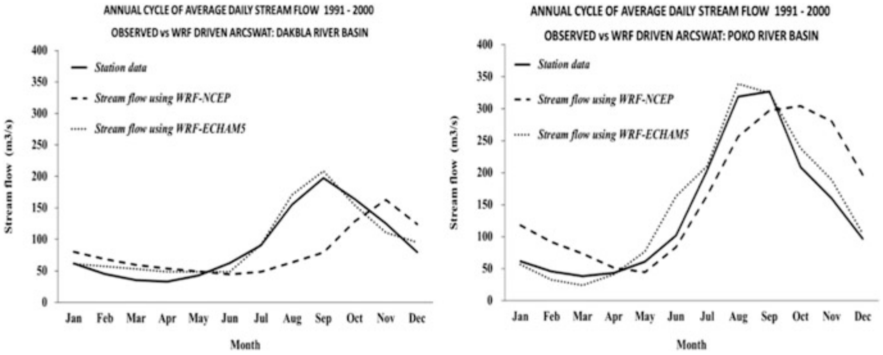


Fig. 17.8 Climatological Annual Cycles of Stream flow (*left*) Kontum observed station—Dakbla river basin (*right*) Trung Nghia observed station—Poko river basin

simulated using WRF NCEP/NCAR reanalysis. This could probably be due to the relatively coarser spatial resolution of the latter (2.5°*3.75°) compared to the former (1.8°*1.8°). The stream flows over the two river basins chosen for study, Poko and Dakbla are shown in Fig. 17.8.

The future climate simulation of the RCM WRF spanned the period 2091–2100 driven by the GCM ECHAM5 T63 under the IPCC SRES A2 emission scenario. This period has been considered in this study as the emission scenario shows clear signal of change towards the end of the century. Just as the simulation for the present-day climate, the RCM WRF derived estimates for future rainfall was used to simulate future stream flow conditions.

For the assessment of future stream flow, the delta factor approach was undertaken where the climate change factor was derived using the estimates of the future climates and present day climates of the WRF-ECHAM5 simulations. This factor is the difference between the future and the present day rainfall estimates. This method is usually practiced by impact modellers as the difference between the future and present day model output cancels the biases in the model output and gives out the clear signal of climate change. Since the best available record of rainfall is the station data, this change factor was added to the station data time series, which in turn represents the changed future conditions of rainfall. The change factor added station rainfall was then used as the input to the SWAT model to simulate future stream flow. Such an application of delta factor for climate change studies have been described by Sushama et al. (2006) and Andersson et al. (2006). A student t-test was done to ensure that the climate signal is true and not a noise. This method entails that a credible estimate of future stream flow could be obtained.

Figure 17.9 shows the stream flow thus derived over the two river basins Poko and Dakbla. For clarity in comparison, the present day stream flow (shown as 'baseline') is overlaid on the future estimated stream flow that used the WRF-ECHAM5 change factor as discussed above. Results show that, both over Dakbla and Poko river basins, the future stream flow is expected to increase, especially during the peak rainfall season. The Dakbla river basin shows a more pronounced

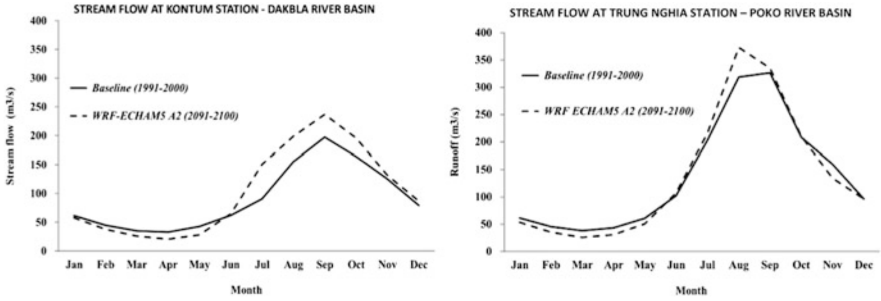


Fig. 17.9 Future stream flow (compared to baseline stream flow) (*left*) Kontum observed station—Dakbla river basin) (*right*) Trung Nghia observed station—Poko river basin

increase in steamflow of about 48 % than Poko river basin that shows about 10 % increase during the peak rainfall season. During the dry season, about 15 % decrease is seen over both Dakbla and Poko basins.

17.4 Conclusions

Existing research studies indicate uncertainties in climate projections stemming from global climate models and different emission scenarios of climate change. Since global climate output have been found insufficient for regional and local impacts, it has been realized that adaptation measures to combat climate change requires high spatial resolution information and hence the use of regional climate models in climate research has become common. Since hydrology is one of the most common impact studies, this chapter highlights the importance of high resolution models in impacts research and the use of sophisticated optimization algorithms when applying hydrological models.

Rainfall derived from climate model has been applied to a hydrological model (SWAT) which was calibrated with SCE-UA algorithm and its simulated discharges were compared with the their observed counterparts. The performance of the model using station data rainfall has been found satisfactory and hence the model derived rainfall were also used to see assess stream flow simulation over the current and future climate. Using the RCM driven by GCM ECHAM5, the present-day and future stream flows were also simulated. Results show that, both over Dakbla and Poko river basins, the future stream flow is expected to increase, especially during the peak rainfall season. The Dakbla river basin shows a more pronounced increase than the Poko river basin.

However, further work is required to improve the confidence in these results. As mentioned before, a higher resolution simulation of the RCM may be required to obtain more credible estimates of future precipitation. In addition, since this result has been obtained only from one run of the RCM, it is recommended to obtain an ensemble estimate of future change in climate by downscaling more GCMs or by using perturbed initial conditions to the RCM to derive multiple estimates of future climate.

References

Allen MR, Ingram WJ (2002) Constraints on future changes in climate and the hydrologic cycle. Nature 419:224–232

Andersson L, Wil J, Todd M, Hughes D, Earl A, Kniveton D, Layberry R, Savenije H (2006) Impact of climate change and development scenarios on flow patterns in the Okavango River. J Hydrol 331(1–2):43–57

Arnell NW (2004) Climate change and global water resources: SRES emissions and socio-economic scenarios. Glob Environ Change 14:31–52

Arnold JG, Srinivasan R, Muttiah RS, Williams JR (1998) Large-area hydrologic modelling and assessment: part I model development. J Am Water Resour Assoc 34(1):73–89

Duan Q, Gupta VK, Sorooshian S (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour Res 28:1015–1031

Duan Q, Sorooshian S, Gupta VK (1994) Optimal use of the SCE-UA global optimization method for calibrating watershed models. J Hydrol 158:265–284

Duan Q, Sorooshian S, Gupta HV, Rousseau HN, Turcotte R (2003) Advances in calibration of watershed models. AGU, Washington, DC

Forest CE, Stone PH, Sokolov AP, Allen MR, Webster MD (2002) Quantifying uncertainties in climate system properties with the use of recent climate observations. Science 295:113–117

Giorgi F (1990) Simulations of regional climate using a limited area model nested in a general circulation model. J Clim 3(9):941–963

Giorgi F, Marinucci MR (1996a) An investigation of the sensitivity of simulated precipitation to model resolution and its implications for climate studies. Mon Wea Rev 124:148–166

Giorgi F, Marinucci MR (1996b) Improvements in the simulation of surface climatology over the European region with a nested modelling system. Geophys Res Lett 23:273–276

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

IPCC (2001) In: Houghton JT et al (eds) Climate change 2001: the scientific basis contribution of working group I to the third assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK/New York, 881 pp

IPCC (2007a) In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge, UK/New York

IPCC (2007b) In: Parry ML, Canziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge, UK/New York

McKay MD (1988) In: Ronen Y (ed) Sensitivity and uncertainty analysis using a statistical sample of input values. CRC Press, Boca Raton, pp 145–186

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2):239–245

Mearns LO et al. (2001) Climate Scenario development, Chap. 13, Climate change-the scientific basis, IPCC TAR, 2001

Melching CS, Yoon CG (1996) Key sources of uncertainty in QUAL2E model of Passaic river. ASCE J Water Resour Plann Manage 122(2):105–113

Morris MD (1991) Factorial sampling plans for preliminary computation experiments. Technometrics 33:161–174

Murray-Hudson M, Wolski P, Ringrose S (2006) Scenarios of the impact of local and upstream changes of climate and water use on hydro-ecology in the Okavango Delta. J Hydrol 331(1–2): 73–84

Nelder JA, Mead R (1965) A simplex method for function minimization. Compt J 7(4):308–313

Pitman WV (1973) A mathematical model for generating monthly river flows from meteorological data in South Africa. Report no 2/73, Hydrological Research Unit, University of Witwatersrand, Johannesburg

Price WL (1987) Global optimization algorithm for a CAD workstation. J Optira Theory Appl 55(1):133–146

Saltelli A, Chan K, Scott EM (eds) (2000) Senstivity analysis. Wiley, New York

Sushama L, Laprise R, Caya D, Frigon A, Slivitzky M (2006) Canadian RCM projected climate-change signal and its sensitivity to model errors. Int J Climatol 26(15):2141–2159

van Griensven A, Meixner T (2004) ParaSol (parameter solutions) PUB-IAHS workshop uncertainty analysis in environmental modelling

van Griensven A, Meixner T (2006) Methods to quantify and identify the sources of uncertainty for river basin water quality models. Water Sci Technol 53(1):51–59

Washington R et al (2006) African climate change: taking the shorter route. Bull Am Meteorol Soc 1355–1366. doi:10.1175/BAMS-87-10-1355

Chapter 18
Entropic Balance Theory and Radar Observation for Prospective Tornado Data Assimilation

Yoshi K. Sasaki, Matthew R. Kumjian, and Bradley M. Isom

Abstract This article reports further theoretical development on the entropic balance theory applied to tornadogenesis (Sasaki 2009, 2010), and the first preliminary application of the theory to radar observations. The entropic balance is a newly found balance, different from the other balance conditions, such as hydrostatic, (quasi-)geostrophic, cyclostrophic, Boussinesq, and anelastic. The entropic balance condition is described as the sole diagnostic Euler-Lagrange equation derived from the Lagrangian of the variational formalism. The entropic balance is most general and involves no additional assumptions other than for the flow with high Reynolds and Rossby numbers estimated as appropriate for supercell storms and tornadoes. The entropic balance theory and the deduced wrap-around mechanism explain well the observations and simulations of tornado, RFD, hook-echo, upward tilting of horizontal vorticity, the vertical in-phase superimposition between upper and lower mesocyclones, and sudden transition from supercell, mesocyclones totornado. In the application, new variables DZ (temporal difference of radar reflectivity) and DZ_{DR} (temporal difference of differential reflectivity) are introduced to compute the entropy anomaly based on the entropic balance theory. The conditions necessary for the transition from supercell to tornado are clarified from the theory and verified from the DZ and DZ_{DR} analyses for a non-tornadic supercell case compared with VORTEX2 tornadic case.

Y.K. Sasaki (✉)
School of Meteorology, University of Oklahoma, Norman, OK 73072, USA
e-mail: yks@ou.edu

M.R. Kumjian
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and National Severe Storms Laboratory, NOAA, Norman, OK, USA

Atmospheric Radar Research Center, University of Oklahoma Norman, OK, USA

B.M. Isom
Atmospheric Radar Research Center, University of Oklahoma Norman, OK, USA

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_18,

Since the entropic balance theory is found to fit well with all analyzed results of tornado and visual observations, it is suggested to use the entropic balance equation as a constraint for variational data assimilation in future development as a challenge.

18.1 Introduction

Tornado data assimilation requires an appropriate dynamical model and observational input data. The dynamical model utilized in current applications is a full set of governing equations of motion, mass continuity, thermodynamics, and cloud-physics. The dynamical model has been tested by tornado simulations. Starting from the numerical simulation of a supercell storm by Klemp and Wilhelmson (1978), many simulations were successful in reproducing supercells and mesocyclones, but not tornadoes. Indeed, Burgess (1997) concluded from his analysis that tornadoes developed from only 20 % of mesocyclones, suggesting that tornadogenesis is still unsolved. Recent advanced observations and successful computer simulations of tornadogenesis (Wilhelmson and Wicker 2001; Noda 2002) clearly suggested super high spatial resolution and the associated temporal resolution are required to solve a full set of governing equations of motion, mass continuity, thermodynamics and cloud-physics by computer. For example, in the first successful simulation of tornadogenesis for a few hours of evolution time, Noda used ARPS (Advanced Regional Prediction System, Version 4.5; Xue et al. 1995) with horizontal grid size of 70 m, not nested, and 45 levels of vertical grid, with 10-m spacing near the ground, with associated time increments on the time split integration scheme (Klemp and Wilhelmson 1978; $\Delta t = 0.03$ s, 0.18 s; the former is for sound wave and the latter for others). The simulation took about 720 h on the IBM Regatta computer of 16 nodes at Tokyo University. It will take several days or weeks of computer execution time to simulate tornado evolution of a few hours by the supercomputers currently available for weather forecasting. These requirements prohibit direct application of the current full simulation model for practical operational use under present computing availability.

Also, recent advanced observations such as phased-array Doppler radar and mobile X-band radars have revealed spatial and temporal details of similar high resolutions that are important for tornadogenesis and should be properly reflected in data assimilation. However, again, the presently-available computing power is not sufficient for practical operational forecasting of tornadoes with numerical models. So, it became the first author's motivation to develop a simple but accurate theory that captures all essential processes of tornadogenesis.

The molecular Reynolds number is extremely large (normally $> 10^7$) for most of atmospheric weather systems, and the molecular viscosity is neglected. It allows us to define an appropriate Lagrangian based on the variational principle instead of use of all governing equations of motion, mass continuity, thermodynamics, and cloud physics. It leads to a sole diagnostic Euler-Lagrange (E-L) equation among all other prognostic E-L equations. The diagnostic E-L equation, first found by Clebsch (1859) and called the Clebsch transformation (Lamb 1932), shows indeed a new balance condition different from the known hydrostatic, (quasi-) geostrophic, cyclostrophic, and anelastic balance conditions, as well as the Boussinesq approximation. It is named as entropic balance because of its analogy to other balance conditions, and found it essential to explain tornadogenesis mechanism (Sasaki 1999, 2009, 2010). The entropic balance theory is based on the variational principle, which is proved valid for the tornado-producing, atmospheric fluid of Reynolds number $\mathrm{R_e} = 10^{8-12}$ and Rossby number $\mathrm{R_o} = 10^{2-4}$. The Euler-Lagrange equations are all prognostic except one that is diagnostic, which was found to play key role in the long-lasting steady state of a tornado. The state is analogous to the attractor of nonlinear dynamics. Based on the entropic balance theory, the wrap-around mechanism is introduced to explain explicitly the nonlinear process of tornadogenesis. The results are consistent with advanced observations and successful tornado simulations of phenomena in tornadic storms, such as overshooting hydrometeors against the upper-level westerlies, the mesocyclone, hook echo, discontinuous transition from supercell to tornadic stages as a transition from baroclinic to barotropic stages, an increase of the relative helicity to one (its maximum value), and the tornado touching the ground in the perpendicular direction.

The wrap-around mechanism is analogous to a nonlinear process, the so-called "baker's transformation," and the transition is discontinuous from baroclinic to barotropic stages by trapping entropic sink core inside the vortex, like a nonlinear attractor (Fig. 18.10). Note that the entropic source and sink are of larger magnitudes, it is baroclinic while they are of smaller magnitudes, it is barotropic. Note also that the wrap-around mechanism is two-dimensional while the baker's transformation is one-dimensional. In the entropic balance theory, the sole diagnostic Euler-Lagrange equation is the key equation of the steady state, long-lasting mesocyclonic and tornadic states, where the entropy anomaly is an essential term. Consequently, to estimate the entropy anomaly from radar reflectivity, dual-polarization radar data, or other observational means is a new challenge for data assimilation and prediction of tornadoes.

The entropic balance theory predicts that the supercell and tornado stages correspond well to the cases of high values of helicity. Helicity has been known as an important index for conservation certain rotation to determine the spin and velocity of air particle or parcel, the swirling direction and magnitude of streamlines in fluid mechanics, and the earth magnetism in the dynamo theory of the earth's core magma convection. It is also an important index of mature tornadoes from data analyses, numerical simulations, and theoretical analyses over recent decades. In meteorology, a positive relation between vertical velocity and vertical vorticity was found by Klemp et al. (1981) and Weisman and Klemp (1982) in their analyses of observations and numerical simulations, which seems to support higher values of helicity in supercell thunderstorms. Theories were developed for the helicity to reach nearly a maximum value at the mature stage of rotating convective storms and tornadoes from dynamical analyses by Lilly (1982, 1986), Davies-Jones (1984), and Davies-Jones et al. (2001). Their theoretical analyses suggested the upward tilting of the horizontal vortex tube for tornadogenesis. Note that fluid mechanics allows either parallel or perpendicular approach of a vortex tube to the ground. These numerical simulations and theories lead unrealistic results of touching the ground in parallel, not perpendicular, direction (Question 7 in Appendix 1). However, in contrary, many visual observations have suggested perpendicular touch down of tornadoes to the ground. To solve this controversy, the dynamic pipe effect (DPE) theory developed by Leslie (1971), Smith and Leslie (1978) was applied by Trapp and Davies-Jones (1997) for tornado touch down perpendicularly to the ground. Their theory is based on dynamical pressure deficit on the barotropic Boussinesq (balance) approximation, not including explicitly any thermo-dynamical term, and is consequently insufficient to explain the transition between baroclinic and barotropic stages.

On the other hand, the wrap-around mechanism developed on the basis of the entropic balance theory (Sasaki 1999, 2009, 2010) seems better to explain the transition as we will discuss it in details in Sect. 18.7. Also, the wrap-around mechanism together with the kinematic boundary condition at the ground better explains the important findings that all observed tornado funnels hit perpendicularly when they touch the ground, contrary to the expectation from the upward tilting of a horizontal vortex tube. The numerical simulation experiments of tornadogenesis by Noda (2002), Noda and Niino (2005, 2010) seem to suggest not tilting, but a vertical coupling mechanism because the helicity (relative or normalized) increases to one at the mature stage, meaning less solenoidal effects in the vorticity equation and parallel alignment of updraft and the vorticity axis in their simulation experiments of tornadogenesis.

Recent VORTEX2 results seem to provide supporting evidence for the entropic balance theory (Markowski et al. 2012a,b) and will be discussed in Sect. 18.9. Methods for estimating the entropy anomaly from radar reflectivity and dual-polarization radar data are presented in Sect. 18.10. Section 18.11 presents preliminary results of the application of dual-polarization WSR-88D radar data (Kumjian et al. 2010) by the second author of the present article, and Sect. 18.12 by the third author of this article for rapid-scanning, mobileX-band radar data. The entropic balance theory (and derived the entropic right-hand rule and the wrap-around mechanism) and other directly related observations, theories and models in references are briefly reviewed in Sect. 18.2.

18.2 Entropic Balance Theory for Tornadogenesis

The entropic balance theory developed for tornadogenesis (Sasaki 1999, 2009, 2010) is briefly summarized in Sects. 18.2, 18.3, 18.4, 18.5, 18.6, and 18.7 with addition of further explanation. A tornado is approximated by inviscid and Coriolis-force free flow because high Reynolds number $\mathrm{R_e}$ with the molecular viscosity of the air and high Rossby number $\mathrm{R_o}$ at the middle latitudes are used,

$$\mathrm{R_e} = 10^{8-12} \quad \text{and} \quad \mathrm{R_o} = 10^{2-4}. \tag{18.1}$$

The entropic balance theory hypothesizes that changes in entropy are a quasi-adiabatic process, that is, the microphysical phase change of a small ensemble of hydrometeor molecules is instantaneous, creating a new entropy level, with adiabatic conditions before and after the phase change. It is hypothesized that this phase change timescale is significantly shorter than the time-scales of convective storms and tornadoes (Hypothesis 1), schematically shown in Fig. 18.15,

$$\Delta \mathrm{t}_{\text{phase change}} << \Delta \mathrm{t}_{\text{supercell, tornado}} \tag{18.2}$$

Variations of the initial entropy levels are small enough and allow us to approximate them by their ensemble means (Hypothesis 2). These hypotheses are further discussed in Sect. 18.10.

The Lagrangian density $\mathscr{L}$ is thus formulated as

$$\mathscr{L} := \rho\,(1/2\mathbf{v}^2 - \mathrm{U}(\rho, \mathrm{S}) - \Phi) - \alpha(\partial_{\mathrm{t}}\rho + \nabla\cdot(\rho\mathbf{v})) - \beta(\partial_{\mathrm{t}}(\rho\mathrm{S}) + \nabla\cdot(\rho\mathbf{v}\mathrm{S})), \tag{18.3}$$

where ρ, U, Φ, S, and $\mathbf{v}$ are density of the air, internal energy, gravitational potential energy, entropy, and flow velocity respectively, and α and β are the Lagrange multipliers to satisfy the constraints of conservation of mass and entropy, respectively. Then, the Lagrangian (action) denoted by L is defined as

$$\mathrm{L} := \int_{\Omega} ;\, \mathscr{L}\mathrm{d}\Omega, \tag{18.4}$$

where Ω represents the temporal and spatial integration domain, and the ensemble of air molecules is represented by the spatial integration.

The first variation of L leads to the Euler-Lagrange (E-L) equations, which, after mathematical manipulation, lead to a full set of dynamical and thermodynamical, nonlinear, equations of the ideal flow (Lamb 1932; Bateman, 1932; Sasaki, 1955; Dutton 1976). The E-L equations are all prognostic except for one that is diagnostic, so-called by Lamb as the Clebsch's transformation (Clebsch 1859) of flow velocity,

$$\mathbf{v} = -\nabla\alpha - \mathrm{S}\nabla\beta. \tag{18.5}$$

Then, the vorticity, ω, equation becomes

$$\omega = (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta). \tag{18.6}$$

The vector relation (18.6) is found to be extremely important to gain clear insight into the development mechanisms of supercells and tornadogenesis. The diagnostic velocity (18.5) is universal for the ideal flow. The vorticity (18.6), derived from (18.5), is demonstrated in convenience by the mutually orthogonal vector relation, similar to the so-called Fleming's right hand rule of electromagnetic fields, called by the author as "entropic right-hand rule", among the orthogonal variables of the spatial three dimensions, the vorticity $\boldsymbol{\omega}$, the entropy gradient (1/S) $\mathbf{\nabla}$S, and

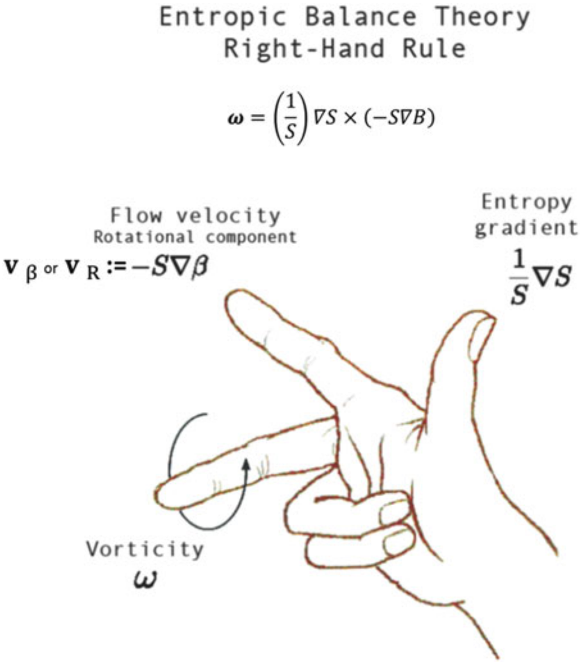

Entropic Balance Theory
Right-Hand Rule
$\boldsymbol{\omega} = \left(\frac{1}{S}\right)\nabla S \times (-S\nabla B)$
Flow velocity
Rotational component
$\mathbf{v}_\beta$ or $\mathbf{v}_R := -S\nabla\beta$
Entropy gradient
$\frac{1}{S}\nabla S$
Vorticity
ω

Fig. 18.1 Entropic right hand rule. The rule shows the mutually orthogonal vector relation, similar to the so-called Fleming's right hand rule of electro-magnetic fields, now we may call as "entropic right hand rule", among the orthogonal variables of spatial three dimensions, entropy gradient (1/S) ∇S, rotational component of flow velocity $\mathbf{v}_\mathrm{R}$ or $\mathbf{v}_\beta := -\mathrm{S}\boldsymbol{\nabla}\beta$ and vorticity ω

the rotational flow velocity component, $-\mathrm{S}\nabla\beta$, denoted by $\mathbf{v}_\beta$ or $\mathbf{v}_\mathbf{R}$, while the divergent component, $-\nabla\alpha$, denoted $\mathbf{v}_\alpha$ or $\mathbf{v}_\mathrm{D}$. These notations are used in the figure illustrations of this article. Figure 18.1 illustrates schematically the entropic right-hand rule.

18.3 Entropic Balance Equation Viewed from Completeness of Solution

Because of the variational principle used in the entropic balance theory, the diagnostic (18.5) should be satisfied always with all other prognostic E-L equations. In the schematic diagram of the solution space (Fig. 18.2), it is shown that the solution subspace DS is expressed as a part of the other solution subspaces, NSS (non-stationary state) and SS (stationary state). Since the helicity becomes nearly maximum at the time of mesocyclone development and tornadogenesis, as will be discussed in Sect. 18.4 (i.e., the local change of vorticity vanishes as will be seen from (18.12)), the long-lasting subspaces, DS (diagnostic state) and SS, are essential.

The solution in the sub-domain covered by DS and SS has a long-lasting property that is similar mathematically to the attractor in nonlinear dynamics. They appear in Fig. 18.2 as the sub-domains of the solution space, DS and SS. Note that

$$\mathrm{DS} \subset \mathrm{SS} \subset \mathrm{NSS}. \tag{18.7}$$

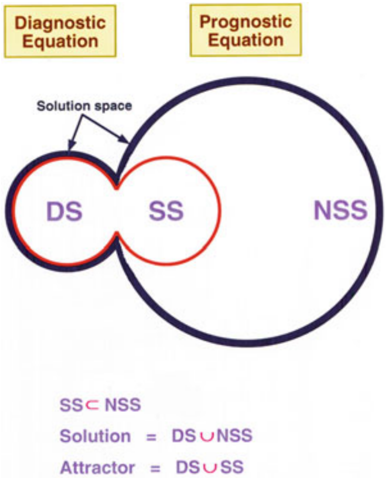


Fig. 18.2 Solution space. The domain of full solution of the Euler Lagrange (E-L) equations is schematically shown in the solution space by the heavy *solid line*. It includes non-stationary state (NSS), stationary state (SS), and the solution of diagnostic E-L equation (DS). The solution in the domain covered by DS and SS has long-lasting property mathematically similar to the attractor

The relationships expressed by (18.7) emphasize the importance of the diagnostic E-L equation (18.5); that is, the transition to a steady state SS or DS from non-steady state NSS must satisfy (18.5). In other words, we can find the necessary conditions for the tornadogenesis and transition among different stages from the entropic balance theory as discussed further in the next sections. The diagnostic balance (18.6) provides insight to a long-lived tornado, presumably by DS and SS steady states, as expressed by (18.7). It is important to note that (18.7) is reached indirectly by a high value of helicity as shown by (18.11) and (18.12) in the next sections.

18.4 Helicity and Tornadogenesis

The helicity, H, is defined as a scalar (inner) product of flow velocity and vorticity,

$$\mathrm{H} := \mathrm{v} \cdot \ \omega, \tag{18.8}$$

where v is flow velocity and ω represents vorticity of the flow $\nabla\times$ v. For fluids of high Reynolds number and high Rossby number, the fluid motion is assumed as an ideal fluid. Without solenoidal effects, the vorticity equation is given by

$$\partial_t \omega = \nabla \times (\mathrm{v} \times \omega). \tag{18.9}$$

The case with solenoidal effects will be shown by (18.25) in the next section. Because of the normal relationship, $\sin^2\theta + \cos^2\theta = 1$, between the scalar product and the vector product, where θ is the angle between two vectors v and ω, we get (Yoshizawa 2001),

$$((\mathrm{v} \times \omega)/\mathrm{D})^2 + ((\mathrm{v} \cdot \omega)/\mathrm{D})^2 = 1, \tag{18.10}$$

where $\mathrm{D}^2 := \mathrm{v}^2\omega^2$, and $(\mathrm{v}\cdot\omega)/\mathrm{D}$ is called relative helicity or normalized helicity, or simply helicity. When the relative helicity approaches unity, (18.10) imposes that

$$(\mathrm{v} \times \omega) \rightarrow 0. \tag{18.11}$$

Then, from (18.9) to (18.10), we get

$$\partial_t\omega \rightarrow 0. \tag{18.12}$$

This means that a steady state of vorticity will be reached when the magnitude of relative helicity increases to unity. Also, it means that the mature stage of a tornado is a long-lasting system, which is similar to the attractor of a nonlinear system (e.g., Lorenz strange attractor of Rayleigh convection). This result agrees with the solution classification in the solution space of the entropic balance theory, as shown as the steady state attractor in Fig. 18.2. It is also clear that (18.11) will be satisfied if the vector v is parallel to the vector ω, and the helicity (18.8) becomes a maximum.

18.5 A Form of Helicity Based on Entropic Balance theory

The entropic balance theory gives further new insight into helicity and entropy. The following E-L equation is the only diagnostic one among all E-L equations obtained from the Lagrangian density of the flow of high Reynolds and Rossby numbers shown by (18.5) as

$$\mathrm{v} = -\nabla\alpha - \mathrm{S}\nabla\beta. \quad \text{same as (18.5)} \tag{18.13}$$

In (18.13), S is entropy, α and β are the Lagrange multipliers of mass conservation and thermodynamics of quasi-adiabatic process, adiabatic with instantaneous phase-change, then entropy change, of microphysics in the Lagrangian density. The Lagrange multipliers α and β are potentials, and they are analogous to the well-known velocity potential usually designated by α^* as follows

$$\alpha^* = \alpha + \mathrm{S}_0\beta, \tag{18.14}$$

where S_0 is a constant along each molecular trajectory and may be determined from the initial condition. Note that S_0 is $\mathrm{S}_0(\mathrm{x}, \mathrm{y}, \mathrm{z})$ at $\mathrm{t} = \mathrm{t}_0$. Determination of S_0 is discussed in Sect. 18.6.

The vorticity is computed from (18.5, or 18.13) and shown by (18.6) as

$$\omega(:= \nabla \times \mathrm{v}) = (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta). \quad \text{same as (18.6)} \tag{18.15}$$

The helicity is calculated from (18.8), (18.5) to (18.6) as

$$\mathrm{H} = (-\nabla\alpha - \mathrm{S}\nabla\beta) \cdot (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta). \tag{18.16}$$

The helicity consists of two parts representing the irrotational and rotational components of **v** in (18.13),

$$\mathrm{H} = \mathrm{H}_\alpha + \mathrm{H}_\beta, \tag{18.17}$$

where the irrotational part is

$$\mathrm{H}_\alpha := (-\nabla\alpha) \cdot (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta), \tag{18.18}$$

and the rotational part is

$$\mathrm{H}_\beta := (-\mathrm{S}\nabla\beta) \cdot (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta), \tag{18.19}$$

where H_α and H_β are used for simplicity instead of $\mathrm{H}_{\mathrm{irrot}}$ and $\mathrm{H}_{\mathrm{rot}}$ which were used in the earlier publications (Sasaki 2009, 2010).

Because $(-\mathrm{S}\nabla\beta) \times (-\mathrm{S}\nabla\beta) = 0$, (18.19) becomes

$$\mathrm{H}_\beta = 0 \tag{18.20}$$

and

$$\mathrm{H} = \mathrm{H}_\alpha. \tag{18.21}$$

Therefore, the helicity is given, using (18.6); $\omega = (1/\mathrm{S})\nabla\mathrm{S} \times (-\mathrm{S}\nabla\beta)$, as

$$\mathrm{H}_\alpha = (-\nabla\alpha) \cdot \omega. \tag{18.22}$$

Comparing the old form of helicity expressed by (18.8), the new form (18.22) shows an important difference, because the rotational term denoted by $(-\mathrm{S}\nabla\beta)$ in the above vector product vanishes. Since ω includes $\nabla\mathrm{S}$ and $\nabla\beta$, two independent thermodynamical parameters for baroclinicity in general, (18.22) becomes

$$H_\alpha = \mathrm{H}_{\alpha,\mathrm{BC}} := \mathrm{u}_\mathrm{D} \cdot \xi + \mathrm{v}_\mathrm{D} \cdot \eta + \mathrm{w}_\mathrm{D} \cdot \zeta. \tag{18.23}$$

where the subscript BC sands for baroclinic, $\mathbf{u}_\mathrm{D}, \mathbf{v}_\mathrm{D}$ and $\mathbf{w}_\mathrm{D}$ (or $\mathrm{u}_\alpha, \mathrm{v}_\alpha$ and w_α respectively) represent the irrotational velocity components $(-\nabla\alpha)$ on Cartesian x,y, and z coordinates, and ξ, η, and ζ are the three-dimensional components of vorticity. This supports Beltrami relation and the tilting of a horizontal vortex tube into the vertical, and a high value of helicity (relative helicity → 1.0) in the supercell stage.

However, the vortex tube at the mature tornadic stage is vertical and hits the ground perpendicularly, so we expect a drastic change from the supercell stage to the mature tornado stage to satisfy the boundary condition of the vortex tube at the ground surface. Therefore, (18.23) becomes drastically different from the tilting process, expressed by,

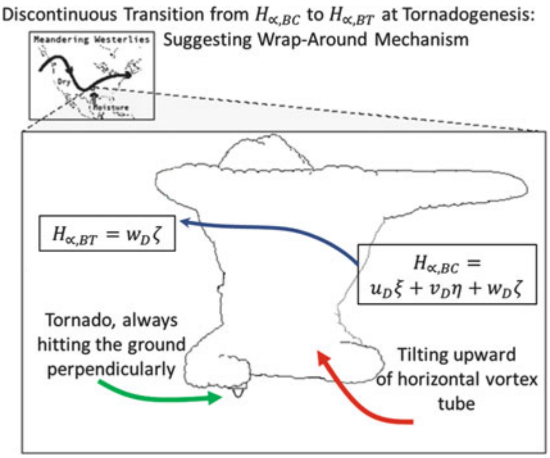

Discontinuous Transition from $H_{\propto,BC}$ to $H_{\propto,BT}$ at Tornadogenesis:
Suggesting Wrap-Around Mechanism
Meandering Westerlies
Dry
Moisture
$H_{\propto,BT} = w_D\zeta$
$H_{\propto,BC} = u_D\xi + v_D\eta + w_D\zeta$
Tornado, always hitting the ground perpendicularly
Tilting upward of horizontal vortex tube

Fig. 18.3 Discontinous transition from supercell storm or mesocyclone to tornado. It is schematically shown by the transition of $\mathrm{H}_\beta \Rightarrow \mathrm{H}_{\alpha,\mathrm{BC}}$ + Wrap-around mechanism $\Rightarrow \mathrm{H}_{\alpha,\mathrm{BT}}$. The divergent velocity $\mathbf{v}_\mathrm{D}$ or $\mathbf{v}_\alpha$ has Cartesian components, u_D, v_D, and w_D

$$\mathrm{H}_\alpha \Rightarrow \mathrm{H}_{\alpha,\mathrm{BT}} := \mathrm{w}_\mathrm{D} \cdot \zeta, \tag{18.24}$$

where the subscript BT satands for barotropic.

The notations of $\mathrm{H}_{\alpha,\,\mathrm{BC}}$ and $\mathrm{H}_{\alpha,\,\mathrm{BT}}$ are used because the former of (18.23) represents the cases where entropic source and sink are of larger magnitudes and the latter of (18.24) does smaller magnitudes. It is valid for the stretching process of tornado and is consistent with the boundary condition of vanishing vertical velocity at the ground surface. The helicity grows up to its maximum near the mature stage of a tornado when the updraft w is intensified due to convective buoyancy, and the vorticity ζ by upward stretching. At and after the mature stage, the updraft changes to a low-magnitude updraft or to a downdraft due to the development of a negative vertical pressure gradient, and the helicity decreases suddenly as demonstrated by Noda (2002) and Noda and Niino (2005, 2010) in numerical simulations of tornadoes. The above discussion suggests that the helicity calculated by the entropic balance theory will vary between $\mathrm{H}_{\beta,\,\mathrm{BC}}$ at the supercell mesocyclone, Lear Frank Downdraft (RFD), hook echo stages, and $\mathrm{H}_{\alpha,\,\mathrm{BT}}$ at the mature tornadic stage. Figure 18.3 is prepared to show schematically the roles of $\mathrm{H}_{\alpha,\,\mathrm{BC}}$ and $\mathrm{H}_{\alpha,\,\mathrm{BT}}$ suggesting sufficient requirement of wrap- around mechanism for tornadogenesis allowing downdraft core of tornado surrounded updraft tornado with high helicity, barotropic surrounding as demonstrated in Fig. 18.4.

The conventional helicity is an index used in tornado research, and, as discussed in the introduction, used to determine how small the term $\nabla \times (\mathrm{v} \times \omega)$ is and when the vorticity becomes stationary, as seen from (18.9) to (18.12). Because of (18.10), the helicity $(\mathrm{v} \cdot \omega)$ is used as an index, although it is indirect. However, the vorticity (18.9) lacks the solenoidal term, which is important. Instead of (18.9), the more accurate vorticity equation is

$$\partial_\mathrm{t}\omega = \nabla \times (\mathrm{v} \times \omega) - \nabla \times ((1/\rho)\nabla \mathrm{p})\,, \tag{18.25}$$

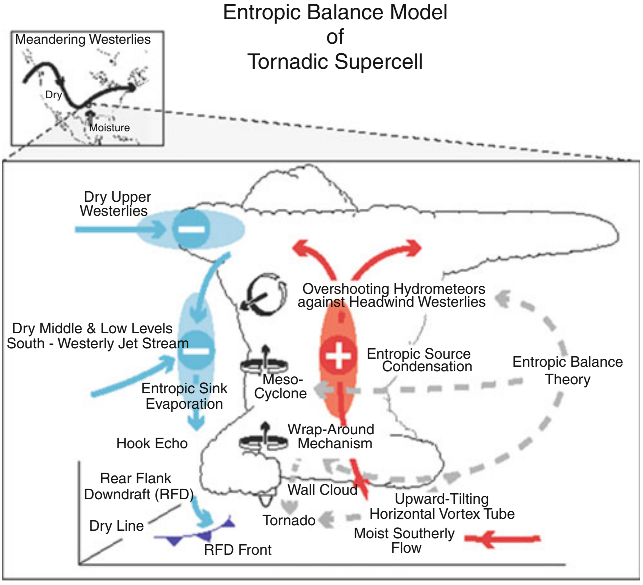


Fig. 18.4 Schematic diagram of tornadogenesis based on the entropic balance theory. Meandering westerlies transport water vapor evaporated from Gulf Stream into the deeper inland of the central US great plain area and meet with the dry air to onset supercell. The moisture of the southerly flow condenses and releases the latent heat to the surrounding air resulting in entropy increase (18.26b) that is shown by circled plus sign, called entropic source, in the figure. The hydrometeors such as raindrops and ice particles created by the condensation are lifted by the updraft of thermal convection of the storm reaching near the cloud top and blow away towards the downstream side, east-side, of the storm. However, the lifted hydrometeors are overshot towards upstream direction against the upper-air westerlies. It is due to the upper air horizontal vortex as shown in this figure where the rotational flow direction of $\boldsymbol{\omega}$ is shown by an *arrow* with *double solid lines*. The overshot hydrometeors will fall down evaporating because of dry air surrounding and cooling the air. The descending hydrometeors with cooled air meet with dry middle-level south-westerly jet and are cooled further to produce the rear frank downdraft. Thus, entropic sink forms nearly at the same altitude of the entropic source at the west of the source. Horizontal spatial gradient of entropy is generated by the pair of the entropic source and sink. A vortex, mesocyclone, is formed and the wrap-around mechanism is organized. The wrap-around mechanism becomes activated by the mesocyclone existed between the entropic source and sink and produces circular belt of entropic source around the sink. The circular system is an ensemble of specific combination of the vorticity under the entropic right-hand rule and its conjugate vorticity. The conjugate vorticity has anti-symmetric entropy gradient and anti-symmetric flow velocity (rotational component), but results in the same vorticity, under nonlinear processes similar to folding of the baker's transformation. The wrap-around mechanism is a nonlinear process, similar to attractor, to generate hook echo, low-level mesocyclone, wall cloud and tornado. These processes are explained by the diagnostic E-L equation, the entropic right-hand rule and wrap-around mechanism which are derived by the entropic balance theory

where ρ is the density of the air, the second term of the right side of (18.25) is the solenoid term. Note that the vorticity diffusion term due to molecular viscosity (υ), $\upsilon\Delta\omega$, is omitted because of the high Reynolds number of the flow. The solenidal effect is significant at the supercell stage, but it will be decreased during the transition period towards the mature stage of the tornado, and the flow becomes barotropic by the wrap-around mechanism as will be discussed in Sect. 18.7.

18.6 Comments on Entropy

Next, we will explore more about entropy. For simplicity, it is assumed that the adiabatic processes are considered to be independent to the diabatic processes, both are added linearly when both are working. Also, because of the high Reynolds number for supercell and tornadic cases, starting with the First law of thermodynamics, d'U = d'Q + d'W where the internal energy U(S, τ) and external heating d'Q is expressed by the following relation, using the entropy S and the specific volume $\tau(= 1/\rho)$, and the work d'W by p$d\tau$ for a dry ideal gas for simplicity as

$$\mathrm{dU = TdS - pd}\tau,\ \ \mathrm{dS = d'Q/T.} \tag{18.26a,b}$$

Hence, we get

$$\mathrm{p} = -\mathrm{U}_\tau \mathrm{T} = \mathrm{U_S}, \mathrm{p}\tau = \mathrm{RT}. \tag{18.27a,b,c}$$

The entropy change dS is expressed by the temperature change and pressure change because the adjustment of d'Q is made due to change of temperature and pressure for the example described in Appendix 4 as, for one mole of gas,

$$\mathrm{dS = d'Q/T = R/2\,dT/T - R\,dp/p} \tag{18.28}$$

where the relations $\mathrm{C_V} = 3/2\,\mathrm{R}$ and $\mathrm{C_P} = 5/2\,\mathrm{R}$ are used.

The internal energy is a function of temperature alone for an ideal gas,

$$\mathrm{U = c_vT}, \tag{18.29}$$

where $\mathbf{c}_\mathrm{v}$ is the specific heat at constant volume. After mathematical manipulation from the above (18.26), (18.27), (18.28), and (18.29), we get

$$\mathrm{S = c_p log(Tp^{\gamma}) + S_0}, \tag{18.30}$$

where the exponent γ is a constant defined as $\mathrm{R/c_p}$ for dry adiabatic processes and its value is adjusted for moist adiabatic processes. Mathematically, $\mathrm{S_0}$ is the arbitrarily determined integral constant, but physically it is discussed as to be determined on the basis of the Third law of thermodynamics, strictly speaking, different from the convention taken in meteorology.

The potential temperature θ is conventionally defined in meteorology as the temperature at a pressure of 1,000 mb after hypothetically moving the particle in adiabatic process to the pressure level,

$$\theta = \mathrm{T(p_{00}/p)^{\gamma}}. \tag{18.31}$$

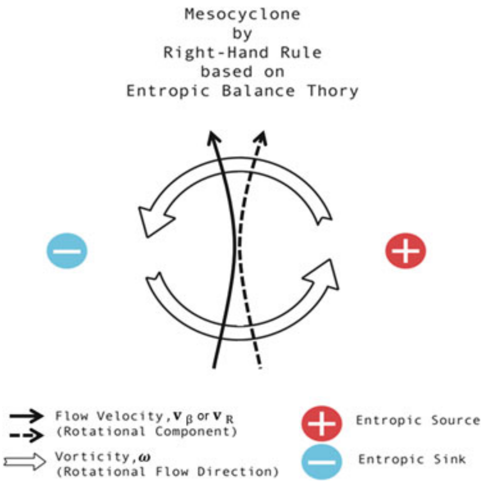

Mesocyclone
by
Right-Hand Rule
based on
Entropic Balance Thory
Flow Velocity, $\mathbf{v}_\beta$ or $\mathbf{v}_R$
(Rotational Component)
Vorticity, $\boldsymbol{\omega}$
(Rotational Flow Direction)
Entropic Source
Entropic Sink

Fig. 18.5 Entropic vortex. Entropic vortex exists by the gradient of entropy in a baroclinic field. The horizontal vortex is formed due to the vertical entropy gradient at the upper levels above the entropic source and overshoots the hydrometeors to upstream against the headwind westerlies. The vertical vortex is formed at the middle levels due to the horizontal gradient between the source and the sink of entropy. The entropic vortex formation is explained by the entropic right-hand rule derived from the entropic balance theory

However, it does not satisfy the Boltszmann's third law of thermodynamics, that is, the entropy S should be zero at the zero absolute temperature T, namely $S = 0$ at $T = 0$, justified by statistical thermodynamics as the entropy defined by $S = k \ln W$, where k is Boltsmann's constant and W is the weight of configuration, and then $S = 0$ for $W = 1$ for perfect configuration and no ambiguity (Atkins and de Paula 2002).

This article uses the third law of the thermodynamics, instead of the conventional potential temperature. An example is shown in Appendix 4 in conjunction with an entropic analysis of tornado.

The baroclinic and barotropic states are viewed also from solenoidal state. The solenoid, σ, is a key term of vorticity generation in (18.25), and it appeared as a vector product of the spatial gradient of specific volume and the pressure gradient, or the spatial gradients of entropy and temperature,

$$\sigma := -\nabla \times (1/\rho)\nabla \mathrm{p} = -\nabla(1/\rho) \times \nabla \mathrm{p}. \tag{18.32}$$

The solenoid defined above is written in terms of temperature and entropy with simplification as

$$\sigma = \nabla \mathrm{T} \times \nabla \mathrm{S}, \tag{18.33}$$

which basically is in agreement with that obtained by Dutton (1976) in a conventional method.

A supercell has properties of baroclinicity, as it is axially asymmetric along the vertical axis (Figs. 18.5 and 18.6), whereas a tornado has confined in entropic sink core, like a singularity, axially-symmetric, surrounded by circular entropic source

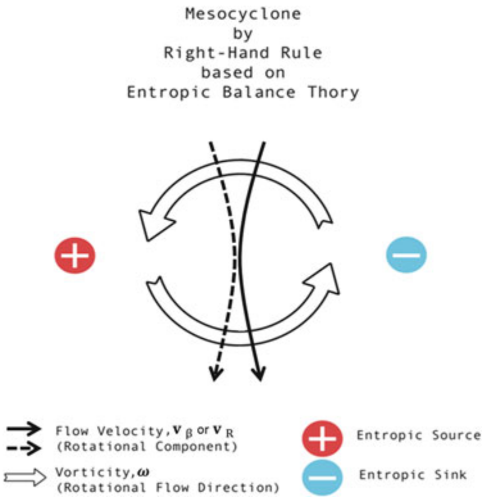

Mesocyclone
by
Right-Hand Rule
based on
Entropic Balance Thory
Flow Velocity, $\mathbf{v}_\beta$ or $\mathbf{v}_R$
(Rotational Component)
Vorticity, $\boldsymbol{\omega}$
(Rotational Flow Direction)
Entropic Source
Entropic Sink

Fig. 18.6 Conjugate entropic vortex. The vortex of the same as Fig. 18.5 is produced under anti-symmetric spatial entropic gradient and anti-symmetric flow velocity (rotational component) under a non-linear super-imposition, with the same vorticity. The conjugate entropic vortex formation is explained by the entropic right-hand rule derived from the entropic balance theory

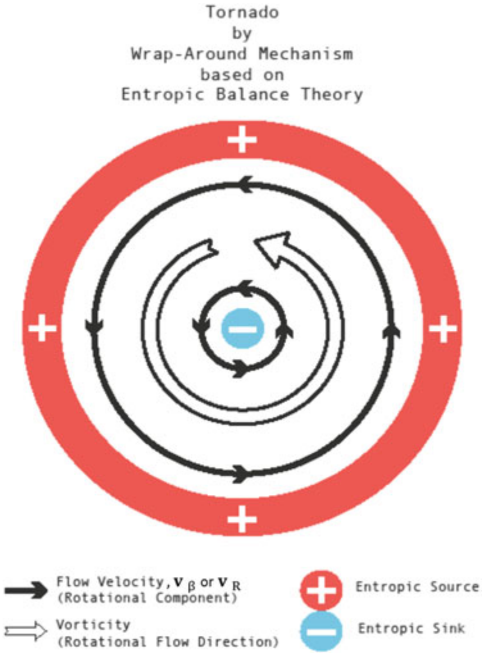


Fig. 18.7 Tornadogenesis. Schematic illustration of tornadogenesis explained with the wrap-around mechanism. The wrap-around mechanism is modeled by the entropic right-hand rule which is derived from the entropic balance theory

environment (Figs. 18.7 and 18.8). The transition from supercell to tornadic stages is physically explained by the proposed wrap-around mechanism (Figs. 18.3, 18.4, 18.9, and 18.10) as explained in the next chapter. Also the entropic balance theory suggests the existence of multiple vortices (Fig. 18.8) as seen from the right hand rule (Fig. 18.1) and the solution space theory (Fig. 18.2).

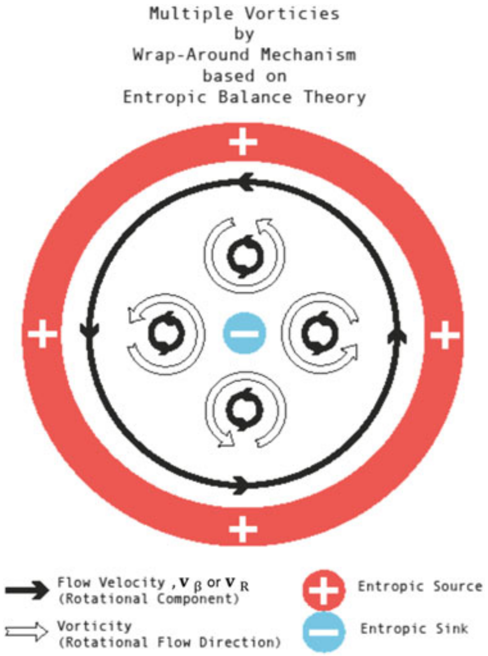


Fig. 18.8 Multi-vortexes. Schematic illustration of formation of four vortexes for an example explained with the wrap-around mechanism. The wrap-around mechanism is modeled by the entropic right-hand rule which is derived from the entropic balance theory

In a successful numerical simulation of tornadoes with horizontal resolution of 75 m, Noda (2002) and Noda and Niino (2005, 2010) clearly showed the particle trajectory which started at the point of about 500 m AGL, 25 km away in the NW direction from the tornado vortex center, moved downward continuously, and converged on the order of 100 m in diameter outside of the tornado vortex. Similar results were shown in other numerical simulations of tornadoes. The helicity (18.17), (18.18), and (18.19) play important roles and provide a theoretical background to explain the above features of the trajectory and unique characteristics of entropy, which seem in agreement with the numerical simulations and detailed analyses of observations. The downdraft agrees with the entropic balance theory, which says intensification of the cyclonic circulation around a tornado is due to the downdraft. Also, the downdraft on the west-side of a tornadic supercell adiabatically transports and converges the entropy into or in the neighborhood of a tornadic vortex with a small area. The entropy from the broader areas outside of 25 km distance from the tornado vortex and several hundreds meters or more above the ground also converges downwards and decreases the gradient of entropy outside of the vortex. The weak baroclinicity in the shallow layer of the atmosphere near the ground that is expected from the entropic theory seems in agreement with the mobile Doppler radar observations and surface observations (Davies-Jones et al. 2001). Also, the entropic balance theory suggests a converged, concentrated, axially (in vertical direction) symmetric, wrap-around entropy field in and near the mature tornado vortex core, similar to a nonlinear attractor, as will be discussed further in the next chapter.

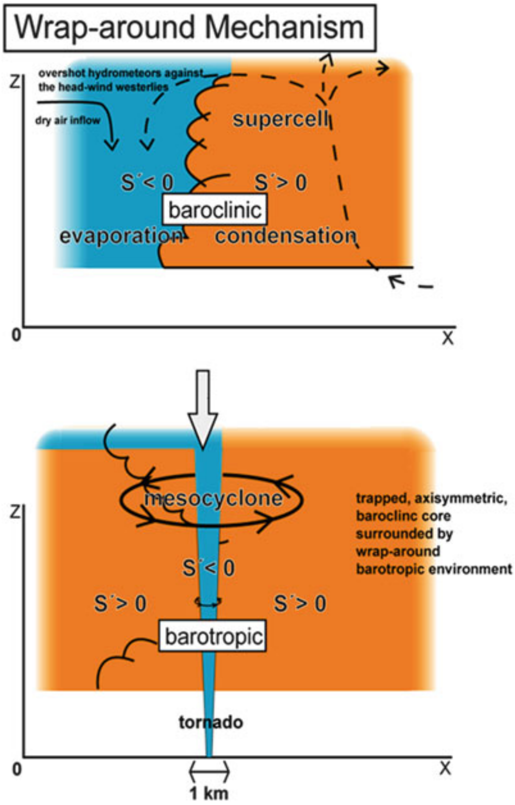

Wrap-around Mechanism
overshot hydrometeors against the head-wind westerlies
dry air inflow
supercell
S'< 0
S'> 0
baroclinic
evaporation
condensation
Z
0
X
mesocyclone
trapped, axisymmetric, baroclinc core surrounded by wrap-around barotropic environment
S'< 0
S'> 0
S'> 0
barotropic
tornado
Z
0
X
1 km

Fig. 18.9 Wrap-around mechanism. Schematic illustration of nonlinear process transforming the baroclinic state (*top*) to barotropic state with baroclinic core (*bottom*)

18.7 Wrap-Around Mechanism

It was suggested in the earlier publications (Sasaki 2009, 2010) that the mature stage of tornado appears almost discontinuously from the parent supercell, like the axially symmetric, nonlinear attractor, and a singular stationary-state vortex, by the proposed process named the “wrap-around mechanism.” The wrap-around mechanism is analogous to the baker’s transformation in nonlinear dynamics although it is two dimensional while the baker’s transformation is one dimensional. The wrap-around mechanism becomes activated by the mesocyclone existing between the entropic source and sink, which is a baroclinic state, and is expected from the entropic right-hand rule (Fig. 18.1). The mesocyclone produces a circular belt of entropic source around the sink. The circular system is an ensemble of specific combinations of the original vorticity and its conjugate (Figs. 18.5 and 18.6). The conjugate vorticity has a conjugate entropy gradient and conjugate flow velocity (rotational component), but with an integrated magnitude of vorticity and direction. Thus, the original vortex and conjugate vortex produce the integrated magnitude of vorticity in the original direction.

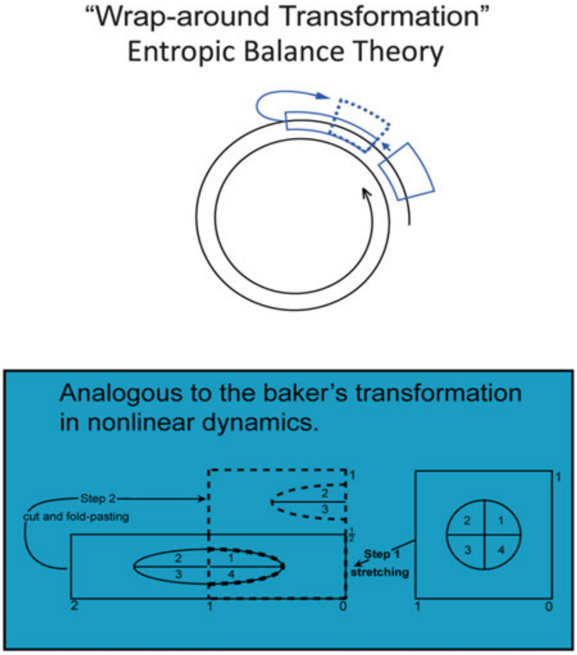

"Wrap-around Transformation"
Entropic Balance Theory
Analogous to the baker's transformation in nonlinear dynamics.
Step 2
cut and fold-pasting
Step 1
stretching
1
2
3
4
2
1
0
1/2
1
0

Fig. 18.10 Wrap-around mechanism (two dimensional) analogous to the nonlinear Baker's transformation (one dimensional)

The wrap-around mechanism explains the observation that all tornadoes hit the ground in the perpendicular direction, and also it is found favorable for the drastic transition from the supercell to tornadic stages. The wrap-around mechanism based on the entropic balance theory creates axially-symmetric structure of a tornado and suggests the transition from supercell to tornado as that from baroclinic to barotropic states. The barotropic state is horizontal due to the axial symmetry along the vertical core axis of tornado. We will further discuss the background reasons for this suggestion.

18.7.1 Tornado Hits the Ground in the Perpendicular Direction

Some theories speculated that a tornado is formed from a horizontally laying vortex tube by tilting upward by a storm updraft (Davies-Jones et al. 2001). However, it is known from many visual observations that a tornado vortex always hits the ground in the perpendicular direction. It is easily understood from fluid mechanics that a vortex tube of finite diameter does so at a wall surface, because the normal component of the flow velocity should vanish at the wall surface.

Therefore, there only exist two cases: (a) the vortex tube lays on the ground in the parallel direction, or (b) the tube hits the ground in the perpendicular direction, but not in a slanted direction.

Accordingly, a tornado core is not formed from upward tilting of a horizontal vortex tube by storm updrafts. Instead, it seems natural to assume that it originates in the storm from mid-levels at an altitude of several hundred meters or a few kilometers above the ground, and with the wrap-around nonlinear mechanism, the tornado vortex tube hits the ground in the perpendicular direction, satisfying the boundary condition of vanishing vertical velocity at the ground surface.

18.7.2 High Relative Helicity and Stationary State

It is supported by successful numerical simulations (Noda 2002; Noda and Niino 2005, 2010) that the relative helicity of a mature tornado is high, near one, implying a stationary state SS (Fig. 18.2) of a relatively long life time for the mature stage of a tornado as discussed in Sects. 18.3 and 18.4 of this article. It should be noted that the helicity is defined as given in (18.8) as the scale product $(\mathbf{v} \cdot \boldsymbol{\omega})$ between the flow velocity $\mathbf{v}$ and the vorticity $\boldsymbol{\omega}$, and it is used to show the stationary state, namely $\partial_t \boldsymbol{\omega} \approx 0$. To do so, it is assumed, as shown in (18.10) of Sect. 18.4, that the magnitude of the tem $(\mathbf{v} \times \omega)$ is sufficiently smaller than that of $(\mathbf{v} \cdot \boldsymbol{\omega})$. Here the solenoid term $\nabla \times ((1/\rho)\nabla \mathrm{p})$ and $\nabla \times (\mathbf{v} \times \omega)$ of (18.23) are neglected, although both play important roles in supercell development stage (baroclinic), but not after the transition to tornadic stage (barotropic), as discussed earlier in Sasaki (2010).

18.7.3 Transition from Supercell to Mature Tornado

In the entropic balance theory, the flow velocity $\mathbf{v}$ is expressed by the diagnostic E-L equation (18.5),

$$\mathbf{v} = -\nabla\alpha - \mathrm{S}\nabla\beta. \quad \text{same as (18.5)} \tag{18.34}$$

The rotational term $(-\mathrm{S}\nabla\beta)$, plays an important role in supercell stages including tornadogenesis as discussed in the author's earlier article (Sasaki 1999, 2009, 2010) and Sect. 18.5 of this article. This can also be been seen in the following vorticity, $\boldsymbol{\omega}$, equation,

$$\omega = (1/\mathrm{S})\nabla\mathrm{S} \times -(\mathrm{S}\nabla\beta). \quad \text{same as (18.6)} \tag{18.35}$$

The helicity, however, uses only the first divergent term (-$\nabla\alpha$) for the flow velocity as shown in (18.22), as

$$\mathrm{H}_\alpha = (-\nabla\alpha) \cdot \omega. \quad \text{same as (18.22)} \tag{18.36}$$

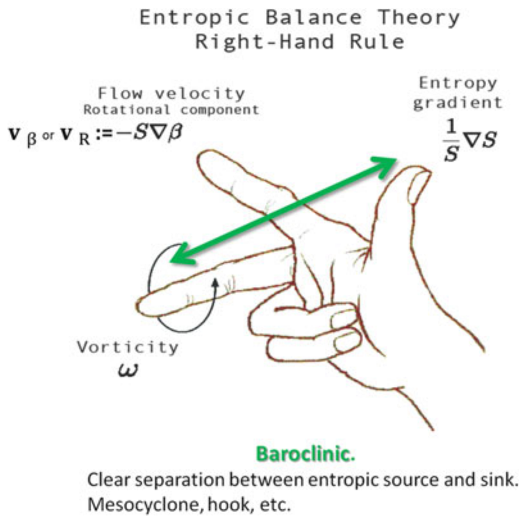


Fig. 18.11 Baroclinic case of the entropic balance expressed in the right hand rule

The transition to the mature stage of a tornado is characterized by the transition from the asymmetric baroclinic stage to the symmetric barotropic stage as discussed in Sect. 18.5,

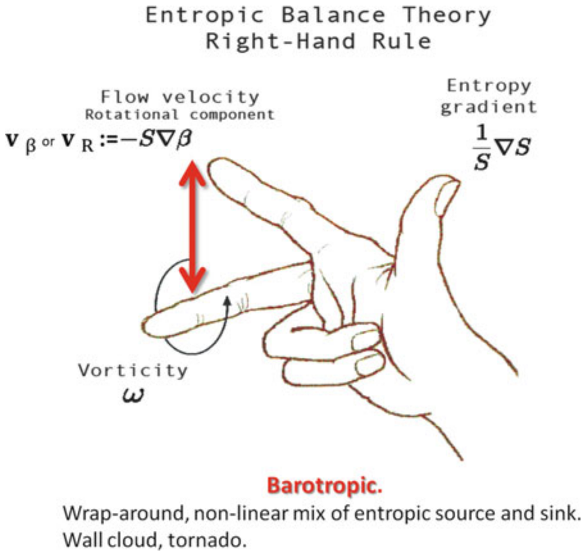
$$\mathrm{H}_\beta \Rightarrow \mathrm{H}_{\alpha,\mathrm{BC}} + \text{Wrap-around mechanism} \Rightarrow \mathrm{H}_{\alpha,\mathrm{BT}}. \tag{18.37}$$

18.7.4 Wrap-Around Mechanism

The wrap-around mechanism is proposed to be responsible for the transition from the supercell stage to the mature tornado stage. In Fig. 18.9, the supercell baroclinic stage is shown at the top, and the mature tornado stage is at the bottom. S' is the entropy anomaly. $\mathrm{S}' > 0$ due primarily to condensation, and $\mathrm{S}' < 0$ due to evaporation in the supercell storm. Tighter wrap-around causes steeper, axially symmetric entropy gradients in and closely around the trapped core of tornado, consequently creating intense vorticity, according to the entropic balance theory. The wrap-around mechanism and the corresponding baker's transformation are schematically shown in Figs. 18.9 and 18.10 respectively.

The supercell stage is baroclinic, $\sigma \neq 0$, created by the axially asymmetric entropy anomaly distribution, due to $\mathrm{S}' > 0$ (condensation in the storm) and $\mathrm{S}' < 0$ (evaporation of the overshot hydrometeors against the head-wind westerlies in the west of storm). The baroclinicity is created by the solenoid. The mature tornadic stage is created by the field of circular band of positive S' wrapping around the tornado core of negative S'. The trapped tornado core and the environment in an small area is like barotropic over all by a nonlinear wrap-around mechanism (Figs. 18.3, 18.4, 18.5, 18.6, 18.7, 18.8, 18.9, 18.10, 18.11, and 18.12).

Fig. 18.12 Barotropic case of the entropic balance expressed in the right hand rule



The wrap-around mechanism developed on the basis of the entropic balance theory provides the flow velocity (18.5), which explicitly includes the thermodynamic terms of entropy S varied by heating d’Q, and the Lagrange multipliers α and β of the constraints of density and entropy, respectively. The wrap-around mechanism and the entropic balance theory seem to explain the transition from supercell to tornado. Also, the wrap-around mechanism, together with the kinematic lower boundary condition, better explain the important findings that all observed tornadoes contact the ground perpendicularly, contrary to the expectation from the upward tilting of a horizontal vortex tube.

18.8 Schematic Entropic Balance Model of Supercell and Tornadogenesis

Figure 18.4 illustrates schematically the entropic balance model of supercell and tornadogenesis under meandering westeries. When a large-amplitude trough develops, more water vapor evaporated from Gulf Stream is transported by the southerly flow deeper inland into the central US Great Plains and meets with the dry air transported by the north-westerly jet-stream. In a developing supercell, the moisture of the southerly flow condenses and releases latent heat into the surrounding air, resulting in an entropy increase (18.26b) that is shown by circled plus sign, called an entropic source, in the figure. The hydrometeors such as raindrops and ice crystals created by the condensation and freezing are lifted by the updraft, reaching near the cloud top, wherein they blow away towards the downstream side, east-side, of the storm in the anvil.

However, some of the lifted hydrometeors are overshot towards the upstream direction of westerlies, against the strong headwind. It is due to the horizontal vortex (represented by the vorticity $\boldsymbol{\omega}$) as shown in Fig. 18.4 where the rotational flow direction of $\boldsymbol{\omega}$ is shown by an arrow with double solid lines. The horizontal vortex is formed within the vertical but slanted, towards head-wind direction, by a dipole of entropic source and sink. The overshot hydrometeors will be evaporating and sublimating in the ambient dry air, cooling the air. The descending hydrometeors with cooled air meet with dry middle-level south-westerly jet and are cooled further and produce rear flank downdraft. Thus, major entropic sink forms nearly at the same altitude as the entropic source, but further west. The horizontal spatial gradient of entropy is generated by this entropic source and sink, (Fig. 18.4) and the mesocyclones are generated. There, the diagnostic E-L equation (18.5) and the entropic right-hand rule play important roles.

The wrap-around mechanism discussed in Sect. 18.7 and shown in schematic Figs. 18.9 and 18.10 is a nonlinear process, similar to the folding process of the baker's transformation of nonlinear dynamics (Fig. 18.10), and produces an axially-symmetric vortex in the vertical axis. This mechanism produces the hook echo, low-level mesocyclone, wall cloud, and tornado. The mechanism is well explained by the diagnostic E-L equation and the entropic right-hand rule, both of which are derived by the entropic balance theory.

18.9 Comparison with a Well Documented VORTEX2 Result

The entropic balance theory is tested with a well-documented casefrom the most recent observational experiment, VORTEX2. Figure 18.13 is a schematic diagram of the supercell and tornadogenesis that occurred on June 5, 2009 in Goshen County, Wyoming (Markowski et al. 2012a, b).

Entropic balance theory implies that the mesocyclone develops in the baroclinic field between the entropic source (primarily due to condensation) and sink (primarily due to evaporation), as shown in Fig. 18.4. It is deduced from the theory that tornado is developed due to the wrap-around of the positive entropic anomaly air around the subsiding negative core at the area of the center of the tornado. Indeed, the subsiding core is shown by DRC (descending reflectivity core) and high value of vertical vorticity in Fig. 18.13. From entropic balance theory, we found that the transition from mesocyclone to tornado is characterized by the transition from baroclinic stage to barotropic stage (Figs. 18.11 and 18.12). The transition is nonlinear, analogous to the baker's transformation, common in nonlinear dynamics, called in this study as the wrap-around mechanism because of its higher dimension than that of the baker's transformation (Fig. 18.10). It is important to note that

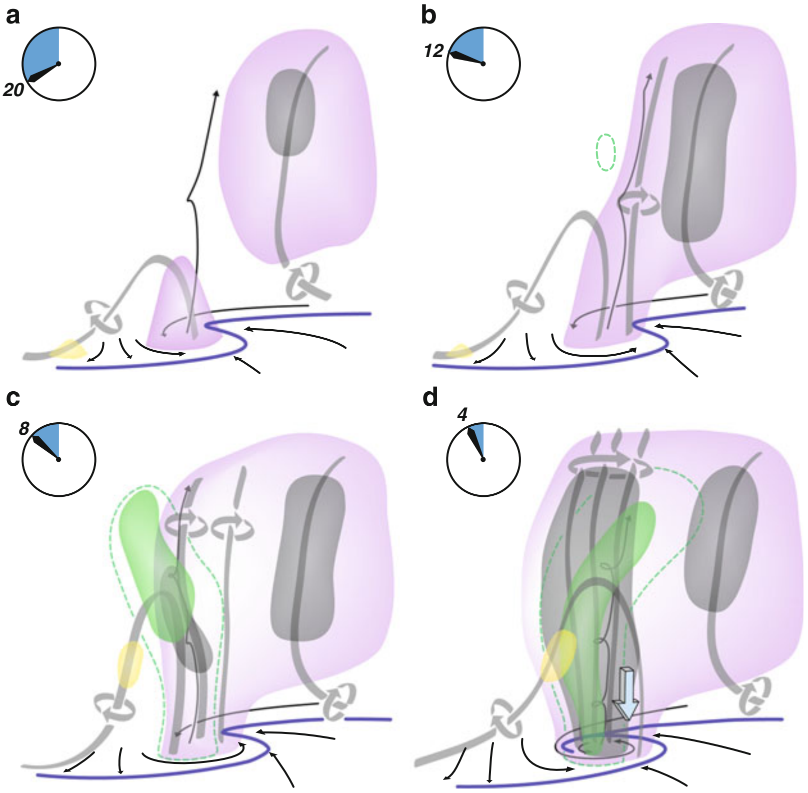


Fig. 18.13 A schematic summary of VORTEX2 data analysis of the Goshen, Wyoming tornadogenesis case. Schematic summarizing the evolution of a tornadic storm during its pre-tornadic phase. The clocks represent the time until tornadogenesis. Regions of significant cyclonic (anticyclonic) vertical vorticity are indicated by the *purple* (*yellow*) shading. *Dark gray* shading encloses even larger cyclonic vertical vorticity. The descending reflectivity core is indicated by the *green* shading. Surface gust fronts are analyzed using *blue lines*. *Streamlines* are shown in *black arrows*. *Vortex lines* are shown in *gray* streamers, with the sense of rotation indicated by the *gray arrows*. In (**d**), the downward-pointing *arrow* indicates the occlusion downdraft (From Markowski et al. (2012a,b))

in Fig. 18.13, the wrap-around mechanism worked actively by the right-hand side vortex in Fig. 18.13 corresponding to the lower and upper mesocyclones in Fig. 18.4. Also, it is noted that the shaded area of high vertical vorticity of Fig. 18.13 covers both areas of the upper-level and the low-level mesocyclones, which may be formed by the baroclinicity generated between entropic source and sink and upward tilting horizontal vortex, respectively. It suggests that vertical superimposition of their phases seems a key of tornadogenesis (Fig. 18.14).

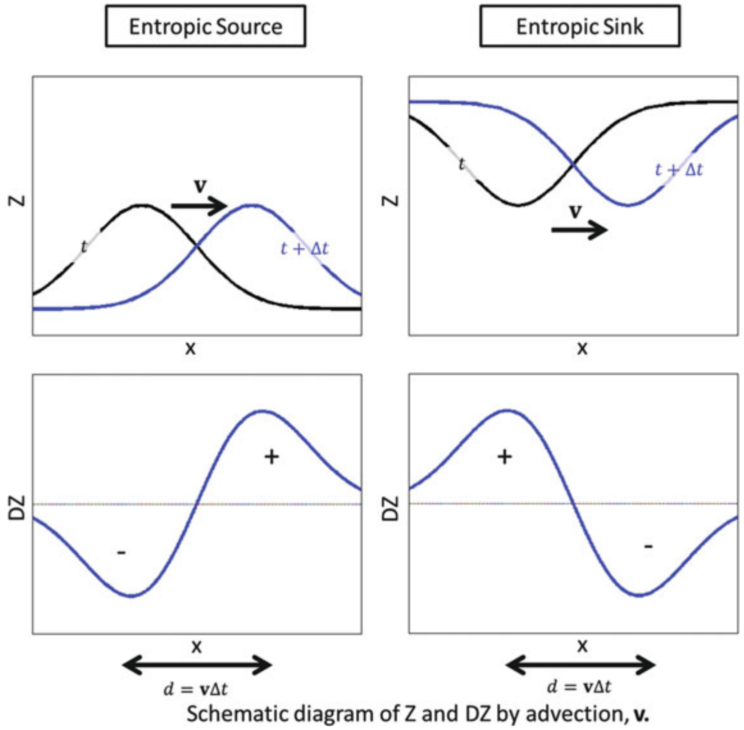


Fig. 18.14 **Change of entropy.** Note that dS is totally differentiable while Δ’Q is not, and $dS = (\partial_t + \mathbf{v} \cdot \nabla)\,S$

18.10 Temporal Discretization of Radar Data for Entropy anomaly

The hypotheses 1 and 2 used for the Lagrangian in Sect. 18.2 are restated here. Hypothesis 1 states that microphysical phase changes of a small ensemble of hydrometeor molecules is instantaneous, creating a new entropy level with adiabatic conditions before and after the phase change, and having a much shorter time-scale than the time-scales of convective storms and tornadoes,

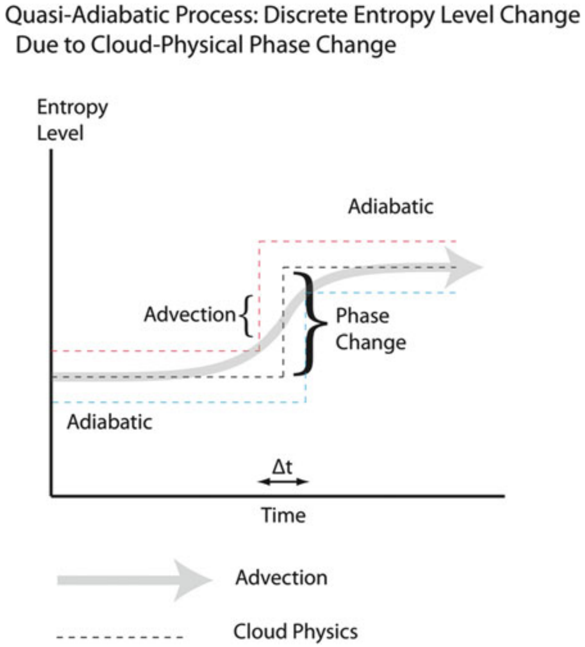
$$\Delta t_{\text{phase change}} << \Delta t_{\text{supercell, tornado}}. \quad \text{same as (18.2)} \tag{18.38}$$

Hypothesis 2 states that variations of the initial entropy levels are small enough to allow us to approximate them by their ensemble means. These hypotheses are shown schematically in Fig. 18.15.

The entropic source and sink are created by diabatic heating and cooling (Fig. 18.4), and (18.26b) is rewritten as

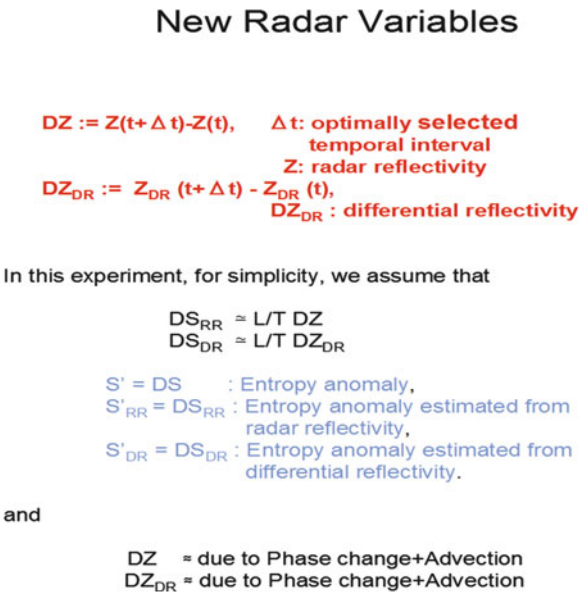
$$(\partial_t S + \mathbf{v} \cdot \nabla \mathbf{S}) = d'Q/T. \quad \text{same as (18.26b)} \tag{18.39}$$

Fig. 18.15 Schematic diagram of molecule ensemble of instantaneous phase change


Quasi-Adiabatic Process: Discrete Entropy Level Change Due to Cloud-Physical Phase Change
Entropy Level
Adiabatic
Advection
Phase Change
Adiabatic
Δt
Time
Advection
Cloud Physics

Using the difference of the time-scales between phase changes of cloud physical processes and supercell processes as shown in (18.38), we can simplify the entropic analysis eliminating the advection effect of not-important dipole of entropic source and sink, because it is merely caused by the term of $\mathbf{v} \cdot \nabla S$ in (18.39), as schematically demonstrated in Fig. 18.15, but focusing on the more important contribution of d'Q/T on the entropy change. It seems to be accomplished by taking smaller value of time interval Δt compared with $\Delta t_{\text{supercell, tornado}}$. An appropriate value of Δt is suggested as $\Delta t < 1$ min.

The latent heat values are much higher with condensation or evaporation, almost five times, than that of freezing or melting at various pressure and temperature conditions of the troposphere, as seen from the three phase diagram of water (Meteorological Glossary, American Meteorological Society 2000; Atkins and de Paula 2002, for pure water; Pruppacher and Klett 1997, for the behavior of water vapor, liquid and solid water states with salt and other condensation nuclei and various environmental conditions; Stensrud 2007, for water-atmosphere parameterization and convective parameterization). It is assumed that there are important roles of condensation and evaporation for tornadogenesis as the first approximation, which occur at lower and middle levels of the troposphere and create a nearly maximum entropy gradient between the entropic source and sink (Fig. 18.4) as explained in Sects. 18.2 and 18.8. It is based on the fact that significantly larger latent heat released by condensation (or removed by evaporation) is expected at low- and middle- levels than from freezing (or melting) at the mid- or upper-levels of the troposphere. This entropic source and sink with neutral stability in the middle and



New Radar Variables

$DZ := Z(t+\Delta t)-Z(t)$, Δt: optimally selected temporal interval
Z: radar reflectivity

$DZ_{DR} := Z_{DR}(t+\Delta t) - Z_{DR}(t)$,
DZ_{DR} : differential reflectivity

In this experiment, for simplicity, we assume that

$DS_{RR} \simeq L/T\ DZ$
$DS_{DR} \simeq L/T\ DZ_{DR}$

$S' = DS$: Entropy anomaly,
$S'_{RR} = DS_{RR}$: Entropy anomaly estimated from radar reflectivity,
$S'_{DR} = DS_{DR}$: Entropy anomaly estimated from differential reflectivity.

and

$DZ \approx$ due to Phase change+Advection
$DZ_{DR} \approx$ due to Phase change+Advection

Fig. 18.16 New radar variables

lower troposphere (Sasaki 2009) seems provide and answer to the question of why the tornado is a low-level phenomenon.

The value of the heating or cooling, d’Q, may be estimated from the temporal change of radar reflectivity, if the time step is taken small enough so that the advection term of reflectivity becomes negligible. Other effects such as radiation seem to be small compared against the latent heat. Because T is nearly constant, 240–273°K for the condensation with super-saturation and evaporation processes leads good estimates of dS from the latent heat release d’Q. It is the purpose of this study to find a clue if we could estimate d’Q from radar reflectivity variations, in spite of not detecting the details for the cause of d’Q. In order to get a feeling on the order of magnitude estimate of flow velocity (rotational component) **v**, vorticity $\boldsymbol{\omega}$ in conjunction of d’Q and ∇S (from the distance between entropic source and sink assuming linear profile of entropy) shown as a numerical example in Appendix 4.

The preliminary results of initial testing are shown in Sects. 18.11 and 18.12 in addition to the comparison with VORTEX 2 (Sect. 18.9). The new notations DZ, DZ_{DR}, S', S'_{RR}, and S'_{DR} are defined as follows, and shown in Fig. 18.16, because of future use of the defined quantities:

$$DZ := Z(t+dt) - Z(t), \text{ with optimally-selected temporal interval dt, and radar reflectivity Z,} \tag{18.40}$$

$$DZ_{DR} := Z_{DR}(t+dt) - Z_{DR}(t) \text{ where } Z_{DR} \text{ is differential reflectivity,} \tag{18.41}$$

$$S' : \text{Entropy anomaly,} \tag{18.42}$$

$$\text{S'}_{\text{RR}} : \text{Entropy anomaly estimated from radar reflectivity,} \tag{18.43}$$

$$\text{S'}_{\text{DR}} : \text{Entropy anomaly estimated from differential reflectivity.} \tag{18.44}$$

In this experiment, for simplicity, we assume that

$$\text{S'}_{\text{RR}} \simeq \text{L/T DZ} \tag{18.45}$$

and

$$\text{S'}_{\text{DR}} \simeq \text{L/T DZ}_{\text{DR}}. \tag{18.46}$$

where L is the latent heat of phase transition of microphysical process, excluding non-phase transition processes such as advection. It will be discussed in Chap. 11 for a selected process. Note that the instantaneous cloud physical phase change (Fig. 18.15) should be captured better by a small temporal interval dt in ((18.40), (18.41), (18.42), (18.43), and (18.44)) because the time scales of the environmental atmospheric flow system, supercell, mesocyclone, and tornado are much larger. However, as we discussed in Sect. 18.7 (D), the advection term of the wrap-around mechanism is the needed important nonlinear process to include for tornadogenesis. However, for simplicity, we focus our initial testing of the entropic balance theory on the diabatic heating and cooling d'Q estimates on a moving coordinates with tornado from radar observations.

18.11 Estimating Entropy from Polarimetric Radar Data

As discussed in previous chapters, the entropic sources and sinks can be created by evaporative cooling or condensational heating:

$$\text{d'Q} = \text{TdS}, \quad \text{same as (18.26b) and (18.39)} \tag{18.47}$$

where d'Q is the heating or cooling. To estimate what changes in entropy dS could look like in radar data, we make use of the evaporation model of Kumjian and Ryzhkov (2010). In this simplified one-dimensional model, raindrops in the 3 km column evaporate as they descend to the surface. Evaporation leads to a decrease in radar reflectivity Z and an increase in the differential reflectivity Z_{DR} (e.g., Li and Srivastava 2001; Kumjian and Ryzhkov 2010). The magnitude of these changes in the radar variables depends on the initial drop size distribution (DSD) aloft as well as the environmental conditions in the model domain. The cooling rate owing to evaporation of liquid water can be expressed as (e.g., Pruppacher and Klett 1997; Bohren and Albrecht 1998):

$$\text{d'Q/dt} = \text{L}_{\text{v}}\text{dm/dt}, \tag{18.48}$$

where L_v is the latent enthalpy of vaporization, and dm/dt is the rate of change of mass owing to evaporation (which is negative, implying a cooling rate). The change in mass of water in the one-dimensional model is calculated based on the change in liquid water content (mass of water per unit volume),

$$M = (\pi/6)\rho_w \int N(D)D^3 dD. \tag{18.49}$$

Here, N(D) is the number concentration (per unit volume) of drops of diameter D, and ρ_w is the density of liquid water. Thus, we can obtain an estimate of the change in entropy per unit volume based on model output:

$$dS \approx (L_v T^*)dM, \tag{18.50}$$

where T* is the average temperature of the model domain. Using a number of different environmental profiles and DSDs, we can estimate the entropy anomalies S'_{RR} and S'_{DR} as a function of the evaporative changes DZ and DZ_{DR}. In general, larger changes in Z and Z_{DR} correspond to larger changes in entropy for a given DSD (not shown).

18.11.1 1 June 2008 case

The temporal difference method is applied to the rapid-scan radar data from the 1 June 2008 case of a cyclic nontornadic supercell in Oklahoma (see Kumjian et al. 2010). Figure 18.17 shows the temporal difference fields of Z and Z_{DR} over the period 0341:36 UTC to 0346:26. At this time, the storm is undergoing cyclic mesocyclogenesis, and the new mesocyclone is developing along the RFD gust front. This time is marked by an increase in the strength of the updraft. Note that the signal of storm advection is evident in each panel (the +/− difference "dipole" is clearly seen in the hook echo at each time). However, meaningful patterns of differences exist. For example, in panels (d), a relatively large region of positive DZ (indicating an increase in Z from one scan to the next) is located across much of the RFD north of the hook echo. At the same time, a large positive DZ_{DR} is located farther downstream along the forward-flank downdraft echo, after several consistently negative differences in the preceding scans. Such changes in behavior of the storm microphysics may be related to changes in entropy (e.g., increased Z could mean more precipitation produced by condensation and accretion aloft, indicating a positive entropy anomaly). A positive DZ_{DR} along the forward flank indicates suddenly larger drops are falling there, as a result of enhanced size sorting or some other process.

It is interesting to note that this case is an excellent example of baroclinicity development at the front edge of RFD, (see Fig. 18.5 for schematic illustration, applying the righthand rule of Fig. 18.1), which is known as favorable for tornadogenesis (Lemon and Doswell 1979). However, a tornado did not develop from this

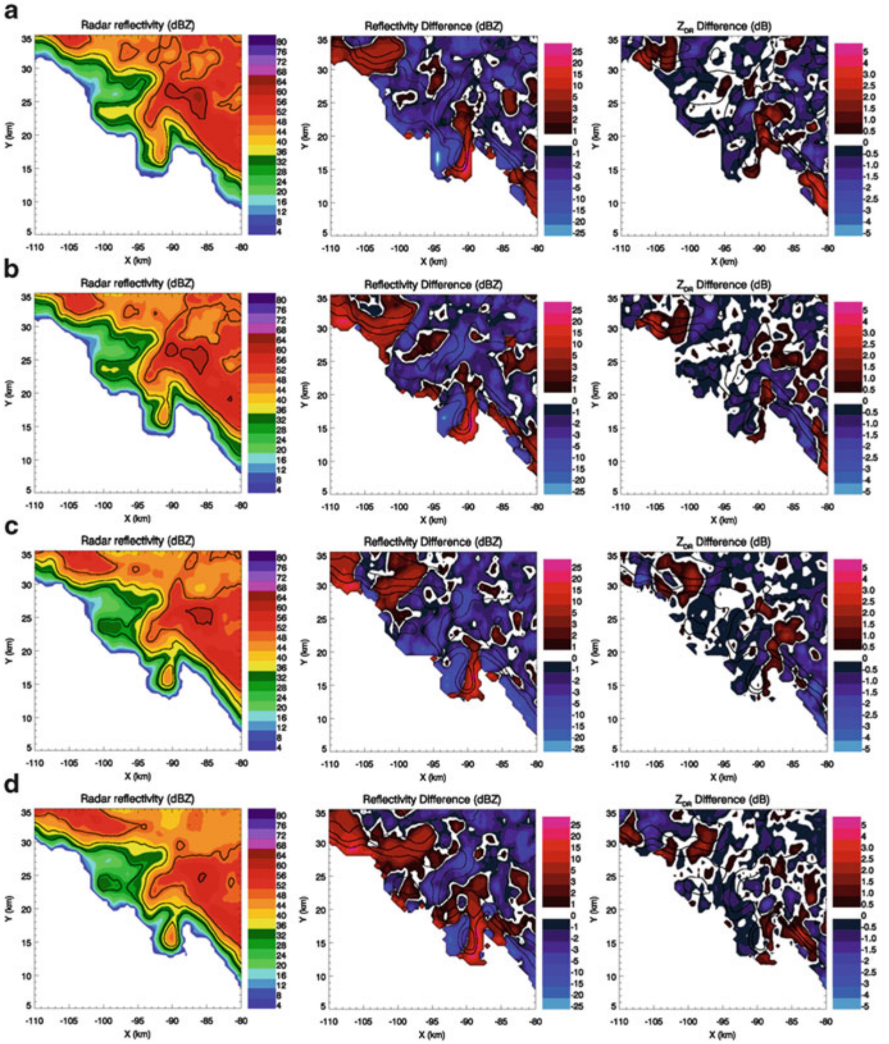


Fig. 18.17 Non-tornadic supercell case. Z: Initial reflectivity field of each section (*left panels*), temporal sequence of DZ (*center panels*), and that of DZ_{DR}(*right panels*) from the case of 1 June 2008. Differences are for (**a**) 0341:36 to 0342:49, (**b**) 0342:49 to 0344:01, (**c**) 0344:01 to 0345:14, and (**d**) 0345:14 to 0346:26 UTC. Overlaid on the difference plots are the 30-, 40-, 50, and 60-dBZ contours of radar reflectivity Z from the most recent of the two times

storm. Why this storm did not produce a tornado even though it seemingly had an environment favorable for tornadogenesis is an important question (listed in the Appendix 1).

The entropic balance theory seems to provide an answer: that tornadogenesis requires the wrap-around nonlinear process (Sect. 18.7), which creates the transition from the baroclinic stageto the barotropic stage (Fig. 18.3). The wrap-around mechanism was apparently missing in this case. It could be because the low-level mesocyclone was incapable of producing the wrap-around process and cyclic symmetry necessary for tornado development. This could be due to the large distance between the major entropic source and sink regions outside of the hook echo (RFD) area, which cause too small of an entropic gradient and thus too weak vorticity.

18.12 Temporal Discretization Necessary for Phase Change Ensemble

The previous case illustrates the potential for the use of entropic balance theory with standard weather radar outputs, Z and Z_{DR}. However, data from this case were collected at approximately 70-s intervals, a rate much higher than is used operationally for the NEXRAD network, which usually receives updates for a particular elevation every 5 min. It is clear from the rapidly evolving scenario presented that, even at this high temporal sampling rate ($\sim$70 s), storm advection produces a bias in the calculation of DZ and DZ_{DR}. Since the parameters used to calculate the system entropy depend on the microphysical changes within a radar resolution volume, it can be assumed that over a necessarily short period of time, the molecular phase state fluctuations will dominate (see Fig. 18.13). The question of how short a time interval is appropriate will be addressed in this section.

The temporal difference method was applied to data collected from the Atmospheric Imaging Radar (AIR), a multi-channel, X-band, mobile imaging weather radar capable of gathering 20° range-height indicator (RHI) scans at approximately 1 s time intervals. A detailed description of the radar and its capabilities can be found in Isom et al. (2011). It should be noted here that this radar is horizontally polarized and thus we cannot calculate the DZ_{DR} parameter. This extremely high temporal resolution made it possible to examine the calculated values of DZ at various intervals and determine an appropriate dt in which the changes in reflectivity are dominated by microphysical processes and not advection.

Three examples of varying interval lengths are given in Fig. 18.18. Data were collected during a squall line that moved through the Norman, OK area on August 9, 2011 at approximately 0200 UTC. RHI scans at a single azimuth angle (no azimuthal scanning) and 1° × 1° angular resolution were used to achieve the high temporal sampling. Range corrected power for 0228:15, 0229:00 and 0231:36 UTC are given in the left column of Fig. 18.18 and DZ/dt for time intervals of 1, 45 and 154 s are given in the right column. Again, several entropic dipoles can be seen throughout the storm cross-sections, especially at the shorter two time intervals. Qualitatively, there is good agreement between the 1 and 45-s DZ calculations, particularly in the convective portions of the storm (4–5 km range) and along the gust front (8–9 km range). The dipole structure has significantly degraded by the 154-s interval, thus indicating that the time-span is dominated by advection. While advection plays a role in the 45-s interval as well, it can be argued that,

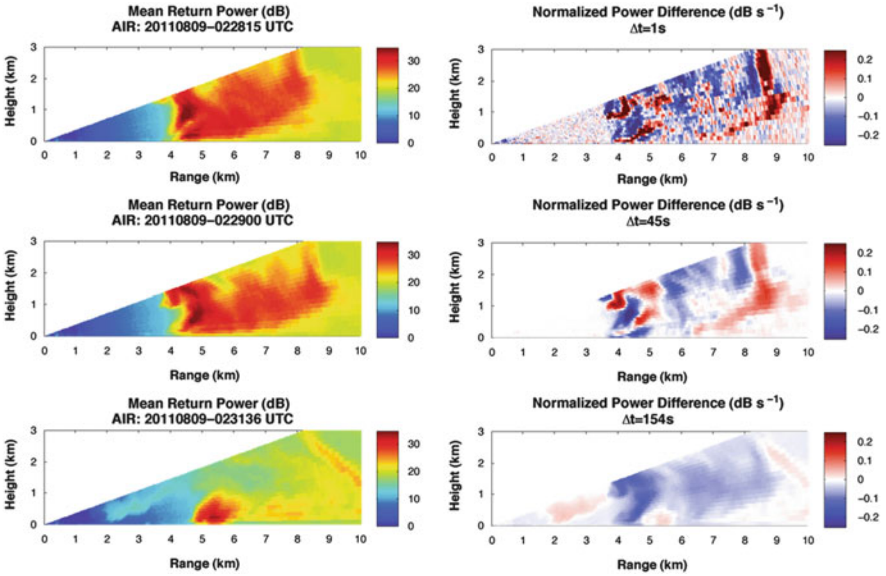


Fig. 18.18 Selection of temporal discretization of Z for entropy anomaly. Range corrected power calculated from the Atmospheric Imaging Radar (AIR) are shown in the *left column* at 0228:15, 0229:00 and 0231:36 UTC. The *right column* shows the calculated DZ values for temporal intervals of 1, 45 and 154 s. Note the dipole structure has degraded in the 154-s interval indicating advection dominates the microphysical phase changes within the radar resolution volume. A temporal resolution of less than 1 min is necessary to reduce the advection bias and obtain measurements appropriate for entropy estimation

since the dipole structure visible in the 1-s interval is still intact, the microphysical information required for the entropy derivation is still present and accurate.

From this experiment, it is determined that high temporal resolution is necessary for meaningful and accurate measurements of DZ, and thus entropy. Revisit times (Δt) of 1 min or less would be appropriate for reflectivity or power measurements and would ensure that environmental advection does not significantly bias the estimates for entropic balance theory.

18.13 Providing a Basis for Tornado Data Assimilation

The entropic balance theory was described and its theoretical applicability for tornadogenesis was shown in the Sects. 18.2, 18.3, 18.4, 18.5, 18.6, 18.7, 18.8, 18.9, and 18.10. Some key questions of tornado and environment (Appendix 1) are answered well by the entropic balance theory (Appendices 2 and 3). All of the governing equations of atmospheric dynamics, thermodynamics and mass continuity for the flow of high Reynolds number can be derived from the Lagrangian

(18.3) of the variational formalism (Sect. 18.2). The Reynolds number is estimated using molecular viscosity as 10^{9-12}. Consequently the molecular viscosity term can be neglected so that the flow is an ideal flow, but the nonlinear terms of the governing equations are fully retained, portion of which is expressed by eddy viscosity.

The entropic balance (18.5) is the sole diagnostic Euler-Lagrange equation and should be always satisfied for the all other prognostic Euler-Lagrange equations as discussed in Sect. 18.3 and in Fig. 18.3 as the completeness of solution. The vorticity (18.6) is derived from the entropic balance (18.6).

Using the variational analysis method (Sasaki 1970) with the constraints (18.5) and (18.6), it may be possible to get entropy, flow velocity, vorticity, the potentials α and β from the conventional and radar observation data. It is noted that the Lagrange multiplier potentials α and β are not easily expressed in terms of conventional meteorological field variables (which can be seen as a weakness of this approach), except in this article they are well interpreted as divergent part and rotational part respectively (Fig. 18.1). In the Sects. 18.11 and 18.12 of this article, preliminary research on uses of the radar data was described. The results are promising, although shown only for one case in each of the chapters. With the new radar variables DZ and DZ_{DR} etc., we may be able to extract cloud microphysical information from the storm. We plan to further test for a number of other cases by this approach to establish a solid basis for tornado data assimilation, because of the applicability of the entropic balance equation as a constraint with the radar observation in variational formulation.

Appendix 1 Some Key Questions on Tornadogenesis

Accurate forecasting of tornadogenesis is one of the unsolved problems, in spite of a great number of observations and research made over many decades. In recent years, significant progress has been made to understand the mechanism of tornadogenesis. However, there still remain key questions and difficult problems in fully understanding tornadogenesis and tornado.

Some of the key questions that need to be answered by any proposed theory of tornadogenesis are:

1. How does the mesocyclone develop? Note that mesocyclone and wall cloud are known as observed with tornadogenesis.
2. Why are hydrometeors able to be overshot against the upper air westerlies of much stronger wind speed than that of low level south-easterly inflow to tornadic storm?
3. Why are the locations of tornado and major precipitation regions spatially separate, not coincident?
4. Why is dry air aloft important for tornadogenesis?
5. Why do multiple vortices sometime develop before and during tornadogenesis?
6. How does the tornado develop in a hook echo, at the south-west corner, not at the center, of a supercell, and why is it associated with wall cloud?

7. Why does the tornado touch down in the perpendicular direction to the ground? Note that tornado should touch down to the ground in a parallel direction if the tornado is generated by the upward tilting theory of horizontal vortex.
8. Why is the tornado a phenomenon of low altitudes (<2–3 km) of the atmosphere?
9. What is the role of upward tilting of low level horizontal vorticity for tornadogenesis ?
10. What are necessary and sufficient conditions to separate tornadogenesis from non-tornadogenesis, in spite of several favorable conditions (such as favorable environmental soundings, supercell, mesocyclones, and RFD)?

These questions seemed to be well answered by the entropic balance theory (Sasaki 2009, 2010).

Appendix 2 Answer to the Questions 1, 2, 3 and 4

Firstly we consider a simple case where the linear slope between the entropic source and sink as shown in Fig. 18.5 and the flow velocity (rotational component) is horizontal and southerly. For this case, the direction of the vorticity $\boldsymbol{\omega}$ is vertical as given by (18.5) that is expressed by the entropic right-hand rule (Fig. 18.1), and is shown in Fig. 18.5 where the rotational flow direction of the vorticity $\boldsymbol{\omega}$ is shown by an arrow with double solid lines. The vorticity $\boldsymbol{\omega}$ represents mesocyclone existed indeed in the linear entropy slope between the entropic source and sink. It may answer the question 1.

The entropic anomaly in the air above the top of convective supercell is negative due to evaporation of the hydrometeors in the dry westerlies and radiative cooling, while the having the positive entropic source below. Accordingly, the maximum spatial gradient of the entropy anomaly between the points just above the supercell ($S' < 0$) and at the convective center of the supercell ($S' > 0$) is directing straight downwards. With the upper air westerlies wind vector and the entropic spatial gradient vector, the vorticity $\boldsymbol{\omega}$ is that of the horizontal vortex tube as shown at the top of the supercell in Fig. 18.5. The upper vortex seems the key to overshoot the hydrometeors generated in the supercell updraft, westwards against the headwind westerly jet. This may answer the question 2. From the above answer to the questions 1, and 2, it is apparent that the question 3 is answered from the above answer to the questions 1 and 2. To answer the both questions 1, 2 and 3, it is noted that the dry air plays the key roles, and it may answer the question 4.

Appendix 3 Answer to the Questions 5, 6, 7, 8, 9 and 10

The “wrap-around mechanism” is explained in a simple way by using the diagnostic E-L equation (18.5) and the entropic right-hand rule (Fig. 18.1) of (18.6). Figure 18.6 is prepared which has opposite sign of spatial entropy gradient and the flow velocity direction (rotational component) and produces the same vorticity of Fig. 18.1. It is called as conjugate vortex (or vorticity) in this article. Combination of the original vortex and its conjugate produces multi-vortexes as well as single vortex. It answers the question 5. The answer for the questions 6 and 8 may be easily found from Fig. 18.4.

The answer for the questions 7 and 8 is discussed in Sect. 18.7. The question 9 may be answered in Sect. 18.5. The answer to the historically long-standing question 10 is a tough one, but may be hinted by the entropic balance theory discussed in this article and in an example shown in the following Appendix 4. It is challenging for continuing research to find a full answer for the question 10.

Appendix 4 Entropy Variation and Tornadogenesis

The entropy variation due to cloud-physical phase change is computed at the altitudes of 1–3 km where condensation and evaporation to provide thermodynamical effects for development of mesocyclones and tornado, and the atmospheric pressure of approximately 750 mb and temperature of 0° C (273°K) as an example. For simplicity for this preliminary investigation, we assume also that $S_0 = 0$ and only consider the diabatic effects of water molecules on S of the surrounding air on a moving coordinates with tornado.

The entropy change Δ_c S of the surrounding air due to water vapor condensation measured at 100 °C and 1,013 mb is estimated as 109.0 J°K^{-1}mol^{-1}, and that of evaporating of water droplet −109 J°K^{-1} mol^{-1}. Since moisture measurement is not considered in this preliminary investigation and insufficient measurement and knowledge on the cloud-physical phase changes of actual cloud, the estimates were made simply based on the measurements of heat in published chemical experiments. Their values are adjusted to the value of 0 °C and 750 mb for representing the altitude of 1–3 km, using the standard adjustment processes (Atkins and de Paula 2002; Watanabe 2003).

The adjustment amount due to the temperature change Δ_T S (100 °C — > 0 °C) = – 16.6 J°K^{-1}mol^{-1} and that due to pressure change Δ_p S (1,013 mb — > 750 mb) = 2.1 J°K^{-1}mol^{-1}.

After the adjustments, the entropy change of the surrounding air due to condensation of water vapor is; Δ_c S = (109.0 − 16.6 + 2.1) J°K^{-1}mol^{-1} = 94.5 J°K^{-1}mol^{-1}, and for that due to evaporation of water droplets is Δ_e S = (−109.0 – 16.6 + 2.1) J°K^{-1}mol^{-1} = −123.5 J°K^{-1}mol^{-1}.

Thus we get the entropy difference between the entropic source and sink separated by the distance d; Δ_d S = (94.5 – (−123.5)) = 218.0 J°K^{-1}mol^{-1}.

Similarly, the absolute entropy S is calculated by adding the entropy changes due to melting of ice, 22.0 J°K^{-1}mol^{-1} and the residual entropy, 0.8 J°K^{-1}mol^{-1} from the Boltzmann's third law of thermodynamics, resulting S = (94.5 + 22.0 + 0.8) J°K^{-1}mol^{-1} = 117.3 J°K^{-1}mol^{-1}.

From (18.6), the vorticity ω is written as

$$\omega = \mathbf{v}_\beta \times (1/\mathrm{S})\,\mathrm{S}. \tag{18.51}$$

where Vrot represents the rotational component of flow velocity.

Usin the estimated values of S and Δ S, (18.51) becomes

$$\omega = \mathrm{V}_\beta \times 1.86(= 273.0/117.3)/\mathrm{d} \quad \mathrm{s}^{-1} \tag{18.52}$$

where d is distance between the entropic source and sink.

For an example of mesocyclone cases, $\mathbf{v}_\beta$ is taken 10 m/s and d as 5 km, then (18.52) leads $\omega = 0.0037\,\mathrm{s}^{-1}$. For tornado by wrap-around mechanism cases, (18.52) with 50 m/s of V_β and 100 m of d leads $\omega = 0.93\,\mathrm{s}^{-1}$. The former and latter seem appropriate order of magnitudes for mesocyclones and tornado respectively.

Acknowledgements The authors would like to express their appreciation for Dr. Paul Markowski and his colleagues of VORTEX2 for allowing us to use their schematic diagram of summarizing their findings in this article (Fig. 18.13). The first author would like to express his appreciation to the encouragement and support given to this research by The Office of Vice President for Research, and National Weather Center (NWC), The University of Oklahoma, USA, and Weathernews, Inc., Japan. Also, the first author would like to express his appreciation to faculty and scientists especially of NWC, The University of Tokyo, Kyoto University, Japan, and Ewha Women's University, South Korea for their useful discussion, and to Vivek Mahale of the university of Oklahoma for his assistance of computer graphics.

References

American Meteorological Society (2000) Glossary of meterorology, 2nd edn. American Meteorological Society, Boston, 855 pp

Atkins P, Paula J (2002) Atkins' physical chemistry, 7th edn. Oxford University Press Inc., New York, 1150 pp

Bateman H (1932) Partial differential equations of mathematical physics. Cambridge Univrsity Press, Cambridge (reprint by Dover Publ., 1944), 522 pp

Bohren CF, Albrecht BA (1998) Atmospheric thermodynamics. Oxford University Press, New York, 402 pp

Burgess DW (1997) Tornado warning guidance. OSB/OTB, Norman, Oklahoma, 28 pp

Clebsch A (1859) Uber rein algemaine transformation der Hydrodynamischen Gleichungen. Crelle J fur Math 54:293–312

Davies-Jones R (1984) Streamwise vorticity: the origin of updraft rotation in supercell storm. J Atmos Sci 41:2991–3006

Davies-Jones R, Trapp J, Bluestein HB (2001) Tornadoes and tornadic storms. In: Doswell CA, III (eds) Severe convective storms, American meteorological society, Bostan, pp 167–222.

Doviak RJ, Zrnic DS (1984) Doppler radar and weather observations. Academic, New York, 458 pp

Dutton JA (1976) The ceaseless wind, an introduction to the theory of atmospheric motion. McGrow-Hill, USA, 579 pp

Isom B, Palmer R, Kelley R, Meier J, Bodine D, Yeary M, Cheong B-L, Zhang Y, Yu T-Y, Biggerstaff M (2011) The atmospheric imaging radar (AIR) for high-resolution observations of severe weather. Radar conference (RADAR), 2011 IEEE pp 627–632, 23–27 May 2011

Klemp JB, Wilhelmson RB (1978) Simulations of right- and left-moving storms produced thunderstorm splitting. J Atmos Sci 35:1097–1110

Klemp JB, Wilhelmson RB, Ray PS (1981) Observed and numerically simulated structure of a mature supercell thunderstorm. J Atmos Sci 20:1558–1580

Kumjian MR, Ryzhkov AV (2010) The impact of evaporation on polatimetric characteristics of rain: theoretical model and practical implication. J Appl Meteor Climatol 49:1247–1267

Kumjian MR, Ryzhkov AV, Melnikov VM, Schuur TJ (2010) Rapid-scan super-resolution observations of a cyclic supercell with a dual- polarization WSR-88D. Mon Wea Rev 138 :3762–3785

Lamb H (1932) Hydrodynamics, 2nd edn. Cambridge University Press/Dover Publication, New York, p 738

Lemon LR, Doswell III CA (1979) Severe thunderstorm evolution and meso-cyclone structure as related to tornadogenesis. Mon Wea Rev 107:1184–1197

Leslie LM (1971) The development of concentrated vortices: a numerical study. J Fluid Mech 49:1–21

Lewis JM, Lakshmivarahan S (2008) Sasaki's pivotal contribution: calculus of variations applied to weather map analysis. Mon Wea Rev 136:3553–3567

Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, Cambridge, 654 pp

Li X, Srivastava RC (2001) An analytical solution for raindrop evaporation and its application to radar rainfall measurement. J Appl Meteor 40:1607–1616

Lilly DK (1982) The development and maintenance of rotation in convective storms. In: Bengtsson L, Lighthill J (eds) Intense atmospheric vortices, Springer, New York, pp 149–160

Lilly DK (1986) The structure, energetics and propergation of rotating convective storms. Part II: helicity and storm stabilization. J Amer Meteor Soc 43:126–140

Markowski P, Richardson Y, Marquis J, Wurman J, Kosiba K, Robinson P, Dowell D, Rasmussen E, Davies-Jones R (2012a) The pretornadic phase of the Goshen County, Wyoming, Supercell of 5 June 2009 intercepted by VORTEX2. Part I: evolution of kinematic and surface thermodynamic fields. Mon Weather Rev 140:2887–2915

Markowski P, Richardson Y, Marquis J, Davies-Jones R, Wurman J, Kosiba K, Robinson P, Rasmussen E, Dowell D (2012b) The pretornadic phase of the Goshen County, Wyoming, Supercell of 5 June 2009 intercepted by VORTEX2. Part II: intensification of low-level rotation. Mon Weather Rev 140:2916–2938

Noda A (2002) Numerical simulation of supercell tornadogenesis and its structure. (Document of Science Dissertation in Japanese). Tokyo University, Tokyo, 93 pp

Noda A, Niino H (2005) Genesis and structure of a major tornado in numerically-simulated supercell storm: Importance of vertical velocity in a gust front. SOLA Meteor Soc Jpn 1: 5–8

Noda A, Niino H (2010) A numerical investigation of a supercell tornado: genesis and vorticity budget. J Meteo Soc Jpn 88:135–159

Pruppacher HR, Klett JD (1997) Microphysics of clouds and precipitation. Academic, The Netherlands, 954 pp

Sasaki Y (1955) A fundamental study of numerical prediction based on the variational principle. J Meteo Soc Jpn 33:262–275

Sasaki Y (1958) An objective analysis based on the variational method. J Meteo Soc Jpn 36:77–88

Sasaki Y (1970) Some basic formalisms in numerical variational analysis. Mon Wea Rev 98:875–883

Sasaki YK (1999) Tornado and hurricane: needs of accurate prediction and effective dissemination of information (in Japanese). J Visualization Soc Jpn 19(74):187–192 (Sasaki Y has been changed to Sasaki YK since 1974)

Sasaki YK (2009) Real challenge of data assimilation for tornadogenesis. In: Park SK, Liang Xu (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, Springer, Berlin/Heidelberg, pp 97–126

Sasaki YK (2010) Entropic balance theory and tornadogenesis. NOVA Science Publishers, New York, 39 pp

Smith RM, Leslie LM (1978) Tornadogenesis. Q J R Met Soc 104:189–199

Stensrud D (2007) Parameterization schemes: key to understanding numerical weather prediction models. Cambridge University Press, Cambridge, UK, 459 pp

Trapp RJ, Davies-Jones R (1997) Tornadogenesis with and without a dynamical pipe effect. J Atmos Sci 54:113–133

Watanabe H (2003) From entropy to chemical potential (in Japanese). Shokabo, Tokyo, 145 pp

Weisman ML, Klemp JB (1982) The dependence of numerically simulated convective storm on vertical wind shear and buoyancy. Mon Wea Rev 110:504–520

Wilhelmson RB, Wicker LJ (2001) Numerical modeling of severe local storms. In: Doswell III CA (ed) Severe convective storms, American Meteorological Society, Boston, pp 123–166

Xue M, Droegemeier KK, Wong W, Shapiro A, Brewster K (1995) ARPS version 4.0 user's guide. The Center for Analysis and Prediction of Storms, University of Oklahoma, p 380

Yoshida Z (2009) Clebsch parameterization-basic properties and remarks on its applications. personal communication, p 9

Yoshizawa A (2001) Fluid dynamics (in Japanese). Tokyo University Press, Tokyo, p 344

Chapter 19
All-Sky Satellite Radiance Data Assimilation: Methodology and Challenges

Milija Zupanski

Abstract Assimilation of satellite radiances is the backbone of today's operational data assimilation. Satellites can cover all parts of globe and provide information in areas not accessible by any other observation type. Of special interest are high-impact weather areas, such as tropical cyclones and severe weather outbreaks, which are mostly covered by clouds. Unfortunately, in current operational practice only clear-sky satellite radiances are assimilated, with only few exceptions. This effectively filters out a potentially useful information from all-sky radiances related to clouds and microphysics, and consequently limits the utility of satellite data. In this paper we will address numerous challenges related to the use of all-sky satellite radiances.

All-sky satellite radiances present a formidable challenge for data assimilation as they relate to numerous technical aspects of data assimilation such as: (1) forecast error covariance, (2) correlated observation errors, (3) nonlinearity and non-differentiability, and (4) non-Gaussian errors. Assimilation of all-sky radiances is also challenging from a dynamical/physical point of view, since observing clouds implies a need for better understanding and ultimately simulation of cloud microphysical processes. Given that a reliable prediction of clouds requires a high-resolution cloud-resolving model, assimilation of all-sky radiancesis also a high-dimensional problem that requires addressing computational challenges.

M. Zupanski (✉)
Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, U.S.A.
e-mail: zupanskim@cira.colostate.edu

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_19,

19.1 Introduction

Satellites, radars, and other remote sensing are the major sources of information about atmosphere and oceans with invaluable implications for weather, climate, and hydrology. Satellites can cover all parts of globe and can provide information in the areas not accessible by any other observation type. It is not surprising that satellite radiance observations are widely used in data assimilation and numerical weather prediction (NWP). Among most relevant data assimilation and NWP applications are for high-impact weather events such as tropical cyclones and severe weather outbreaks, which are mostly covered by clouds. Unfortunately, in current operational practice only clear-sky satellite radiances are commonly assimilated, with rare exceptions. This effectively filters out potentially useful information from cloud and precipitation affected radiances and consequently limits the utility of satellite data. In this paper we will address several challenges related to the use of all-sky (i.e. combined clear-sky and cloud/precipitation affected) radiances in data assimilation.

The relevance of all-sky radiance assimilation is widely recognized and discussed (e.g., Errico et al. 2007a, b; Auligne et al. 2011; Bauer et al. 2011). Challenges of all-sky satellite radiances for data assimilation and prediction can be all traced back to clouds and precipitation. Cloud microphysical processes are highly nonlinear, discontinuous, and characterized by very small spatial scales of the order of hundreds of meters to a kilometer, requiring high-resolution and complex modeling. Consequently, data assimilation of clouds and precipitation is also nonlinear and high-resolution. High spatiotemporal resolution of cloud processes has a direct consequence on computations, posing an additional challenge in practical applications.

In this paper we focus on addressing the data assimilation challenges of all-sky satellite radiance assimilation related to variational and ensemble data assimilation, since they represent the most relevant methodologies used today. In Sect. 19.2 we present the current status of data assimilation in reference to all-sky radiance assimilation, and discuss several major challenges in detail in Sect. 19.3. We summarize the issues and look at the future of all-sky satellite radiance assimilation in Sect. 19.4.

19.2 Current Status

In this section we introduce a more formal overview of data assimilation methodologies currently used in research and operations, and also describe current status of data assimilation with respect to all-sky radiance assimilation.

19.2.1 Data Assimilation Overview

We assume that reader is at least partially familiar with data assimilation, thus we will only describe data assimilation methodologies in general. This section will also serve to introduce the notation.

Data assimilation (DA) may be referred to as a mathematical algorithm that provides optimal combination of observations and model prediction. DA produces optimal estimates of the model state vector (e.g., analysis), as well as its uncertainty, typically represented by the analysis error covariance matrix. "Optimal" is commonly defined as a minimum variance or a maximum likelihood estimate (e.g., Jazwinski 1970). Current data assimilation methodologies generally rely on the use of Bayes formula for describing conditional probability density function (pdf). A detailed overview of data assimilation methods can be found in books by Daley (1993), Kalnay (2003), Lewis et al. (2006), and Evensen (2009).

We briefly describe two major DA methodologies, variational and ensemble. Variational DA is commonly used in operational weather centers, while ensemble DA is mostly used in research and is making progress towards operational use. Also, there are a variety of sub-methods and hybrid methods that combine the two methodologies.

Following Lorenc (1986), for Gaussian probability distribution one can derive a cost function by taking a negative logarithm of the Bayes formula for posterior probability

$$J(x) = \frac{1}{2}[x - x^f]^T \; P_f^{-1}[x - x^f] + \frac{1}{2}\,[y - h(x)]^T \; R^{-1}\,[y - h(x)] \qquad (19.1)$$

Where x denotes the state vector, y is the observation vector, the superscript f denotes forecast, h is a nonlinear observation operator and P_f and R are the forecast and observation error covariances, respectively. Although this cost function is typically mentioned in variational methods, it is important to note that this function is also relevant for Kalman filter based methodologies, including ensemble Kalman filter (Li and Navon 2001). As shown in Jazwinski (1970), minimization of the cost function (19.1) defined for linear observation operator produces the Kalman filter analysis solution formally obtained using Newton method for minimization of quadratic function (e.g., Luenberger 1989). This apparent equivalence between the minimum variance and maximum likelihood estimates is ultimately a consequence of using Gaussian pdfs for which the mean and the mode are identical.

Variational and ensemble data assimilation methods can be applied sequentially, in which case they are referred to as filters, or in batch mode, in which case they are referred to as smoothers. Variational filtering method is called the three-dimensional variational (3d-Var) DA, while when it is used as a smoother it is referred to as the four-dimensional variational (4d-Var) DA. Ensemble data assimilation methodologies are mostly applied sequentially, although there is a possibility to apply them in a smoother framework (e.g., Evensen and van Leeuwen 2000).

The main assumption of variational methods is related to the forecast error covariance, which is modeled, and thus generally static without time dependence. Another characteristic of variational methods is that they are developed around an iterative minimization algorithm, typically of unconstrained type (e.g., conjugate-gradient, quasi-Newton), making them suitable for nonlinear processes and operators.

Ensemble DA methods use ensemble of forecast models to define a time-dependent forecast error covariance, however for the price of being a reduced rank approximation. They are algorithmically simpler than variational methods, thus easier to develop and maintain. Although ensemble DA can handle very well nonlinearities of the forecast model, their straightforward application is not very good for addressing nonlinearities of observation operator because the analysis is based on using the Kalman filter linear solution.

19.2.2 Assimilation of All-sky Radiances

Most operational weather centers are currently using variational DA methods, although they actively investigate ensemble and hybrid variation-ensemble methods. On the other side, ensemble and hybrid variational-ensemble methodologies are typically developed at research laboratories and universities. However, this distinction is not that clear and there are several research data assimilation algorithms being tested for operational use. Even though there is a wealth of information that all-sky radiances could bring to the prediction system, only a limited research effort to assimilate such observations exists. Most efforts include the use of variational methods (e.g., Vukicevic et al. 2004; Bauer et al. 2006, 2010; Geer et al. 2010; Geer and Bauer 2010; Polkinghorne and Vukicevic 2011) with some applications within ensemble and hybrid variational-ensemble methods (e.g., Zupanski et al. 2011a, b; Zhang et al. 2012).

Especially relevant is the pioneering work by satellite research group at the European Centre for Medium Range Weather Forecast (ECMWF) (e.g., Bauer et al. 2010; Geer et al. 2010) leading to the first operational assimilation of all-sky radiances, since 2009. There are similar efforts to assimilate all-sky satellite radiances in the United States at the National Centers for Environmental Prediction (NCEP), and other centers will likely follow.

19.3 Challenges

In general, challenges of all-sky satellite radiance assimilation originate due to their relation to clouds. Observing and simulating clouds is challenging in it own right. This is magnified in data assimilation, being a method that combines information from observations and from prediction models. Although problems related to accurate and efficient assimilation of all-sky radiances are fundamentally related to each other, one could try to distinguish the challenges related to (1) data assimilation, (2) simulation and prediction, and (3) computation. Simulation and prediction of clouds is related to the ability of prediction models to represent clouds, and the complexity of the employed microphysics. Although this clearly impacts data assimilation, it is typically assumed an input to data assimilation and thus it will not be discussed here. Computational requirements for all-sky radiance assimilation we refer to are caused by high spatiotemporal resolution of cloud microphysical processes, as well as by a necessity to include cloud scattering processes in forward radiative transfer model. Computational restrictions will impact the choices one could have regarding methodology and algorithms used in data assimilation. In this paper we will focus on data assimilation issues related to all-sky satellite radiances, and discuss model prediction and computational issues only in context of data assimilation.

Data assimilation challenges of all-sky radiance assimilation are defined as the aspects of data assimilation that are especially exposed by assimilation of all-sky radiances and related cloud-resolving scales. They can be all traced back to clouds, and range from methodological to computational: (1) forecast error covariance, (2) correlated observation errors, (3) nonlinearity and non-differentiability, and (4) non-Gaussian errors.

19.3.1 Forecast Error Covariance

Forecast error covariance is typically used as a measure of uncertainty of the forecast, and could be defined as

$$P_f = \left\langle (x^f - x^t)(x^f - x^t)^T \right\rangle \tag{19.2}$$

Where x^f and x^t are the first-guess forecast and the (unknown) truth, respectively, $\langle \cdot \rangle$ denotes mathematical expectation and the superscript T denotes a transpose. It also represents one of the main differences between variational and ensemble based data assimilation methodologies: P_f is modeled in variational methods, while computed from a forecast ensemble in ensemble methods. This implies time-independent covariance in variational methods, while ensemble methods produce a flow-dependent structure. Time-dependence is in principal an advantage for applications at cloud-scales with characteristic fluctuating dynamical processes. One can think of cloud microphysical processes in a hurricane, or in severe storm outbreaks. However, one should also be aware that ensemble data assimilation at such high resolution implies a low-rank approximation to the forecast error covariance, with the number of ensembles much smaller than the state dimension. Although this issue can be considerably improved by error covariance localization techniques (e.g., Hamill et al. 2001; Houtekamer and Mitchell 2001), this is still a limitation.

We now describe how the choice of static, but full-rank, versus flow-dependent, but reduced-rank error covariance could impact all-sky radiance assimilation.

Forecast error covariance has a fundamental role in data assimilation as it defines the subspace where the analysis correction can be defined (Appendix 1). Following the relations (19.26) and (19.30) from Appendix 1, one can represent a generic analysis increment as

$$x^a - x^f = \sum_i \beta_i u_i \tag{19.3}$$

where β_i is a coefficient equal to γ_i and η_i defined in Appendix 1, and u is a singular vector. Therefore, an arbitrary analysis increment can be represented as linear combination of forecast error covariance singular vectors.

The implication of (19.3) is that a well-defined forecast error covariance is critical for successful data assimilation. Therefore, the quality of data assimilation can be assessed by examining the structure of forecast error covariance used in assimilation. In weather, climate, hydrology, and other geoscience applications the structure of true forecast error covariance can be very complex, since it incorporates relations between various state variables. Of special interest for all-sky radiance assimilation is the structure of forecast error covariance with respect to cloud microphysical variables since cloud variables are input to radiative transfer model. In general, there are various processes that imply cross-correlation between cloud variables, approximately described by the cloud microphysics component of a forecast model. One can also anticipate correlations between cloud and standard dynamical variables, such as temperature, pressure and wind. It is convenient to use a block matrix form to represent forecast error covariance

$$P_f = \begin{bmatrix} P_{dd} & P_{cd}^T \\ P_{cd} & P_{dd} \end{bmatrix} \tag{19.4}$$

where index d refers to dynamical variables, and index c to cloud variables (e.g., cloud ice, snow, rain, etc.).We also use the fact that P_f is symmetric matrix, thus $P_{cd}^T = P_{dc}$. For simplicity, assuming that dynamical variables include temperature, pressure and wind, and that cloud variables include cloud ice, snow and rain, the block matrices in (19.4) are

$$P_{dd} = \begin{bmatrix} P_{T,T} & P_{T,p} & P_{T,v} \\ P_{T,p} & P_{p,p} & P_{p,v} \\ P_{T,v} & P_{p,v} & P_{v,v} \end{bmatrix}, \tag{19.5}$$

$$P_{cc} = \begin{bmatrix} P_{ice,ice} & P_{ice,snow} & P_{ice,rain} \\ P_{ice,snow} & P_{snow,snow} & P_{snow,rain} \\ P_{ice,rain} & P_{snow,rain} & P_{rain,rain} \end{bmatrix}, \tag{19.6}$$

$$P_{cd} = \begin{bmatrix} P_{ice,T} & P_{ice,p} & P_{ice,v} \\ P_{snow,T} & P_{snow,p} & P_{snow,v} \\ P_{ice,T} & P_{snow,p} & P_{rain,v} \end{bmatrix}. \tag{19.7}$$

The diagonal blocks (19.5) and (19.6) are symmetric matrices, while the off-diagonal block matrix (19.7) is not symmetric.

The forecast error covariance structure defined by (19.4)–(19.7) is indicative of the complexity of relations that this matrix represent. Given that these matrices represent the uncertainty of model variables, one can quickly realize that elements of these matrices are fundamentally time dependent. The natural formation and decay of clouds will have a profound impact on the elements of matrices (19.6) and (19.7). In clear skies these matrices have all elements essentially equal to zero. When clouds begin forming, the block matrices describe how various variables impact each other in the process. Unfortunately, there is a limited capability of current data assimilation methodologies to accurately address the structure (19.4)–(19.7).

Variational methods include modeling of forecast error covariance and typically do not represent time-dependent information in its definition. One should note that 4d-Var method includes time-dependence through tangent linear and adjoint model integration, which does have some impact on the uncertainties at the end of assimilation interval. However, the forecast error covariance defined at initial time of assimilation is modeled as in 3d-Var data assimilation. For dynamical variables one can identify simplified relations such as hydrostatic, geostrophic, and similar balance constraints that are commonly used in modeling cross-variable interactions (e.g., Parrish and Derber 1992). Unfortunately, this approach is much more difficult to apply at cloud scales due to poorly known or unknown balance constraints. This apparently creates a difficulty for variational data assimilation to represent cloud variable cross-correlations in (19.6), as well as dynamical-cloud correlations in (19.7). Although in principle it may be possible to successfully model cross-variable correlations, this still has not been done for cloud variables. A more feasible solution applicable to current variational methods is to assume a regular (i.e., isotropic and homogeneous) correlation for the diagonal blocks in (19.6), i.e. to pre-define correlation function for $P_{ice,ice}$, $P_{snow,snow}$, and $P_{rain,rain}$ and thus avoid modeling more complex cross-correlations. In this case the matrices P_{cd} and P_{cc} become

$$P_{cc} = \begin{bmatrix} P_{ice,ice} & 0 & 0 \\ 0 & P_{snow,snow} & 0 \\ 0 & 0 & P_{rain,rain} \end{bmatrix}, \tag{19.8}$$

$$P_{cd} = \begin{bmatrix} 0\;0\;0 \\ 0\;0\;0 \\ 0\;0\;0 \end{bmatrix}. \tag{19.9}$$

On the other hand, ensemble-based forecast error covariance has a potential to capture all inter-variable correlations, and is also inherently time-dependent. However, even with error covariance localization, the low-rank limitation of ensemble error covariance does not allow accurate representation of all cross-correlations between variables. Fortunately, there are still inter-variable correlations that can be represented well in ensemble-based methods. Typically, a correlation between variables at near-by points is well represented even by a limited size ensemble. This property essentially allows ensemble forecast error covariance to include all terms defined by (19.4)–(19.7), however with reduced accuracy. The practical problem is how to distinguish between "good" and "bad" correlations, and the answer is not yet clear. More aggressive localization may have an appearance of better controlling cross-variable correlations, but it does potentially impact dynamical balance of the analysis and would prevent some important correlations in vertical for well-developed cloud systems. Alternatively, one could introduce other localizing functions that would selectively apply localization to off-diagonal matrix blocks depending on the variable.

In addition to algebraic representation of all-sky assimilation issues with respect to forecast error covariance described above it is also instructive to visually examine its structure. The structure can be inspected by plotting columns of the forecast error covariance that is also related to "single-observation" data assimilation experiments. Let define a vector with all zero elements except for the i-th element with the value one

$$z_i = \left[0_1 \cdots 0_{i-1} \; 1_i \; 0_{i+1} \cdots 0_{Ns} \right]^T \tag{19.10}$$

where the index refers to a grid point and a variable (e.g., index of the state vector), and N_S is the dimension of state vector. After multiplying vector z_i by matrix P_f one obtains the i-th column of the forecast error covariance matrix

$$c_i = P_f z_i = \left[f_1^i \rightleftharpoons f_{i-1}^i \;\; f_i^i \;\; f_{i+1}^i \rightleftharpoons f_{Ns}^i \right]^T \tag{19.11}$$

with f_j^i representing the i-th column value at location j. Note that location refers to a grid point and variable. Following Thepaut et al. (1996) and Huang et al. (2009), one can derive the analysis increment for single observation at i-th point

$$x^a - x^f \;\propto\; P_f \left[y - h(x^f) \right]_i \tag{19.12}$$

Applying the matrix–vector product in (19.12), and using (19.11)

$$x^a - x^f \;\propto\; \left[y - h(x^f) \right]_i c_i \tag{19.13}$$

i.e. the analysis increment is simply the i-th column of forecast error covariance scaled by the observation increment. This result, also expected on the basis of (19.3), allows us to interpret a column of the forecast error covariance as analysis response, and thus give a physical meaning to the structure of forecast error covariance.

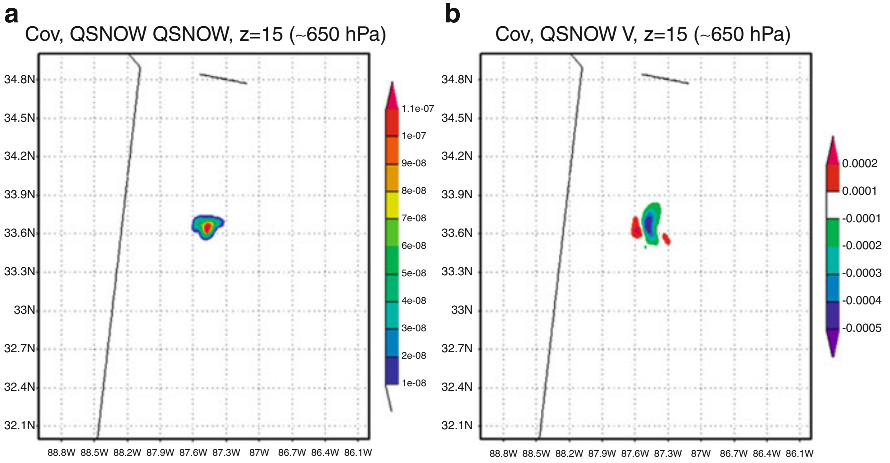


Fig. 19.1 Horizontal analysis response to a single observation of cloud snow at 650 hPa on 09 September 2012 at 18 UTC: (**a**) cloud snow analysis increment, and (**b**) north–south wind analysis increment. The results are shown for the inner nest at 3 km horizontal resolution

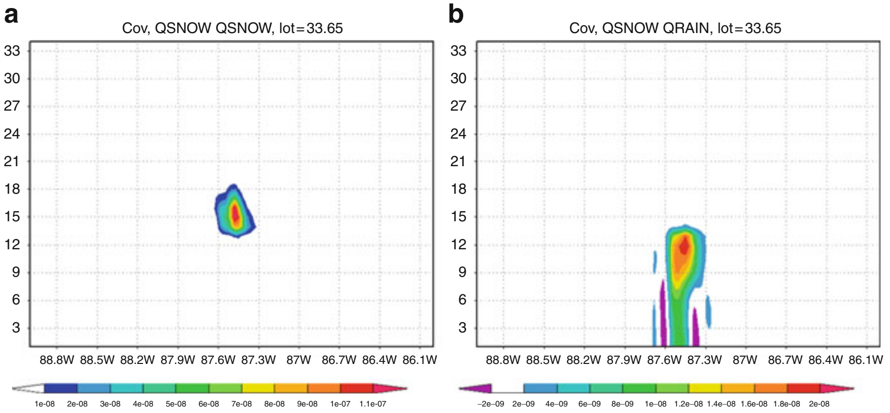


Fig. 19.2 Vertical analysis response to a single observation of snow at 650 hPa on 09 September 2012 at 18 UTC for: (**a**) cloud snow analysis increment, and (**b**) rain analysis increment. The results are shown for the inner nest at 3 km horizontal resolution

One such example using ensemble error covariance is shown in Figs. 19.1 and 19.2, where the analysis response to a single observation of cloud snow at 650 hPa is plotted. This corresponds to a hypothetical observation of high-frequency microwave all-sky satellite radiance that is sensitive to cloud snow and ice. The results are obtained using the Weather Research and Forecasting (WRF) model (e.g., Skamarock et al. 2005) at 9 km/3 km resolution and the Maximum Likelihood Ensemble Filter (MLEF) data assimilation algorithm (e.g., Zupanski 2005; Zupanski et al. 2008). The control variables include dynamical variables (e.g., perturbation pressure, perturbation height, perturbation potential temperature, and winds) as well as cloud variables (e.g., cloud ice, cloud snow, cloud water, graupel, and water vapor). In Fig.19.1a, b we show a horizontal map of analysis increment for cloud snow and for the north–south wind component at the level of the cloud snow observation. One can see that snow analysis has a strong positive response to snow observation (Fig.19.1a), as expected. It is also interesting to note that cloud snow observation impacts wind (Fig.19.1b), a dynamical variable, corresponding to $P_{snow,v}$ component of the forecast error covariance from (19.7). This is important since it indicates that, with adequate forecast error covariance structure, one can impact dynamical variables by all-sky radiance observations. One can also note a relatively regular response that resembles modeled error covariance structure (e.g., Parrish and Derber 1992; Wu et al. 2002), possibly suggesting that such covariance components can be successfully modeled in variational methods.

In Fig.19.2 we show cloud snow and rain analysis responses in the vertical, corresponding to the components $P_{snow,snow}$ and $P_{snow,rain}$. One can see a well-defined cloud snow response centered at the observation location (Fig.19.2a). The response is confined to few levels above and below the observation, again suggesting that modeling this covariance component may be possible. However, the response of cloud rain (Fig.19.2b) exposes a potential difficulty in modeling cross-variable correlations such as snow-rain. The first problem is to create a non-centered response of rain to cloud snow observation. Although this may be mathematically possible (e.g., Gaspari and Cohn 1999), it has not been done in practice and opens several new problems. One such problem is to know what exactly needs to be modeled, because there is a very limited knowledge about cloud-variable correlation statistics. The most difficult problem may be related to flow-dependence of these correlations. It is clear that the existence of cloud rain and snow depends on the current temperature conditions that change with time and thus require additional flow-dependent parameters to be introduced to the modeling function and eventually estimated.

Therefore, forecast error covariance can have very different structure depending on the methodology used. Even within same methodology one can choose different parameters related to decorrelation length of correlation function in variational methods, or to the covariance localization in ensemble methods, effectively implying a large number of possible choices for forecast error covariance. This apparent variety of possible choices for the forecast error covariance creates a problem since in light of (19.3) it implies a non-unique analysis solution. The "optimal" choice of forecast error covariance may be good for overall data assimilation performance, but may not adequately address all-sky radiance observations since they are generally confined to a smaller local area of intense dynamical development and thus their global impact is relatively small. One possible way to address the problem of non-uniqueness of forecast error covariance is to reduce the number of additional parameters, or at least to include their estimation in the data assimilation algorithm. Another possibility may be to develop a new methodology that will be less dependent on forecast error covariance and include fewer undefined parameters.

19.3.2 Correlated Observation Errors

The number of satellite radiance observations can dramatically increase when cloudy radiances are included. This immediately opens several new data assimilation issues such as observation error correlations and computational overhead. Consider a commonly used observation equation

$$y = h(x) + \varepsilon \tag{19.14}$$

Where ε is a Gaussian random variable $N(0, R)$ and

$$R = \left\langle \varepsilon\varepsilon^T \right\rangle \tag{19.15}$$

The observation error covariance matrix R is typically defined as diagonal, implying uncorrelated observation increments. This assumption greatly simplifies data assimilation that requires the inverse of R, and is also relatively accurate if observations are not very close to each other. However, when observations are densely distributed the uncorrelated observation error assumption may not be justified. The ultimate consequence of correlated observation errors is that the information content of near-by observations is reduced compared to their independent information. This intuitive conclusion can be formalized using mathematical framework of Shannon information theory (Shannon and Weaver 1949; also Appendix 2). Let Y_1 and Y_2 represent random observation errors for two near-by observations. Using a general relationship between entropy H and mutual information I (e.g., Cover and Thomas 2006)

$$I(Y_1; Y_2) = H(Y_1) + H(Y_2) - H(Y_1, Y_2) \tag{19.16}$$

as well as (19.34) from Appendix 2

$$I(Y_1; Y_2) = I(Y_1; Y_1) + I(Y_2; Y_2) - H(Y_2, Y_2). \tag{19.17}$$

Since by definition $H(Y_1, Y_2) \geq 0$ for arbitrary random variables Y_1 and Y_2, we have

$$I(Y_1; Y_2) \leq I(Y_1; Y_2) + I(Y_2; Y_2). \tag{19.18}$$

The relation (19.18) states that mutual information of dependent variables is smaller than mutual information of independent variables. Since in Gaussian framework the notion of dependence is directly related to correlations, one can say that correlated observations bring less information than uncorrelated observations. This conclusion implies a need to account for correlated observation in all-sky satellite radiance assimilation.

There are several possible ways one could address observation error correlations in data assimilation:

1. *Increase observation errors*: If observation density is high, reduce the impact of dense observations by increasing the observation error.

2. *Observation thinning*: If observation density is high, thin observations by selecting every n-th observation. The reduced number of observations could keep the original error, or have some error adjustment.
3. *Cut-off based selection* (Fertig et al. 2007): Based on an empirical estimate of observation correlations one can design an algorithm to select which radiance observations to assimilate.
4. *Eigenvalue decomposition* (Parrish and Cohn 1985; Anderson 2003): For non-diagonal R work in eigenvectors space where the errors are diagonal, i.e. $R = S\Lambda S^T$. Introduce the change of variable $S^T\,[y - h(x)] = S^T\,\varepsilon$ to obtain transformed error as $R_s = \left\langle (S^T\varepsilon)(S^T\varepsilon)^T \right\rangle = S^T\,\left\langle \varepsilon\varepsilon^T \right\rangle S = S^T RS = \Lambda$. Note that new observation error is diagonal, thus a standard data assimilation algorithm with diagonal observation covariance can still be applied.
5. *Direct application of the inverse*. This can be done directly by calculating the matrix inverse, or indirectly by using a matrix–vector product. In both cases one needs to assume the correlation properties, since there is insufficient statistical information available from data. Given large number of observations the former approach may be computationally prohibitive. The latter approach is computationally feasible and can be described as follows. Here we assume that $R = DCD$ where D is the diagonal matrix of observation errors, and C is the correlation matrix. One can decompose $C = EE^T$ using the unique symmetric square root E. The inverse square root of correlated observation error covariance is $R^{-1/2} = (DE)^{-1} = E^{-1}D^{-1}$. Since the inverse of a symmetric positive definite matrix is symmetric and positive definite, E^{-1} can be modeled using a simple correlation matrix such as Toeplitz (e.g., Golub and van Loan 1989) and thus avoid the calculation of the inverse E^{-1}. In practice the method is applied using a matrix–vector product such as $R^{-1/2}\,[y - h(x)]$, which makes the approach feasible even for large number of observations.

Note that the approaches (1) and (2) never assume non-diagonal R, they only adjust the observation errors (1), or the number of observations (2) to match the desired observation impact. However, if observations errors are correlated, the approach (1) is implicitly using a top-hat function instead of a true correlation function. The approach (2) is implying that near-by observations have similar information content (i.e. homogeneity) which may not be true for observations of clouds and precipitation given that the quality of radiance observation depends on the scan angle, for example. The approach (3) implicitly assumes non-diagonal R but it employs this information only to select radiance observations to be assimilated, still using the original diagonal-based assimilation framework.

The approaches (4) and (5), assume non-diagonal (e.g., correlated) observation errors. The approach (4) is mathematically more general than (5), since it can be applied to an arbitrary R, while the approach (5) requires simplified R correlations in order to be practical. However, the approach (4) also needs an assumption about an eigenvalue threshold. For example, the inverse square root of $R = S\Lambda S^T$ is $R^{-1/2} = S\Lambda^{-1/2}S^T$. Calculation of $\Lambda^{-1/2}$ requires defining a threshold value in order to avoid the division by zero. Unfortunately, the smallest values of Λ are exactly those that are most important for the inverse, making the decision about the threshold difficult.

Additional computational issues arise due to the use of radiative transfer operator for all-sky radiances. Inclusion of cloud and precipitation scattering processes required for all-sky radiance calculations adds considerably to the computational cost of data assimilation (e.g., Stephens 1994). Coupled with a significant increase of the number of radiance observations, the cost of all-sky radiance calculations can be much larger than the cost of clear-sky radiances. This directly impacts the cost of data assimilation and needs to be taken into account.

19.3.3 Nonlinearity and Non-Differentiability

Nonlinearity of cloud microphysical processes and the radiative transfer operator for all-sky radiances is a well-known issue (e.g., Errrico et al. 2007b; Steward et al. 2012). The approach to address nonlinearity may be to choose fundamentally different methodology, such as particle filters (e.g., Gordon et al. 1993; Xiong et al. 2006; van Leeuwen 2009), or to improve minimization conditioning (e.g., Axelsson and Barker 1984; Axelsson 1994; Zupanski et al. 2008). Also, there are so-called linear channels (e.g., frequencies) that do not have strong nonlinearity and thus can be treated using linear or weakly nonlinear methods. Variational methods are generally equipped to address nonlinearity using iterative minimization of the cost function. Standard ensemble Kalman filtering methods do not address the observation nonlinearity specifically, which prompted a development of hybrid variational-ensemble Ensemble methods (e.g., Zupanski 2005; Wang et al. 2007), ensemble iterative Kalman filters (e.g., Gu and Oliver 2007), or a refinement of the ensemble Kalman filter (e.g., Evensen 2003).

Since majority of practical data assimilation algorithms today use iterative minimization to solve nonlinear problems, we will discuss this approach in more detail. These minimization algorithms are typically unconstrained algorithms, most often a nonlinear conjugate-gradient algorithm or quasi-Newton algorithms (e.g., Luenberger 1989). One of the minimization algorithm components most relevant for all-sky satellite radiance assimilation is Hessian preconditioning (e.g., Axelsson and Barker 1984; Axelsson 1994; Yang et al. 1996; Zupanski 1995; Steward 2012). Its general role is in speeding up minimization by a change of variable that effectively reduces the condition number of Hessian matrix (e.g., second derivative of the cost function). The ideal impact of preconditioning is illustrated in Fig. 19.3, which shows a change of a quadratic cost function from an elongated ellipse to a circle. Starting minimization from an arbitrary point will lead to numerous minimization iterations for the original ellipsoidal cost function (Fig. 19.3a), whilst a single iteration will be sufficient for a preconditioned minimization problem (Fig. 19.3b).

Another role of preconditioning, of special importance in practical applications, is to provide a "balanced" reduction of cost function. This means that minimization should produce a change of all control variables that is in agreement with actual

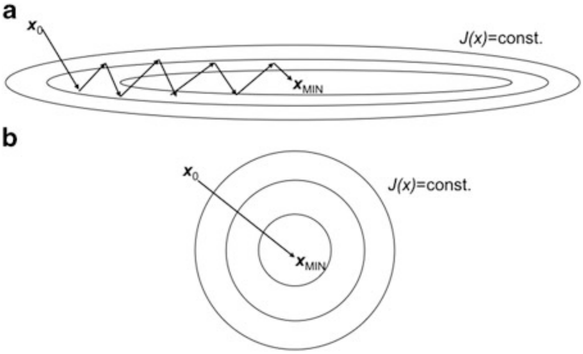


Fig. 19.3 Impact of Hessian preconditioning on minimization: (**a**) physical space, and (**b**) preconditioned space. In this example of a quadratic cost function it is shown how an ideal preconditioning can change the cost function so that the minimum is reached in single minimization iteration

weather situation. In principle, one would like to achieve a similar percentage of adjustment for each control variables, with the idea that although minimization may not have sufficient time to reach mathematical convergence it will still produce an acceptable physical solution. Consider temperature and wind as an example. Let the initial guess have dynamically balanced fields, which is generally true given that it is produced by a forecast. If the temperature component of the cost function is perfectly preconditioned, while the wind component is not preconditioned at all, the analysis after first minimization iteration would have fully adjusted temperature but practically unchanged wind. Since wind and temperature were in dynamical balance before minimization, the produced solution after first iteration would be unbalanced and eventually create noise in the ensuing forecast. This situation can be visualized from Fig. 19.3 with temperature converging according to Fig. 19.3b, while the wind slowly converging as shown in Fig. 19.3a. This situation also illustrates the impact of Hessian preconditioning on the utility of observations: temperature observations will be efficiently used, while wind observations would have a marginal impact. The important point is that this potential problem can be resolved by adequate Hessian preconditioning.

Let now consider the impact of Hessian preconditioning on all-sky satellite radiance assimilation. Assimilation of all-sky satellite radiance would be most beneficial if cloud variables were defined as control variables since they have the strongest impact on the radiative transfer operator. However, if the cloud variable component of the cost function is not adequately preconditioned, all-sky radiances will not be well utilized, eventually producing an unbalanced analysis. Assuming that other dynamical variables were well preconditioned, it is likely that the ensuing forecast will get rid of clouds and precipitation created by the analysis, simply as a consequence of inadequate preconditioning. This problem is real and it can have dire consequences for all data assimilation methods.

In case of variational methods one should recall that forecast error covariance is used to precondition minimization. Unfortunately, since there are no practical ways to include cross-variable correlations for cloud variables, while error covariance of dynamical variables has a relatively well-defined cross-variable structure, it is clear that the overall preconditioning will be off balance. This can also happen in

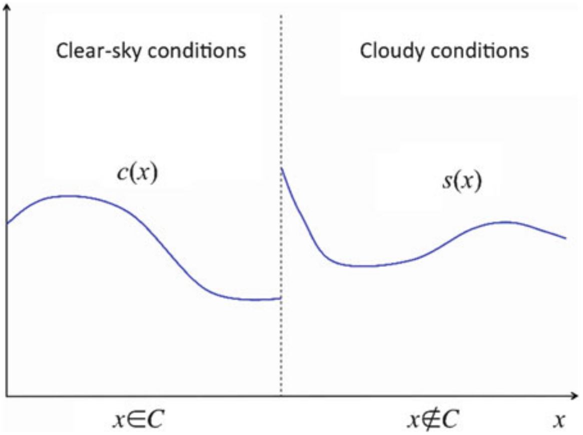


Fig. 19.4 Discontinuous all-sky radiative transfer operator defined by (19.19). The left area of the figure represents clear-sky conditions and the right area corresponds to cloudy conditions. In principle, the function value and its derivatives all can experience a discontinuity

hybrid variational-ensemble methods since often the variational component is used for adjustment of dynamical variables, while the ensemble component is primarily used for cloud variables. In this situation cloud variables will have a much better preconditioning than dynamical variables eventually creating unbalanced analysis that can also act to remove adjusted clouds and precipitation in the forecast. The above examples illustrate the important role of Hessian preconditioning in assimilation of all-sky radiances, and indirectly suggest that preconditioning method should be related to dynamics.

Although nonlinearity of all-sky radiance operator has been generally acknowledged, non-differentiability is typically not discussed and thus requires attention. The definition of observation operator h becomes an issue in the case of all-sky radiances. This is because the radiative transfer operator has an on-off switch to decide if it should go through the cloudy branch that normally includes scattering effects, or not in the case of clear-sky radiances. Since this decision depends on the atmospheric parameters such as cloud mixing ratios and temperature, the forward radiative transfer operator has a discontinuity in the function value and derivative implying discontinuity of the cost function. Therefore, one can write the radiative transfer operator for all-sky radiances as

$$h(x) = \begin{cases} c(x) & x \in C \\ s(x) & x \notin C \end{cases} \tag{19.19}$$

where C represents the state subspace corresponding to clear-sky conditions, c is the clear-sky component and s is the cloudy component of the observation operator. The point where the state can cross between clear and cloudy conditions is the discontinuity point, and thus the operator $h(x)$ has two branches. This is visualized in Fig. 19.4, indicating that the function value and all its derivatives can have a discontinuity.

Discontinuity of observation operator creates an obvious problem for variational methods since they are commonly using gradient-based minimization (e.g., Nocedal 1980; Navon 1986). A non-existent gradient at the point of transition from clear-sky to cloudy conditions prevents correct performance of minimization, eventually resulting in incorrect minimizing solution. Since KF can be described in terms of gradient-based minimization, (e.g., Jazwinski 1970), it has similar problems as variational methods. This implies that non-differentiability of all-sky observation operator is a problem of data assimilation in general.

There are many ways to deal with discontinuity, most obvious being: (1) neglect it, (2) apply smoothing, (3) use non-differentiable minimization algorithms. Although the first option may not be mathematically correct, it does not require any additional effort. Since the discontinuity point is in the area of transition from cloudy to clear-sky conditions, the discontinuity problem may be confined to only those geographical areas, allowing minimization to perform well in the rest of the domain. However, since discontinuity also impacts the line-search algorithm (i.e. finding the optimal step-size), its influence can be more pronounced. Therefore, neglecting the discontinuity problem may be acceptable in some situations, but not in general and definitely not in operational practice. The option (2) has been successfully applied within 4d-Var methods in cases of parameterization schemes (e.g., Zupanski and Mesinger 1995). The approach is to change the original operator by introducing a smooth function in the place of an on-off switch, thus preventing code branching. In choosing the adequate smoothing function and parameters it is important to maintain approximately the same skill and accuracy of the original operator. This could be difficult and it requires an extensive preparation of the code. The third option (3) is the most correct approach, since it does not alter the original observation operator and it addresses the true problem, which is the minimization algorithm performance. There are numerous minimization algorithms that can address non-differentiability, some of those developed as an extension of gradient-based algorithms (Haarala et al. 2004; Karmitsa et al. 2012; Steward et al. 2012). Encouraging results obtained using this approach in data assimilation have been reported by Steward et al. (2012).

It is likely that the choice of approach will depend on the actual assimilation problem and the amount of work required to implement the changes. Important message from this section is that non-differentiability of all-sky observation operator should not be overlooked. Once the problem is identified, one can proceed to solutions (1)–(3), or take an alternative approach.

19.3.4 Non-Gaussian Errors

Current data assimilation methodologies are generally designed to address only Gaussian errors. It is also understood that there are numerous applications with non-Gaussian errors and that a non-Gaussian data assimilation framework may be necessary (e.g., Abramov and Majda 2004; Fletcher and Zupanski 2006a, b; Bocquet et al. 2010). Satellite radiance observation error statistics indicates a skewness for some instruments and channels that may be attributed to non-Gaussian pdfs (e.g., Okamoto and Derber 2006; Errrico et al. 2007b; Bauer et al. 2010), and thus implies that data assimilation of all-sky satellite radiances may not perform correctly in such cases. One should also be aware of ways to mitigate non-Gaussianity in data assimilation (e.g., Simon and Bertino 2009; Bocquet et al. 2010).

One can distinguish several possible approaches to deal with non-Gaussian errors of all-sky radiances: (1) neglect the problem, (2) apply Gaussian assumption, but introduce bias correction, (3) apply change of variable to convert from non-Gaussian to Gaussian framework, and (4) use non-Gaussian data assimilation framework.

The option (1) is the simplest, and thus the easiest. If one chooses to assimilate only channels with approximately Gaussian observation errors, it may be still possible to use the original Gaussian data assimilation framework. However, this approach may leave important observation information not assimilated. It also requires a good knowledge of the observation error statistics by channels, which could take time to accumulate.

Option (2) is commonly used in operations (e.g., Harris and Kelly 2001; Dee and Uppala 2009). Satellite bias is typically defined to include predictors, defined to include satellite geometry (e.g., viewing angle) and atmospheric precursors (e.g., thickness, skin temperature, surface wind speed)

$$b(\varphi, x) = \sum_i \varphi_i r_i(x) \tag{19.20}$$

where r is predictor and φ is regression coefficient. Parameters of such formed regression are added as control variables to minimization thus creating an augmented control variable and cost function. Although used with great success, there are channels that are not well controlled using this technique. Also, current operational practice includes mostly clear-sky, not all-sky radiance assimilation, and so does the bias correction. This means that atmospheric precursors used for clear-sky may not be adequate for cloudy conditions. Even if adequate atmospheric precursors are found, it is clear that this approach requires a lot of experimenting and fundamental knowledge of interactions between clouds and satellites. In addition, the augmented control variable in minimization may be technically difficult to implement, depending on the existing minimization setup. As suggested in several papers (e.g., Errrico et al. 2007b), if the actual observation error has skewed pdf that resembles lognormal distribution, then one could pose the problem in terms of logarithmic variable which would then be Gaussian. Although this is a feasible solution, it was shown to be non-unique (e.g., Fletcher and Zupanski 2008). The option (4) may be the most complete since it addresses the true problem of having non-Gaussian errors in data assimilation. It was shown that ensemble data assimilation could be defined in terms of non-Gaussian errors within hybrid variational-ensemble methodology (e.g., Fletcher and Zupanski 2006a), or within particle filters (e.g., van Leeuwen 2009). In either case, implementing new methodology is a slow process and the ultimate decision about the approach for handling non-Gaussian errors will depend on desired application and developmental time constraints. Addressing non-Gaussian all-sky radiance observation errors inherently implies a need for better handling of nonlinearity. Therefore, if the available data assimilation algorithm is not very good for nonlinear operators, it is probably good to avoid introducing non-Gaussian errors.

19.4 Summary and Future

Data assimilation of all-sky satellite radiances is a difficult problem that puts to test data assimilation methodology and algorithmic solutions that are used today. Since the information from all-sky radiances is potentially very valuable, using "short-cut" solutions is not acceptable. We discussed several critical aspects of successful all-sky radiance data assimilation, with emphasis on forecast error covariance, Hessian preconditioning, non-differentiability, and correlation of observation errors. Also, the focus of our presentation was on how variational and ensemble data assimilation can handle these challenging problems. In conclusion, both methods have their advantages and disadvantages and likely best approach is to develop hybrid variational-ensemble methods that can selectively choose the better option. One can also adopt other methodologies that can possibly address better the difficulties arising in variational and ensemble methods.

Although we did not describe in detail all issues related to all-sky radiances, such as verification, or algorithmic details related to a specific methodology, they also have to be taken into account. There may be research issues that we are not aware of at present, but will be eventually addressed as all-sky radiance assimilation research becomes widespread.

One important implication of presented challenges is that development of new data assimilation methodology that is better suited for all-sky satellite radiance assimilation has to be comprehensive. For example, solving nonlinear issues cannot be properly done without addressing non-Gaussian errors or without utilizing the full power of Hessian preconditioning. Similarly, better definition of forecast error covariance will not be fully beneficial unless it is combined with superior Hessian preconditioning that can maximize the benefit of nonlinear minimization. As suggested here, it may not be always necessary to develop most complex algorithms to solve the challenges of all-sky radiance assimilation. There are applications that may accept simple solutions to some of the issues, but it is important not to dismiss an issue before its impact is well understood. For example, although one can opt for uncorrelated observation errors at the end, it first needs to evaluate the potential impact of correlations or to investigate statistics of observations errors.

Development of hybrid variational-ensemble and other new methodologies is an ongoing effort and will likely produce an improved all-sky radiance assimilation methodology capable of extracting maximum information from this valuable data.

Acknowledgements This work was supported by the National Science Foundation Collaboration in Mathematical Geosciences Grant ATM-0930265, and by NOAA NESDIS Grant

NA10NES4400012.We would also like to acknowledge high-performance computing support provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Appendix 1

Analysis Increment and Forecast Error Covariance

Forecast error covariance is one of the major components of the analysis update equation. In this section we derive the analysis update as a function of forecast error covariance. It is convenient to begin by defining the singular value decomposition (SVD) of a square root forecast error covariance (e.g., Golub and van Loan 1989)

$$P_f^{1/2} = \sum_i \sigma_i u_i v_i^T \tag{19.21}$$

where $\{u_i\}$ and $\{v_i\}$ are singular vectors and $\{\sigma_i\}$ are singular values. This leads to eigenvalue decomposition (EVD) in the form

$$P_f = P_f^{1/2} P_f^{T/2} = \sum_i \lambda_i u_i u_i^T \tag{19.22}$$

with $\lambda_i = \sigma_i^2$.

1 Kalman Filter and Related Methods

The analysis update is given by the KF analysis equation

$$x^a - x^f = P_f H^T \left(H P_f H^T + R\right)^{-1} \left(y - h(x^f)\right) \tag{19.23}$$

where the superscript a denotes analysis, R is the observation error covariance, and h and H are the nonlinear observation operator and its Jacobian, respectively. After using (19.22) in (19.23), and denoting

$$\gamma_i = \lambda_i u_i^T \left(H^T [H P_f H^T + R]^{-1} \left[y - h(x^f)\right]\right), \tag{19.24}$$

the KF analysis update (19.23) becomes

$$x^a - x^f = \sum_i \gamma_i u_i \tag{19.25}$$

i.e. it can be represented as a linear combination of the forecast error covariance singular vectors.

2 Variational Methods

Successful and efficient minimization of the cost function ((19.1) from main text) requires the so-called Hessian preconditioning (e.g., Axelsson and Barker 1984; Zupanski 1993, 1995). The square-root forecast error covariance is commonly used for this purpose in variational methods, introduced as a change of variable

$$x - x^f \; = \; P_f^{1/2} w \tag{19.26}$$

Where w is the preconditioned control variable. The iterative minimization solution w^a is obtained as a limit of sequence $\{w_k \; = \; w_{k-1} + \alpha_{k-1} d_{k-1}; k \; = \; 1, 2, \ldots\}$ where index k is the iteration index, α is step-size, and d is descent direction. After substituting the minimization solution w^a in (19.26) one obtains the analysis solution in terms of the physical state variable

$$x^a \; - x^f \; = \; P_f^{1/2} w^a. \tag{19.27}$$

After substituting (19.21) in (19.27), and denoting

$$\eta_i \; = \; \sigma_i v_i^T w^a \tag{19.28}$$

The variational method solution (19.27) becomes

$$x^a \; - x^f \; = \; \sum_i \eta_i u_i. \tag{19.29}$$

Therefore, the variational solution can also be represented as a linear combination of forecast error covariance singular vectors.

Since majority of currently used data assimilation algorithms are based on KF and/or variational methods, one can see from (19.25) to (19.29) that analysis correction $x^a - x^f$ lies in the space defined by the forecast error covariance singular vectors.

Appendix 2

Entropy and Mutual Information

We follow Cover and Thomas (2006) to quantify the information content of observations, based on Shannon information theory (Shannon and Weaver 1949) and relative entropy (Kullback and Leibler 1951). Entropy of a random variable X is defined as a non-negative measure of uncertainty

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x). \tag{19.30}$$

Where p is a pdf. One can also define joint entropy

$$H(X, Y) = -\sum_{x \in \chi} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{19.31}$$

and relative entropy (e.g., Kullback-Leibler distance) between probabilities $p(x)$ and $q(x)$

$$D(p \parallel q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)}. \tag{19.32}$$

The mutual information $I(X; Y)$ is defined as the relative entropy between the joint distribution and the product distribution

$$I(X;\, Y) = \sum_{x \in \chi} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{19.33}$$

and it represents a reduction of uncertainty due to information sharing between variables. The mutual information is non-negative and becomes zero for independent variables. One can also see that

$$I(X; X) = H(X) \tag{19.34}$$

since there is no new information that can reduce uncertainty.

References

Abramov RF, Majda AJ (2004) Quantifying uncertainty for non-Gaussian ensembles in complex systems. SIAM J Sci Comput 26:411–447

Anderson JL (2003) A local least squares framework for ensemble filtering. Mon Wea Rev 131:634–642

Auligne T, Lorenc A, Michel Y, Montmerle T, Jones A, Hu M, Dudhia J (2011) Toward a new cloud analysis and prediction system. Bull Am Meteorol Soc 92:207–210

Axelsson O (1994) Iterative solution methods. Cambridge University Press, Cambridge, p 668

Axelsson O, Barker VA (1984) Finite-element solution of boundary-layer problems. Theory and computations. Academic Press, Orlando, p 432

Bauer P, Geer A, Lopez P, Salmond D (2010) Direct 4D-Var assimilation of all- sky radiances. Part I: implementation. Q J Roy Meteorol Soc 136A:1868–1885

Bauer P, Lopez P, Salmond D, Benedetti A, Saarinen S, Moreau E (2006) Implementation of 1D+4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. II: 4D-Var. Q J Roy Meteorol Soc 132:2307–2332

Bauer P, Ohring G, Kummerow C, Auligne T (2011) Assimilating satellite observations of clouds and precipitation into NWP models. Bull Am Meteorol Soc 92:ES25–ES28

Bocquet M, Pires CA, Wu L (2010) Beyond Gaussian statistical modeling in geophysical data assimilation. Mon Wea Rev 138:2997–3023

Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken, 776 pp

Daley R (1993) Atmospheric data analysis. Cambridge University Press, Cambridge, 472 pp

Dee DP, Uppala S (2009) Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. Q J Roy Meteorol Soc 135:1830–1841

Errico RM, Ohring G, Bauer P, Ferrier B, Mahfouf J-F, Turk J, Weng F (2007a) Assimilation of satellite cloud and precipitation observations in numerical weather prediction models: introduction to JAS special collection. J Atmos Sci 64:3737–3741

Errrico RM, Bauer P, Mahfouf JF (2007b) Issues regarding the assimilation of cloud and precipitation data. J Atmos Sci 64:3785–3798

Evensen G (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn 53:343–367

Evensen G (2009) Data assimilation: the ensemble Kalman filter. Springer, Berlin/Heideberg, p 307

Evensen G, van Leeuwen PJ (2000) An ensemble Kalman smoother for nonlinear dynamics. Mon Wea Rev 128:1852–1867

Fertig EJ, Hunt BR, Ott E, Szunyogh I (2007) Assimilating non-local observations with a local ensemble Kalman filter. Tellus 59A:719–730

Fletcher SJ, Zupanski M (2006a) A hybrid normal and lognormal distribution for data assimilation. Atmos Sci Lett 7:43–46

Fletcher SJ, Zupanski M (2006b) A data assimilation method for log-normally distributed observational errors. Q J Roy Meteorol Soc 132:2505–2519

Fletcher SJ, Zupanski M (2008) Implications and impacts of transforming lognormal variables into normal variables in VAR. Met Zeit 16:755–765

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J Roy Meteorol Soc 125:723–757

Geer A, Bauer P, Geer A, Lopez P (2010) Direct 4D-Var assimilation of all- sky radiances. Part II: assessment. Q J Roy Meteorol Soc 136A:1886–1905

Geer A, Bauer P (2010) Enhanced use of all-sky microwave observations sensitive to water vapour, cloud and precipitation. ECMWF Tech Memorandum 620:41

Golub GH, and van Loan CF (1989) Matrix computations. The John Hopkins University Press, Baltimore, p 642

Gordon NJ, Salmond DJ, Smith AFM (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc 140:107–113

Gu Y, Oliver DS (2007) An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. SPE J 12:438–446

Haarala M, Miettinen K, Makela MM (2004) New limited memory bundle method for large-scale nonsmooth optimization. Optim Meth Softw 19:673–692

Hamill T, Whitaker J, Snyder C (2001) Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Mon Wea Rev 129:2776–2790

Harris BA, Kelly G (2001) A satellite radiance-bias correction scheme for data assimilation. Q J Roy Meteorol Soc 127:1453–1468

Houtekamer P, Mitchell H (2001) A sequential ensemble Kalman filter for atmospheric data assimilation. Mon Wea Rev 129:123–136

Huang X-Y, et al (2009) Four-dimensional variational data assimilation for WRF: formulation and preliminary results. Mon Wea Rev 137:299–314

Jazwinski AH (1970) Stochastic processes and filtering theory. Academic, San Diego, 376 pp

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, 341 p

Karmitsa N, Bagirov A, Makela MM (2012) Comparing different nonsmooth minimization methods and software. Optim Meth Softw 27:131–153

Kullback S, Leibler RA (1951) On information and sufficiency. Annals Math Stat 22:79–86

Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, Cambridge, 680 p

Li Z, Navon IM (2001) Optimality of 4D-Var and its relationship with the Kalman filter and Kalman smoother. Q J Roy Meteorol Soc 127:661–684

Lorenc AC (1986) Analysis methods for numerical weather prediction. Q J Roy Meteorol Soc 112:1177–1194

Luenberger DL (1989) Linear and non-linear programming. Addison-Wesley, Reading, 491 pp

Navon IM (1986) A review of variational and optimization methods in meteorology. In: Sasaki YK (ed) Variational methods in geosciences. Developments in Geomathematics, vol 5. Elsevier Science Publishers, Amsterdam, pp 29–35

Nocedal J (1980) Updating quasi-Newton matrices with limited storage. Math Comp 35:773–782

Okamoto K, Derber JC (2006) Assimilatioon of SSM/I radiances in the NCEP global data assimilation system. Mon Wea Rev 134:2612–2631

Parrish DF, Cohn SE (1985) A Kalman filter for a two-dimensional shallow- water model: formulation and preconditioning experiments. Office Note 304, National Meteorological Center, Washington, DC

Parrish DP, Derber JC (1992) The national meteorological center's spectral statistical interpolation analysis system. Mon Wea Rev 120:1747–1763

Polkinghorne R, Vukicevic T (2011) Data assimilation of cloud-affected radiances in a cloud-resolving model. Mon Wea Rev 139:755–773

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana, 144 pp

Simon E, Bertino L (2009) Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. Ocean Sci 5:495–510

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG (2005) A description of the advanced research WRF version 2. NCAR Techinal Note 468, 88 pp. http://www.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf

Stephens GL (1994) Remote sensing of the lower atmosphere. Oxford University Press, New York, p 544

Steward JL (2012) On a unifying interpretation of empirically-determined square-root background error covariance matrices for variational and ensemble data assimilation. JPL Earth Science Seminar, Pasadena

Steward JL, Navon IM, Zupanski M, Karmitsa N (2012) Impact of non-smooth observation operators on variational and sequential data assimilation for a limited-area shallow-water equation model. Q J Roy Meteorol Soc 138:323–339

Thepaut J-N, Courtier P, Belaud G, Lemaitre G (1996) Dynamical structure functions in a four-dimensional variational assimilation: a case study. Q J Roy Meteorol Soc 122:535–561

van Leeuwen PJ (2009) Particle filtering in geophysical systems. Mon Wea Rev 137:4089–4114

Vukicevic T, Greenwald T, Zupanski M, Zupanski D, VonderHaar T, Jones AS (2004) Mesoscale cloud state estimation from visible and infrared satellite radiances. Mon Wea Rev 132:3066–3077

Wang X, Hamill TM, Whitaker JS, Bishop CH (2007) A comparison of hybrid ensemble transform Kalman filter-OI and ensemble square-root filter analysis schemes. Mon Wea Rev 135**:**1055–1076

Wu W-S, Purser RJ, Parrish DF (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. Mon Wea Rev 130:2905–2916

Xiong X, Navon IM, Uzunoglu B (2006) A note on the particle filter with posterior Gaussian resampling. Tellus 58A:456–460

Yang W, Navon IM, Courtier P (1996) A new Hessian preconditioning method applied to variational data assimilation experiments using adiabatic version of NASA/GEOS-1 GCM. Mon Wea Rev 124:1000–1017

Zhang S, Zupanski M, Hou A, Lin X, Cheung S (2012) Assimilation of precipitation- affected radiances in a cloud-resolving WRF ensemble data assimilation system. Mon Weather Rev. doi:10.1175/MWR-D-12-00055.1 (in press)

Zupanski M (1993) A preconditioning algorithm for large scale minimization problems. Tellus 45A:578–592

Zupanski M (1995) A preconditionin algorithm for four-dimensional variational data assimilation. Mon Wea Rev 124:2562–2573

Zupanski M (2005) Maximum likelihood ensemble filter: theoretical aspects. Mon Wea Rev 133:1710–1726

Zupanski D, Mesinger F (1995) Four-dimensional variational assimilation of precipitation data. Mon Wea Rev 123:1112–1127

Zupanski M, Navon IM, Zupanski D (2008) The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. Quart J Roy Meteorol Soc 134:1039–1050

Zupanski D, Zupanski M, Grasso LD, Brummer R, Jankov I, Lindsey D, Sengupta M, DeMaria M (2011a) Assimilating synthetic GOES-R radiances in cloudy conditions using an ensemble-based method. Int J Remote Sens 32:9637–9659

Zupanski D, Zhang SQ, Zupanski M, Hou AY, Cheung SH (2011b) A prototype WRF-based ensemble data assimilation system for downscaling satellite precipitation observations. J Hydromet 12:118–134

Chapter 20
Development of a Two-way Nested LETKF System for Cloud-resolving Model

Hiromu Seko, Tadashi Tsuyuki, Kazuo Saito, and Takemasa Miyoshi

Abstract A two-way nested Local Ensemble Transform Kalman Filter (LETKF) system has been developed to improve the accuracy of numerical forecasts on local heavy rainfalls. In this system, mesoscale convergence which drives local heavy rainfalls, is first reproduced by the LETKF with a grid interval of 15 km (Outer LETKF) which assimilates conventional data. The convection cells associated with the local heavy rainfall are then reproduced by the higher resolution LETKF with a grid interval of 1.875 km (Inner LETKF) which assimilates local data. The boundary conditions of the Inner LETKF are given by the forecast of the Outer LETKF. To consider the upward cascade effect from storm scale to mesoscale, the forecast results of the Inner LETKF are reflected into the Outer LETKF every 6 h.

This system was applied to a thunderstorm that caused a local heavy rainfall event on the Osaka Plain on 5th September 2008. The rainfall distributions similar to the observed ones were reproduced in a few ensemble members of the Inner LETKF, although the observed scattered convection cells were expressed as weak rainfall regions in the Outer LETKF. When the precipitable water vapor or slant-path water vapor data obtained by GPS and horizontal wind or radial wind data observed by Doppler radars were assimilated in the Inner LETKF, the number of ensemble forecasts, which reproduced the local heavy rainfall, increased. The experiments on the small-scale disturbances in the initial seeds of the Inner LETKF and on the initial conditions produced by the no-cost smoother showed that these improvements might enhance the accuracy of local heavy rainfall forecasts.

H. Seko (✉) · T. Tsuyuki · K. Saito
Meteorological Research Institute, 1-1 Nagamine, Tsukuba, Ibaraki, 305-0052, Japan,
e-mail: hseko@mri-jma.go.jp

T. Miyoshi
RIKEN Advanced Institute for Computational Science, Kobe, 650-0047, Japan

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_20,

20.1 Introduction

In the last decade, local heavy rainfalls that developed in urban areas (e.g. Tokyo Metropolitan area) in summer have influenced urban functions, and have occasionally caused urban flash floods (e.g. Nerima heavy rainfall in 1999, Itabashi heavy rainfall in 2005). To mitigate damages from local heavy rainfalls, accurate forecasts are desired. A few experiments on local heavy rainfalls have been performed so far with variational data assimilation systems. For instance, Kawabata et al. (2007) reproduced the Nerima heavy rainfall that occurred in the Tokyo Metropolitan area by assimilating GPS, (Global Positioning System) precipitable water vapor, (PWV, amount of water vapor in a column) and radial wind of Doppler radars with JMANHM (Japan Meteorological Agency Non-hydrostatic Model)-4DVAR (4 dimensional variational data assimilation system), and pointed out that low-level convergence of water vapor is essential to reproduce local heavy rainfalls.

Besides variational data assimilation methods, ensemble Kalman filters (EnKFs) can provide initial conditions that are close to actual fields by assimilation of observation data. Some previous studies have used the EnKF for mesoscale applications and have obtained promising results (e.g., Snyder and Zhang 2003; Zhang et al. 2004, 2006; Dowell et al. 2004; Tong and Xue 2005; Xue et al. 2006; Meng and Zhang 2008; Seko et al. 2011; Miyoshi and Kunii 2012). In addition to accurate initial conditions, EnKFs used as assimilation systems provide the probability of heavy rainfalls and a number of rainfall forecasts. Especially in the forecast of local heavy rainfalls, horizontal convergence in which local heavy rainfalls are generated is generally relatively weak and the predicted distribution of local heavy rainfalls is widely spread. Thus, it should be considered that the predicted distribution is a part of the fields that have probability density distributions. Because of these merits, Local Ensemble Transform Kalman Filter (LETKF, Hunt et al. 2007) based on the JMANHM (Saito et al. 2006), known as the NHM-LETKF (Miyoshi and Aranami 2006), was used as the data assimilation system in this study.

As mentioned before, local heavy rainfalls most often are generated in mesoscale convergences. Even if convergence is relatively weak, mesoscale convergences need to be reproduced by assimilation. In this study, mesoscale convergences were produced by the LETKF system with a grid interval of 15 km (Outer LETKF). Besides the position of convergence, rainfall intensity of local heavy rainfalls is also important. Then, the LETKF systems with a grid interval of 1,875 km (Inner LETKF), which reproduce positions and intensities of intense convection cells, are deployed within the domain of the Outer LETKF.

As pointed out in Kawabata et al. (2007), convergence of low-level water vapor is indispensable in reproducing local heavy rainfalls. Because GPS-derived PWV or slant water vapor (SWV, water vapor amount along paths from GPS satellites to GPS receivers) and horizontal wind or radial wind observed by Doppler radars provide information about low-level convergence of water vapor, this data is expected to improve the rainfall forecast. A number of data assimilation experiments on radial wind from Doppler radars and GPS-PWV using the EnKFs have been reported so far (e.g. Xue et al. 2006; Seko et al. 2011). The impacts of the GPS-SWV and the synergistic effect of simultaneous assimilation of Doppler radar data and GPS-water vapor data (PWV or SWV) have not been investigated so far with EnKFs, though they were shown in the data assimilation experiments in which the JMA meso-4dvar assimilation system was used (e.g. Seko et al. 2004). In this study, these impacts are investigated with EnKFs, in addition to the impacts of GPS-PWV data and Doppler radar data.

In Sect. 20.2, a local heavy rainfall which the nested LETKF system was applied to is explained. Section 20.3 briefly explains the nested LETKF system. Results of assimilation of conventional data are described in Sect. 20.4. Impacts of GPS data and Doppler radar and their synergistic effect are shown in Sect. 20.5. Section 20.6 is the conclusion of this study.

20.2 Thunderstorm of 5th September 2008 Developed on the Osaka Plain

As the target of experiments in this study, a local heavy rainfall generated on the Osaka Plain on 5th September 2008 was adopted, because Doppler radar data of the Osaka and Kansai international airports can be used.

On 5th September 2008, a high pressure system widely covered Japan and a small low pressure system was seen in the south of western Japan (Fig. 20.1a). Because large-scale disturbances, such as synoptic fronts or typhoons did not exist near Japan's main islands, scattered convection cells were generated not only near Osaka, but also over western Japan (Fig. 20.1b). Figure 20.2b shows the radar echo distribution observed by conventional radars of JMA (Japan Meteorological Agency) from 1400 JST to 1600 JST (Japan Standard Time; 0900 JST corresponds to 0000 UTC; Universal Time Coordinate). At 1400 JST, there were scattered convection cells in mountainous areas east and south of the Osaka Plain (indicated by circles in Fig. 20.2a). These convection cells developed there by 1500 JST and other intense convection cells were generated on the Osaka Plain. The intense convection cells on the Osaka Plain organized, and then produced a rainfall amount exceeding 90 mm from 1450 JST to 1600 JST at Sakai City (not shown). This rainfall region extended northwestward, maintaining its rainfall intensity until 1600 JST. The horizontal wind and the sea level temperature and pressure at 1400 JST are shown in Fig. 20.3. A high temperature region existed on the Osaka Plain (Fig. 20.3b) and a thermodynamic low-pressure system was generated there (Fig. 20.3c). The southerly flow from the Kii-channel and the northeasterly flow over the Osaka Plain were converged near Sakai City (Fig. 20.3a). This low-level convergence is one of the reasons why the intense convection cells generated on the Osaka Plain.

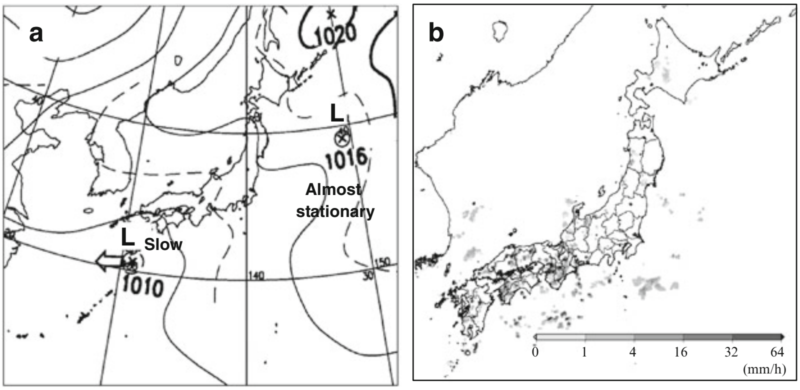


Fig. 20.1 (**a**) Surface weather chart at 15 JST 5th September 2008. (**b**) Rainfall distribution observed by operational radars of JMA at 15 JST 5th September 2008

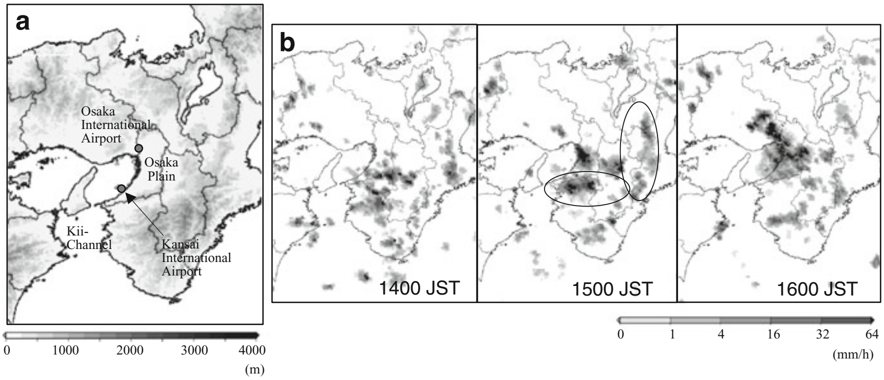


Fig. 20.2 (**a**) Topography in and around the Osaka Plain. (**b**) Rainfall distribution observed by operational radars of JMA from 1400 JST to 1600 JST 5th September 2008. *Circles* in (**b**) indicate the rainfall regions generated in mountainous areas

20.3 Outlines of the Nested LETKF System

Figure 20.4a shows the schematic illustration of the nested LETKF system. This data assimilation system was composed of two LETKF systems: the Outer and Inner LETKFs. The vertical layer structure was common in both LETKFs. Namely, the number of vertical layers was 50 and the depth of the vertical layers was increased from 40 to 880 m as the height increased. The height of the domain's top was 22.6 km. The number of ensemble members was 12. Other parameters, such as the horizontal grid interval, microphysical processes and so on depend on the LETKFs.

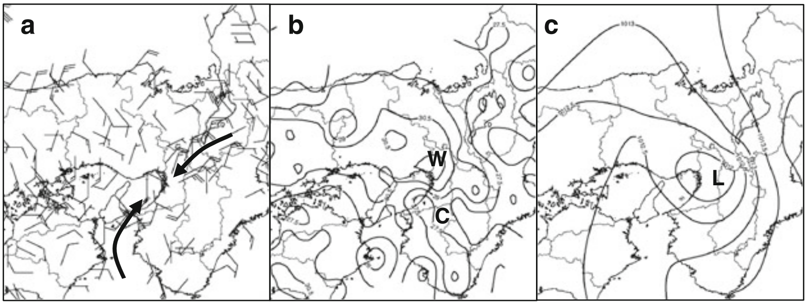


Fig. 20.3 Surface meteorological data observed by AMeDAS and the Meteorological Observatories of JMA at 1400 JST 5th September 2008. (**a**) Horizontal wind, (**b**) Sea level temperature (˚ C), (**c**) Sea level pressure (hPa)

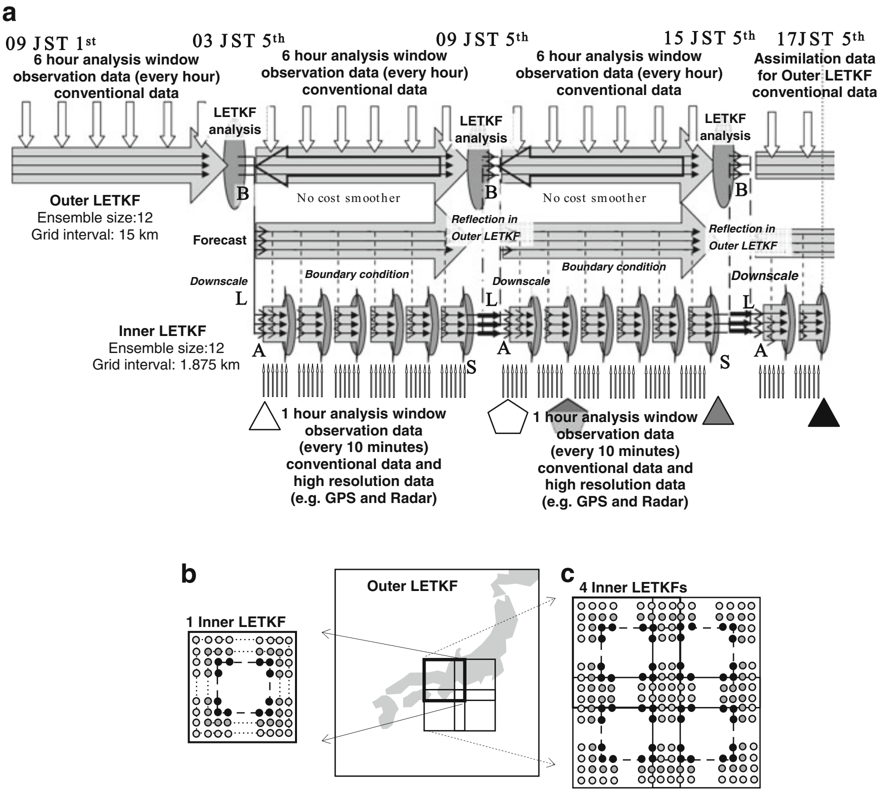


Fig. 20.4 (**a**) Schematic illustration of the nested LETKF system. (**b**, **c**) Schematic illustration of the blending weights of the analyzed values of the Outer and Inner LETKFs. One Inner LETKF is deployed in (**b**), and 4 Inner LETKFs are deployed side by side in (**c**). *Darker circles* indicate the points where the weights of the Inner LETKF are larger

As for the Outer LETKF system, it was used to reproduce mesoscale distributions including convergence lines. The grid interval of the Outer LETKF is 15 km and the grid number in the horizontal directions was 80 × 80. The Kain-Fritch parameterization scheme was adopted. The ensemble forecast started at 0900 JST 1st September 2008 and the initial seed of the Outer LETKF was obtained from the JMA mesoscale analysis fields from 29th to 31st August. The boundary condition from 1st to 5th September was also produced from the JMA mesoscale analysis. The data assimilation window (1 cycle) was 6 h and the conventional data (surface and upper sounding data), which was used in the JMA mesoscale analysis, were assimilated every hour.

The Inner LETKF was used to reproduce the intense convection cells of local heavy rainfalls. The grid interval of the Inner LETKF was as small as 1.875 km to resolve small convection cells. The microphysical process, in which the mixing ratio of cloud, rain, ice crystals, graupel and the number density of ice crystals were predicted, was used. The boundary conditions and first initial seed of the Inner LETKF (indicated by an open triangle in Fig. 20.4a) were produced from the forecast of the Outer LETKF. The data assimilation window (1 cycle) is 1 h, and three sets of 6 cycle experiments were performed from 03 JST 5th. In addition to the conventional data, GPS water vapor data and radar wind data were assimilated every 10 min.

To reflect the analysis of the Inner LETKF in the Outer LETKF, the analyzed value of the Outer LETKF was replaced by that of the Inner LETKF every 6 h at the end of the assimilation windows of the Outer LETKF (namely, 09 JST and 15 JST of 5th), at which time both LETKFs produced the analyses. To reduce the inconsistencies between the Inner LETKF and the Outer LETKF, the values of the Outer LETKF at one and two grids inside from the boundary of the Inner LETKF were produced by blending with those of the Outer LETKF (Fig. 20.4b). The weights used in blending are determined with linear interpolation.

If an Inner LETKF has to cover a wide area, huge computer resources are needed. In this study, the wide area is divided into a number of small domains and the Inner LETKFs are executed at each divided domain. When a number of Inner LETKFs are deployed in the domain of the Outer LETKF, there might be overlapped regions (Fig. 20.4c). As the first step of the multi Inner LETKF, the values in the overlapped regions were determined by averaging the analyzed values produced by the aforementioned procedure. For instance, when N Inner LETKFs ($x_{i1} \sim x_{iN}$) were used, the values of the Outer LETKF (x_o) were obtained with the following equation;

$$x_o = (w_{o1}x_o + w_{i1}x_{i1} + w_{o2}x_o + w_{i2}x_{i2} + \cdots + w_{oN}x_o + w_{iN}x_{iN}) \,/\, N. \quad (20.1)$$

This method allows the Inner LETKFs to be deployed flexibly. This method might be more robust, because the replaced values of the Outer LETKF in the regions where the domains of the Inner LETKFs are overlapped include the analyzed values of the Outer LETKF.

20.4 Assimilation Results of Conventional Data and Other Improvements of Nested LETKF System

20.4.1 Assimilation Results of Conventional Data

Figures 20.5a, b show the ensemble mean and spread of the rainfalls reproduced by the Outer LETKF at 15 JST (the end of the second sets of the Inner LETKF experiments, indicated by a gray triangle in Fig. 20.4a), just before the occurrence of the local heavy rainfall. As shown in Sect. 20.2, the small intense convections scattered over the western Japan were observed. Due to the large grid interval of the Outer LETKF (15 km), the reproduced rainfalls were expressed as weak rainfall regions (Fig. 20.5a). Although the rainfall intensities were much weaker than the observed ones, the distribution of rainfall regions was roughly similar to that of the observed one. In this experiment, any deviation, such as deviations produced from ensemble forecasts of Global models, was not added to the boundary conditions of the Outer LETKF. Due to the fixed boundary conditions, the ensemble spreads near the boundary were small (Fig. 20.5b). However, the spread around the center of the domain where the Osaka Plain is located was relatively large, and one Inner LETKF was deployed around the center of the Outer LETKF.

Next, the ensemble mean distribution of rainfalls analyzed by the Inner LETKF was compared with the observed one (Figs. 20.5c and 20.2b). Although the rainfall regions were more widely distributed and their rainfall intensities were smaller due to the averaging procedure, the positions of the reproduced rainfall regions were similar to the observed regions at 1500 JST (Figs. 20.5c and 20.2b). Namely, the reproduced rainfall regions more than 1 mm/h roughly corresponded to the observed regions of which rainfall intensities were more than 4 mm/h. Because these well-reproduced rainfall regions were located in mountainous areas (indicated by circles in Fig. 20.5c), these scattered rainfalls at 15 JST are likely related to orographic effects. As for the rainfall that developed into the local heavy rainfall at Sakai City (indicated by an arrow in Fig. 20.5c), it was too small and its intensity was too weak.

As mentioned, the rainfall intensity of the ensemble mean was smaller than that of each ensemble member due to the averaging procedure. Because the rainfall intensity, as well as the position of the rainfalls, is important from the point of view of disaster prevention, the rainfall distributions of each ensemble members are shown in the following sections. Figure 20.6b is the rainfall distributions at 17 JST (indicated by a solid triangle Fig. 20.4a) reproduced by the Inner LETKF to show the development of the rainfall at Sakai City. The conventional data was assimilated in the Outer LETKF and Inner LETKF from 1510 JST to 1700 JST. In four ensemble members (#001, #004, #007 and #008), the intense rainfall regions that were developed at Sakai City extended northwestward (indicated by circles in Fig. 20.6b). This feature of the analyzed distributions, which was the same as the observed one at 16 JST, indicates that the local heavy rainfall at Sakai City were

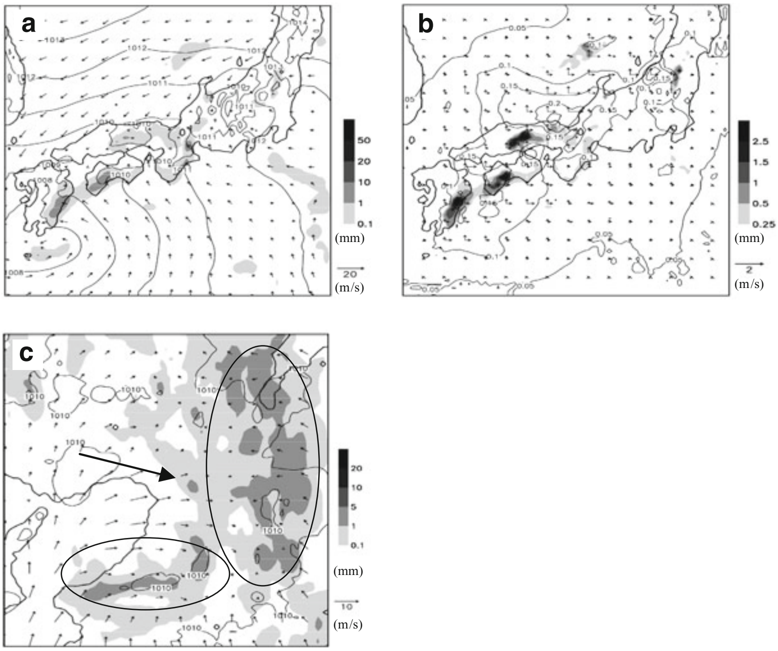


Fig. 20.5 (**a**) Ensemble mean and (**b**) spread of 1 h rainfall (*shaded regions*), horizontal wind at the height of 20 m (*vectors*) and sea level pressure (*contours*) at 15 JST 5th September 2008 reproduced by the Outer LETKF. (**c**) Same as (**a**) except ensemble mean by the Inner LETKF. *Circles* in (**c**) indicate the rainfall regions generated in mountainous areas. An *arrow indicates* the rainfall region that developed into the thunderstorm of Sakai City

well reproduced in one third of the ensemble members, though there was a time lag of 1 h. When these rainfall regions were traced backward in time, these reproduced intense rainfall regions were originated from the small rainfall regions near Sakai City at 15 JST (indicated by arrows in Fig. 20.6a). As explained in the comparison with the ensemble mean distribution, these small rainfall regions were much weaker than the observed one. To increase the accuracy of initial conditions, in other words, to increase the number of the ensemble members in which the local heavy rainfall was reproduced, more high-resolution data, such as GPS water vapor data or Radar wind data, would be helpful. The impacts of this high-resolution data are explained in Sect. 20.5.

20.4.2 Feedback to Outer LETKF and Convection Cells Near the Boundary of Inner LETKF

In this nested system, the analyzed values of the Outer LETKF within the regions of the Inner LETKF were replaced with those of the Inner LETKF every 6 h. It is

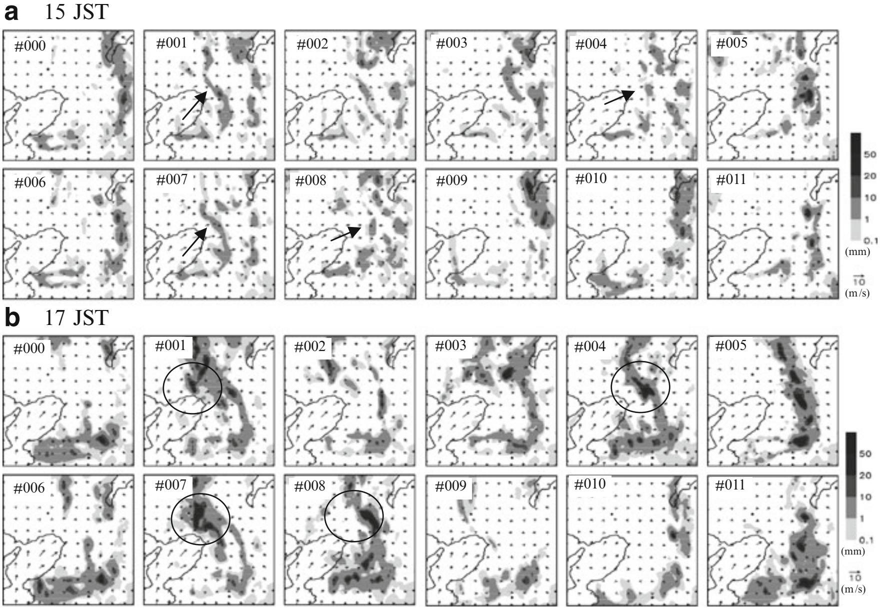


Fig. 20.6 Horizontal distributions of 1 h rainfall (*shaded regions*) and horizontal wind at the height of 20 m (*vectors*) reproduced by the Inner LETKF at (**a**) 15 JST and (**b**) 17 JST. *Circles* in (**b**) indicate the intense rainfall regions which extended northwestward. *Arrows* in (**a**) indicate the small rainfall regions that developed into the intense rainfall regions at 17 JST

expected that analyzed values of the Outer LETKF inside and around the domain of the Inner LETKF were modified by this procedure. To investigate the impacts of this procedure, the spread distributions of rainfall and sea level pressure were compared with ones obtained by performing the assimilation without the Inner LETKF (Fig. 20.7). In this case, the rainfall distribution of the Outer LETKF was similar to that obtained by performing the assimilation without the Inner LETKF (Figs. 20.5a and 20.7a). However, the spread distributions of the Outer LETKF region were affected by the Inner LETKF. Namely, the spread of rainfall in the Inner LETKF became larger and that of sea level pressure became smaller, when the Inner LETKF was used (indicated by arrows in Figs. 20.7c, d). These small spreads of sea level pressure expanded outside of the domain of the Inner LETKF (Figs. 20.5b and 20.7b). This comparison indicates that the Inner LETKF influences the analysis of the Outer LETKF, though the impact was not large in this study.

Next, the convection cells near the boundaries of the Inner LETKF are described. Because the analyzed values of the Outer LETKF within the domain of the Inner LETKF were replaced with the analyzed values of the Inner LETKF, this procedure might lead to the generation of unrealistic convection cells near the boundary of the Inner LETKF, especially when the analyzed horizontal winds between the Inner and Outer LETKFs are greatly different. Figure 20.8 shows the rainfall and horizontal

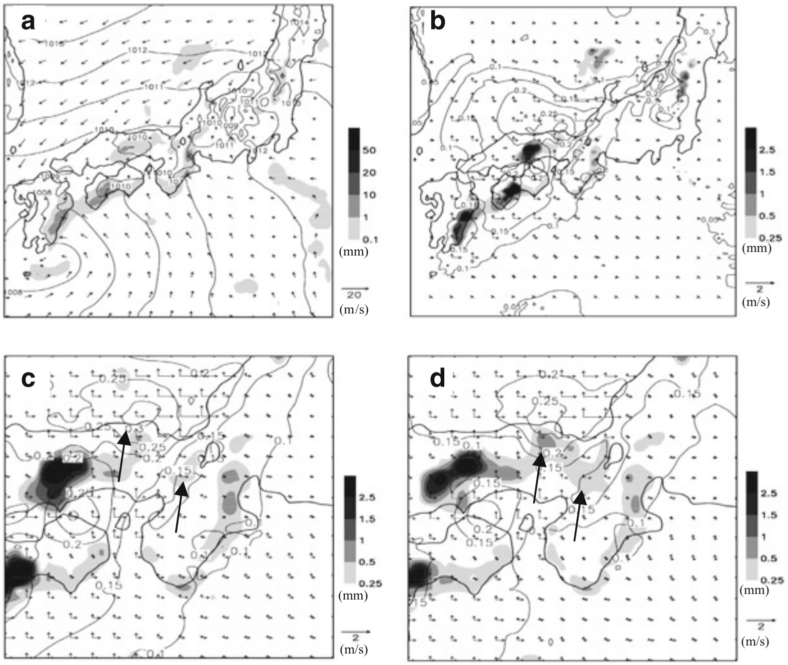


Fig. 20.7 (**a**) Ensemble mean and (**b**) spread of 1 h rainfall (*shaded regions*), horizontal wind at the height of 20 m (*vectors*) and sea level pressure (*contours*) at 15 JST 5th September 2008 obtained by performing the assimilation without the Inner LETKF. (**c**, **d**) Enlarged distribution of spreads that are shown in Fig. 20.7b (without the Inner LETKF) and 20.5b (with the Inner LETKF). *Arrows* in (**c**) and (**d**) indicate the regions in which spreads of sea level pressure and rainfall were larger, respectively

wind distributions of the ensemble member #005 that were obtained by deploying 4 Inner LETKFs side by side. In this experiment, the configuration of the Inner LETKFs was the same as the aforementioned one except that their horizontal grid number was changed to 121 × 121. In the reproduced distributions, the rainfall regions and horizontal wind varied smoothly near the boundary of the Inner LETKFs and the unrealistic convection cells were not generated there. It is deduced that the blending of the analyzed values of the Outer and Inner LETKFs at the boundaries of Inner LETKFs reduced the generation of unrealistic convection cells.

20.4.3 Impact of No-Cost Smoother

Next, other improvements of the nested system are explained in this section. The aforementioned experiment of the nested system will be called “CNTL” hereinafter. In CNTL, the initial seeds and boundary conditions of the Inner LETKF were

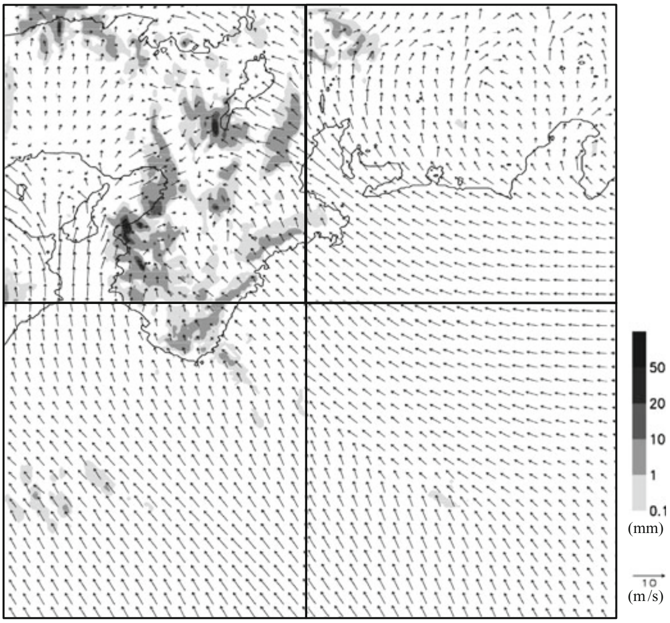


Fig. 20.8 Horizontal distributions of 1 h rainfall (*shaded regions*) and horizontal wind at the height of 20 m (*vectors*) at 15 JST 5th September 2008 obtained by deploying 4 Inner LETKFs side by side

produced by the spatial interpolation of the forecast of the Outer LETKF. Namely, the boundary conditions of the Inner LETKF in CNTL did not have the information of the Outer LETKF's assimilation data, even if the assimilation data existed just outside of the Inner LETKF. To reflect the assimilation data of the Outer LETKF into the Inner LETKF, the analyzed fields obtained by the no-cost smoother (Kalnay et al. 2007; Yang et al. 2009), which is equivalent to the Ensemble Kalman Smoother (Evensen 2003), were used in producing the initial seeds of the Inner LETKF (indicated by thick open arrows in the Outer LETKF in Fig. 20.4a). The boundary conditions of the Inner LETKF were also obtained from the forecast from the analyzed fields of the no-cost smoother. Figure 20.9 shows the ensemble mean distributions of rainfalls at 17 JST that were obtained with the initial seeds and boundary conditions from the Outer LETKF's forecast (CNTL) and from the analyzed fields of the no-cost smoother. When the analyzed fields of the no-cost smoother were used, the rainfall regions, which were not generated in CNTL, were generated at the northwestern part of the Inner LETKF's domain (indicated by an arrow in Fig. 20.9). In the no-cost smoother experiment, the conventional data that was used in Outer LETKF was assimilated again in the Inner LETKF. However, it is considered that the conventional data can be used in the Inner LETKF, because this data provides the small-scale information through the small scale localization in the Inner LETKF. Since these rainfall regions that were generated in the no-cost

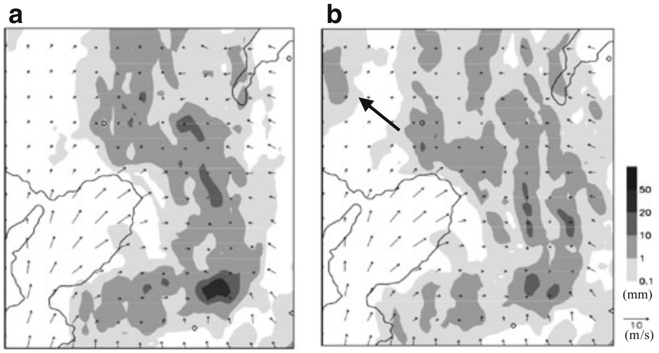


Fig. 20.9 Ensemble mean distributions of rainfall (*shaded regions*) and horizontal wind at the height of 20 m (*vectors*) reproduced by the Inner LETKF at 17 JST obtained from (**a**) the initial conditions of the Outer LETKF's forecast (CNTL) and (**b**) the analyzed fields of the no-cost smoother. An *arrow* in (**b**) indicates the rainfall region that was not generated in CNTL and generated in the no-cost smoother experiments

smoother experiments were close to the boundary of Inner LETKF, it is deduced that the observation data just outside of the Inner LETKF's domain modified the analysis of the Inner LETKF through the initial seeds and boundary conditions of the Inner LETKF. This result indicates that the no-cost smoother can improve the rainfall distribution of the Inner LETKF. However, the whole observation data in the assimilation window period (6 h from the initial time of assimilation in this experiment) needs to be waited for if the no-cost smoother is used. If real-time analysis is required, using the initial seeds and boundary conditions produced from the Outer LETKF's forecast might be a more realistic approach.

Next, the initial seeds of the experiments using the no-cost smoother are explained. When the initial seeds are produced by the interpolation of the analysis of the Outer LETKF, the small-scale disturbances that cannot be resolved by the grid interval of the Outer LETKF are not included in the initial seeds. To reproduce local heavy rainfalls, small-scale disturbances should be included in the initial seeds because they are expected to affect the generation of the convection cells. To show the impact of small-scale disturbances in the initial seeds, the experiments were performed with the Inner LETKF of the grid number of 121 × 121 by using the interpolated fields of the Outer LETKF (**L** at 9 JST in Fig. 20.4a) and the last analyzed fields of the Inner LETKF (**S** at 9 JST in Fig. 20.4a), as the initial seeds at 9 JST (indicated by an open pentagon in Fig. 20.4a). Figure 20.10 shows the spread fields of rainfalls at 11 JST, after 2 cycles of the Inner LETKFs (indicated by a gray pentagon in Fig. 20.4a). During 2 cycles of the Inner LETKF, the boundary conditions of the Inner LETKF that were produced by assimilation of conventional data in the Outer LETKF were used, and the conventional data was assimilated from 0910 JST to 1100 JST. These procedures from 0900 JST to 1100 JST are common to two experiments. Because the mesoscale convergence was reproduced in both seeds, the rainfall regions were similar to each other (not shown). However, relatively large

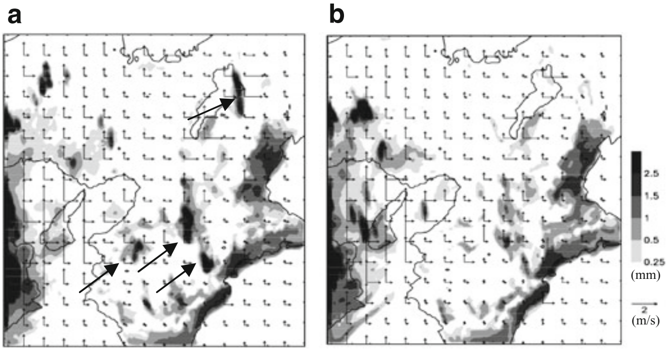


Fig. 20.10 Spread distributions of 1 h rainfall (*shaded regions*) and surface wind at the height of 20 m (*vectors*) reproduced by the Inner LETKF at 11 JST of which initial seeds were produced by using (**a**) the interpolated fields of the Outer LETKF (**L** in Fig. 20.4a) and (**b**) the last analyze fields of the Inner LETKF (**S** in Fig. 20.4a). *Arrows* in (**a**) indicate the rainfall regions of which spreads became smaller when the last analyzed fields of the Inner LETKF were used as the initial seeds

spreads that were seen at several points became smaller, when the analysis fields of the Inner LETKF (**S**) were used (indicated by thin arrows in Fig. 20.10a). Because small-scale convergences might influence the occurrence of local heavy rainfalls, initial seeds which include small-scale disturbances are desired.

In the CNTL experiment, the initial seeds of the second or later sets of the Inner LETKF (**A** in Fig. 20.4a) should be close to the initial conditions of the Outer LETKF (**B** in Fig. 20.4a). Because the initial conditions (**B**) are originated from the last analyzed fields of the Inner LETKF (**S** in Fig. 20.4a), the smoothed distributions of the last analyzed fields of the Inner LETKF (**S**) are similar to the interpolated distributions (**L** in Fig. 20.4a) of the Outer LETKF's initial conditions (**B**). Therefore, the last analyzed fields of the Inner LETKF (**S**) are used directly as the initial seeds of the Inner LETKF in the next assimilation set (**A**). In the experiments in which the no-cost smoother is adopted in producing the initial conditions of the Inner LETKF, however, the analyzed fields of the no-cost smoother at the initial time of assimilation are different from the smoothed distributions of the last analyzed fields of the Inner LETKF (**S**). Therefore, the initial seeds of the Inner LETKF should be produced by adding the small disturbances, which were extracted from the last analyzed fields of the Inner LETKF (**S**), to the analyzed fields of the no-cost smoother.

20.5 Impact of Doppler Radar Data and GPS Water Vapor Data

As mentioned in Sect. 20.2, convergence of low-level water vapor is essential to reproduce local heavy rainfalls. In this section, impacts of the GPS data and Doppler radar data, which provide information of convergence of water vapor, are shown.

20.5.1 Impact of GPS Water Vapor Data

Due to frequent occurrence of earthquakes, movements of clusters are monitored by more than 1,200 GPS receivers that have been deployed by the Geospatial Information Authority of Japan. Because signals of radio waves transmitted from GPS satellites are delayed by water vapor in the atmosphere, the delays of signals are estimated as well as GPS receivers' positions. In this study, PWV and SWV that were estimated from the delays (Shoji et al. 2004), which have information of water vapor, were used as assimilation data. PWV and SWV are the integrated value of water vapor in the column or along the paths from GPS satellites to GPS receivers. For this study, we produced intermediate profiles of relative humidity from observations and statistical data from outputs of the ensemble forecast, and these profiles were assimilated by LETKF. In the estimation of intermediate profiles, the following two assumptions were used; (1) Differences between intermediate profile and first guess are proportional to the spread of relative humidity. Due to position errors of rainfall regions and large dispersion of relative humidity distributions caused by small ensemble size, area-mean relative humidity profile and area-maximum spread profiles of relative humidity within the areas from 18 km from GPS receivers were used as the first guess and spread profiles of relative humidity (detailed procedures were explained in Seko et al. 2011). (2) The intermediate profile is produced at the layers where the correlation among the ensemble members between relative humidity of each layer and PWV exists (Fujita et al. 2011). In this study, the correlation of 0.3 was used as the threshold. Namely, relative humidity was increased where the correlation was larger than 0.3, and decreased where the correlation was smaller than –0.3. The assimilation method of SWV was the same as that of PWV, except for the slant paths and the small areas that were used in producing the ensemble mean and maximum spread profiles of relative humidity. Because SWV is water vapor between GPS satellites and receivers, SWV data provides water vapor values as well as its direction. If large areas were used in producing the ensemble mean and maximum spread profiles, the direction would become ambiguous because large areas dilute this information. To exploit this advantage of SWV, areas used in producing ensemble mean and maximum spread were reduced from 18 to 3 km.

Figure 20.11 shows the rainfall regions at 17 JST that were obtained by assimilation of PWV and SWV data. In addition to convectional data, PWV and SWV data from 9 to 15 JST were assimilated in the Inner LETKF. When the PWV data was added, the number of ensemble members in which the rainfall regions were extending northwestward increased from 4 to 7 (#001–#004, #008, #009 and #011) (Figs. 20.6b and 20.11a). Because the assimilation of PWV data modified the water vapor that was supplied into the rainfall region, the intensities of the rainfall regions are expected to be improved. In this experiment, the position and intensities of rainfall regions were improved. When SWV data was added to assimilation data, the intense rainfall regions were generated at the northwestern side of Sakai City, where the intense rainfall region was observed (16 JST of Fig. 20.2b), in most of

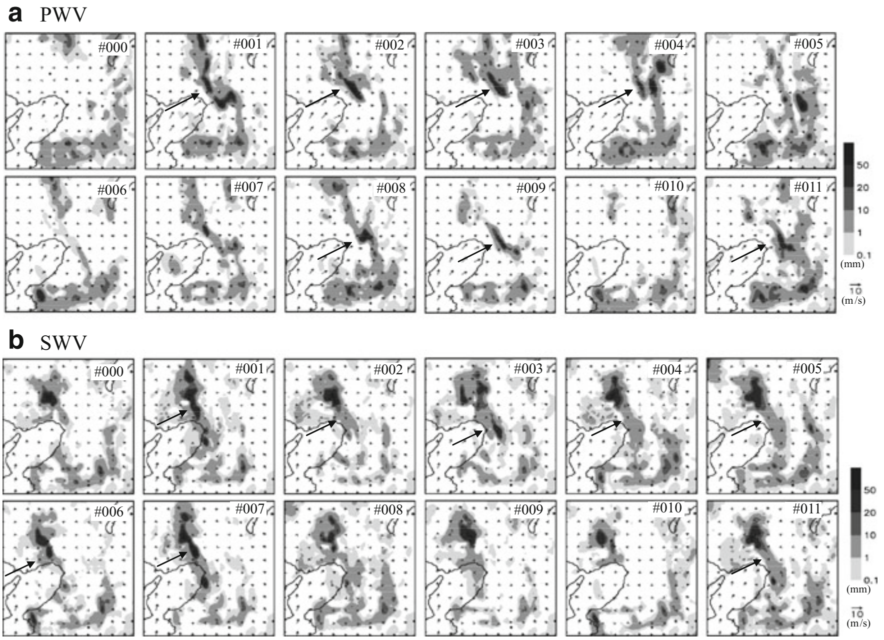


Fig. 20.11 Horizontal distributions of 1 h rainfall (*shaded regions*) and horizontal wind at the height of 20 m (*vectors*) reproduced by the Inner LETKF at 17 JST obtained by the assimilation of (**a**) PWV data and (**b**) SWV data. *Arrows* indicate the rainfall regions that extended northwestward

the ensemble members. It is deduced that some paths from GPS receivers to GPS satellites penetrated the humid regions that generated the convection cells northwest of Sakai City. The rainfall regions extending northwestward were reproduced in 9 ensemble members (#000–#007, #011) (Fig. 20.11b). The number of ensemble members in which the intense rainfall was reproduced was increased by assimilation of GPS water vapor data. This means that GPS water vapor data had a strong influence on the reproduction of the heavy rainfall. The ensemble spread is expected to be smaller because the rainfall regions became closer to the observed ones by data assimilation. In the case that the rainfall regions with small spreads are greatly different from the observed ones, the ensemble spreads should be wider by removing the causes of the small spreads before data assimilation. On the other hand, the analyzed rainfall regions in this study were close to the observed ones. These distributions indicate that the ensemble forecasts were performed appropriately.

20.5.2 Impact of Doppler Radar Data and the Synergistic Effect of GPS Water Vapor Data and Radial Wind Data

As for the Doppler radar data, two kinds of data were assimilated. The first was the horizontal wind obtained by the dual analyses of the radial wind of Kansai and

Osaka international airports (indicated by two circles in Fig. 20.2a) (Tanaka and Suzuki 2000). Although horizontal wind obtained by dual analyses of radial wind can be assimilated directly as assimilation data, the regions where the horizontal wind is estimated are smaller than those of radial wind, because the horizontal wind can be obtained in the overlapping regions of two Doppler radars. The second kind of data is radial wind of Doppler radar. This wind data are expected to be more effective to improve the rainfall forecasts because it provides the information from a wider area.

Figures 20.12a, c are the rainfall distributions at 17 JST that were obtained by the assimilation of the horizontal wind and the radial wind. The horizontal and radial winds from 1400 to 1500 JST were assimilated, because the rainfall regions were fewer and smaller before 1400 JST. When the horizontal winds were added to assimilation data, the number of ensemble members in which the intense rainfall regions extended northwestward was increased from 4 to 7 (#001–004, #007–#009). When the radial winds were assimilated, the intense rainfall regions that extended northwestward were reproduced in 9 ensemble members (#001–#008 and #011), though the rainfall intensity remained relatively weak. These results indicate that the wind data, especially radial wind, can improve rainfall forecasts. This impact of wind data was the same as that of variational data assimilation methods (Kawabata et al. 2007).

To show the synergistic effects of water vapor data and wind data, simultaneous assimilation of the horizontal wind and PWV was performed. Figure 20.12b shows the rainfall distributions of 17 JST. When both data were added to assimilation data, the rainfall forecasts in the ensemble members #007 and #011, in which the intense rainfall region was not reproduced by the individual assimilation of the PWV and horizontal wind, were improved (Figs. 20.11a and 20.12a, b). The improvements of rainfall forecasts in the ensemble members #002 and #003 became obscure, compared with those in which the PWV or horizontal wind was assimilated separately (Figs. 20.11a and 20.12a, b). However, the rainfall distributions of #002 and #003 remained better than those obtained from the assimilation of conventional data (Figs. 20.6a and 20.12b). These results indicate that the simultaneous assimilation is useful for increasing the number of members in which local heavy rainfalls are reproduced. This result is consistent with that of Seko et al. (2004). In this study, the synergistic effects of water vapor data and wind data were investigated by the combination of the PWV and horizontal wind. Further improvements are expected when the SWV and radial wind data are assimilated simultaneously, because the impacts of the SWV and radial wind were more significant than those of the PWV and horizontal wind. The combinations of water vapor data and wind data except PWV and horizontal wind is to be investigated in the future study.

20.6 Summary and Future Plan

The nested LETKF system was developed to reproduce local heavy rainfalls. In the experiments in this study, the convection cells of local heavy rainfalls were well reproduced in one third of ensemble members of the Inner LETKF, when the

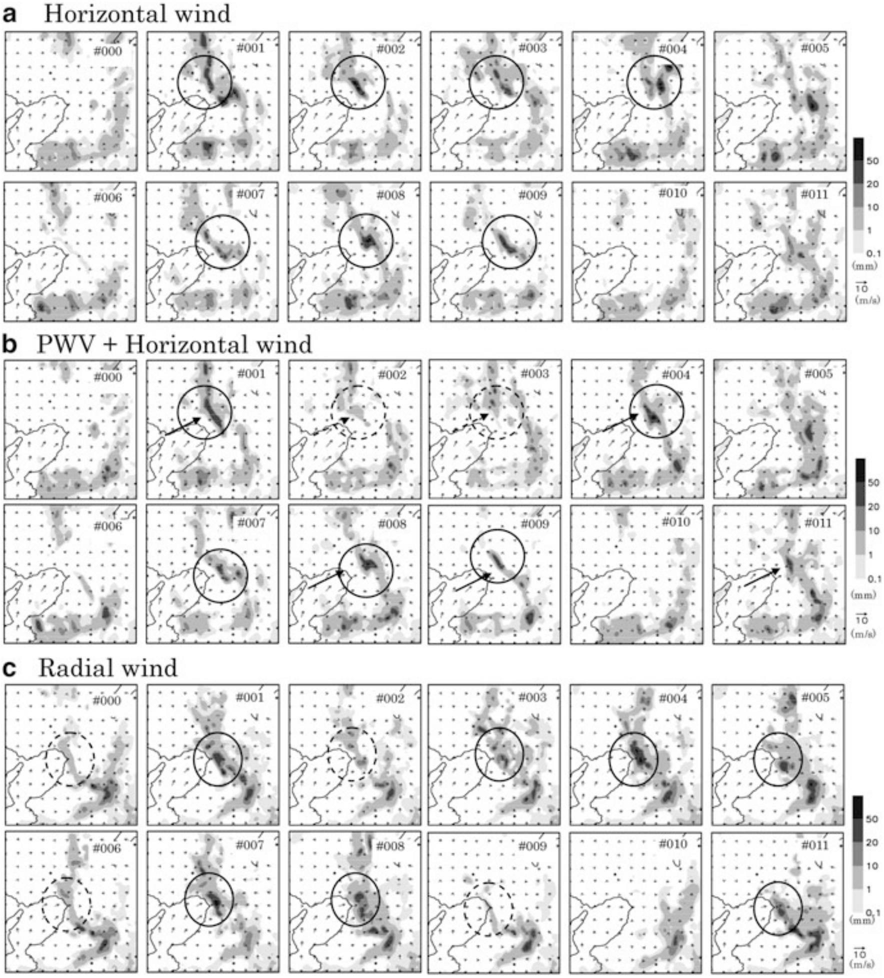


Fig. 20.12 Horizontal distributions of 1 h rainfall (*shaded regions*) and horizontal wind at the height of 20 m (*vectors*) reproduced by the Inner LETKF at 17 JST obtained by the assimilation of (**a**) horizontal wind, (**b**) horizontal wind and PWV, and (**c**) radial wind. *Circles* and *arrows* indicate the rainfall regions that were reproduced by the assimilation of wind data and of GPS water vapor data, respectively. *Circles* and *arrows* of *broken lines* indicate the rainfall regions where the rainfalls were reproduced but their intensities were weaker

conventional data from JMA was assimilated. Assimilations of GPS water vapor data and Doppler radar data further increased the number of ensemble members in which local heavy rainfalls were reproduced. These results indicate that the nested LETKF system has the potential to be used as a data assimilation system for local heavy rainfalls. In general, high-resolution ensemble forecasts using high-resolution data (e.g. Radar data) have a huge computational cost. The nested LETKF system has the merit that it can be executed with limited computational resources by cutting out small regions from a whole domain. Some modifications, such as the no-cost smoother that was used in producing the initial seeds and boundary conditions of the inner LETKF, were also investigated. To apply this nested system to various phenomena (e.g. synoptic fronts and typhoons), rainfall systems that extend wider than the domain of the Inner LETKF should be investigated.

To increase the accuracy of rainfall forecast, the high-resolution data that provides small-scale disturbances is indispensable. The project "Tokyo Metropolitan Area Convection Study for Extreme Weather Resilient Cities (TOMACS)" in which local heavy rainfalls developed in the Tokyo Metropolitan area were observed by Ku-band radar, surface dense network and Doppler lidars etc. was started in 2011. The observation data of TOMACS will be used as assimilation data of the nested LETKF system.

Because of the 'Tohoku-earthquake' that occurred over Eastern Japan on 11th March 2011, computer resources were limited so the number of ensemble members and the horizontal grid of LETKFs became 12 and 80 × 80 (121 × 121). However, results of this study show the potential of the nested LETKF system. The results shown in this study will be further investigated using the high-performance super computer'Kei', which will be provided by the 'High Performance Computing Infrastructure' project.

Acknowledgements The authors would like to express their gratitude to Dr. Yoshinori Shoji of Meteorological Research Institute, Dr. Tadashi Fujita of Numerical Prediction Division/JMA, Mr. Yuji Esaki of Osaka District Meteorological Observatory and anonymous reviewers, who provided the PWV and SWV data and their useful comments. The authors' gratitude extends to the Geospatial Information Authority of Japan and the Osaka District Meteorological Observatory of JMA, which provided the GPS data and Doppler radar data. The improvements of severe weather forecasts (i.e. local heavy rainfalls), which were achieved by the assimilations of Doppler radar data, will contribute to aviation safety and the mitigation of damages of other urban functions. The "Kasaneru 3D tool" developed by Tokyo District Meteorological Observatory of JMA was used to generate the rainfall distribution graphics observed by the operational radars and the surface meteorological data distribution graphics. This study was performed by the "Studies on formation process of line-shaped rainfall systems and predictability of rainfall intensity and moving speed" and the "Study of advanced data assimilation and cloud resolving ensemble technique for prediction of local heavy rainfall" projects.

References

Dowell DC, Zhang F, Wicker LJ, Snyder C, Crook NA (2004) Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, Supercell: ensemble Kalman filter experiments. Mon Wea Rev 132:1982–2005

Evensen G (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn 53:343–367

Fujita T, Tohru K, Origuchi S, Seko H, Saito K (2011) Development of Meso-LETKF. In: Proceedings of the autumn conference of meteorological society Japan B213:(in Japanese)

Hunt BR, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230:112–126

Kalnay E, Li H, Miyoshi T, Yang S-C, Ballabrera-Poy J (2007) Response to the discussion on "4-D-Var or EnKF?" by Nils Gustafsson. Tellus 59A:778–780

Kawabata T, Seko H, Saito K, Kuroda T, Tamiya K, Tsuyuki T, Honda Y, Wakazuki Y (2007) An assimilation and forecasting experiment of the nerima heavy rainfa11 with a cloud-resolving nonhydrostatic 4-dimensional variational data assimilation system. J Meteor Soc Jpn 85: 255–276

Meng Z, Zhang F (2008) Test of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: comparison with 3DVar in a real-data case study. Mon Wea Rev 136: 522–540

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

Miyoshi T, Kunii M (2012) The local ensemble transform Kalman filter with the weather research and forecasting model: experiments with real observations. Pure Appl Geophys 169:321–333

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 Wea Rev 134:1266–1298

Seko H, Kawabata T, Tsuyuki T, Nakamura H, Koizumi K, Iwabuchi T (2004) Impacts of GPS-derived water vapor and radial wind measured by Doppler radar on numerical prediction of precipitation. J Meteor Soc Jpn 82:473–489

Seko H, Miyoshi T, Shoji Y Saito K (2011) Data assimilation experiments of precipitable water vapor using the LETKF system: intense rainfall event over Japan 28 July 2008. Tellus 63A: 402–412

Shoji Y, Nakamura H, Iwabuchi T, Aonashi K, Seko H, Mishima K, Itagaki A, Ichikawa R, Ohtani R (2004) Tsukuba GPS dense net campaign observation: improvement in GPS analysis of slant path delay by stacking one-way postfit phase residuals. J Meteor Soc Jpn 82:301–314

Snyder C, Zhang F (2003) Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. Mon Wea Rev 131:1663–1677

Tanaka Y, Suzuki O (2000) Development of radar analysis software "Draft". In: Proceedings of the spring conference of meteorological society Japan vol 303:(in Japanese)

Tong M, Xue M (2005) Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSSE experiments. Mon Wea Rev 133:1789–1807

Xue M, Tong M, Droegemeier KK (2006) An OSSE framework based on the ensemble square root Kalman filter for evaluating the impact of data from radar networks on thunderstorm analysis and forecasting. J Atmos Oceanic Tech 23:46–66

Yang S-C, Corazza M, Carrassi A, Kalnay E, Miyoshi T (2009) Comparison of local ensemble transform kalman filter, 3DVAR, and 4DVAR in quasigeostrophic model. Mon Wea Rev 137:693–709

Zhang F, Snyder C, Sun J (2004) Impacts of initial estimate and observation availability with an ensemble Kalman filter. Mon Wea Rev 132:1238–1253

Zhang F, Meng Z, Aksoy A (2006) Test of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part I: perfect model experiments. Mon Wea Rev 134:722–736

Chapter 21
Observing-System Research and Ensemble Data Assimilation at JAMSTEC

Takeshi Enomoto, Takemasa Miyoshi, Qoosaku Moteki, Jun Inoue, Miki Hattori, Akira Kuwano-Yoshida, Nobumasa Komori, and Shozo Yamane

Abstract Recent activities on ensemble data assimilation and its application to observing-system research at the Japan Agency for Marine-Earth Science and Technology are reviewed. A revised version of an ensemble-based data assimilation system for global atmospheric data has been developed on the second-generation Earth Simulator. This system assimilates conventional atmospheric observations and satellite-based wind data into an atmospheric general circulation model using the local ensemble transform Kalman filter (LETKF), a deterministic ensemble Kalman filter algorithm that is extremely efficient with parallel computer architecture. The updated system incorporates improvements to the previous system in the forecast model, data assimilation algorithm and input data. Using the LETKF system,

T. Enomoto (✉)
Disaster Prevention Research Institute, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology, Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan
e-mail: eno@dpac.dpri.kyoto-u.ac.jp

T. Miyoshi
RIKEN Advanced Institute for Computational Science, 7-1-26, Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD 20742, USA

Q. Moteki · J. Inoue · M. Hattori
Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushimacho, Yokosuka, Kanagawa 237-0061, Japan

A. Kuwano-Yoshida · N. Komori
Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology, Showamachi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan

S. Yamane
Department of Environmental Systems Science, Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe 610-0394 Kyoto, Japan

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_21,

observations taken during field campaigns are evaluated by data assimilation experiments involving adding or removing observations. The results of these observing-system experiments successfully demonstrate the value of the observations and are highly useful for exploring the predictability of atmospheric disturbances.

21.1 Introduction

The Japan Agency for Marine-Earth Science and Technology (JAMSTEC) conducts observations of the climate system over ocean and land in various parts of the globe. JAMSTEC operates eight research vessels (Chikyu, Natsushima, Kaiyo, Yokosuka, Mirai, Kairei, Hakoho-maru and Tansei-maru), Triangle Trans-Ocean Buoy Network (TRITON) in the western tropical Pacific and eastern Indian Oceans, Polar Ocean Profling System (POPS) in the Arctic Ocean and hundreds of Argo floats. In addition, JAMSTEC conducts observations of the hydrological cycle of the cryosphere and field campaigns to collect data on the atmosphere and ocean. These various data sources provide a number of measurements daily, contributing to the monitoring and investigation of the climate system.

JAMSTEC has one of the world's largest super computer systems devoted to the earth sciences, the Earth Simulator. The Earth Simulator has been used to conduct large-scale simulations in various areas in solid earth, ocean and atmospheric sciences. The observational and computational capabilities at JAMSTEC provide enormous opportunity for data assimilation. Ocean observations have already been used to produce a number of data sets. A monthly global ocean analysis of Argo, TRITON and CTD (conductivity, temperature and depth), called MOAA (Monthy Objective Analysis using the Argo data) GPV (grid point value), was produced by the optimal interpolation (Hosoda et al. 2008). Argo was also used to construct G-YoMaHa, with an objectively mapped velocity at 1,000 dvar (Katsumata and Yoshinari 2010) and MILA (Mixed Layer data set of Argo) GPV (Hosoda et al. 2010). The four-dimensional variational algorithm has been successfully applied to an ocean general circulation model (Masuda et al. 2003) and a global coupled atmosphere–ocean model (Sugiura et al. 2009). Atmospheric observations, however, have not been fully utilized, indicating the potentials for data assimilation and observing-system research.

A collaborative project was conducted from fiscal year (starting in April) 2006 to 2008 to develop an ensemble data assimilation system among the Japan Meteorological Agency (JMA), JAMSTEC and the Chiba Institute of Science. The system, named ALEDAS (AFES–LETKF ensemble data assimilation system), is composed of the atmosphere general circulation model (AGCM) for the Earth Simulator (AFES) as a forecast model (Numaguti et al. 1997; Ohfuchi et al. 2004; Enomoto et al. 2008) and the local ensemble transform Kalman filter (LETKF) as an assimilation algorithm (Hunt et al. 2007; Miyoshi and Yamane 2007). Observations were prepared from those used for numerical weather prediction at JMA except for satellite radiances. An experimental analysis data set, called ALERA

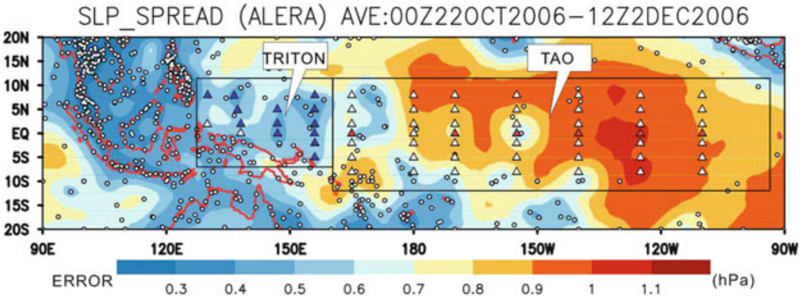


Fig. 21.1 The analysis ensemble spread of the sea-level pressure averaged between 0 UTC October and 12 UTC 2 December 2006 over the equatorial Pacific. *Circles* indicate surface observations and *triangles* indicate buoys. *Filled triangles* denote buoys equipped with a barometer. TRITON and TAO regions are marked by *rectangles*

(AFES–LETKF experimental ensemble reanalysis), with 40 members was produced for approximately one and a half years beginning in May 2005 (Miyoshi et al. 2007a). These data are freely available from the Earth Simulator Center. The quality of ALERA is comparable to existing long-term reanalysis data sets in the troposphere even though it does not use satellite radiance observations. In addition, ALERA provides a flow-dependent analysis ensemble spread, which can be regarded as a measure of the analysis error.

Figure 21.1 shows the analysis ensemble spread of the sea-level pressure averaged for approximately 40 days in the Tropical Atmosphere Ocean (TAO) and TRITON regions. In general, the analysis ensemble spread is small over land owing to the dense observational network. Over the equatorial Pacific, there is a longitudinal asymmetry between the TRITON and TAO regions. Most of the TRITON buoys are equipped with a barometer. These barometers contribute to a spread smaller than 0.6 hPa. In contrast, only TAO buoys along the equator have a barometer. Pressure observations at 170 and 155 W contribute to the local minima, which are likely achieved with surrounding observations over land. The distribution of the analysis ensemble spread implies the importance of two-dimensional coverage to reduce analysis error.

Because the analysis ensemble spread is dependent on observation density, ALEDAS can be used to evaluate observations. The impact of observations can be quantified by the reduction in the analysis ensemble spread. The first observing-system experiments (OSEs) using ALEDAS were conducted to evaluate the impact of additional dropsonde observations in the western Pacific during the PALAU (the Pacific area long-term atmospheric observation for understanding of climate change) 2005 field campaign (Moteki et al. 2007). These researchers found the influence of the temperature and humidity observations over the western Pacific reaches as far as Japan through the southerly peripheral flow of the Pacific anticyclone. Moteki et al. (2011) further investigated the impact of observations in the tropics. Moteki et al. conducted OSEs with additional sondes at three locations in the Indian Ocean during the MISMO project (Yoneyama et al. 2006) and found that the reduction in analysis error in terms of the difference of the vertically integrated analysis ensemble spread of the geopotential height (OSE–ALERA) extends eastward in Kelvin waves and westward in Rossby waves. Moteki et al. also found signals indicating that the tropical cyclogenesis is affected by the Kelvin wave. In the Arctic Ocean, Inoue et al. (2009) conducted data denial experiments to evaluate the surface pressure observations by drifting buoys. The analysis ensemble spread, not only in the Beaufort Sea, where buoys are densely deployed, but also throughout the whole Arctic Ocean, increased without the surface pressure observations north of 70 N. The influence of the buoys reaches 700 hPa. The relationship between the number of buoys and accuracy is confirmed with consistency among the different reanalysis data sets.

The analysis is not static but varies in time under the influence of atmospheric disturbances. Enomoto et al. (2010) investigated the relationship between the analysis ensemble mean and the spread of ALERA in various atmospheric phenomena. These researchers found an increase in the analysis ensemble spread prior to the westerly bursts in the tropical eastern Indian Ocean, in the onset of the monsoon westerlies in southern Vietnam, and even in the stratospheric sudden warming. Because the error growth implies instability in the linear perturbation theory, it is anticipated that the analysis ensemble spread, which is an estimate of the analysis error, contains some precursory signals. The actual error growth is, however, nonlinear owing to the finite amplitude and complex physical processes and is not fully understood. Further investigations into the precursory signals contained in the analysis ensemble spread require the variables related to physical processes, which, unfortunately, are not included in ALERA.

Motivated by the success of the observing-system experiments and predictability studies, JAMSTEC formed a research team called OREDA (Observing-System Research and Ensemble Data Assimilation development research team). This article reports the activities of OREDA related to the development of the ensemble data assimilation system of the global atmospheric data and OSEs. The updated version of the ensemble data assimilation system (ALEDAS2) is described in Sect. 21.2. OSEs were conducted with ALEDAS2. Preliminary results are shown in Sect. 21.3. Finally, the summary and our plan for future research are given in Sect. 21.4.

21.2 The Ensemble Data Assimilation System

ALEDAS2 is composed of improved versions of AFES and LETKF. In the forecast step, an ensemble forecast is conducted with AFES to propagate the mean and covariance to the next time level. In the analysis step, LETKF, one of the deterministic Kalman filters that assimilate observations to the ensemble mean, is used. This section summarizes the updates of the forecast model and analysis scheme and describes the forecast–analysis cycle of ALEDAS2. With this system, we are producing ALERA2, a successor to ALERA. The configurations for ALERA2 are described in comparison with those for ALERA.

Table 21.1 The root-mean square difference of the 5-day geopotential height forecast at 500 hPa averaged over the Northern Hemisphere for the different versions of AFES forecast from the JMA analysis

Version	Resolution	RSMD m	Remarks
1.22	T159L48	55.4	Original
2.2	T159L48	52.1	ALERA
2.7	T119L48	51.3	
2.7 with MATSIRO	T119L48	49.5	
3.6	T119L48	48.9	ALERA2

21.2.1 The Forecast Model

AFES integrates the primitive equations of winds, temperature, specific humidity, cloud water and surface pressure using the spectral transform method and Eulerian advection, and it has physics schemes common to many forecast and climate models. As in the version of AFES used in ALERA, the radiative fluxes are parameterized using mstrnX (Sekiguchi and Nakajima 2008), and the cumulus convection is represented by the Emanuel scheme (Emanuel 1991; Emanuel and Živković-Rothman 1999; Peng et al. 2004) without discrimination between shallow and deep convection. Updated physics schemes of AFES include cloud (Kuwano-Yoshida et al. 2010) and land-surface schemes (Takata et al. 2003). The new cloud scheme with moist turbulence improves the representation of the boundary-layer clouds in the eastern oceanic basins. The land-surface scheme MATSIRO improves the modeling of the hydrological cycle.

Reduction in the forecast error of AFES has been achieved continuously. Table 21.1 shows the geopotential height error averaged in the Northern Hemisphere (> 30 N) in the forecast experiments for August 2004 with different versions of AFES. We conducted a 5-day forecast from the analysis at 12 UTC on each day in August 2004. The root-mean square difference from the JMA analysis was regarded as an error. AFES 2.2 used in ALERA has an error of 52.1 m, a 3.3 m (approximately 6 %) decrease from the original version (Ohfuchi et al. 2004). Better estimations of the cloud water (Bony and Emanuel 2001) and of the saturation specific humidity allow the error to decrease from AFES 2.2 to 2.7 despite the use of a somewhat coarser resolution. The introduction of the new land-surface and cloud schemes contributed to a further reduction in the forecast error. The RSMD of AFES 3.6 used in ALERA2 is reduced by approximately 6 % from AFES 2.2 used in ALERA and by more than 10 % from the original version.

21.2.2 Analysis Scheme

ALEDAS2 uses the LETKF for analysis as in ALEDAS but with distance-based covariance localization. Before illustrating this improvement, the formulation is briefly described following Hunt et al. (2007). LETKF is a deterministic ensemble

Kalman filter, and the unperturbed observations are assimilated to update the ensemble mean. The analysis ensemble mean $\bar{\boldsymbol{x}}^{\mathrm{a}}$ is calculated from the forecast ensemble mean $\bar{\boldsymbol{x}}^{\mathrm{f}}$ and the linear combination of the ensemble forecast perturbation matrix $\mathbf{X}^{\mathrm{f}}$ with the weight $\bar{w}^{\mathrm{a}}$.

$$\bar{\boldsymbol{x}}^{\mathrm{a}} = \bar{\boldsymbol{x}}^{\mathrm{f}} + \mathbf{X}^{\mathrm{f}}\bar{\boldsymbol{w}}^{\mathrm{a}}, \tag{21.1}$$

where superscripts a, f, and o denote analysis, forecast and observation, respectively, and over bars indicate the ensemble mean. The i th column (i th member) of $\mathbf{X}^{\mathrm{f}}$ is $\boldsymbol{x}_i^{\mathrm{f}} - \bar{\boldsymbol{x}}^{\mathrm{f}}$. The weight is determined by

$$\bar{\boldsymbol{w}}^{\mathrm{a}} = \tilde{\mathbf{P}}^{\mathrm{a}} \left(\mathbf{Y}^{\mathrm{f}}\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\boldsymbol{y}^{0} - \bar{\boldsymbol{y}}^{\mathrm{f}}\right) \tag{21.2}$$

where $\mathbf{R}$ the observation error covariance matrix and $\mathbf{Y}^{\mathrm{f}}$ is a forecast perturbation matrix in the observation space, whose i th column is $\boldsymbol{y}_i^{\mathrm{f}} - \bar{\boldsymbol{y}}^{\mathrm{f}}$. The forecast observation vector $\boldsymbol{y}_i^{\mathrm{f}}$ is calculated by

$$\boldsymbol{y}_i^{\mathrm{f}} = H\left(\boldsymbol{x}_i^{\mathrm{f}}\right). \tag{21.3}$$

where H is the observation operator. The analysis covariance is computed as

$$\tilde{\mathbf{P}}^{\mathrm{a}} = \left[(k-1)\mathbf{I} + (\mathbf{Y}^{\mathrm{f}})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}^{f}\right]^{-1}. \tag{21.4}$$

The analysis ensemble perturbation matrix $\mathbf{X}^{\mathrm{a}}$ is obtained from the forecast ensemble perturbation matrix $\mathbf{X}^{\mathrm{f}}$ using the transform matrix $\mathbf{W}^{\mathrm{a}}$:

$$\mathbf{X}^{\mathrm{a}} = \mathbf{X}^{\mathrm{f}}\mathbf{W}^{\mathrm{a}} \tag{21.5}$$

where

$$\mathbf{W}^{\mathrm{a}} = \left[(k-1)\tilde{\mathbf{P}}^{\mathrm{a}}\right]^{1/2} \tag{21.6}$$

(Bishop et al. 2001). Analysis in LETKF is performed in a local subspace of the model, and different linear combinations of ensemble members in different regions (21.3) can be chosen. In this way, localization acts to reduce the sampling error by making the global dimensions larger than the ensemble size, and to remove spurious correlations between distant locations (Hunt et al. 2007).

In the previous version of the LETKF code used in ALERA (Miyoshi and Yamane 2007), analysis is performed with a square-shaped local patch in the model grid space. The zonal distance of a local patch decreases toward the poles because of the convergence of meridians. As a result, significant discontinuity exists in ALERA, especially in the analysis ensemble spread in the polar regions. In ALEDAS2 the error covariance is localized by physical distance rather than by model-space local patches (Miyoshi et al. 2007b). The weight w for covariance localization diminishes with distance r from each grid point for analysis to an observation in the form

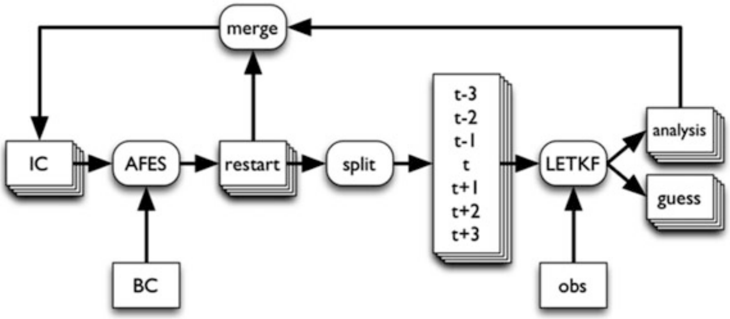


Fig. 21.2 The data flow chart of ALEDAS2. *Rectangles* represent data, and *round rectangles* represent processes

$$w(r) = \exp\left[-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2\right] \tag{21.7}$$

where σ is the localization length parameter. In this modified algorithm, the analysis is conducted at each model grid point. The reduced matrix size contributes to two to threefold gains in speed (Miyoshi et al. 2007b).

21.2.3 The Forecast–Analysis Cycle

Figure 21.2 shows the data flow chart of ALEDAS2. In the forecast step, each ensemble member is integrated in time with AFES from the initial conditions input from an IC file to produce a `restart` file. AFES is capable of running multiple ensemble members with a single MPI execution. The `restart` contains hourly forecasts at 3–9 h (± 3 h from analysis time t) from the initial time. Each restart is split into seven files for input into LETKF. In the analysis step, observations (`obs`) are assimilated into the forecast in a 6-h window by LETKF to produce analysis (`analysis`). Analysis is performed locally in parallel with MPI processes. The `guess` files from LETKF represent the forecast at analysis time t in the `restart` files. Finally, the `analysis` files from LETKF replace the model forecast in the part of the restart files for analysis time t to produce the next initial condition.

21.2.4 Configurations of ALERA2

ALEDAS2 is used to produce the control run for ALERA2 and OSEs. Table 21.2 summarizes the configurations of the two systems. AFES in ALEDAS2 uses a slightly coarser horizontal resolution and the same vertical resolution, but it produces a better forecast because of the improved physics described in Sect. 21.2.1.

Table 21.2 A comparison of the configurations of ALEDAS and ALEDAS2

	ALEDAS	ALEDAS2
AFES version	2.2	3.6
Resolution	T159L48	T119L48
Ensemble size	40	$63 + 1$
Covariance localization	$21 \times 21 \times 13$	400 km/0.4ln p
Spread inflation	0.1	
Observations compiled by	JMA	NCEP

The ensemble size is increased from 40 to 63. Control runs are conducted from the analysis ensemble mean. The localization length is 400 km in the horizontal and 0.4 ln p in the vertical. The spread inflation of 0.1 has not been changed.

For ALERA, the observations used for the global NWP at JMA were provided under the collaboration. For ALERA2, a less restricted PREPBUFR compiled by the National Centers for Environmental Prediction (NCEP) and archived at the University Corporation for Atmospheric Research (UCAR) is used. The number of the observations is reduced as follows. For weather balloons (ADPUPA), the data used are at 1,000, 925, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50 and 10 hPa. Reports from aircraft and satellite retrievals are trimmed to one in every four values and wind profilers to one in every three levels (St-James and Laroch 2005).

NOAA (National Oceanic and Atmospheric Administration) daily $1/4°$ OISST (optimal interpolation sea-surface temperature) version 2 (Reynolds et al. 2007) is used to provide ocean boundary conditions with a higher resolution in both time and space. The initial conditions are prepared from the AMIP (Atmospheric Model Intercomparison Project)-type integration for 20 years. The integration is performed with AFES at the same resolution (T119L48). Ensemble members are arbitrarily chosen to be the atmospheric states of a particular date and nearby dates from different years.

The second generation of the Earth Simulator (ES2) updated in March 2009 is used to conduct ALERA2 and OSEs. Each ensemble forecast is performed on a single vector processor. Each node hosts 8 processes (members), and a total of 64 processes are used on eight nodes of ES2.

21.3 Data Assimilation Experiments

Currently, three streams have been conducted as ALERA2:

- Stream 2003: from 1 June 2003,
- Stream 2008: from 0 UTC 1 January 2008,
- Stream 2010: from 1 August 2010.

Stream 2003 aims to produce a data set that is as long as possible temporally. Stream 2003 began at the earliest continuously available time in the PREPBUFR archive at UCAR[1]. Stream 2008 was integrated to 0 UTC 25 February 2009 to cover the periods of the following field campaigns: PALAU 2008 in the western Pacific, the Mirai Arctic Ocean Cruise 2008, and the summer and winter T-PARC (THORPEX Pacific Asia Regional Campaign; THORPEX: the observing-system research and predictability experiment). Stream 2010 was initiated separately to quickly evaluate field campaigns conducted in 2010: the Mirai Arctic Ocean Cruise 2010 and the Vietnam–Philippines Rain Fall Experiment (VPREX) 2010. Stream 2010 continues to integrate toward the present. In this section, preliminary results are shown for ALERA2, and OSEs are shown for the latter two campaigns.

21.3.1 ALERA2

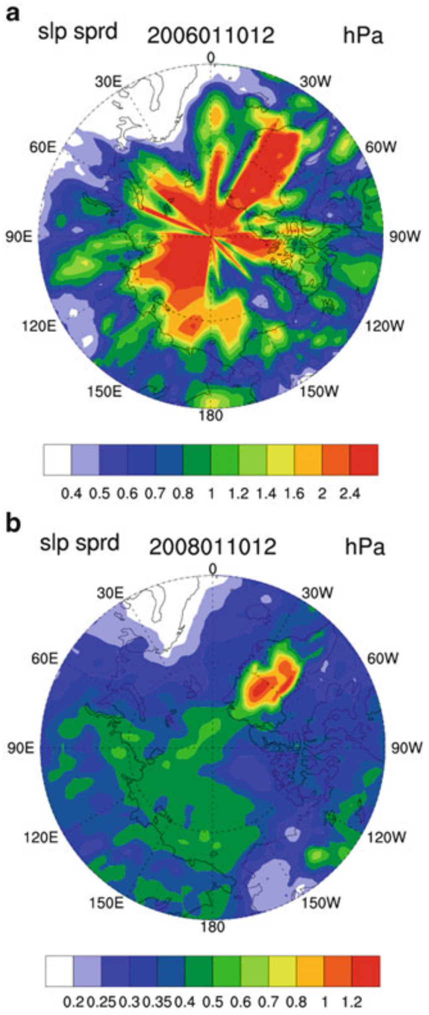
ALERA2 provides smoother analysis ensemble spreads than those in ALERA. Figure 21.3 depicts the analysis ensemble spread of sea-level pressure in the Arctic. Note that the date is arbitrarily chosen to be the same, but the year is different. Discontinuities found in ALERA are absent in ALERA2. Another advantage of ALERA2 over ALERA is the richness of diagnostic variables produced from 6-h forecast. Here, we compare ALERA2 precipitation with the Global Precipitation Climatology Project (GPCP) analysis (Adler et al. 2003) as an example. The forecast (Fig. 21.4a) agrees well with the satellite-based analysis (Fig. 21.4b) at peaks that represent disturbances both in the mid-latitudes and in the tropics. Figure 21.4c shows the ensemble spread of the precipitation. Convective precipitation contributes most of the variability, causing weak precipitation in the subtropics and tropics. Large-scale precipitation is responsible for the variability in the higher latitudes. A large ensemble spread may also be interpreted as uncertainty of the mean.

ALERA2 even represents some fine details (Fig. 21.5). On 18 July 2003 torrential rainfalls occurred in northern Kyushu. As a result of this disaster, 23 human lives were lost; 51 houses were destroyed; and thousands of houses were flooded. ALERA2 reproduces features found in the GPCP analysis: a local maximum over northern Kyushu owing to a precipitation band running from southwest to northeast, and a secondary band in the south of Shikoku and the Kii peninsula. Both GPCP and ALERA2, however, fail to reproduce the intensity observed by radar and gauge observations, likely because the precipitation bands organized at the meso scales cannot be resolved at the coarse resolution.

The forecast has excessive weak precipitation. It is known that reanalysis produces a larger global average than satellite-based estimates (Onogi et al 2007). Uncertainties remain in the satellite-based estimates despite recent improvements in rain/no rain classification methods (Kida et al. 2009). In addition, larger biases exist in forecast model models. Precipitation is unrealistically predicted in the convective parameterization over the subtropical ocean, where convective inhibition should act

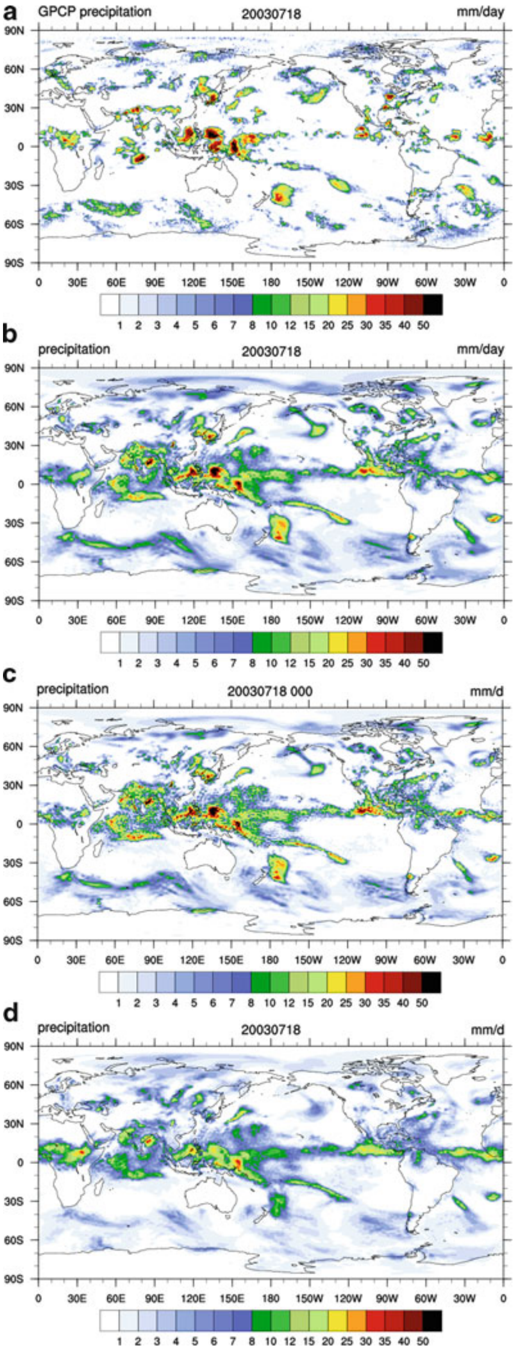
[1] Currently PREPBUFR observations are available continuously from 1 May 1997.

Fig. 21.3 The analysis ensemble spread of sea-level pressure in the Arctic (**a**) on 10 January 2006 in ALERA and (**b**) on 10 January 2008 in ALERA2



to suppress precipitation. Recent convective parameterization schemes are designed to suppress such precipitation by including sensitivity to environmental humidity (Emori et al. 2001; Betchold et al. 2008; Chikira and Sugiyama 2010). We also modified the Emanuel scheme this way, but it has not yet been adopted in ALEDAS2. There is another reason for the excessive weak precipitation: precipitation is a highly non-Gaussian process. Under a conditionally unstable environment, a small perturbation can ignite convective parameterization. If precipitation is generated with only a few members with modest intensity, the ensemble mean precipitation will be weak rather than non-existent. With a convectively unstable profile, almost all members yield precipitation, and the distribution will be closer to Gaussian. Consequently, the ensemble mean agrees well with observations in the regions with

Fig. 21.4 The daily precipitation (mm d^{-1}) on 18 July 2003 from (**a**) GPCP analysis, (**b**) the ensemble mean, (**c**) control and (**d**) ensemble spread of ALERA2 6-h forecast



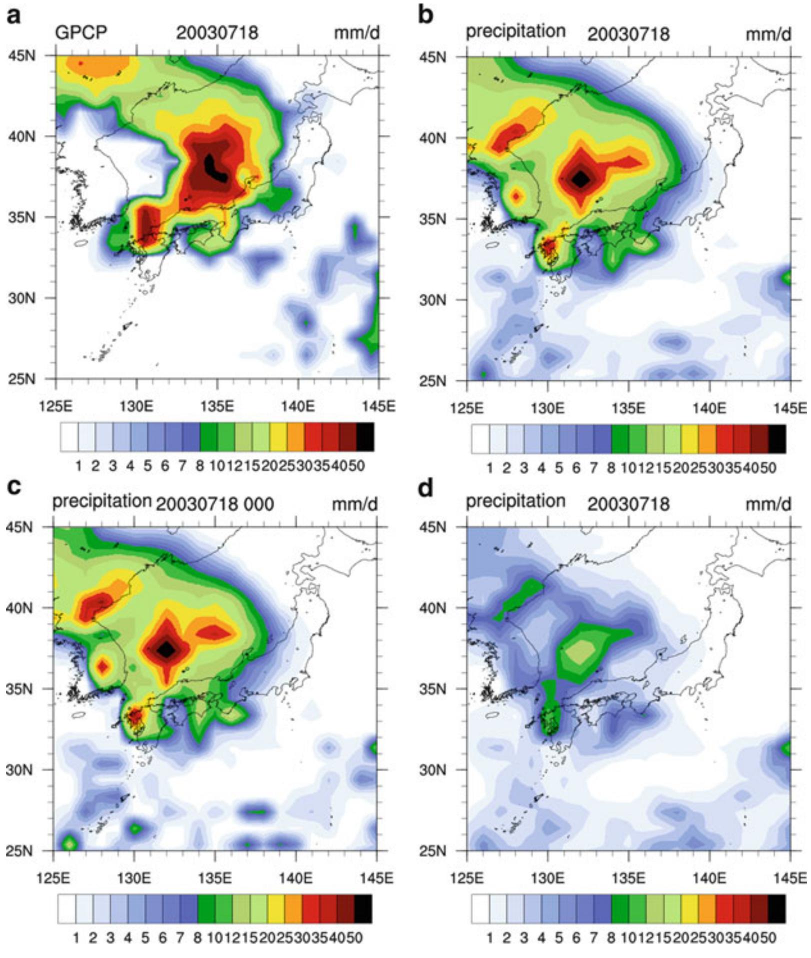


Fig. 21.5 As in Fig. 21.4 but near Japan

strong precipitation. The surrounding regions have weak precipitation because not all the members necessarily precipitate. Another likely state is given by the control run or the forecast from the analysis ensemble mean (Fig. 21.4c). The control run has less weak precipitation. The excess of weak precipitation in the ensemble mean can also be understood statistically. If the probability distribution of precipitation is lognormal, the mean is larger than the mode.

The overestimation of weak precipitation in ALERA2 is also evident in the histogram (Fig. 21.6). The almost clear ($< 1 \mathrm{mmd}^{-1}$) bin holds 68 % in GPCP but only 38 % in ALERA2. In the control run, the first bin holds 47 %, and the percentages in the bins between 2 and 16 mm d^{-1} are reduced by 1–3 %. The control run indicates that weak precipitation is partly caused by statistics. A local minimum at 1–2 mm d^{-1} bin of GPCP might imply the limited performance of detecting weak precipitation.

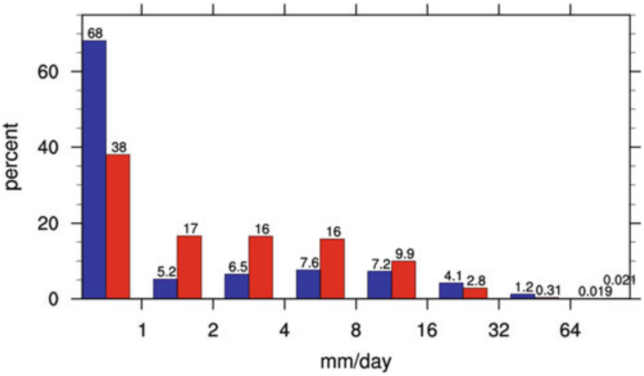


Fig. 21.6 A histogram of precipitation in July 2003. *Blue and red bars* represent GPCP and ALERA2, respectively. Numbers on the bars indicate the percentage in a bin

21.3.2 The Mirai Arctic Ocean Cruise 2010

Research vessel Mirai conducted a cruise in the Arctic Ocean from 2 September to 16 October 2010. One of the major meteorological outcomes of this cruise is an in-situ observation of an Arctic cyclone (Inoue and Hori 2011). An OSE was conducted to evaluate the impact of balloon observations from Mirai during the cyclogenesis. Figure 21.7 shows the analysis ensemble spread of temperature at 250 hPa with shading and winds in vectors. During the cyclogenesis, the reduction in error in the lower troposphere may be attributable to information over land carried by southerly winds. In contrast, the balloon observations are the major contributor to error reduction in the upper troposphere. In fact the analysis ensemble spread along the track of Mirai is smaller than the surrounding region. The result of the OSE is under detailed investigation.

21.3.3 VPREX 2010

A field campaign focusing on heavy precipitation in central Vietnam was conducted from September 2010 to January 2011. The frequency of radio sonde observations is doubled from twice to four times a day at Da Nang, Vietnam. Special balloon observations are conducted twice a day at Mactan, Philippines. These observations were transmitted to GTS and are included in PREPBUFR. Figure 21.8 shows the westward migration of disturbances produced from ALERA2. The analysis ensemble spread of the meridional wind correlates well with vorticity. A data denial experiment was conducted to further investigate the impact of balloon observations at a station at Da Nang, Vietnam and six locations in the Philippines, including

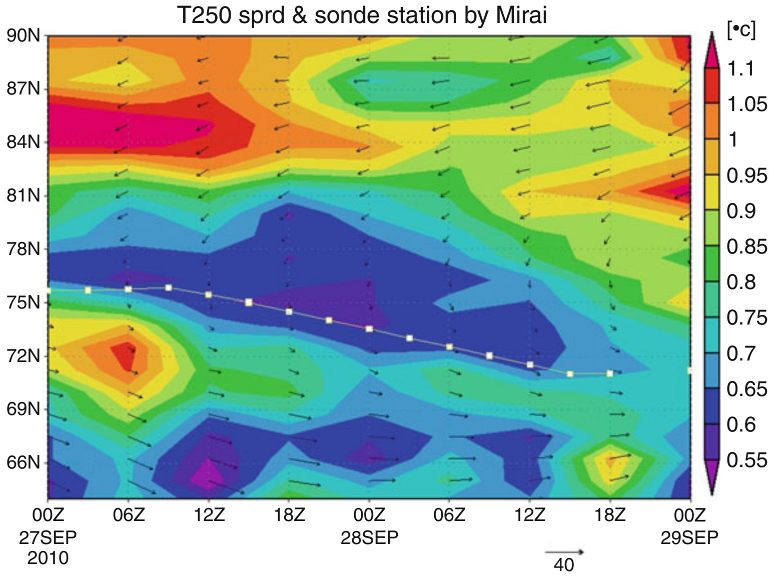


Fig. 21.7 A time–latitude cross-section of the analysis ensemble spread of temperature (*shading*) and winds (*vectors*) during 0 UTC 27 and 0 UTC 29 September 2010. The *white line* and *squares* indicate the track of Mirai and the observation locations, respectively

Mactan. The balloon observations contribute to the reduction in the analysis error at $\sigma = 0.7$ by 25 % over the Philippines. An error reduction of 5–25 % extends from the Indochina peninsula to the southern coast of Japan. Details will be reported elsewhere.

21.4 Concluding Remarks

We have constructed an ensemble data assimilation system of global atmospheric observations, called ALEDAS2, on the Earth Simulator 2. An ensemble reanalysis data set called ALERA2 is being produced with ALEDAS2. ALERA2 provides a smooth analysis ensemble mean and spread by replacing patches of covariance localization with weights based on the distances from observations. The output of ALERA2 includes various variables such as precipitation, radiative and surface fluxes and land surface variables. ALERA2 reproduces the intensity and location of intense precipitation associated with mid-latitude and tropical disturbances. ALERA2, however, produces excessive weak precipitation in comparison with the satellite-based estimation. In addition to the uncertainties in the model and satellite retrieval algorithm, the ensemble mean operation contributes to the generation of weak precipitation. Using ALERA2 as a reference, we have conducted observing-system experiments (OSEs) to evaluate the atmospheric observations collected at

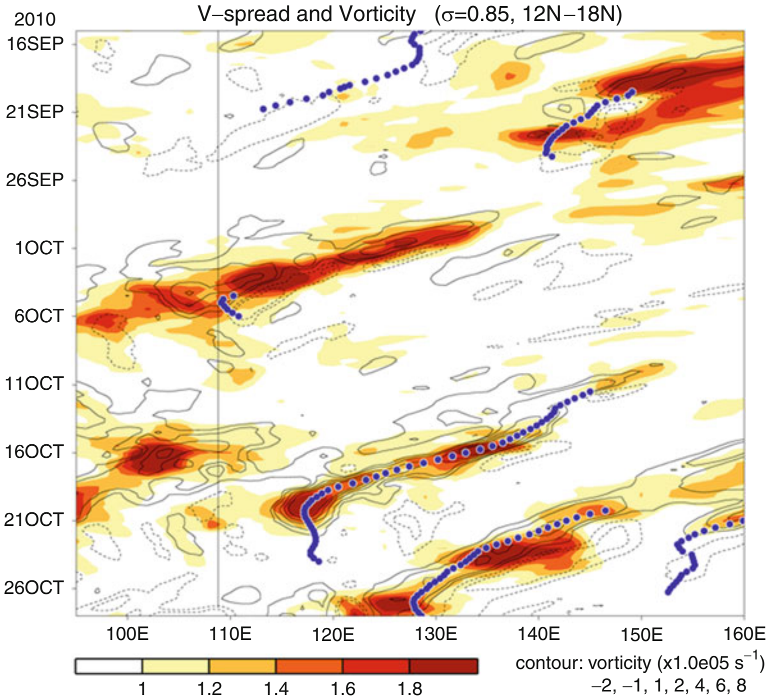


Fig. 21.8 A longitude–time section of the analysis ensemble spread of meridional winds (*shading*) and vorticity (*contours*) at $\sigma = 0.85$ averaged between 12 and 18 N. *Blue dots* indicate the locations of the center of a tropical cyclone

during field campaigns such as PALAU 2008, the Mirai Arctic Ocean Cruise 2010 and VPREX 2010. These experiments show the importance of additional observations in data-sparse regions over the ocean. Details will be reported elsewhere.

Our next step is to incorporate the variability of the ocean. With the coupled atmosphere–ocean model for the Earth Simulator (CFES) (Komori et al. 2008), we began the development of data assimilation system called CLEDAS. In this system, the ocean boundary conditions are replaced by the ocean general circulation model for the Earth Simulator (OFES) (Pacanowski and Griffies 2000; Masumoto et al. 2004) with the sea-ice process (Komori et al. 2005). Currently, only atmospheric observations can be assimilated. Preliminary tests indicate that the replacement of AFES with CFES contributes to an increased ensemble spread in the lower troposphere. In the future, CLEDAS will allow ocean and land observations to be assimilated. We plan to use CLEDAS as a test bed for the a study of the problem of assimilating data with different spatiotemporal scales and degrees of nonlinearity. CLEDAS may also be used to evaluate and design the observation systems and field campaigns conducted by JAMSTEC.

List of Acronyms

AFES: AGCM for the Earth Simulator
ALEDAS: AFES–LEKTF Ensemble Data Assimilation System
ALERA: AFES–LETKF Experimental Ensemble Reanalysis
AMIP: Atmospheric Model Intercomparison Project
CFES: Coupled Atmosphere–Ocean Model for the Earth Simulator
CTD: Conductivity, Temperature and Depth
ES: The Earth Simulator
GPCP: Global Precipitation Climatology Project
GPV: Grid Point Value
GTS: Global Telecommunication System
JAMSTEC: Japan Agency for Marine-Earth Science and Technology
JMA: Japan Meteorological Agency
LETKF: Local Ensemble Transform Kalman Filter
MATSIRO: Minimal Advanced Treatments of Surface Interaction and RunOff
MILA: Mixed Layer data set of Argo
MISMO: Mirai Indian Ocean Cruise for the Study of MJO Onset
MOAA: Monthly Objective Analysis using the Argo data
NCEP: National Centers for Environmental Prediction
NOAA: National Oceanic and Atmospheric Administration
OISST: Optimal Interpolation Sea-Surface Temperature
OREDA: Observing System Research and Ensemble Data Assimilation Development Research Team
PALAU: Pacific Area Long-Term Atmospheric Observation for Understanding of Climate Change
POPS: Polar Ocean Profiling System
T-PARC: THORPEX Pacific Asia Regional Campaign
TAO: Tropical Atmosphere Ocean
THORPEX: The Observing-System Research and Predictability Experiment
TRITON: Triangle Trans-Ocean Buoy Network
UCAR: University Corporation for Atmospheric Research

References

Adler RF, et al (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitaion analysis (1979–present). J Hydrometeor 4:1147–1167

Betchold P, Köhler M, Jung T, Doblas-Reyes F, Leutbecher M, Rodwell MJ, Vitart F, Balsamo G (2008) Advances in simulating atmospheric variability with the ECMWF model: from synoptic to decadal time-scales. Q J R Meteor Soc 134:1337–1351. doi:10.1002/qj.289

Bishop CH, Etherton J, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. Mon Wea Rev 129:420–436. doi:10.1175/1520-0493(2001)129 < 420:ASWTET > 2.0.CO;2

Bony S, Emanuel KA (2001) A parameterization of the cloudiness associated with cumulus convection; evaluation using TOGA COARE data. J Atmos Sci 58:3158–3183

Chikira M, Sugiyama M (2010) A cumulus parameterization with state-dependent entrainment rate. Part I: description and sensitivity to temperature and humidity profiles. J Atmos Sci 67:2171–2193. doi:10.1175/2010JAS3316.1

Emanuel KA (1991) A scheme for representing cumulus convection in large-scale models. J Atmos Sci 48, 2313–2329

Emanuel KA, Živković-Rothman M (1999) Development and evaluation of a convection scheme for use in climate models. J Atmos Sci 56, 1766–1782

Emori S, Nozawa T, Numaguti A, Itsushi U (2001) Importance of cumulus parameterization of precipitation simulation over east Asia in June. J Meteor Soc Jpn 79:939–947. doi:10.2151/jmsj.79.939

Enomoto T, Kuwano-Yoshida A, Komori N, Ohfuchi W (2008) Description of AFES 2: improvements for high-resolution and coupled simulations. In: Hamilton K, Ohfuchi W (eds) High resolution numerical modelling of the atmosphere and ocean, Springer, New York, pp 77–97

Enomoto T, Hattori M, Miyoshi T, Yamane S (2010) Precursory signals in analysis ensemble spread. Geophys Res Lett 37. doi:10.1029/2010GL042723

Hosoda S, Ohira T, Nakamura T (2008) A monthly mean dataset of global oceanic temperature and salinity derived from Argo float observations. JAMSTEC Rep Res Dev 8:47–59

Hosoda S, Ohira T, Sato K, Suga T (2010) Improved description of global mixed-layer depth using Argo profiling floats. J Oceanogr 66:773–787. doi:10.1007/s10872-010-0063-3

Hunt B, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local transform Kalman filter. Physica D 230:112–126

Inoue J, Hori ME (2011) Arctic cyclogenesis at the marginal ice zone: a contributory mechanism for the temperature amplification? Geophys Res Lett 38. doi:10.1029/2011GL047696

Inoue J, Enomoto T, Miyoshi T, Yamane S (2009) Impact of observations from Arctic drifting buoys on the reanalysis of surface fields. Geophys Res Lett 36. doi:10.1029/2009GL037380

Katsumata K, Yoshinari H (2010) Uncertainties in global mapping of Argo drift data at the parking level. J Oceanogr 66:553–569. doi:10.1007/s10872-010-0046-4

Kida S, Shige S, Kubota T, Aonashi K, Okamoto K (2009) Improvement of rain/no-rain classification methods for microwave radiometer observations over ocean using the 37-GHz emission signature. J Meteor Soc Jpn 87A:165–181. doi:10.2151/jmsj.87A.165

Komori N, Kuwano-Yoshida A, Enomoto T, Sasaki H, Ohfuchi W (2008) High-resolution simulation of the global coupled atmosphere-ocean system: description and preliminary outcomes of CFES (CGCM for the earth simulator). In: Hamilton K, Ohfuchi W (eds) High resolution numerical modelling of the atmosphere and ocean, Springer, New York, pp 241–260

Komori N, Takahashi K, Komine K, Motoi T, Zhang X, Sagawa G (2005) Description of sea-ice component of Coupled Ocean–Sea-Ice Model for the Earth Simulator (OIFES). J Earth Sim 4:31–45

Kuwano-Yoshida A, Enomoto T, Ohfuchi W (2010) An improved statistical cloud scheme for climate simulations. Q J R Meteor Soc 136. doi:10.1002/qj.660

Masuda S, Awaji T, Sugiura N, Ishikawa Y, Baba K, Horiuchi K, Komori N (2003) Improved estimates of the dynamical state of the North Pacific Ocean from a 4 dimensional variational data assimilation. Geophys Res Lett 30. doi:10.1029/2003GL017604

Masumoto Y, et al (2004) A fifty-year eddy-resolving simulation of the World Ocean—preliminary outcomes of OFES (OGCM for the earth simulator). J Earth Simul 1:35–56

Miyoshi T, Yamane S (2007) Local ensemble transform Kalman filtering with an AGCM at a T159/L48. Mon Wea Rev 135:3841–3861

Miyoshi T, Yamane S, Enomoto T (2007a) The AFES-LETKF experimental ensemble reanalysis: ALERA. SOLA 3:45–48. doi:10.2151/sola.2007-012

Miyoshi T, Yamane S, Enomoto T (2007b) Localizing the error covariance by physical distances within a local ensemble transform Kalman filter (LETKF). SOLA 3:89–92. doi:10.2151/sola.2007-023

Moteki Q, et al (2007) The impact of the assimilation of dropsonde observations during PALAU2005 in ALERA. SOLA 3:97–100

Moteki Q, et al (2011) The influence of observations propagated by convectively coupled equatorial waves. Q J R Meteor Soc 137:641–655. doi:10.1002/qj.779

Numaguti A, Takahashi M, Nakajima T, Sumi A (1997) Description of CCSR/NIES atmospheric general circulation model. In: Study on the climate system and mass transport by a climate model, CGER's supercomputer monograph report, vol. 3. National Institute for Environmental Sciences, Tsukuba, pp 1–48

Ohfuchi W, et al (2004) 10-km Mesh Meso-scale resolving simulations of the global atmosphere on the earth simulator—preliminary outcomes of AFES (AGCM for the earth simulator). J Earth Simul 1:8–34

Onogi K, et al (2007) The JRA-25 reanalysis. J Meteor Soc Jpn 85:369–432

Pacanowski RC, Griffies SM (2000) The MOM 3.0 manual. National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, 680pp

Peng SM, Ridout JA, Hogan TF (2004) Recent modifications of the Emanuel convective scheme in the Navy operational global atmospheric prediction system. Mon Wea Rev 132:1254–1268

Reynolds RW, Chunying L, Smith TM, Chelton DB, Schlax MG, Casey KS (2007) Daily high-resolution-blended analyses for sea surface temperature. J Climate 20:5473–5496

Sekiguchi M, Nakajima T (2008) A k-distribution-based radiation code and its computational optimization for an atmospheric general circulation model. J Quant Spectrosc Radiat Trans 109:2779–2793. doi:10.1016/j.jqsrt.2008.07.013

St-James JS, Laroch S (2005) Assimilation of wind profiler data in the Canadian meteorological centre's analysis systems. J Atmos Oceanic Technol 22:1181–1194

Sugiura N, Awaji T, Masuda S, Toyoda T, Igarashi H, Ishii M, Kimoto M (2009) Potential for decadal predictability in North Pacific region. Geophys Res Lett 36. doi:10.1029/2009GL039787

Takata K, Emori S, Watanabe T (2003) Development of the minimal advanced treatments of surface interaction and runoff. Glob Planet Change 38:209–222

Yoneyama K, Masumoto Y, Kuroda Y, Katsumata M, Mizuno K (2006) Mirai Indian Ocean cruise for the Study of the MJO-convection Onset. CLIVAR Exch 11:8–10

Chapter 22
Data Assimilation of Weather Radar and LIDAR for Convection Forecasting and Windshear Alerting in Aviation Applications

Wai Kin Wong and Pak Wai Chan

Abstract In this paper, variational data assimilation techniques to retrieve 3-dimensional wind fields from weather radars and LIDAR are discussed. The retrieved wind field from the 3-dimensional variational (3DVAR) technique applied to the weather radar data are found useful to delineate the mesoscale features leading to the convective development in a rainstorm event that brought significant lightning and thunderstorms near the Hong Kong airport and heavy precipitation over the territory. Impacts in improving analysis and forecast of a non-hydrostatic NWP model are also obtained through the data assimilation of wind retrieval data as additional observations in the model analysis. To capture the low-level windshear due to complex wind flow around the Hong Kong airport, 3DVAR and 4DVAR techniques are applied to LIDAR data. The performance of the wind retrieval algorithms and results of case studies will be illustrated. It is found that the wind fields obtained are useful to depict salient features of terrain-induced airflow disturbances at HKIA, such as mountain waves and vortices in a gustnado event.

22.1 Introduction

Ground-based remote sensing platforms including radars, LIDARs (Light Detection and Ranging), wind profilers and GPS (Global Positioning System) provide valuable observations for monitoring and forecasting of the development of mesoscale weather systems, significant convection, thunderstorm, windshear and turbulence. In this paper, a brief summary on the use of radar data in the data assimilation system of the non-hydrostatic NWP modeling system of Hong Kong Observatory will first be given. Using data from multiple weather radars, 3-dimensional variational data

W.K. Wong (✉) · P.W. Chan
Hong Kong Observatory, 134A, Nathan Road, Kowloon, Hong Kong, China
e-mail: wkwong@hko.gov.hk; pwchan@hko.gov.hk

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_22,

assimilation algorithm is applied to retrieve the horizontal and vertical wind fields that are found to provide useful source of data in showing the mesoscale circulation characteristics in a heavy rain event. Furthermore, the retrieved wind data show impacts on improving the NWP model analysis and forecast of thunderstorms and convective systems.

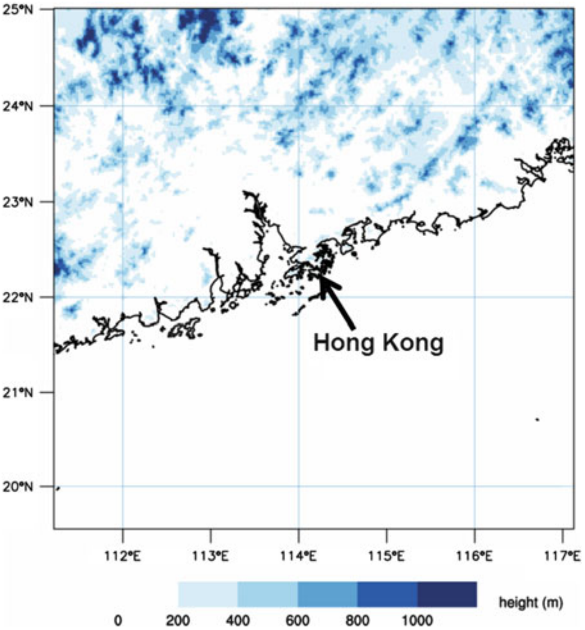
In monitoring the occurrence of low-level windshear and turbulence around the Hong Kong airport due to complex wind flow over the Lantau Island, LIDAR data have been put into operational and as a key component in the windshear and turbulence alerting system to capture the wind and their changes at high spatial and temporal resolution. To further investigate the use of LIDAR to depict a complete view of three dimensional wind flow over the complex terrain around the airport area, the variational data assimilation techniques are utilized and found to be very useful to analyze the wind flow and capture the small-scale features. The formulation of the variational methods will be discussed in this paper and their performance will be illustrated through some case studies.

22.2 An Overview of Data Assimilation of Weather Radar Data in Operational Non-Hydrostatic Mesoscale NWP Model in HKO

In June 2010, Hong Kong Observatory (HKO) started to operate a new generation of mesoscale NWP model suite called the Atmospheric Integrated Rapid-cycle (AIR) Forecast Model System (Wong 2010) based on the Non-Hydrostatic Model (NHM) of the Japan Meteorological Agency (JMA) (Saito et al. 2006). In brief, the new NWP system contains two domains called Meso-NHM and RAPIDS-NHM with horizontal resolution of 10 and 2 km respectively to provide forecast up to 72 and 15 h ahead respectively. With increased in model resolution, use of 3-dimensional variational data assimilation (3DVAR) system and better representation of physical processes like cloud microphysics and convective parameterization, benefits in forecasting of severe weather phenomena are obtained to support aviation applications of AIR/NHM (Wong et al. 2011).

In particular, to capture the fast evolving convective systems its associated mesoscale circulation features, RAPIDS-NHM is executed every hour to provide storm-scale prediction over Hong Kong and its nearby Guangdong region (Fig. 22.1). 3DVAR analysis at full model resolutions and vertical levels, and the boundary conditions from Meso-NHM forecast in one-way nesting configuration are used. Due to short observation cut-off time (∼35 min) in RAPIDS-NHM 3DVAR, the numbers of conventional observations from synoptic surface and upper-air stations, ships, buoys and aircrafts (AMDAR—Aircraft Meteorological Data Relay) are usually limited depending on availability in real-time. The observations ingested in 3DVAR of RAPIDS-NHM are mostly from mesoscale observation networks in Hong Kong (HK) and the Guangdong Province, including data from

Fig. 22.1 Domain coverage and terrain height of RAPIDS-NHM



automatic weather stations, wind profilers, total precipitable water vapour from the Global Positioning System (GPS), radar Doppler velocity and radar wind retrieval (see next section).

In NHM-3DVAR, the model optimal analysis is calculated from the best linear unbiased estimate of the control variables representing the model states that minimize the following cost function:

$$J(\mathrm{x}) = J_b + J_o = \frac{1}{2}(\mathbf{x} - \mathbf{x_B})\mathbf{B}(\mathbf{x} - \mathbf{x_B})^T + \frac{1}{2}(\mathbf{y} - \mathbf{Hx})\mathbf{R}(\mathbf{y} - \mathbf{Hx})^T \qquad (22.1)$$

where $\mathbf{x}$, $\mathbf{x_B}$ are respectively control variable vector and model background field. The control variables of NHM-3DVAR include horizontal wind components, pressure, potential temperature and pseudo relative humidity in terms of the ratio of specific humidity of water vapour to its saturation value. $\mathbf{y}$ represents a state vector containing observation data and $\mathbf{H}$ is the observation operator. In (22.1), $\mathbf{B}$ and $\mathbf{R}$ are respectively background and observation error covariance matrices where model error represented in the $\mathbf{B}$ matrix is estimated using the NMC method (Parrish and Derber 1992).

In RAPIDS-NHM, Doppler velocity data from the two S-band weather radars in Tai Mo Shan and Tates' Cairn in Hong Kong are used. Radar radial velocity data on selected CAPPI levels (at altitudes from 1 to 3 km above sea levels) are thinned to separate the wind data into about three grid-point spacing (5–6 km) in order to reduce the correlation between them. The radial velocity data are passed

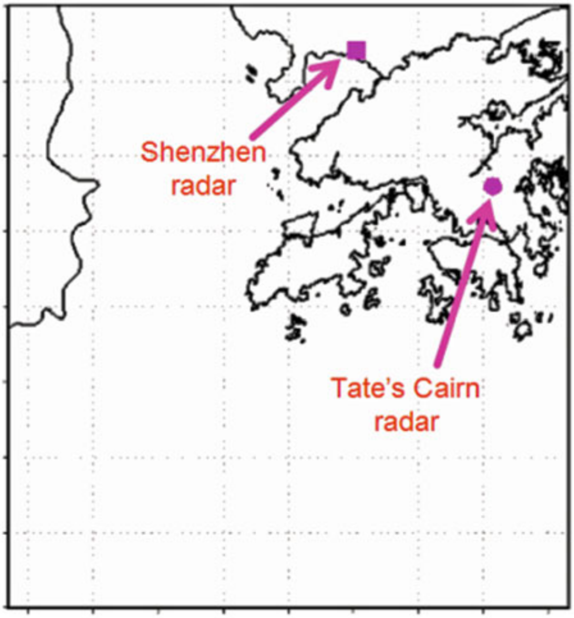


Fig. 22.2 Domain of 3D wind retrieval computation and locations of radars in Shenzhen and Hong Kong

to quality control procedure to filter out those observations with large departure from the model first-guess. Additionally, to better adjust the moisture content in the model analysis, pseudo-observations of humidity near to the saturation values at that height level are created in case the relative humidity in the first-guess is low at the grid point corresponding to the observed radial velocity and non-zero reflectivity. The radial wind data and the relative humidity observations are then assimilated together into the 3DVAR together with other available observations.

22.3 Wind Retrieval Technique of Doppler Weather Radar Data

22.3.1 Doppler Weather Radars and Wind Retrieval Algorithm

While the mesoscale features of convective systems could be delineated from the radar radial velocity field, it would be better if the 3-dimensional (3D) wind components can be estimated to facilitate real-time diagnosis, nowcasting applications as well as for ingestion into mesoscale NWP models to capture the flow characteristics. In this study, the weather radar data from Hong Kong and Shenzhen (Fig. 22.2). They are both S-band radars and complete one volume scan in every 6 min.

Prior to the wind retrieval calculation, the weather radar data are pre-processed by a couple of steps. In the first step, velocity de-aliasing is performed with the radial velocity data. A simple one-dimensional method is used, namely, for each radial, the adjacent velocity data pair is compared starting from near range to far range of the radar. If the adjacent data pair is found to have large difference in the velocity value, the velocity data at the further range would be de-aliased based on the corresponding data at the nearer range. This method appears to be simple, but it turns out to be quite effective in de-aliasing the radial velocity in most of the situations.

After de-aliasing, the velocity data are medium-filtered to remove speckles and large variations to help improve the quality of the radial velocity data. In the present study, the template for medium-filtering has a grid size of 3 by 3, and the filtering is performed for 3 times. The template dimensions and the number of filtering are determined based on several trials. The radar data are interpolated into a 3D grid with a size of 90 (in east–west direction) ×90 (north–south) and 21 (vertical levels). The corresponding resolution is 1 × 1 × 0.5 km. The 3D wind field is then obtained by variational method. Details of this method could be found in Shimizu et al. (2008) and a summary is described in the next paragraph. Using upper level winds collected from wind profilers over Hong Kong, it was found that the 3D radar retrieval winds are quite consistent with these observations and the root-mean-square errors of the wind speed and wind direction are about 2 m/s and 20° respectively.

The wind retrieval method aims at minimizing a cost function J defined as:

$$J = J_O + J_B + J_D + J_S \tag{22.2}$$

where:

1. J_O considers the difference between the measured radial velocity and the retrieved radial velocity, namely,

$$J_O = \frac{1}{2} \sum_{i,j,k} \lambda_m \left(\mathrm{P}V_{rm} - V_{rm}^{\mathrm{rob}}\right)^2 \tag{22.3}$$

   and,

$$\mathrm{P}V_{rm} = \frac{(x - x_m)\,u + (y - y_m)\,v + (z - z_m)\,(w + w_T)}{\sqrt{(x - x_m)^2 + (y - y_m)^2 + (z - z_m)^2}} \tag{22.4}$$

   is the projection of the retrieved 3D wind on the radial direction, w_T is the terminal velocity of the rain drop that depends on the intensity of radar reflectivity, V_{rm}^{rob} is the observed radial velocity of the m-th radar (after interpolation) and i, j, k are the indices of the 3D grid.

2. J_B is the difference between the retrieved 3D wind field (u, v, w) and the background 3D wind field $(u_b,\ v_b,\ w_b)$ with the following equation:

$$J_B = \frac{1}{2} \left[\sum_{i,j,k} \lambda_{ub} (u - u_b)^2 + \sum_{i,j,k} \lambda_{vb} (v - v_b)^2 + \sum_{i,j,k} \lambda_{wb} (w - w_b)^2 \right] \tag{22.5}$$

3. J_D controls the retrieved wind field to follow weak anelastic mass constraint:

$$J_D = \frac{1}{2}\sum_{i,j,k}\lambda_D\left(\frac{\partial\bar{\rho}u}{\partial x}+\frac{\partial\bar{\rho}v}{\partial y}+\frac{\partial\bar{\rho}w}{\partial z}\right)^2 \tag{22.6}$$

where $\bar{\rho}$ is the average atmospheric density; and

4. J_S controls the smoothness of the retrieved wind field:

$$J_S = \frac{1}{2}\left[\sum_{i,j,k}\lambda_{us}\left(\nabla^2 u\right)^2+\sum_{i,j,k}\lambda_{vs}\left(\nabla^2 v\right)^2+\sum_{i,j,k}\lambda_{ws}\left(\nabla^2 w\right)^2\right] \tag{22.7}$$

where ∇^2 is the Laplacian operator.

The weights of the various quantities are taken following the consideration of the order-of-magnitude of the respective terms and experiments with the actual radar data:

$$\lambda_1=\lambda_2=1,\ \lambda_{ub}=\lambda_{vb}=\lambda_{wb}=1\times10^{-4},\ \lambda_D=1\times10^5,\ \lambda_{us}=\lambda_{vs}=\lambda_{ws}=100$$

22.3.2 Illustration of Radar Wind Retrieval Data in a Significant Convection Event on 28 July 2010

The case of intense convective weather occurred in the afternoon of 28 July 2010 over Hong Kong where more than 100 mm of rainfall were recorded over part of the territory between 0600 UTC and 1000 UTC (1400–1800 HKT where HKT = UTC + 8 h). About 4,000 numbers of cloud-to-ground lightning strokes were registered over Hong Kong, nearly half of them occurred in the region of the Hong Kong International Airport (HKIA) and Lantau Island (Fig. 22.3). On the side of aviation weather service, the Airport Thunderstorm Lightning Alerting System (ATLAS) operated by the Observatory at HKIA issued red alert for two and a half hours, i.e. cloud-to-ground lightning occurred or was expected to occur within 1 km from the boundary of HKIA, so that the ground personnel had to operate their operation and should look for shelters. This is the longest period for the issuance of red alert since the operation of ATLAS in March 2008. During the period when red alert was in force, there were 290 times of cloud-to-ground lightning strokes within 1 km of the boundary of HKIA. The thunderstorms also brought about significant windshear and microburst at the airport area, leading to significant disruption of air traffic at HKIA.

Synoptically, there was a trough of low pressure over the surface starting from 27 July 2010, bringing showers and thunderstorms to south China coast. Active southwesterly jets on 850 and 700 hPa levels were observed over the southern China that favoured transport of warm and moist airstream towards the Pearl River Delta region. In the middle troposphere (500 hPa level), a trough was found over southeastern coast of China. Upward motion could be analyzed over south China

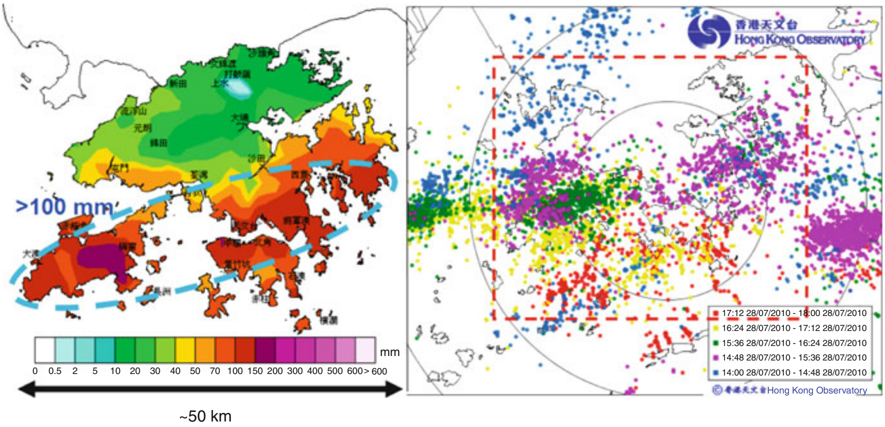


Fig. 22.3 Total accumulated rainfall (color in mm) from 0600 to 1000 UTC on 28 July 2010 (*left*). Distribution of cloud-to-ground lightning strokes during the same period (colored according to different record time) (*right*)

coast. Further aloft on 200 hPa level, the coast of Guangdong was located between a deep westerly trough to the north and an east–west oriented ridge axis to the south where divergence could be analyzed over Hong Kong and adjacent regions. The supply of moist air at low level, the upward motion in the middle level and the divergence at the upper level were all favourable to the occurrence of heavy rain over the coastal regions. Convective unstable environment was also revealed from radiosonde ascent in Hong Kong at 0000 UTC on 28 July 2010 in which K-index of 40 K and CAPE of around 3,000 J/kg were found together with a saturated condition between 900 and 500 hPa.

With the aforementioned dynamic and thermodynamic setup, two mesoscale convective systems (MCS) developed over the east of Pearl River Estuary and another over western Guangdong (satellite image omitted) during the late morning on 28 July 2010. Moreover, isolated thunderstorms developed over the seas just south of Hong Kong. Merging of thunderstorms could be readily observed in the radar imageries (Fig. 22.4). At about 05 UTC on 28 July 2010, three areas of thunderstorms could be analyzed in the vicinity of Hong Kong, namely, an east–west oriented band to the east of Hong Kong (labelled "A"), a north–south oriented band to the west of Hong Kong (labelled "B") and isolated thunderstorms over the seas to the southwest of Hong Kong (labelled "C"). The band "A" was basically quasi-stationary, whereas thunderstorms in "B" and "C" were moving towards Hong Kong in the next couple of hours. The dense network of anemometers on the surface of Hong Kong shows that there are mesoscale shear lines over the territory in this heavy rain event. Such shear lines appear to be associated with the convergence between outflow from the thunderstorms and the background southwesterly winds. Around 0730 UTC, the thunderstorms merged and there was a "X" shape in the areas of intense convection, namely, the intersection of a basically east–west oriented band of heavy rain, and another band with north to south-southwest orientation. In the

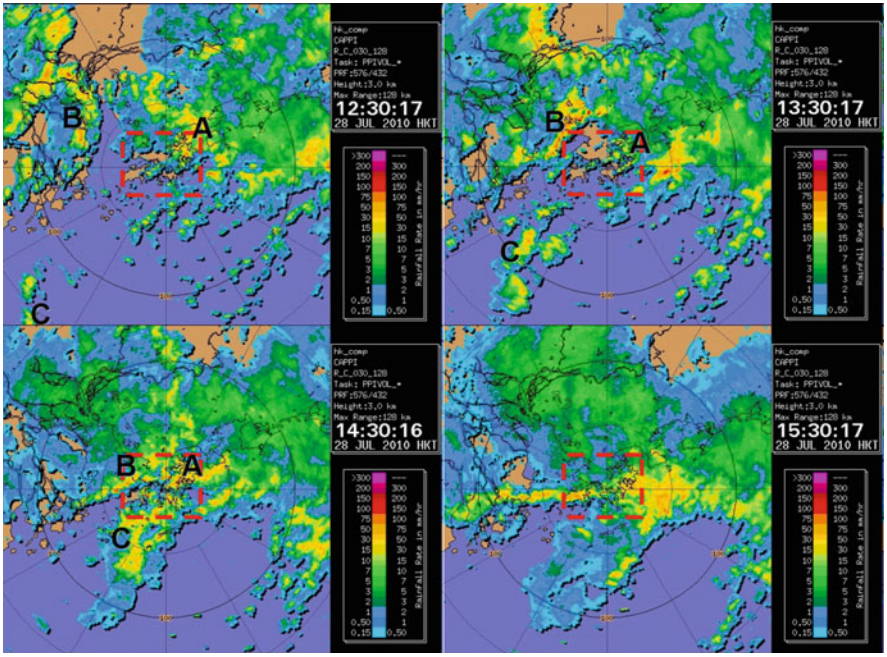


Fig. 22.4 Radar CAPPI reflectivity imagery at 0430, 0530, 0630, 0730 UTC on 28 July 2010 (1230, 1330, 1430 and 1530 HKT respectively). Area of HK is marked in *red dashed box*

following hour or so, the former band remained nearly stationary over Hong Kong, and the latter area moved eastwards gradually. The intersection of the two areas of thunderstorms resulted and the quasi-stationary nature of the east-oriented rain band brought about heavy downpour of rain over Hong Kong. Between 09 and 10 UTC, the east–west oriented rainband and the surface convergence line moved south gradually and rain over the territory weakened gradually.

Frequently occurring short waves could be analyzed in the westerly airflow in the lower to middle troposphere as shown in the wind fields retrieved from the multiple radars. The retrieved winds at a height of 2 km above mean sea level could be found in Fig. 22.5. Between 07 and 09 UTC (15 and 17 HKT) of 28 July 2010 when the rain was the heaviest over HKIA, the east–west oriented rain band was found to have a number of short waves in the westerly at 2 km level. These waves were expected to trigger and sustain the occurrence of intense convection over Hong Kong. To the south of the westerly waves, there was active southwesterly flow bringing moisture from the South China Sea towards the coast of Guangdong.

Similar wavy activity could also be observed from the radar-retrieved winds at a height of 5 km above mean sea level. Therefore, the westerly waves in the middle and lower troposphere were conductive to the intense convective developments. During 09 UTC, the winds over Pearl River Estuary changed from southwesterly to west-northwesterly. There was a deeper wave passing along the coastal areas of

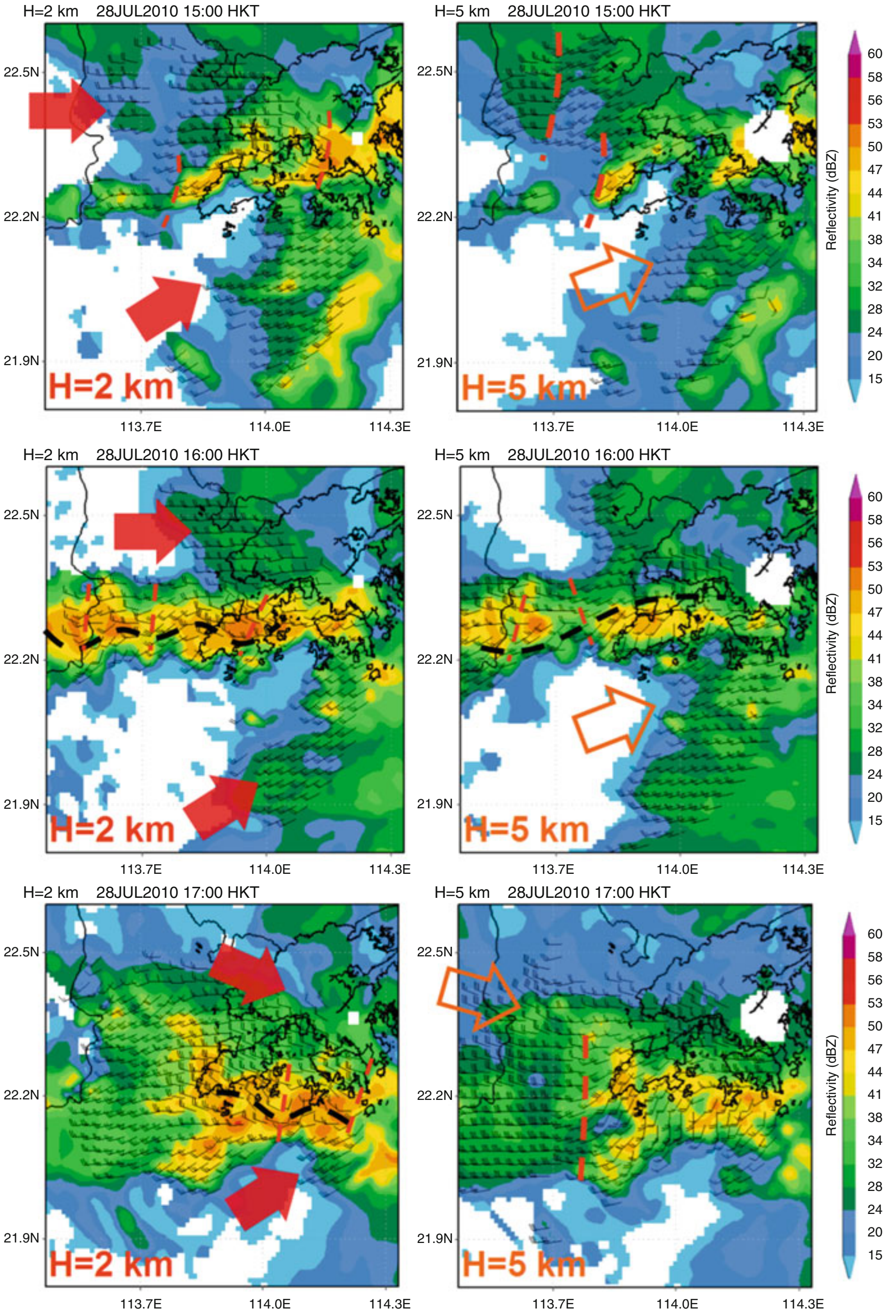

H=2 km 28JUL2010 15:00 HKT
H=5 km 28JUL2010 15:00 HKT
H=2 km 28JUL2010 16:00 HKT
H=5 km 28JUL2010 16:00 HKT
H=2 km 28JUL2010 17:00 HKT
H=5 km 28JUL2010 17:00 HKT
H=2 km
H=5 km
22.5N
22.2N
21.9N
113.7E
114.0E
114.3E
Reflectivity (dBZ)
60
58
56
53
50
47
44
41
38
34
32
28
24
20
15

Fig. 22.5 Radar retrieved winds on 2 km (*left column*) and 5 km (*right*) levels overlaid on the radar reflectivity at 0700 UTC (*top*), 0800 UTC (*middle*), and 0900 UTC (*bottom*) on 28 July 2010

Guangdong. As a result, the areas of convective developments were “pushed” to the south over the coastal waters of Guangdong, moving away from Hong Kong. As a result, the rain over the territory weakened gradually.

22.3.3 Data Assimilation and Forecast Experiments Using RAPIDS-NHM

Numerical experiments on the data assimilation and forecast using radar retrieval winds in RAPIDS-NHM are conducted in 0300 UTC run. As the vertical velocity is not a control variable in the 3DVAR system of RAPIDS-NHM, only the horizontal components of radar retrieval winds are ingested. Data thinning is applied to the retrieved radar wind data in horizontal direction using a grid separation of about 4 km. Quality control procedure is then used to remove those suspicious radar retrieval wind data that are opposite in direction to the wind vectors in model first guess of 3DVAR. To study the impact of retrieved wind data, control experiments (CNTL) are used in which all available observations except radar retrieval winds are ingested in 3DVAR.

Figure 22.6 shows the analyzed wind and relative humidity (WIND + RH) on 500 hPa and 700 hPa levels using the radar retrieval wind (upper panel) and CNTL (lower). Radar retrieval winds and reflectivity at 5 and 3 km of altitude are shown in same figure. It can be seen that the retrieval winds are effectively assimilated in the 3DVAR, resulting in enhancement in southwesterly flow over the coastal waters of Guangdong and generate a short-wave westerly disturbance on 500 hPa level. In CNTL, only moist westerly flow is found in the analysis field. Convergence in lower troposphere is also enhanced as the southwesterly winds are analyzed over the coastal waters using the radar retrieval winds data, thus resi;tomg in improvements in the forecast cyclonic shear on 850 hPa is shifted to the vicinity of HK and Pearl River Estuary where it is located more to the north in CNTL (Fig. 22.7). The main cyclonic shear is located over the northern part of HK corresponding to the intense convection development areas labelled as A and B in Fig. 22.4. In CNTL, the low-level cyclonic vorticity is forecast over the northeastern part of HK only, in accordance with actual active development area (Label A in Fig. 22.4).

Figure 22.8 depicts the 5 h forecast equivalent reflectivity (derived from RAPIDS-NHM forecasts of specific humidity of cloud hydrometeors) and the actual radar imagery. The east–west oriented reflectivity band is better forecast with the assimilation of radar wind retrieval data, whereas the intensity of simulated reflectivity is appreciably weaker in CNTL. However, the model forecast reflectivity field shows a time lag by about 1 h as compared to the actual radar image. Another difference from the actual radar image is found for the convection over the coastal sea areas, in which both experiments do not show any sign of the storm development over the region. To investigate whether this discrepancy could be alleviated through the use of hourly-update cycle of RAPIDS-NHM, the numerical experiment is repeated for analysis and forecast at 0400 UTC. The first guesses of 3DVAR in

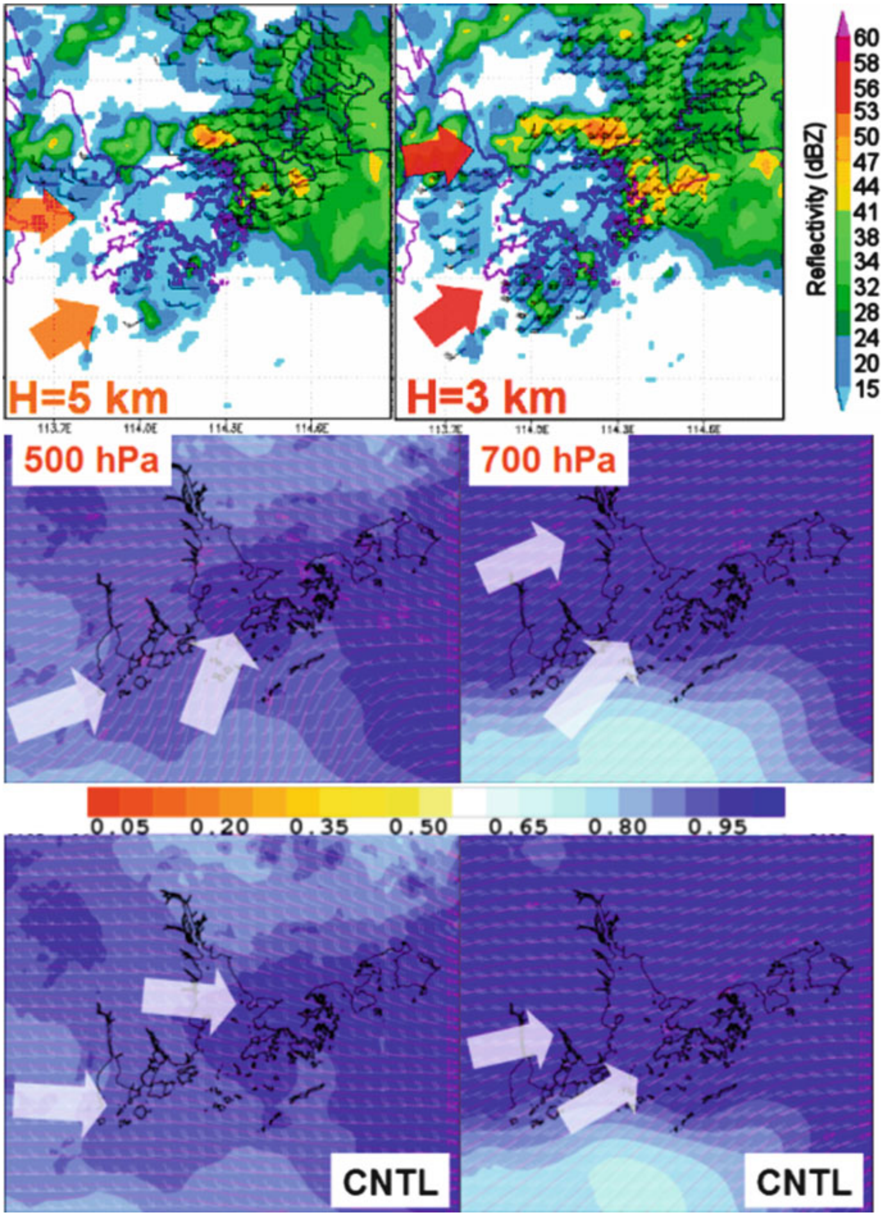


Fig. 22.6 Radar retrieved winds and reflectivity on 5 km and 3 km at 0300 UTC on 28 July 2010 (*top*). RAPIDS-NHM analysed wind and relative humidity at 0300 UTC (1100 HKT) on 500 hPa and 700 hPa using radar retrieval winds (*middle*) and the analysis in the control experiment (CNTL, *lower*)

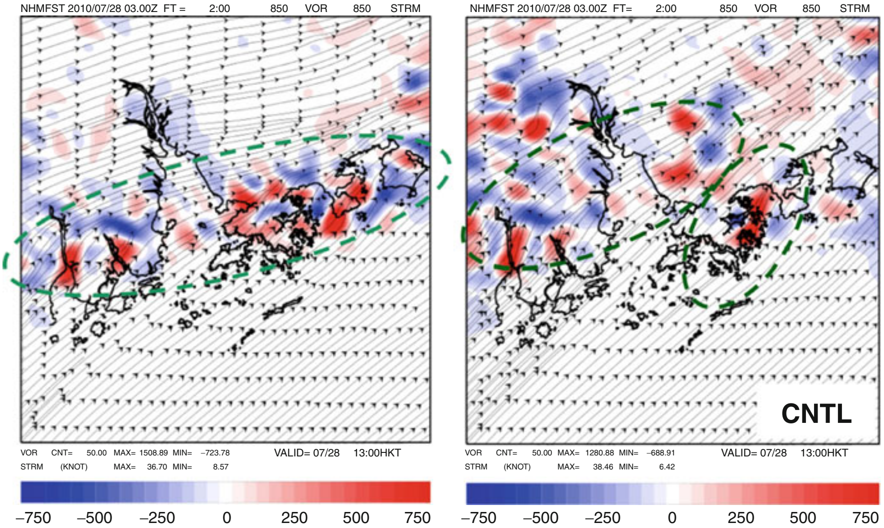

NHMFST 2010/07/28 03.00Z FT = 2:00 850 VOR 850 STRM
VOR CNT= 50.00 MAX= 1508.89 MIN= −723.78
STRM (KNOT) MAX= 36.70 MIN= 8.57
VALID= 07/28 13:00HKT
NHMFST 2010/07/28 03.00Z FT= 2:00 850 VOR 850 STRM
VOR CNT= 50.00 MAX= 1280.88 MIN= −688.91
STRM (KNOT) MAX= 38.46 MIN= 6.42
VALID= 07/28 13:00HKT
CNTL
−750 −500 −250 0 250 500 750
−750 −500 −250 0 250 500 750

Fig. 22.7 T + 2 h forecast relative vorticity on 850 hPa level (positive/cyclonic vorticity in *red* and vice versa for *blue*) using radar retrieval winds (*left*) and CNTL

both experiments are based on 1 h forecast of corresponding 0300 UTC run. It is found that radar retrieval winds and Doppler velocity data corresponding to convection over the coastal waters (Area C in Fig. 22.3) are assimilated in RAPIDS-NHM to generate disturbances in the southwesterly flow and enhanced the cyclonic vorticity (Fig. 22.9). A north–south oriented reflectivity band is thus be forecast through the use of radar retrieval data (Fig. 22.10), although the timing for the arrival and merging of two echo bands over HK remain lagged behind by about 1 h. Figure 22.11 shows 3-h accumulated rainfall forecasts ending at 0830 UTC (1630 HKT) from the two experiments. With radar retrieval winds in the initial condition, rainfall generally over Hong Kong and Lantau Island is forecast to exceed 100 mm in 3 h that is similar to the actual condition. In RAPIDS-NHM forecast, the merging of rainband and hence the maximum rainfall areas are found over southeastern part of territory, while in actual they are located mainly to the east of Hong Kong.

22.4 Application of LIDAR Data and Wind Retrieval Using Variational Techniques

22.4.1 Doppler LIDAR for Windshear Alerting in Hong Kong International Airport

For the alerting of low-level windshear and turbulence in clear air, non-rainy weather conditions, two Doppler *LI*ght *D*etection *A*nd *R*anging (LIDAR) systems

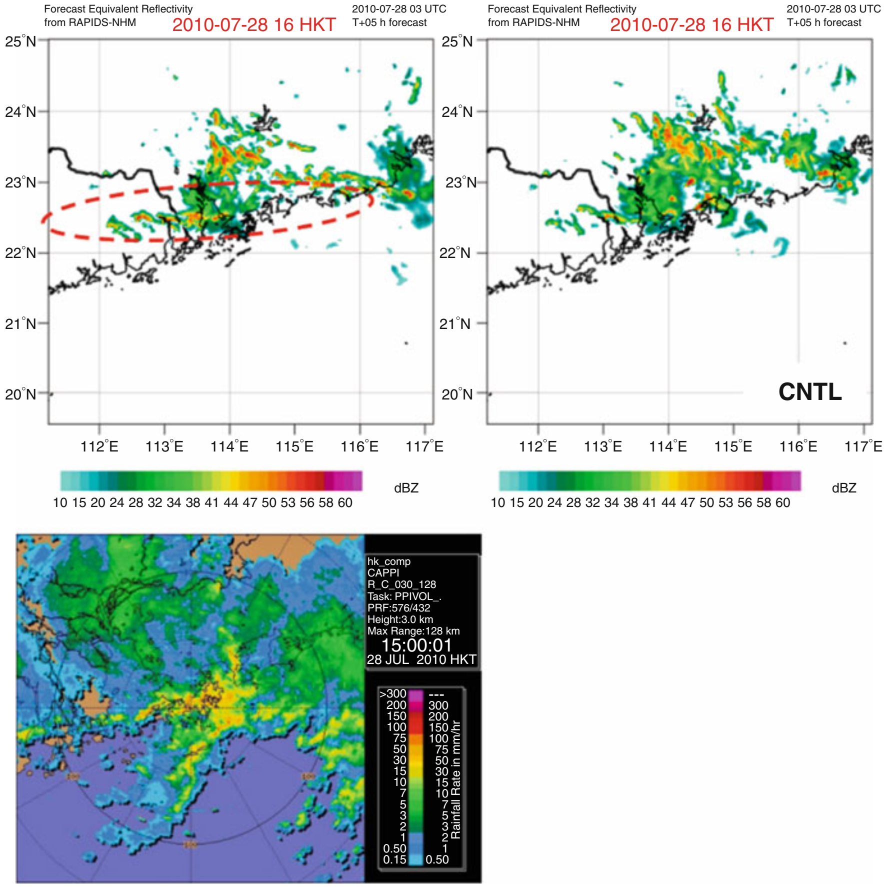


Fig. 22.8 T + 5 h forecast of equivalent reflectivity at 0800 UTC using radar retrieval winds and CNTL. Radar CAPPI reflectivity at 0700 UTC is shown in *lower panel*

are operated by the Hong Kong Observatory (HKO) at the Hong Kong International Airport. Locations of the LIDAR systems are shown in Fig. 22.12. They are coherent LIDARs using a 2-μm laser beam with pulse energy of 2 mJ. The measureable range reaches 10 km and the range resolution is about 100 m.

Since the majority of windshear at HKIA occurs in clear-air weather conditions such as terrain-disrupted airflow (70 % of pilot reports of windshear encounter) and sea breeze (20 % of the reports), LIDAR turns out to be well suited for windshear detection. The remaining windshear events (10 % of the reports) are mostly associated with convective weather like gust front and microburst. They are monitored by the Terminal Doppler Weather Radar (TDWR) operated by HKO (location in Fig. 22.12). A network of ground-based anemometers and weather buoys has also been set up inside and around HKIA for windshear detection.

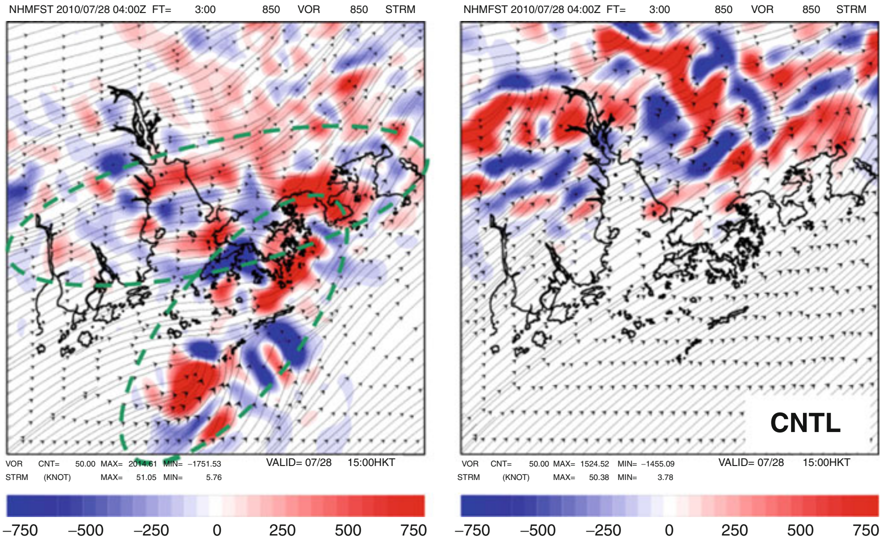

NHMFST 2010/07/28 04:00Z FT= 3:00 850 VOR 850 STRM
VOR CNT= 50.00 MAX= 2014.61 MIN= −1751.53 VALID= 07/28 15:00HKT
STRM (KNOT) MAX= 51.05 MIN= 5.76
−750 −500 −250 0 250 500 750
NHMFST 2010/07/28 04:00Z FT= 3:00 850 VOR 850 STRM
CNTL
VOR CNT= 50.00 MAX= 1524.52 MIN= −1455.09 VALID= 07/28 15:00HKT
STRM (KNOT) MAX= 50.38 MIN= 3.78
−750 −500 −250 0 250 500 750

Fig. 22.9 T + 3 h forecast relative vorticity on 850 hPa level (positive/cyclonic vorticity in *red* and vice versa for *blue*) using radar retrieval winds (*left*) and CNTL from 0300 UTC model run

For wind monitoring and windshear alerting, the LIDAR at HKIA has employed a special scan strategy, comprising the following scans:

(a) Plan-position Indicator (PPI) scans (or conical scans) to provide the weather forecasters with an overview of the wind condition in the vicinity of HKIA—There are three PPI scans, namely, with the elevation angles of 1.4°, 3° and 6° in the current implementation. The former two scans are mainly used for monitoring the wind along the arrival glide paths (which have smaller elevation angles, viz. 3° from the ground), and the last one dedicated to the departure glide paths (which have larger elevation angles, viz. ≧ 6° from the ground). The PPI scans are blocked by the Air Traffic Control Tower to the north.

(b) Range-height Indicator (RHI) scans (or vertical-slice scans) to measure the vertical structure of the windshear features, e.g. interaction between sea breeze and the background flow, hydraulic jump in cross-mountain airflow, etc.

(c) Glide-path scans to focus on the wind conditions along the glide paths for operational windshear alerting—The LIDAR measures the headwind profile to be encountered by the aircraft and significant wind changes in the profile are detected automatically (Shun and Chan 2008).

22.4.2 3DVAR Analysis of LIDAR's Radial Velocity Data

To better visualize the complex airflow around HKIA in the assessment of low-level windshear, the present seection studies the possibility of retrieving the 2D wind

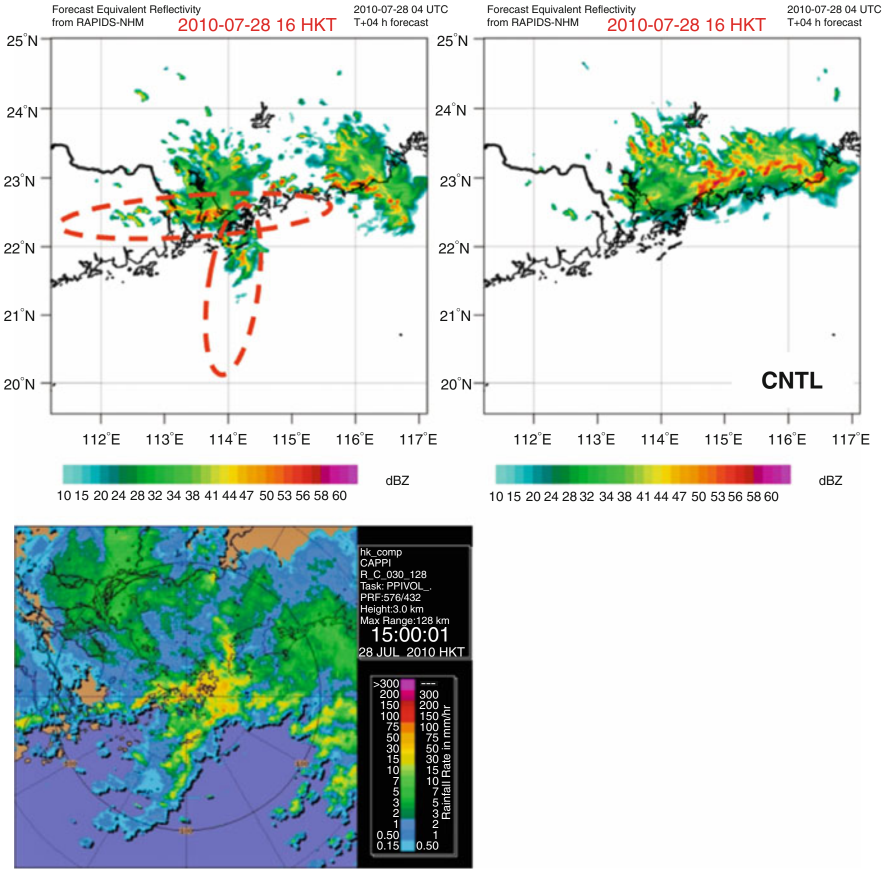


Fig. 22.10 T + 4 h forecast of equivalent reflectivity at 0800 UTC using radar retrieval winds and CNTL. Radar CAPPI reflectivity at 0700 UTC is shown in *lower panel*

field based on the radial velocity data from the PPI scans of the LIDAR. To study the possibility of real-time monitoring of the wind field based on LIDAR measurements, the computationally more efficient approach of parameter identification (PI) method is adopted here

The cost function in the variational method is defined as:

$$J(u, v) = J_1 + J_2 + J_3 + J_4 + J_5 + J_6. \tag{22.8}$$

The first term in (22.8) is the background term as given by:

$$J_1 = \sum_{i,j} W_1[(u - u_B)^2 + (v - v_B)^2]. \tag{22.9}$$

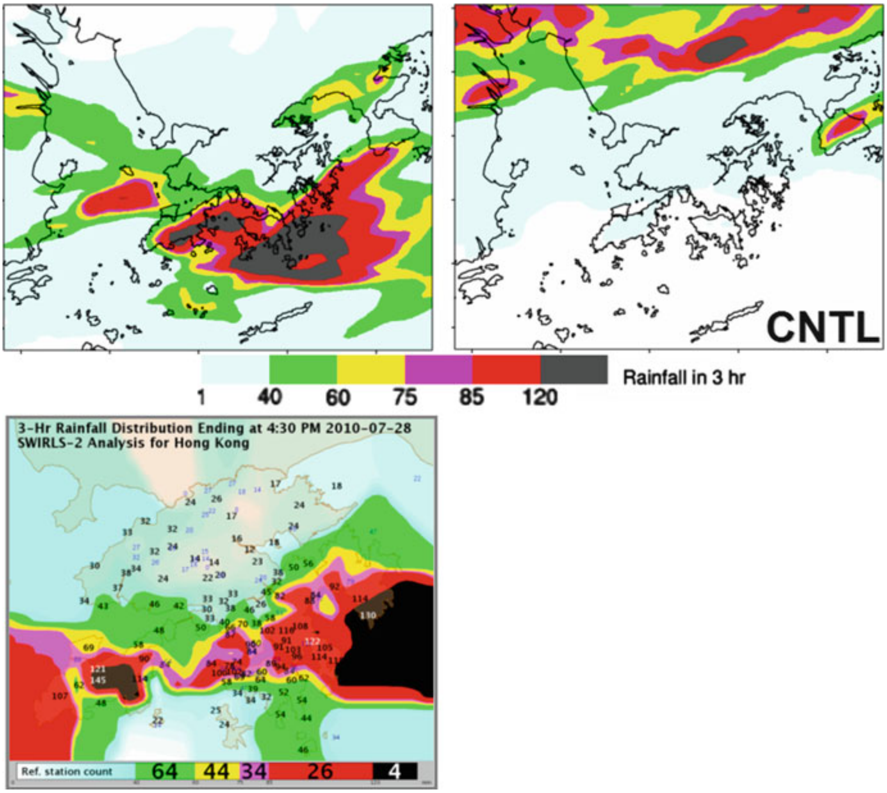

CNTL
1
40
60
75
85
120
Rainfall in 3 hr
3-Hr Rainfall Distribution Ending at 4:30 PM 2010-07-28
SWIRLS-2 Analysis for Hong Kong
Ref. station count
64
44
34
26
4

Fig. 22.11 Forecast 3-h accumulated rainfall ending at 0830 UTC using radar retrieval winds and CNTL; Actual 3-h accumulated rainfall over Hong Kong based on analysis of raingauge data is given in lower panel. The numbers in the color bar denote the amount of raingauge over Hong Kong with 3-h rainfall exceeding the respective thresholds

The summation is made over the grid points (i, j). (u_B, v_B) is the background velocity to be described later, and (u, v) is the velocity to be retrieved. W's are the weighting coefficients.

The second term in (22.8) is a measure of the difference between the observed (v_r^{obs}) and the retrieved (v_r) radial velocities as given by:

$$J_2 = \sum_{i,j} W_2 (v_r - v_r^{obs})^2. \tag{22.10}$$

The third, fourth and fifth terms in (22.8) are the smoothing terms involving the divergence, vorticity and Laplacian of the retrieved velocity field respectively. They are given by:

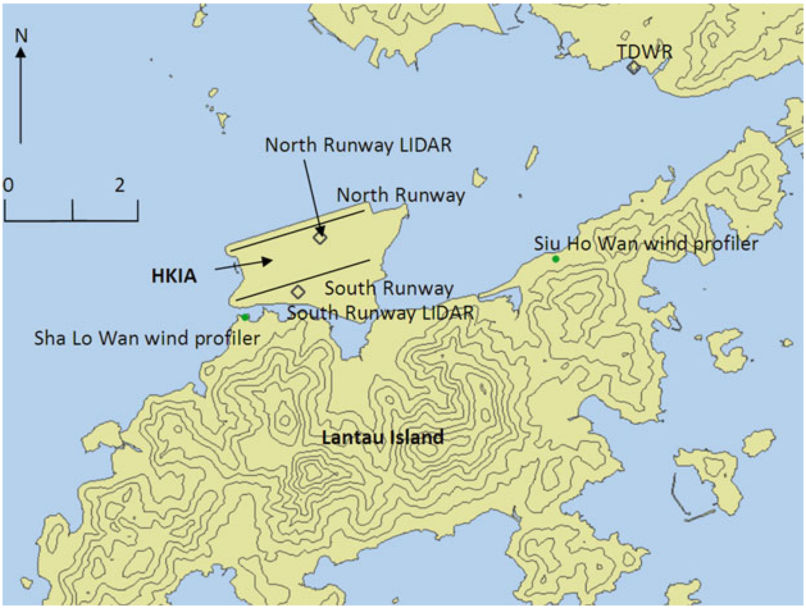


Fig. 22.12 The geographical setup in the vicinity of the Hong Kong International Airport (HKIA) with height contours in 100 m. The locations of the meteorological equipment considered in the present study are also shown

$$J_3 + J_4 + J_5 = \sum_{i,j} [W_3(\Delta x)^2 \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)^2 + W_4(\Delta x)^2 \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)^2 + W_5(\Delta x)^4(\nabla^2 u + \nabla^2 v)^2]. \quad (22.11)$$

Here $\Delta x = \Delta y = 100$ m is the grid size in the retrieval domain.

The last term in (22.8) is the conservation constraint. In Qiu et al. (2006), the conservation of precipitation content based on Radar reflectivity is used. In the present paper, conservation of the observed radial velocity is employed because the LIDAR's backscattered power data appear to have beam-to-beam variability arising from the fluctuations of the output power. The conservation equation is given by:

$$J_6 = \sum_{i,j} \sum_{n} \left[W_6 \left(\frac{\partial v_r^{obs}}{\partial t} + u \frac{\partial v_r^{obs}}{\partial x} + v \frac{\partial v_r^{obs}}{\partial y} \right)^2 \right] \quad (22.12)$$

where n is the time index. It is involved in the time derivative of the observed radial velocity. In the current scanning strategy of the LIDAR, the PPI scans are performed every 2 min or so. Three consecutive scans are used in the 2D wind retrieval in this paper. Equation (22.12) is the approximate form of the conservation of momentum for the radial velocity component. It is to ensure that the retrieved velocity field (u, v) could observe the conservation of momentum approximately.

The weighting coefficients are taken as: $W_1 = 0.1(1/\mathrm{m}^2\,\mathrm{s}^{-2})$, $W_2 = 1(1/\mathrm{m}^2\mathrm{s}^{-2})$, $W_3 = W_4 = W_5 = 0.1(1/\mathrm{m}^2\mathrm{s}^{-2})$ and $W_6 = 10^4$ $(1/\mathrm{m}^2\mathrm{s}^{-4})$. They are chosen empirically in this paper to ensure that the constraints have proper orders of magnitude.

Following Qiu et al. (2006), the background velocity field is determined by expanding it in terms of second-order Legendre polynomials:

$$u_B(x, y) = \sum_{nx=0}^{2} \sum_{ny=0}^{2} a_{nx,ny} P_{nx}(x) Q_{ny}(y),$$

$$v_B(x, y) = \sum_{nx=0}^{2} \sum_{ny=0}^{2} b_{nx,ny} P_{nx}(x) Q_{ny}(y). \tag{22.13}$$

$P_{nx}(x)$ and $Q_{ny}(y)$ are the orthonormal functions (Legendre polynomials). The background field is then fully determined by the expansion coefficients $a_{nx,ny}$ and $b_{nx,ny}$, which are the retrieved variables in this step. The cost function for retrieval is similar to (22.8), except that the first term vanishes (i.e. setting $W_1 = 0$).

Before performing the retrieval, the radial velocity data are quality-controlled to remove the outliers due to, for instance, reflection from clutters (Shun and Chan 2008). The main source of clutter is the moving aircraft in the sky and the clutter does not occur very frequently (in the order of a few per day). Such outliers could be detected by mimicking visual inspection to compare each piece of radial velocity with the data points around, and replaced by a median-filtered value if the difference between them is larger than a pre-defined threshold. The threshold is determined from the frequency distribution of velocity difference between adjacent range/azimuthal gates of the LIDAR over a long period of time. The quality-controlled radial velocity in the range-azimuth coordinate system is then interpolated to a Cartesian grid with resolution of 100 m using Barnes scheme. According to Chan and Shao (2007), the root-mean-square errors of the retrieved wind components (u and v) were about 2 m/s when compared with the anemometer measurements (Table 2 and Fig. 2 in Chan and Shao (2007)).

One application of the 3DVAR retrieved 2D wind field is the identification of coherent structure in the airflow at HKIA, which may be related to the low-level windshear and turbulence to be encountered by the aircraft. The monitoring of airflow near HKIA using the LIDAR's radial velocity data is a kind of Eulerian descriptions of the flow field. It has recently been established that, such descriptions, inefficient and somewhat arbitrary at best, could lead to serious flaws as instantaneous streamline sketches is not an objective representation of actual particle motion in an unsteady flow. Lagrangian analyses, however, provide frame-independent description when the flow field is not evolving too quickly, and certain trajectories of an unsteady flow persist with coherent motion over some period of time. The method analyzes the relative motion of fluid particles in the Lagrangian frame. In this framework, the Lagrangian coherent structures (LCSs) are identified as distinguished sets of fluid particle trajectories that attract or repel nearby trajectories

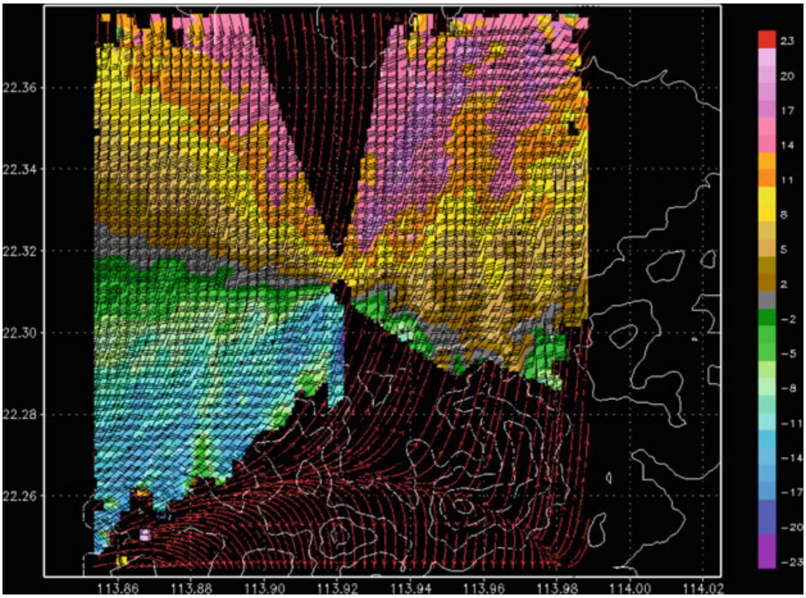


Fig. 22.13 Wind vectors and streamlines based on the retrieved velocity at 14:29 UTC overlaid on the LOS velocity (*color shades*) measured by the LIDAR. Positive values indicate LOS velocity away from LIDAR. *White contours* show the mountainous topography near HKIA at 100 m intervals

at locally the highest rate in the flow. Practically, they are identified using finite-domain finite-time Lyapunov exponents (FDFTLE) method. Technical details of the method can be found in Tang et al. (2011).

Figure 22.13 shows the 3DVAR retrieved wind overlaid on the line of sight (LOS) velocity at 14:29 UTC 19 April 2008. In the airport region, sometimes a long and distinct ridge of updraft is persistent as an organizing structure. We show the evolution of this updraft between 14:36 UTC and 14:41 UTC, 19 April 2008, in Fig. 22.13, at 150 second intervals. Discussion and illustrations on the results of 3DVAR retrieved wind can be referred to Tang et al. (2011). This ridge of updraft originates downwind of Lin Fa Shan, a mountain on Lantau Island to the south of HKIA. The ridge could correspond to the merging of gap flows on the two flanks of the mountain peak, leading to convergence and updraft when they meet. Unlike other coherent structures which either stay in the vicinity of mountain topography or move with the background flow and quickly dissipate, this ridge is larger in scale and stay longer in time. More importantly, this ridge is transversal to the runway corridor, where many flights passed through. In Fig. 22.14a–c, the LOS velocity is shown and its magnitude is represented by the color shades. Not much of the velocity structures can be directly inferred from these plots, though locations of strong wind convergence could be associated with gradient of LOS velocity. The FDFTLE plots in Fig. 22.14e, f, g, however, reveal the the locations of LCS corresponding to the updraft structure. In addition, we plot the Hovmoller diagram of the LOS velocity

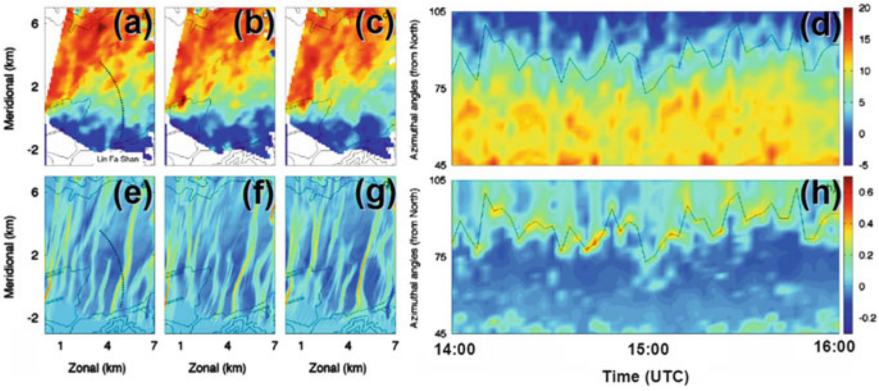


Fig. 22.14 Ridge of updraft identified to the east of the airport, during the episode of spring tropical cyclone. (**a**), (**b**), and (**c**) are the LOS velocity output from LIDAR. It is not apparent that a *ridge structure* is present. (**e**), (**f**), and (**g**) are the backward-time FDFTLE. A long ridge of FDFTLE maxima is seen persistent over time, trailing Lin Fa Shan. The different times, from left to right for each pair of plots, are 14:36 UTC, 14:39 UTC and 14:41 UTC. (**d**) Hovmoller diagram of the LOS velocity at 5 km range between 14:00–16:00 UTC. The coverage is shown as the *arc of black dots* in (**a**). (**h**) Hovmoller diagram of the backward-time FDFTLE between 14:00–16:00 UTC. The FDFTLE maxima (on the persistent ridge) is connected by the *black curve*. This curve is also plotted in (**d**). It is seen that the ridge correspond to a rather strong change in LOS velocity

(Fig. 22.14d) and the backward-time FDFTLE (Fig. 22.14h) at 5 km range between 45° and 105° azimuth and 14:00–16:00 UTC to study the relation between LCS and LOS velocity for this specific updraft. Since the updraft structure is transversal to the arc 5 km from LIDAR, we locate its time evolution in terms of the change in azimuthal angles where the ridge appears. We plot the evolution of the azimuthal angle in black in both Hovmoller diagrams. It is seen that this curve corresponds to a sharp transition of LOS velocity at 5 km range from the LIDAR. Above the curve, the flow is to the right of the ridge, and move faster towards the LIDAR. Below the curve, the flow is to the left of the ridge and move slower. As such, the converging flow gives rise to the persistent ridge in our analyses.

22.4.3 Retrieval of 3-Dimensional Winds from LIDAR Using 4DVAR

The use of 4DVAR analysis in retrieving the three wind components and thermodynamic fields from LIDAR radial velocity data has been investigated by researchers in recent years. Fundamentals of the 4DVAR include a forward large-eddy-simulation (LES) and a backward adjoint integration. The adjoint formulation is particularly complicated due to the required estimation of the gradients of the cost function with respect to all control variables. Two major approaches in constructing 4DVAR have been developed by Chai et al. (2004) and Newsom and Banta (2004). The main difference between the two approaches lies in the number of control variables employed. In Newsom and Banta (2004), the subgrid-scale fluxes of momentum and heat are modeled through theoretical assumptions for turbulent eddy viscosity and thermal diffusivity estimations, rather than treating directly the viscosity and diffusivity as control variables. The advantage of using theoretical subgrid-scale model is that the reduced number of control variable may improve the efficiency of 4DVAR calculation. However, the use of theoretical sub-grid scale model may not be sufficient for resolving the turbulent eddy structures. This is due to the drawback of the inability of using the subgrid-scale model to represent the turbulent field correctly with a single universal constant, especially in strong shear, rotating flow, near topography or transitional regimes (Germano et al. 1991). In order to create a computationally efficient analysis for our purposes, we have followed similar approach as developed by Newsom and Banta (2004). For ensuring the correctness of the retrieved eddy structures, the subgrid-scale model coefficients need to be properly preset before performing LES.

The fundamental idea of the 4DVAR is to fit the prognostic/forward model to the observations. This would rely on the estimation of the cost function to tell whether the “fitting” is good enough. In our case, the cost function is given as follows.

$$J = Jr + Jd + Js \tag{22.14}$$

The first term in (22.14), *Jr*, is the difference between forward model predicted radial velocity and the LIDAR observations within the specified time window ($\sim$ 3 min in our cases). *Jd* is the divergence penalty term used for suppressing the divergence in the initial field. *Js* is the smoothing penalty term and it helps to smooth the output a little for easily identifying any possible eddy structures in the retrieved wind field.

Jr and *Jd* have the forms as taken from Newsom and Banta (2004) whereas the *Js* is given by

$$Js = \frac{1}{2}\sum_{i,j,k}[w_u(\nabla^2 u)+w_v(\nabla^2 v) + w_w(\nabla^2 w)] \tag{22.15}$$

The weighting factors w_u and w_v are normally set to 0.001 and w_w is set to 0.5. These are guess values for the time being. Further tests are required for determining these weightings empirically. The governing equations and adjoint derivations are summarized in the following subsections (the formulation is similar to Newsom and Banta (2004)).

22.4.3.1 Governing Equations

The governing equations are the Boussineq equations for a shallow atmospheric boundary layer:

$$\frac{\partial u_i}{\partial t} + \frac{\partial (u_i u_j)}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \delta_{i3} g \frac{(\theta - \langle\theta\rangle)}{\Theta_{ref}} - \varepsilon_{ijk} f_j u_k - \frac{\partial \tau_{ij}}{\partial x_j}, \tag{22.16}$$

$$\frac{\partial \theta_1}{\partial t} + \frac{\partial(\theta_1 u_j)}{\partial x_j} + w\frac{\partial \theta_0}{\partial z} = -\frac{\partial \gamma_\theta}{\partial x_j}, \tag{22.17}$$

$$\frac{\partial u_j}{\partial x_j} = 0, \tag{22.18}$$

Here x_i's are the Cartesian components of the position vector $\mathbf{x} = [x, y, z]$, u_i's the Cartesian components of the velocity vector $\mathbf{u} = [u, v, w]$, δ_{ij} the Kronecker delta, ε_{ijk} the permutation tensor, f_j the Coriolis parameter, g the acceleration due to gravity and angled brackets represent averaging on horizontal planes. The pressure p is the non-hydrostatic component of the pressure normalized by the reference density ρ_{ref}. Virtual potential temperature θ is decomposed as

$$\theta(x, y, z, t) = \theta_0(z) + \theta_1(x, y, z, t), \tag{22.19}$$

where subscript 0 refers to the initial base state profile and subscript 1 the dynamic perturbations about the base state. Θ_{ref} is a reference virtual potential temperature and is set to be equal to the virtual temperature at the reference level, namely, the ground. τ_{ij} and γ_θ are the turbulent fluxes of momentum and temperature respectively.

The anisotropic component of the turbulent momentum flux is modeled as

$$\tau_{ij} - \frac{1}{3}\delta_{ij}\tau_{kk} = -2K_m D_{ij}, \tag{22.20}$$

where D_{ij} is the strain rate tensor,

$$D_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) - \frac{1}{3}\delta_{ij}\frac{\partial u_k}{\partial x_k}. \tag{22.21}$$

The eddy viscosity K_m can be calculated using a number of different models. There are models based on Troen and Mahrt (1986) and Smagorinsky (1963). The former model is used in this paper. The isotropic component of the turbulent momentum flux $\frac{1}{3}\delta_{ij}\tau_{kk}$ is absorbed into the pressure term.

Similarly, the turbulent flux of virtual potential temperature is modeled as

$$\gamma_\theta = -K_h\frac{\partial \theta}{\partial x_j}. \tag{22.22}$$

Here the eddy diffusivity K_h is given by

$$K_h = \frac{K_m}{P_{r_t}} \tag{22.23}$$

where the turbulent Prantle number P_{r_t} is typically set to 0.4.

The requirement that the velocity field remains divergence free, as implied by (22.17), is enforced either using a pressure correction method or a Poisson pressure equation. In the pressure correction method, the momentum equations are integrated first giving an estimate of the new velocity field u_i^*. This velocity field will in general not be divergence free. The divergence becomes the source term in the pressure correction equation, which is written

$$\frac{\partial^2 p'}{\partial x_i^2} = \frac{\partial}{\partial x_i}\left(\frac{u_i^*}{\Delta t}\right). \tag{22.24}$$

This is an elliptic equation, which is solved using the BiCGstab matrix equation solver (Nocedal 1980; Liu and Nocedal 1989). The resulting pressure correction fields are then used to correct the pressure and velocity fields.

The pressure Poisson equation is written as

$$\frac{\partial^2 p}{\partial x_i^2} = \frac{\partial}{\partial x_i}\left(\frac{u_i}{\Delta t} - \frac{\partial(u_i u_j)}{\partial x_j} + \delta_{i3} g \frac{(\theta - \langle\theta\rangle)}{\Theta_{ref}} - \varepsilon_{ijk} f_j u_k - \frac{\partial \tau_{ij}}{\partial x_j}\right) \tag{22.25}$$

This equation is solved before the momentum equations, to give a pressure field, which, when used to calculate the pressure gradient in the momentum equations ensures that the velocity field at the end of the next time step remains divergence free. The pressure Poisson equation is used in the 4DVAR wind retrieval for the following selected cases in this paper.

22.4.3.2 Adjoint Model Equations

The 4DVAR procedure uses an adjoint method to minimize the cost function J. The adjoint equations are derived by requiring that the first variation of the Lagrangian L with respect to all variables vanishes for $t > 0$. For conciseness, we present the adjoint equations for the first-order Adam-Bashforth time integration scheme (Shampine and Gordon 1975) as an example. For the first-order in time scheme, the Lagrangian is defined as

$$\begin{aligned} L = J + \sum_{n=0}^{N-1}\sum_{r} [\tilde{u}_i^{n+1}(u_i^{n+1} - u_i^n - \Delta t F_i^n) + \tilde{\theta}^{n+1}(\theta^{n+1} - \theta^n - \Delta t G^n) \\ + \Delta t\, \tilde{p}^{n+1} P^n] \end{aligned} \tag{22.26}$$

Here $\tilde{u}$, $\tilde{\theta}$ and $\tilde{p}$ are the adjoint variables corresponding to u, θ and p respectively, and Δt is the time step. The functions F^n, G^n and P^n are essentially the right hand sides of the forward model equations with all variables evaluated at the n^{th} time step. For details of the adjoint formulations, reference may be made to the appendix in Newsom and Banta (2004). The current 4DVAR uses the adjoint equations for the second-order Adam-Bashforth scheme that are derived analogously as for the first order scheme.

As an example of the application of the 4DVAR method, the analysis of a gustnado event is presented here. On 6 September 2004, a trough of low pressure associated with Typhoon Songda affected the coast of southern China. Troughing flow could be analyzed on the surface up to 850 hPa level (not shown). In the upper troposphere (e.g. 200 hPa level), divergence could be analyzed in the region in association with a broad anticyclone (not shown). The atmosphere was humid and unstable as revealed in the radiosonde data on that day. The K index was 35 at 0000 UTC (0800 HKT) and rose to 41 at 1200 UTC (2000 HKT). Under the unstable atmosphere, intense convective development was triggered over inland areas of southern China due to solar heating during the day and the thunderstorms so produced moved south towards the coast in the evening.

Starting from about 08 UTC (16 HKT) on that day, the radarscope of Hong Kong showed that there were isolated thunderstorms along the south China coast. Among them, one storm appeared to the west of HKIA at about 0930 UTC (1730 HKT) and brought westerly winds to the western part of the airport. The westerly was basically rain-free and appeared to be the gust front associated with the thunderstorm. It spread eastwards and converged with the background east to southeasterly winds over the eastern part of HKIA. The tornado developed over the convergence zone of the two airstreams.

The LIDAR's PPI scan image at the climax of the tornado (about 09:54 UTC, 6 September 2004) is shown in Fig. 22.15. The tornado under study is encircled in red. It could be found in the southern side of HKIA at which the cargo apron is located. 4DVAR analysis is carried out with a horizontal resolution comparable with the radial resolution of the LIDAR, viz. 100 m.

Figures 22.16 and 22.17 show the 4DVAR-analyzed horizontal wind vectors (as arrows) and perturbation horizontal velocities (u -<u> and v -<v>, in colour contours) at a height of 200 m above sea level. The anticyclonic flow associated with the tornado is captured successfully by the 4DVAR-retrieved wind field. The perturbation horizontal velocities also show a couple of features:

(a) From the plot of u -< u >, the perturbed easterly wind at the southern part of the tornado is generally stronger than the perturbed westerly wind at the northern part of the system. The former reaches about 8 m/s. Similarly, from the plot of v -< v >, the perturbed northerly wind at the eastern part of the tornado is generally stronger than the perturbed southerly wind at the western part of the system. The former reaches about 10 m/s;
(b) A series of small-scale cyclones and anticyclones (each with a size of several hundred metres) is analyzed at the convergence zone between the westerly flow associated with the gust front and the background east to southeasterly flow at the eastern part of HKIA; and
(c) Even with the mean wind subtracted, the perturbed v-component still shows the southerly jet emerging from a gap of Lantau Island at the lower right corner of the analysis domain (Fig. 22.17).

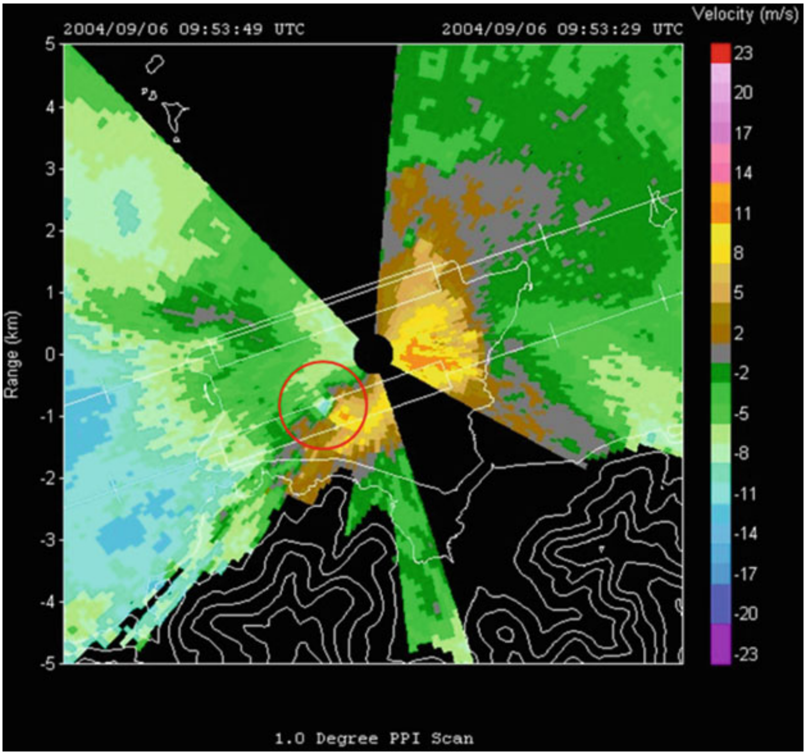


Fig. 22.15 The LIDAR's PPI scan image at the climax of the tornado at about 09:54 UTC, 6 September 2004

Apart from the horizontal winds, the 4DVAR analysis gives the vertical velocity and the pressure, which are not measured directly by the LIDAR. The perturbation vertical velocity is shown in Fig. 22.18. In general, downward motion up to about –3 m/s is analyzed for the westerly wind over the airport island. At the same time, at the eastern and southern edges of the westerly flow, there are areas of upward motion with a vertical velocity of +3 m/s. As such, there is vertical circulation along the periphery of the gust front. The circulation may be tilted to give rise to a tornado. This is one possibility for the occurrence of the gustnado.

Moreover, as revealed from 4DVAR analysis (Fig. 22.18), the tornado itself has rather complicated pattern of vertical velocity. There is upward motion at its core. At the eastern and western sides of the core, the motion is generally downward. An area of significant upward motion (with vertical velocity of +3 m/s) is also identified to the northeast of the core. The origin of this upward motion is not certain, and it may be a perturbation in the vertical circulation associated with the southern periphery of the gust front.

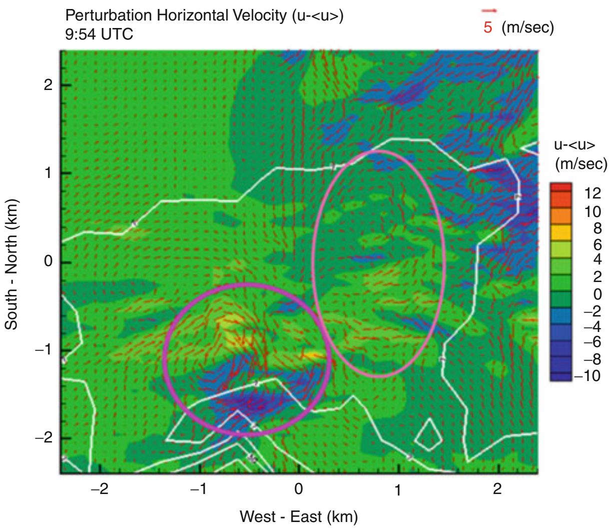


Fig. 22.16 4DVAR-analyzed horizontal wind vectors (as *arrows*) and perturbation horizontal velocities (u -< u >, in *colour contours*) at a height of 200 m above sea level. Locations of small-scale cyclones and anticyclones are enclosed in *pink ellipses*

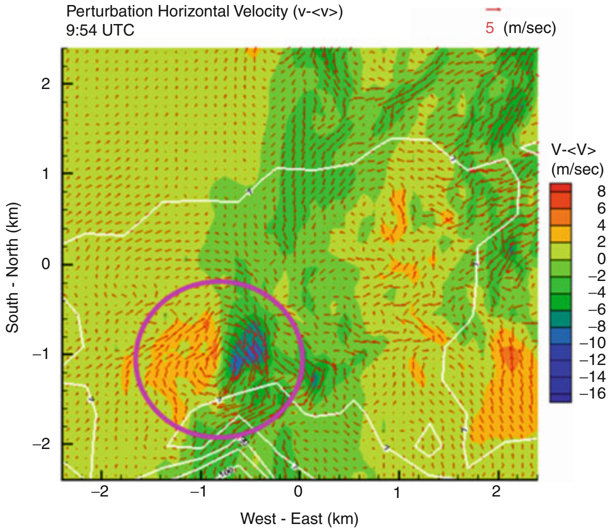


Fig. 22.17 4DVAR-analyzed horizontal wind vectors (as *arrows*) and perturbation horizontal velocities (v -< v >, in *colour contours*) at a height of 200 m above sea level

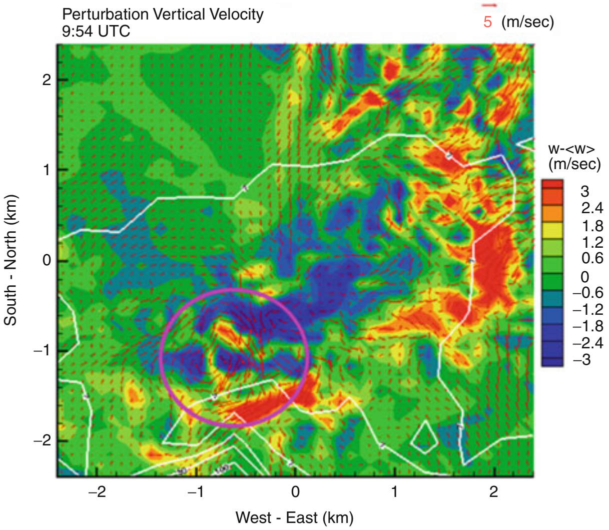


Fig. 22.18 4DVAR-analyzed perturbation vertical velocity with the horizontal wind vectors overlaid

22.5 Concluding Remarks

In this chapter, variational data assimilation algorithms to retrieve the 3-dimensional wind field from weather radars and LIDAR have been discussed. The retrieval winds from weather radars show benefits in the analysis of dynamical condition conducive to intense development of convective storms. The retrieved wind data are also found to provide positive impacts in resolving the wind flow over low to mid-tropospheric levels in the non-hydrostatic NWP model (RAPIDS-NHM). Improvements are made in model simulation of the mesoscale features and dynamics of a severe convective storm. Furthermore, the novel variational minimization algorithms (3DVAR and 4DVAR) applied to the LIDAR data are found to be very useful to retrieve the 3-dimensional wind flow over the complex terrain near HKIA, for example in study of low-level windshear effects and even for a gustnado event.

References

Chai T, Lin CL, Newsom RK (2004) Retrieval of microscale flow structures from high-resolution Doppler data using an adjoint model. J Atmos Sci 61:1500–1520

Chan PW, Shao AM (2007) Depiction of complex airflow near Hong Kong International Airport using a Doppler LIDAR with a two-dimensional wind retrieval technique. Meteorol Z 16(5):491–504

Germano M, Piomelli U, Moin P, Cabot WH (1991) A dynamic subgrid-scale eddy viscosity model. Phys Fluids A 3(7):1760–1765

Liu D, Nocedal J (1989) On the limited memory method for large scale optimization. Math Program B 45(3):503–528

Newsom RK, Banta RM (2004) Assimilating coherent Doppler lidar measurements into a model of the atmospheric boundary layer. Part I: algorithm development and sensitivity to measurement error. J Atmos Ocean Technol 21:1328–1345

Nocedal J (1980) Updating quasi-Newton matrices with limited storage. Math Comput 35:773–782

Parrish DF, Derber JC (1992) The national meteorological center's spectral statistical-interpolation analysis system. Mon Weather Rev 120:1747–1763

Qiu CJ, Shao AM, Liu S, Xu Q (2006) A two-step variational method for three-dimensional wind retrieval from single Doppler radar. Meteorol Atmos Phys 91:1–8

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

Shampine LF, Gordan MK (1975) Computer solution of ordinary differential equations: the initial value problem. W. H. Freeman, San Francisco, 318 pp

Shimizu S, Uyeda H, Moteki Q, Maesaka T, Takaya Y, Akaeda K, Kato T, Yoshizaka M (2008) Structure and formation mechanism on the 24 May 2000 supercell-like storm developing in a moist environment over the Kanto Plain, Japan. Mon Weather Rev 136:2389–2407

Shun CM, Chan PW (2008) Applications of an infrared Doppler Lidar in detection of wind shear. J Atmos Ocean Technol 25:637–655

Smagorinsky J (1963) General circulation experiments with the primitive equations, I. The basic experiment. Mon Weather Rev 91:99–164

Tang W, Chan P, Haller G (2011) Lagrangian coherent structure analysis of terminal winds detected by Lidar. Part I: turbulence structures. J Appl Meteorol Climatol 50:325–338

Troen I, Mahrt L (1986) A simple model of the atmospheric boundary layer: Sensitivity to surface evaporation. Bound–Layer Meteorol 37:129–148

Wong WK (2010) Development of operational rapid update non-hydrostatic NWP and data assimilation systems in the Hong Kong Observatory. In: Proceedings of the 3rd international workshop on prevention and mitigation of meteorological disasters in Southeast Asia, Beppu, March 2010, pp 1–4

Wong WK, Chan PW, Ng CK (2011) Aviation applications of a new generation of mesoscale numerical weather prediction system of the Hong Kong Observatory. In: Proceedings of the 24th conference on weather and forecasting/20th conference on numerical weather prediction, American meteorological society, Seattle, Jan 2011, pp 24–27

Chapter 23
Ensemble Adaptive Data Assimilation Techniques Applied to Land-Falling North American Cyclones

Brian C. Ancell and Lynn A. McMurdie

23.1 Introduction

Adaptive data assimilation is becoming an increasingly important aspect of numerical weather prediction. Traditional data assimilation involves combining a set of *routine* observations with a first-guess field provided by a numerical weather prediction model to produce an analysis of the atmospheric state. These analyses subsequently serve as the initial conditions for extended forecasts. There are three primary modern data assimilation methods that assimilate routine observations at operational centers around the world and within a number of research applications: (1) three-dimensional variational (3DVAR) systems, (2) four-dimensional variational (4DVAR) systems, and (3) ensemble Kalman filter (EnKF) systems. Each of these techniques are based on the assumption that the errors of both the first-guess, or background, variables and the observations are distributed normally, and aim to identify the most likely atmospheric state within the statistical framework of Bayes' Theorem (overview provided in Kalnay 2002).

Adaptive data assimilation allows the consideration of observational impact in some way beyond the aggregate effects of a set of routine observations. There are two primary types of adaptive data assimilation: (1) observation impact, and (2) observation targeting. Observation impact methods estimate the relative impact of each assimilated observation, or any subset of assimilated observations, on a chosen forecast metric. In turn, these techniques are able to identify which observations are important, and which are redundant, with regard to a number

B.C. Ancell (✉)
Department of Geosciences, Texas Tech University, Lubbock, TX 79409, USA
e-mail: Brian.Ancell@ttu.edu

L.A. McMurdie
Dept. of Atmospheric Sciences, University of Washington, Box 351640,
Seattle WA 98195-1640, USA

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_23,

of different forecast aspects. The benefit of observation impact schemes is that they perform the assimilation of numerous observations only once to estimate the impact of each observation, removing the need to perform a large number of experiments (assimilating different observations each time) to achieve the same goal. Observation targeting methods estimate the impact from hypothetical observations that could be taken beyond an initial set of assimilated observations, revealing the locations where additional observations should be taken to produce the most benefit to a chosen forecast metric. In this way, targeting methods can be used to indicate the optimal placement of additional observational platforms. One attractive aspect of both observation impact and targeting approaches is that they easily allow the consideration of a specific forecast metric that diagnoses different sensible and high-impact weather events, such as localized wind speed or regional precipitation amount. Thus, these methods will likely have important applications in the future to answer a key question: what is the best way to observe the atmosphere to improve forecasts of specific severe weather phenomena?

This chapter reviews some of the leading observation impact and targeting methods today, gives a discussion of their evolution from older techniques, and applies one such targeting approach within an ensemble framework to a particular high-impact weather event: land-falling mid-latitude cyclones on the west coast of North America. Through this application, a variety of basic observation targeting characteristics of a specific data assimilation/forecasting system can be learned with regard to a specific, high-impact weather event, and include (1) the most important observation type to target, (2) whether targeting regions occur in the same location for different events or if they span a wide range of horizontal and vertical locations, and (3) if the relative impacts of targeted observations depend on the specific nature of the event (e.g. deepening or decaying cyclones) for which one is trying to improve the forecast. The application portion of this chapter addresses each of these three characteristics for Pacific land-falling North American cyclones.

23.2 The Evolution of Adaptive Data Assimilation Techniques

Early objective adaptive data assimilation techniques focused mostly on observation targeting, and addressed primarily the dynamical growth of forecast errors. Adjoint sensitivity (overview provided by Errico 1997) or singular vector methods (summarized in Kalnay 2002) were both employed to understand where analysis errors would grow rapidly, regardless of the data assimilation procedure used in creating the analyses. Atmospheric adjoint sensitivity was first derived in LeDimet and Talagrand (1986), and can be represented with the following equation:

$$\partial \mathrm{R}/\partial \mathbf{x}_{\mathbf{o}} = \mathrm{M}_{\mathrm{t,to}}^{\mathrm{T}} * \partial \mathrm{R}/\partial \mathbf{x}_{\mathrm{t}} \tag{23.1}$$

where $\mathrm{M}^{\mathrm{T}}_{\mathrm{t,to}}$ is the transpose of the tangent-linear operator matrix obtained by linearizing the forcing terms of the full nonlinear forecast model equations, and $\partial \mathrm{R}/\partial \mathbf{x}_t$ is the differentiated response function R with respect to the atmospheric state at forecast time. The response function R can be any differentiable function of the forecast state variables, and is typically chosen to diagnose a specific aspect of the atmospheric state, such as low-level wind speed or localized precipitation amount. The term $\partial \mathrm{R}/\partial \mathbf{x}_o$ exists at every model grid point and represents the adjoint sensitivity of R with respect to the initial-time atmospheric state. For large sensitivity values, small perturbations to the initial-time atmospheric state will result in large perturbations to the forecast response function R. On the other hand, very large initial-time perturbations hardly influence R where sensitivity values are very small. In turn, adjoint sensitivity reveals regions where analysis errors would grow rapidly to cause large errors in the forecast response function R, revealing areas where it would be undesirable to have initial condition error.

Singular vectors (SVs) are similar to adjoint sensitivity in that they also utilize the tangent linear propagator matrix $\mathrm{M}_{\mathrm{t,\,to}}$. Gelaro et al. (1999) provide an overview of how SVs can be obtained by calculating the eigenvectors of the eigenvalue/eigenvector problem:

$$\left(\mathrm{M}^{\mathrm{T}}_{\mathrm{t,to}} * \mathrm{M}_{\mathrm{t,to}}\right) * \mathrm{u_i} = \sigma_{\mathrm{i}}^2 * \mathrm{u_i} \tag{23.2}$$

where u_i are the orthogonal initial-time SVs of $\mathrm{M}_{\mathrm{t\,,to}}$ (or eigenvectors of $\mathrm{M}^{\mathrm{T}}_{\mathrm{t,\,to}} {*} \mathrm{M}_{\mathrm{t,\,to}}$) with growth rates σ_{i}. The SVs with largest growth rates are the fastest growing perturbations with respect to the Euclidean norm $\left(\mathrm{u}_{\mathrm{i}}^{\mathrm{T}} * \mathrm{u_i}\right)^{1/2}$. Gelaro et al. (1999) show how the fastest growing perturbations with respect to more sophisticated norms, such as the dry total energy norm can be found, which adds additional weighting terms to equation (23.2) and presents a new eigenvalue/eigenvector problem that must be solved. In any case, the leading SVs reveal where errors would grow most rapidly with regard to a specified norm, and like adjoint sensitivity reveal areas where analysis error is undesirable with regard to the predictability of a specified aspect of the forecast state. For both adjoint sensitivity and SV applications, perturbation growth is measured about a previously run forecast. Both methods possess errors associated with the assumption of linear perturbation growth and the lack of a tangent-linear propagator containing the linearization of certain complex physics that exist in the full nonlinear model.

Perhaps motivated by studies that supported the notion of key analysis errors in regions of large adjoint sensitivity and leading SVs being most detrimental to forecasts (Rabier et al. 1996; Klinker et al. 1998), early observation targeting techniques were based on these locations. The basic idea was that by reducing errors where they would grow rapidly is the most effective way to improve forecasts. Buizza and Montani (1999), Gelaro et al. (1999), Langland et al. (1999), and Liu and Zou (2001) all found that by ingesting targeted observations in areas of leading SVs or large adjoint sensitivity, significant forecast error reductions (from 10 % to 50 %) were produced. These studies revealed the usefulness and value of SV and sensitivity-based targeting for improving forecasts, and similar methods are still in use today with regard to high-impact weather events such as tropical cyclones (Reynolds et al. 2009).

Early efforts were also made to account for analysis uncertainty in addition to dynamical error growth through SV or adjoint sensitivity techniques. Taking into account analysis uncertainty is important because if observations are taken and assimilated in regions based on leading SVs, for example, they would have little impact if the background uncertainty was very small because the data assimilation system would essentially ignore the targeted observations. In turn, other locations with less-amplifying SVs may produce larger forecast impacts if large uncertainty and larger analysis increments were produced there, even if the dynamical error growth rates of those perturbations were smaller. Barkmeijer et al. (1998) addressed this issue with Hessian SVs, which are calculated with a norm based on analysis uncertainty provided by a 3DVAR system at initial time. Bishop and Toth (1999) developed the ensemble transform method, which accounts for uncertainty within the framework of an ensemble.

A major step forward in observation targeting techniques came with the realization that the characteristics of the data assimilation system used to assimilate the targeted observations should be considered. Data assimilation systems not only provide background uncertainty estimates at initial time, but also include observation error estimates, and contain the exact procedure that would be used to assimilate targeted observations. In turn, by considering both the assimilation characteristics and a way to estimate error growth (such as through SVs or adjoint sensitivity), more appropriate observation targeting techniques can be formulated that estimate more accurately how hypothetical observations would impact forecasts in a specific assimilation system. Both Berliner et al. (1999) and Langland (2005) elaborate on the necessity to include error evolution dynamics, analysis uncertainty, observation errors, and the specific assimilation system in formulating observation targeting schemes. This holistic approach to targeted observing laid the groundwork for modern adaptive data assimilation techniques using variational and ensemble methods.

Modern adaptive data assimilation was marked by the extension of initial condition sensitivity into observation sensitivity, and was first described in Baker and Daley (2000) in the context of a 3DVAR system. Observation sensitivity describes not how perturbations to initial conditions would change the forecast (as adjoint sensitivity does), but how an assimilated observation would change the forecast, and can be written as:

$$\partial \mathrm{R}/\partial \mathbf{y_o} = \partial \mathrm{R}/\partial \mathbf{x_o} * \partial \mathbf{x_o}/\partial \mathbf{y_o} \tag{23.3}$$

where $\partial \mathrm{R}/\partial \mathbf{y_o}$ is the observation sensitivity, which is a function of the adjoint sensitivity and the change to the analysis given observations ($\partial \mathbf{x_o}/\partial \mathbf{y_o}$). For data assimilation systems that assume Gaussian statistics to achieve a most-likely state, the term $\partial \mathbf{x_o}/\partial \mathbf{y_o}$ simply becomes the Kalman gain matrix. Equation (23.3) is a form of observation targeting as described in Baker and Daley (2000) as it allows one to estimate the impact on a response function R due to innovations ($\varDelta \mathbf{y_o}$) with the calculation:

$$\varDelta \mathrm{R} = \partial \mathrm{R}/\partial \mathbf{y_o} * \varDelta \mathbf{y_o} \tag{23.4}$$

The only drawback of this method is that innovations associated with hypothetical observations are not known prior to obtaining the observations, although the technique is still very useful for understanding relative forecast impacts from a fixed innovation anywhere in the model domain.

Langland and Baker (2004) derived the observation impact methodology directly from (23.3) and (23.4), noting that $\varDelta$R is composed of a sum of terms, each term containing a coefficient representing a different innovation (and thus a different observation). In this way, the contribution from each observation or any subset of observations to $\varDelta$R can be easily calculated, and the impact of different, assimilated observations or subsets of observations can easily be produced. Conceptually, this is equivalent to the analysis increment produced from a specific set of assimilated observations projected onto the adjoint sensitivity field, yielding an estimate of $\varDelta$R. This technique is the foundation for a number of observation impact studies (Langland and Baker 2004; Tremolet 2008; Gelaro and Zhu 2009; Gelaro et al. 2010), although these studies expand the observation impact method to account for nonlinear terms in the definition of the response function. Errico (2007) and Gelaro et al. (2007) discuss the accuracy of the expanded higher-order methodology, and also offer a more in depth interpretation of equation (23.4) noting that cross-correlations appear in each observational term that sum to produce $\varDelta$R. The important issue these studies address through the observation impact technique is to understand which types of observations, such as those at different heights or those associated with different observational platforms, contribute to reducing forecast error and which do not. These results are crucial toward designing the most effective routine observational networks for operational assimilation/forecasting systems.

Significant observation impact and targeting developments were also made using ensemble data assimilation systems. Bishop et al. (2001) developed an ensemble transform Kalman filter (ETKF) observation targeting method based on the ensemble transform technique of Bishop and Toth (1999). The ETKF method is able to estimate the reduction in forecast variance due to hypothetical observations. A similar method was provided by Ancell and Hakim (2007a) within an EnKF assimilation system that also estimates the reduction in forecast variance of a chosen response function R due to hypothetical observations. This method is based on ensemble sensitivity which can be calculated in the following way (Ancell and Hakim 2007a):

$$\partial \mathrm{R}/\partial \mathbf{y_o} = \mathrm{Cov}(\mathrm{R}, \mathbf{y}_{\mathrm{o}}) * \mathrm{D}^{-1} \tag{23.5}$$

where $\partial \mathrm{R}/\partial \mathbf{y_o}$ is a row vector representing the ensemble sensitivity of R with respect to each analysis variable, Cov(R, $\mathbf{y_o}$) is a row vector representing the covariance between the response function R and each analysis variable, and D is a diagonal matrix containing the variance of each analysis variable. Ancell and Hakim (2007a) explain that ensemble sensitivity allows one to estimate the perturbation to the response function R resulting from the temporal evolution of an initial-time perturbation spread spatially and into other variables through the background error covariance relationships of the ensemble (similar to (23.4)). Since observational information is spread in a similar manner within the EnKF analysis procedure, and since the temporal evolution of perturbations can be represented with adjoint sensitivity, Ancell and Hakim (2007a) exploit the relationship between ensemble and adjoint sensitivity to derive an expression for the reduction in the variance of R due to a single observation within an EnKF:

$$\mathit{\Delta}\mathrm{Variance}_{\mathrm{R}} = (\mathrm{D_i} * \partial\mathrm{R}/\partial\mathrm{y_i})^2/(\mathrm{D_i} * \mathrm{O_i}) \qquad (23.6)$$

where $\mathrm{D_i}$ represents the variance of a single anlaysis variable, $\partial\mathrm{R}/\partial\mathrm{y_i}$ is the ensemble sensitivity with respect to the same analysis variable, and $\mathrm{O_i}$ represents the observation error variance associated with a targeted observation. This calculation can be quickly made with respect to each observable analysis variable to reveal the estimated variance reduction from a single additional, hypothetical observation anywhere on the model domain. An advantage of these ensemble-based methods is that they rely not on actual observation values, but on observation error variance which exists prior to hypothetical observations being taken. They also allow the estimation of forecast variance reduction of additional targeted observations conditioned on the simultaneous assimilation of the initial targeted data. Ancell and Hakim (2007a) also derive an ensemble version of the observation impact developed in Langland and Baker (2004) without the use of an adjoint model. Liu and Kalnay (2008) discuss yet another ensemble-based observation impact technique that requires no adjoint model.

In summary, adaptive data assimilation techniques have evolved from those that consider only dynamical error growth to those that consider all aspects of the data assimilation system used to assimilate routine and targeted observations. In turn, modern adaptive data assimilation methods provide estimates of the impacts from assimilated or additional hypothetical observations with regard to a specific assimilation system such as 3DVAR or an EnKF. It should be noted that nearly all observation impact/targeting techniques are based on the assumption that error evolution is linear over the duration of the forecast, an assumption that doesn't always hold. Furthermore, both modern data assimilation systems and forecasting models are not perfect, and present another source of error for observation impact and targeting schemes. Langland (2005) provides an excellent review of the potential impacts these issues cause, and discusses the performance of different adaptive data assimilation methods during a variety of recent field programs. As computational resources are constantly improving, investigating adaptive assimilation techniques at very high resolution (grid spacing of a few kilometers) is now becoming possible. In turn, a major research focus in the coming years will likely be on the application of adaptive data assimilation systems at different scales.

23.3 Application of EnKF Observation Targeting to Land-Falling Mid-Latitude Cyclones

We now apply the EnKF observation targeting methodology of Ancell and Hakim (2007a) to understand the nature of the impacts of hypothetical observations beyond those of routine data for land-falling mid-latitude cyclones on the west coast of North America. Land-falling cyclones routinely produce heavy precipitation and high winds in coastal regions, and their predictability characteristics are thus important to understand. Mass and Dotson (2010) describe some of the most intense cyclones to strike the west coast of North America, and discuss the major societal impacts they produced. McMurdie and Mass (2004) and Wedam et al. (2009) describe how short-term forecast errors by deterministic operational numerical models can be large for storms impacting the west coast. The purpose of this study is to demonstrate how adaptive assimilation techniques can be applied retrospectively to cases of land-falling cyclones as a research tool to investigate the role of additional observations within a specific assimilation system and observing network.

23.3.1 Details of the EnKF and the Forecast Model

The EnKF used in this study is an ensemble square-root filter that assimilates observations serially (Whitaker and Hamill 2002) and was created at the University of Washington (Torn and Hakim 2008). The 80-member EnKF runs on a 6-h update cycle on the modeling domain shown in Fig. 23.1 at 36-km grid spacing with 37 vertical levels. The routine observations that are assimilated are cloud-track wind (typically from 1,000 to 4,000 total), acars aircraft wind and temperature (typically from 1,000 to 4,000 total), radiosonde wind, temperature, and relative humidity (typically around 1,500 total), and surface wind, temperature, and altimeter data (typically from 7,000 to 10,000 total). The EnKF uses both Gaspari-Cohn horizontal localization (Gaspari and Cohn 1999) and posterior inflation to address sampling error and to avoid filter divergence (Anderson and Anderson 1999). The inflation and localization parameters used in this study are the same as those in Torn and Hakim (2008) which were tuned over a similar domain to produce appropriate spread and minimum ensemble mean errors. Boundary conditions were perturbed around the Global Forecasting System (GFS) analyses and forecasts using the fixed covariance perturbation method of Torn et al. (2006).

The EnKF was cycled for 6 months from 0000 UTC October 1, 2009 to 1800 UTC March 31, 2010, and extended ensemble forecasts were run to 24-h forecast time to capture a number of wintertime land-falling cyclones. The forecast model used here is the Advanced Research Weather Research and Forecasting (WRF-ARW) model Version 3.0.1.1 (Skamarock et al. 2008). The WRF physics used are the Mellor-Yamada-Janjic (MYJ) planetary boundary layer scheme (Janjic 1990, 1996, 2002), the Kain-Fritsch cumulus parameterization (Kain and Fritsch 1990,

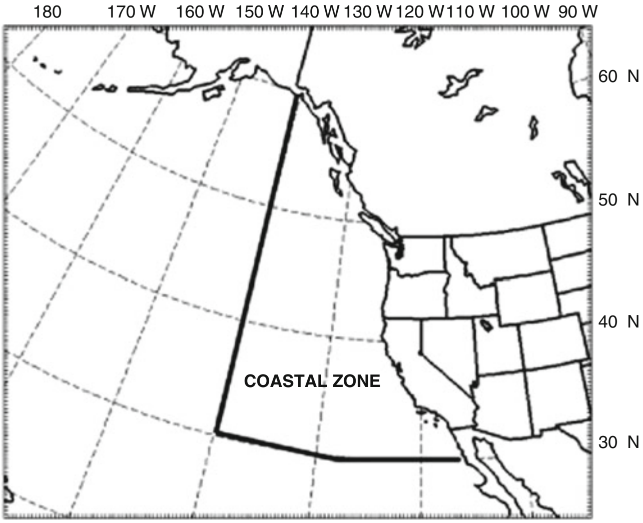


Fig. 23.1 The EnKF domain used in this study. The coastal zone used to identify land-falling mid-latitude cyclones is outlined by the *thick black lines* and the North American coastline

1993), the Noah land surface model (Chen and Dudhia 2001), WRF Single-Moment 3-class microphysics (Hong et al. 2004), the Rapid Radiative Transfer Model (RRTM) longwave radiation scheme (Mlawer et al. 1997), and the Dudhia shortwave radiation scheme (Dudhia 1989).

23.3.2 Description of the Response Function and Case Selection

The response function used in these experiments to diagnose land-falling cyclones is the average sea-level pressure in a 216 km by 216 km box surrounding the 24-h forecast ensemble mean cyclone center. Only cyclones that could be identified as a local minimum and tracked back for the 24-h forecast period were included. Observation targeting calculations based on the methodology of Ancell and Hakim (2007a) are performed for every 24-h forecast cyclone that was found in the coastal zone (outlined in Fig. 23.1) over the 6-month duration of this study. A total of 27 storms were found to impact the coastal zone over this time, which is a typical frequency for wintertime land-falling cyclones. However, each storm lasted for several days and its position at the 24-h forecast time would be within the coastal zone over several consecutive forecast runs. Therefore, targeting calculations were made several times for each storm, resulting in a total of roughly 200 cases. Table 23.1 characterizes each forecast run of these storms with regard to their deepening rate and direction of coastal approach, two aspects that are analyzed later in this section. It should be noted that it was not possible to characterize each of the 27 individual storms as deepening or decaying as a particular event may

Table 23.1 The total number of cyclones as well as those counted as deepening or decaying, or coming from the north, northwest, west, southwest, or south in this study

Cyclone characteristic	Number
Deepening	37
Decaying	87
From the north	4
From the northwest	19
From the west	37
From the southwest	48
From the south	16
All cyclones	198

have forecast runs early in its lifetime when it deepens and forecast runs later in its lifetime when it decays. In the subsequent discussion, we will use 'storms' to refer to unique cyclones (i.e. the 27 storms) and 'cyclones' to refer to the individual forecast runs (i.e. one of the 200 samples).

The observation targeting calculations indicate the estimated variance reduction to the response function due to the assimilation of hypothetical temperature, wind, and pressure observations at analysis time *beyond* the assimilated routine data. In turn, the largest variance reduction values reveal the locations where an initial-time observation would reduce the uncertainty of the 24-h response function the most.

23.3.3 Characteristics of Observation Targeting for a Single Cyclone

Figure 23.2 shows the 00-h, 12-h, 18-h, and 24-h ensemble mean forecast initialized at 0600 UTC November 9, 2009 that depicts one particular cyclone that made landfall on the west coast of North America. This cyclone decays from 986-hPa central pressure in the analysis to 993-hPa central pressure when it makes landfall on the Canadian coast at 24-h forecast time. The targeting regions for winds, temperature, and pressure valid at analysis time are shown in Fig. 23.3, and are plotted at the level where the maximum value of estimated variance reduction was found. The targeting regions based on temperature and winds are localized and mesoscale in nature, which is generally the case for most land-falling cyclones during the 2009/2010 winter season (not shown). The targeting regions based on pressure are more typically characterized by synoptic-scale features, which is the case in Fig. 23.3. The largest targeting regions based on winds and pressure for this specific cyclone are found in the lower troposphere (from roughly 880 to 750 hPa for winds, 930 hPa for pressure), and near the tropopause (roughly 380 hPa) for temperature. Targeting regions based on all four observation types reveal areas in the immediate vicinity of the incipient system at analysis time, with wind and temperature targets aloft flanking the central position of the 500-hPa geopotential height minimum, and the primary pressure targets positioned just over the cyclone center at the surface. Dynamically, the primary zonal and meridional wind targets exist north and south (for zonal wind) and east and west (for meridional wind) of the cyclone center aloft, suggesting the effects of observations there would beneficially

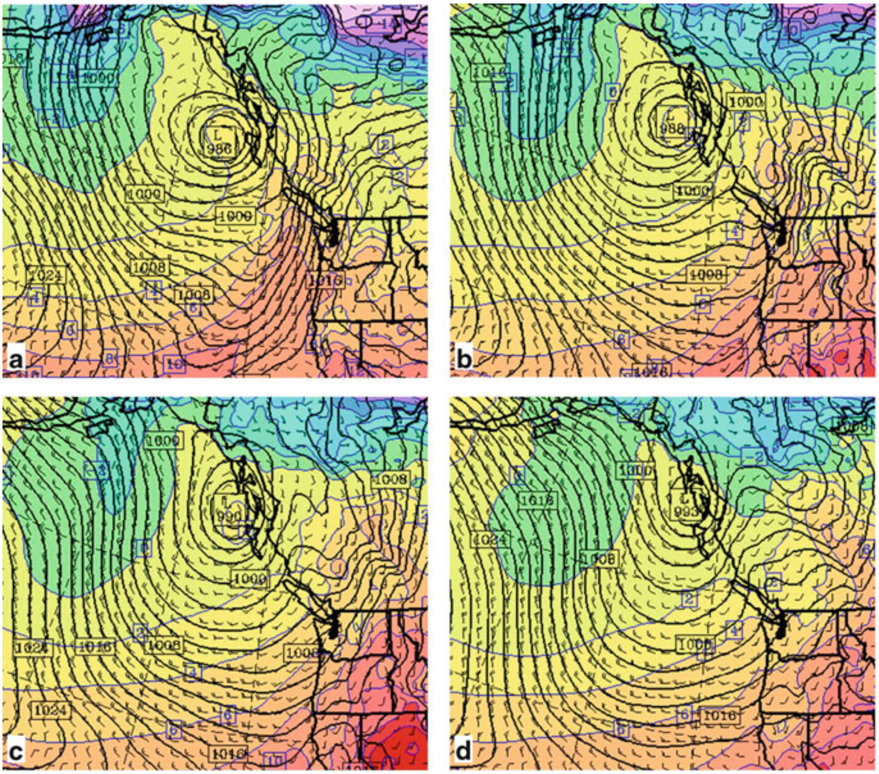


Fig. 23.2 Ensemble mean forecast initialized at 0600 UTC November 9, 2009 of a decaying mid-latitude cyclone making landfall on the North American coastline valid at (**a**) 00-h, (**b**) 12-h, (**c**) 18-h, and (**d**) 24-h. Black contours represent sea-level pressure (contour interval is 2 hPa), *blue contours* represent 925-hPa temperature (contour interval is 2°C), and *wind barbs* represent 10-m winds

alter the cyclonic wind field flowing around the cyclone. Magnitudes of the variance reduction field are largest for pressure (reaching just over 1.4 hPa^2), are slightly less for the wind field (reaching about 1.2 hPa^2), and are smallest for the temperature field (reaching about 0.9 hPa^2).

One interesting and unique feature of the targeting regions based on pressure in Fig. 23.3 is that they are less localized and show some impact away from the center of the system. Although values in these more distant regions are not as large as those in the immediate vicinity of the cyclone, they clearly highlight features in the flow at analysis time. Both the frontal trough near 40°N, −130°W and the large oceanic region of high pressure in the western portion of the domain are shown to be relatively important. In tune with the discussion of Ancell and Hakim (2007a), these features are highlighted as targets because they reveal areas where analysis increments would project substantially onto regions of large dynamical sensitivity (a quantity estimated through adjoint sensitivity analysis). This in turn reveals a defining characteristic of observation sensitivity over that of adjoint sensitivity—

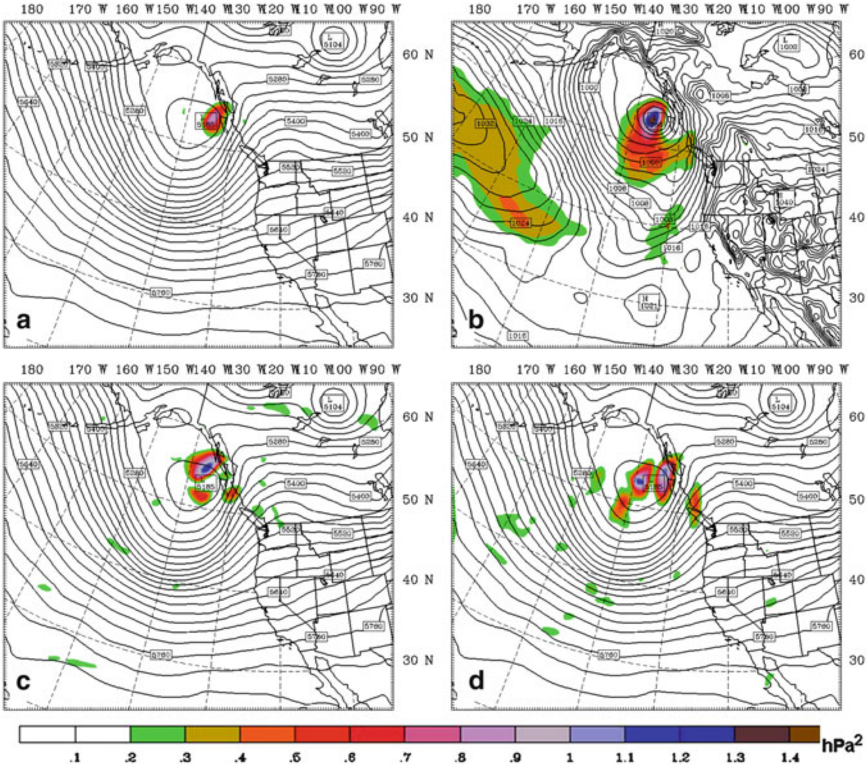


Fig. 23.3 Variance reduction (*shaded*, hPa^2) of the response function estimated from a single observation at the level of the maximum variance reduction for (**a**) temperature (∼ 380 hPa), (**b**) pressure (∼ 930 hPa), (**c**) zonal wind (∼ 750 hPa), and (**d**) meridional wind (∼ 880 hPa) valid at 0600 UTC November 9, 2009. Ensemble mean sea-level pressure (panel b, *black contours*, contour interval is 2 hPa) and 500-hPa geopotential height (panels a, c, and d, *black contours*, contour interval is 30 m) are also shown

observation targets can exist in relatively distant areas from the regions of large adjoint sensitivity, sometimes indicating larger impacts than the regions of large adjoint sensitivity itself. An important consequence of this characteristic are that targeting regions based on observation sensitivity can differ strongly from those based on adjoint sensitivity, indicating the importance of observation sensitivity for adaptive data assimilation techniques as discussed in Sect. 23.2. As pointed out, Fig. 23.3 shows the Pacific high surface pressure to be an important targeting region, even though it is likely far upstream from the area of large adjoint sensitivity (typical adjoint sensitivity fields associated with cyclones are shown by Ancell and Mass 2006; Ancell and Hakim 2007a, b in similar experimental configurations). Although perturbing this area of high pressure itself would do little to the 24-h forecast of the land-falling cyclone, information spread during assimilation of observations of the area of high pressure into regions of large dynamic sensitivity downstream would act to significantly influence the forecast of the cyclone.

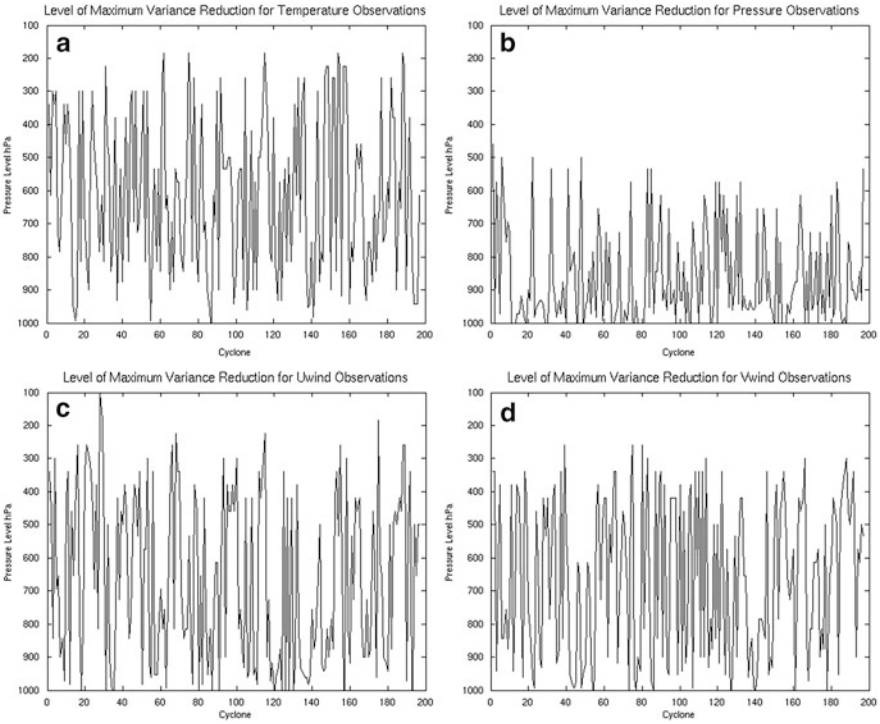


Fig. 23.4 Level of maximum variance reduction of the response function for all cyclones for observations of (**a**) temperature, (**b**) pressure, (**c**) zonal wind, and (**d**) meridional wind

The fact that targeting regions are clearly co-located with features in the flow is likely a unique feature of ensemble targeting techniques. Following the discussion above, the impacts from hypothetical observations within an EnKF highlight specific flow features through their relationship to dynamically sensitive areas. In turn, these relationships depend on how the specific features in each ensemble member covary with the dynamically sensitive regions, and are thus strongly linked to the flow dependence present in the atmospheric state at any given time. Consequently, covariances that do not possess such flow dependence are unlikely to capture these relationships, and targeting regions based on 3DVAR systems will probably differ to some degree from those based on ensemble methods. This further stresses the importance of directly accounting for the specific assimilation system when calculating the impacts of targeted observations.

23.3.4 Characteristics of Observation Targeting for all 2009/2010 Cyclones

Figure 23.4 represents the level where the maximum variance reduction exists for each cyclone for pressure, temperature, and wind observations. For both temperature

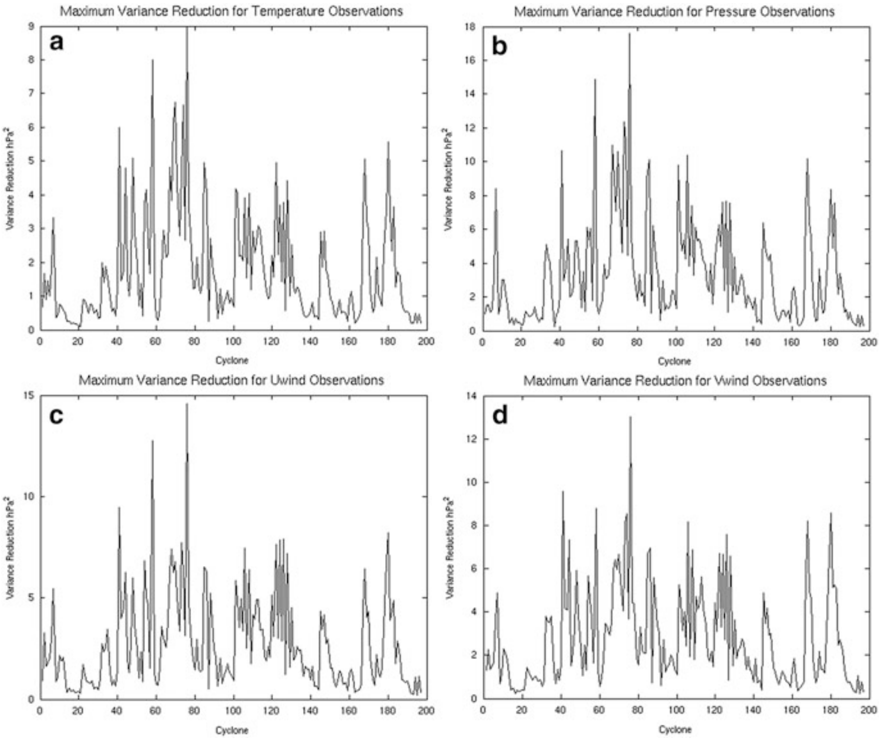


Fig. 23.5 Value of maximum variance reduction (hPa^2) of the response function for all cyclones for observations of (**a**) temperature, (**b**) pressure, (**c**) zonal wind, and (**d**) meridional wind

and wind, the level of maximum values exists throughout the troposphere, whereas for pressure the maximum levels are confined to the lower half of the troposphere below 500 hPa. Furthermore, the level of maximum variance reduction is near the surface for pressure observations for a large number of cyclones. This suggests a substantial benefit might be gained by assimilating scatterometer sea-level pressure retrievals, such as those that used to be provided by the QuickSCAT satellite, with regard to forecasts of North American land-falling mid-latitude cyclones.

Figure 23.5 depicts the maximum values of variance reduction for each cyclone, regardless of vertical level, for each observation type. As shown in Fig. 23.3, the largest values are associated with pressure observations. For all observations, significant variability exists with these maximum values as they range from near zero to about 15 hPa^2 for winds, to about 9 hPa^2 for temperature, and to roughly 18 hPa^2 for pressure. Interestingly, the maximum values for all observation types follow the same general trend. When a cyclone exhibits large maximum variance reduction for one type of observation (e.g. pressure), it also exhibits large maximum variance reduction for the other types of observations (wind and temperature). This property reveals that the largest impacts of targeted observations for the cyclones in this study are independent of observation type.

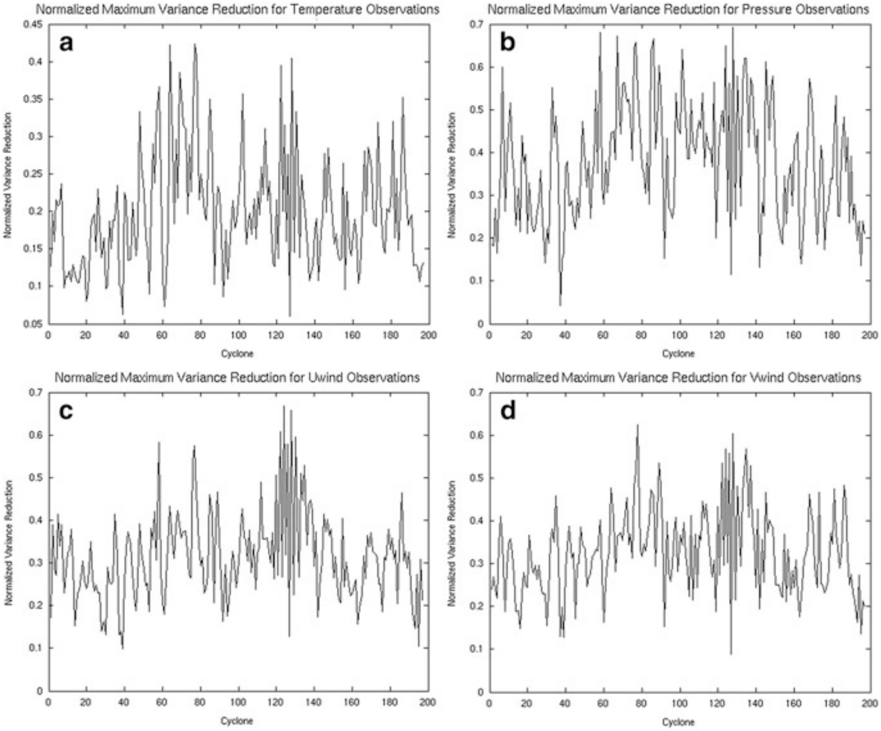


Fig. 23.6 Value of maximum variance reduction of the response function normalized by the response function variance for all cyclones for observations of (**a**) temperature, (**b**) pressure, (**c**) zonal wind, and (**d**) meridional wind

It is also useful to describe targeting impacts in terms of the estimated variance reduction relative to the response function variance. This can be done by dividing the maximum variance reduction values by the original response function variance resulting in what is referred to here as normalized variance reduction. In this way, it is possible to reveal cases where observations might reduce a substantial fraction of the original response function variance even if the actual variance reduction values themselves (as shown in Fig. 23.5) are quite small. Figure 23.6 shows the normalized maximum values of estimated variance reduction, and reveals the percentage of response function variance that could be reduced through the assimilation of observations. In general, pressure observations show an estimated 34 % variance reduction averaged over all cyclones, which is larger than both wind (23 %) and temperature (12 %). As with the absolute values in Fig. 23.5, the trend among all variables remains the same. Figure 23.7 depicts both the normalized and absolute maximum values of variance reduction with regard to temperature for all cyclones. The trend for both the normalized and absolute values is generally similar for all cyclones, although there are localized differences in the plots. For example, the peaks in the normalized and non-normalized values between cyclone number 30

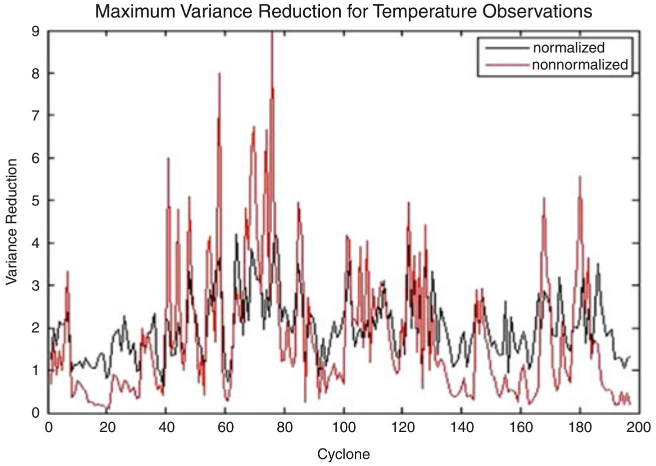


Fig. 23.7 Both normalized (nondimensional) and non-normalized (hPa^2) values of maximum variance reduction of the response function for all cyclones for observations of temperature. Normalized values are multiplied by 10 for ease of comparison with non-normalized values

and 40 are offset. This indicates that whereas a larger absolute variance reduction is estimated from temperature observations at the peak of the non-normalized values, the response function variance must be somewhat larger for that cyclone such that the estimated fraction of response function variance is smaller. This indicates that for specific cases, the impacts of EnKF targeted observations can be viewed with differing degrees of importance depending on whether these impacts are determined by the total or the fraction of estimated response function variance reduction.

Another interesting result that can be found by analyzing targeted observations for many cases is how the location of the most significant targeting locations vary in time. For a specific high-impact weather event, if the targeting regions remained constant over many cases, strong support would exist for taking routine observations in those locations. If targeting regions were not constant, the degree to which they vary would provide crucial information toward how to best design an adaptive observing network. Figure 23.8 shows the mean estimated maximum variance reduction calculated over all cyclones throughout the vertical. Interestingly, pressure observations show roughly a constant impact throughout the troposphere. This reveals that although the maximum estimated targeting values tend to occur in the lower atmosphere for pressure observations (Fig. 23.4), values are nearly constant within the entire troposphere. In turn, there are no preferred targeting locations in the vertical with regard to pressure observations to improve land-falling cyclone forecasts. Wind and temperature observation targeting regions, however, show two distinct peaks in the vertical. Wind targeting regions show peaks near 400 hPa and the surface, whereas temperature targeting regions reveal the 250-hPa and the 600-hPa levels to be most important. It seems reasonable that the important temperature targeting locations near 250-hPa are due to large variance near the

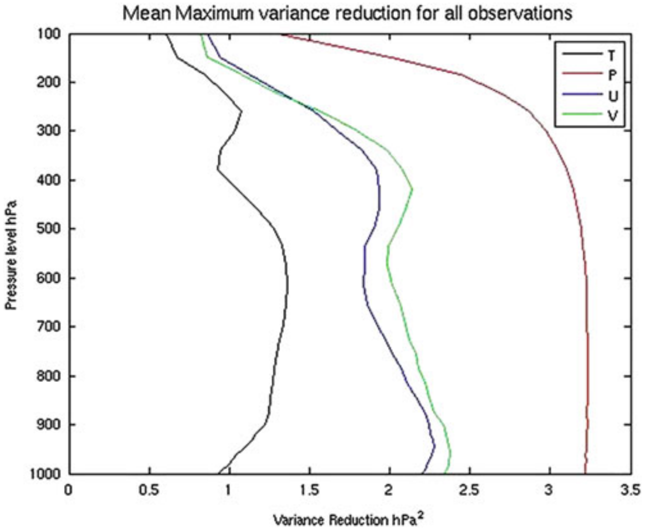


Fig. 23.8 Value of maximum variance reduction of the response function averaged over all vertical model levels for all cyclones for all observation types

tropopause involved with the general temperature minimum at that level. It is not obvious why the 600-hPa level is important, or why wind targeting locations are important near the surface and at 400 hPa. Either large ensemble-based sensitivity or analysis variance values would contribute to large estimated variance reduction values, and it is thus likely one of these two quantities is consistently larger at the levels where the peaks are evident in Fig. 23.8. In any case, these levels indicated preferred locations where on average, supplemental temperature and wind observations would be most beneficial.

Figure 23.9 shows the horizontal locations of the maximum estimated variance reduction for temperature observations organized by cyclones that approach the coastal zone from the northwest, west, southwest, and south. Cyclones along these tracks have been binned over a 45° swath centered on each direction listed. The locations are shown relative to the 24-h forecast position of the ensemble mean cyclone. It is clear that for all cyclone tracks there is significant variability in the horizontal location of the maximum variance reduction values, varying up to about 40° both longitudinally and latitudinally. Interestingly, a number of the maximum variance reduction locations occur at or downstream of the 24-h forecast position of the cyclone, indicating the ability of the EnKF to spread observational information upstream into regions where large dynamical sensitivity is likely to exist. Nonetheless, it seems the highest priority targeting regions rarely exist in the same location relative to the forecast position of the cyclone. Although there is some clustering of maximum variance values to the southwest of cyclones that approach from the southwest, there is still a large spread in the location of maximum variance reduction such that the chance that a single location would provide consistently large

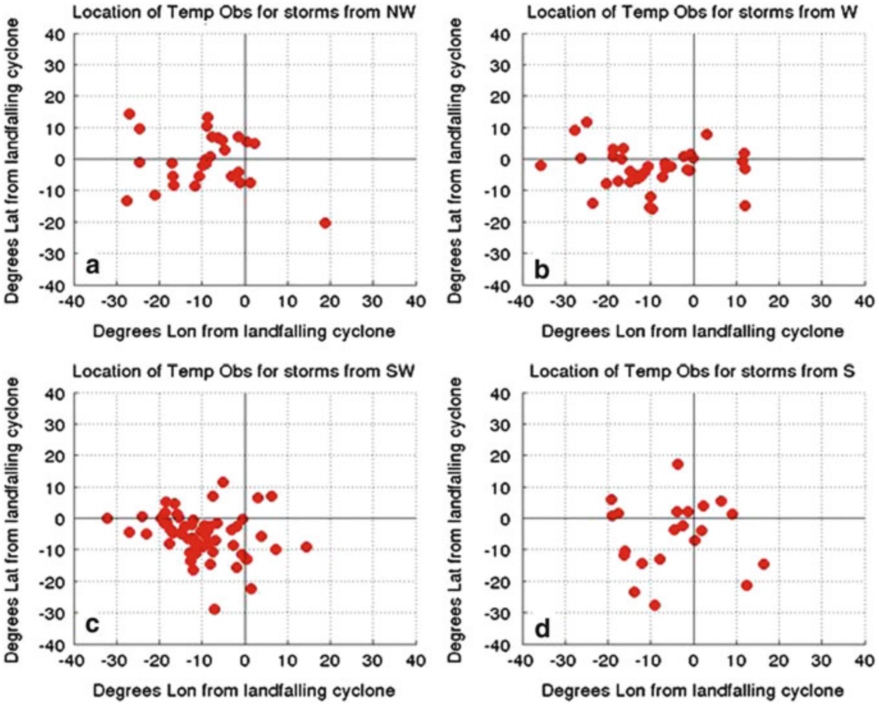


Fig. 23.9 Horizontal location relative to the 24-h forecast cyclone position of the maximum variance reduction of the response function for temperature observations for cyclones approaching the coastal zone from the (**a**) northwest, (**b**) west, (**c**) southwest, and (**d**) south

positive forecast benefits is unlikely. This is especially true since these locations are relative to the forecast position of the cyclone, and less clustering would be evident when considering the actual positions of the maximum variance reduction values within the modeling domain. Very similar results are found regarding the horizontal location of targeting sites for pressure and wind observations (not shown).

Figure 23.10 depicts the average maximum variance reduction for all observation types segregated by whether the cyclone was deepening or decaying over the 24-h forecast period. The error bars represent the 95 % confidence interval. For each observation type, deepening cyclones are associated with larger variance reduction values on average, implying that observation targeting is more effective for deepening cyclones. This result is not reproduced when considering cyclone track, as the average variance reduction values in Fig. 23.11 are essentially indistinguishable at the 95 % confidence level among different cyclone tracks. These results demonstrate how targeting impacts can relate to certain characteristics of the high impact event in question (in this case the deepening rate of land-falling cyclones).

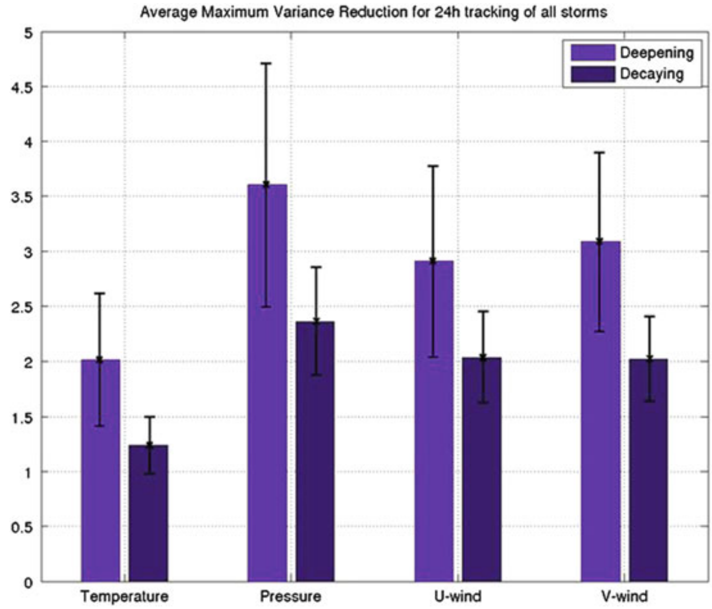


Fig. 23.10 Average value of the maximum variance reduction of the response function for all observation types for both deepening (*light purple*) and decaying (*dark purple*) cyclones

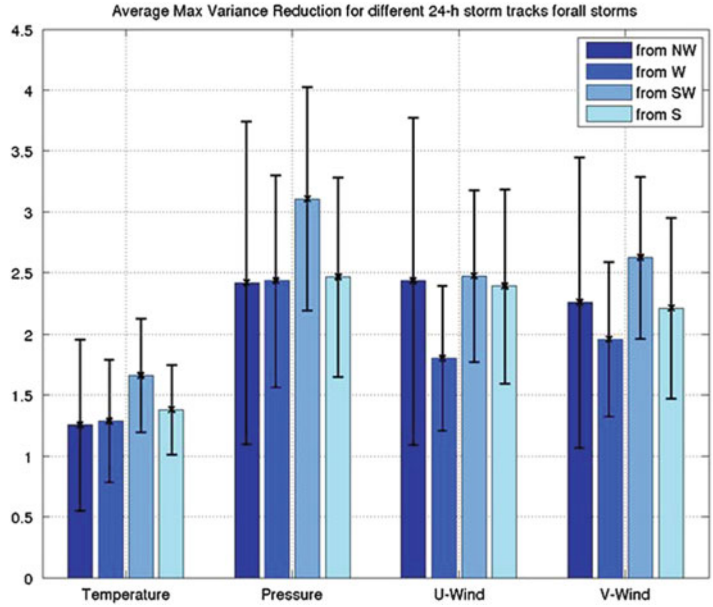


Fig. 23.11 Average value of the maximum variance reduction of the response function for all observation types for cyclones approaching the North American coast from the northwest, west, southwest, and south

23.3.5 Summary and Concluding Remarks

Observation targets of pressure, winds, and temperature within an EnKF for land-falling Pacific cyclones on the west coast of North America were examined for a 6-month wintertime synoptic period in 2009/2010. These targets represented estimates of where assimilated hypothetical observations beyond assimilated routine observations would produce the largest reduction in the uncertainty of 24-h cyclone forecasts around the time of landfall. It was found that temperature and wind targets were mesoscale in nature, whereas pressure targets were more prominent on the synoptic scale. Furthermore, pressure observations produced the largest positive impacts on the uncertainty of cyclone forecasts of the four observation types examined. The most important targeting regions in the vertical for winds and temperature varied substantially throughout the troposphere when considering all cyclones, but there was an indication of preferred regions in the mid- and upper-troposphere for temperature and the upper-troposphere and near the surface for winds. Although the largest benefits from pressure observations existed near the surface, similar benefits existed throughout the troposphere with no clear preferred level. In the horizontal, there was significant variability in the most important targeting areas, showing no clear location where a routine observation would be consistently beneficial to land-falling cyclone forecasts. Lastly, it was found that targeted observations are more beneficial to forecasts of deepening cyclones than to decaying systems as they approach the coast. This result was not found when considering the directions along which the cyclones track as cyclones from all directions showed similar benefits from targeted observations.

It is important to note that the best way to view the results presented here is in a relative sense. Specifically, these experiments have provided an understanding of how impacts vary within an EnKF among the different observation types of pressure, winds, and temperature for land-falling mid-latitude cyclones. Whether these results extend to other assimilation systems and different high-impact events is unclear. Furthermore, the estimated variance reductions in this study are based on ensemble sensitivity, and thus the particular variance reduction values would need to be compared with experiments that actually assimilate targeted observations to understand the relationship between estimated and actual forecast impacts. The effects of nonlinearity, inflation, and localization may all play a role in any discrepancy. Nonetheless, this study has provided a unique perspective on how targeting techniques might be designed to best benefit forecasts of land-falling Pacific cyclones. Lastly, as assimilation and forecasting systems at higher and higher resolution become more feasible in the coming years, gaining an understanding of the effects of targeted observations across multiple scales will be an intriguing endeavor.

References

Ancell BC, Hakim GJ (2007a) Comparing adjoint and ensemble sensitivity analysis with applications to observation targeting. Mon Weather Rev 135:4117–4134

Ancell BC, Hakim GJ (2007b) Interpreting adjoint and ensemble sensitivity toward the development of optimal observation targeting strategies. Met Zeitschrift 16:635–642

Ancell BC, Mass CF (2006) Structure, growth rates, and tangent linear accuracy of adjoint sensitivities with respect to horizontal and vertical resolution. Mon Weather Rev 134:2971–2988

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

Baker N, Daley R (2000) Observation and background sensitivity in the adaptive observation-targeting problem. Q J Roy Meteorol Soc 126:1431–1454

Barkmeijer J, Bouttier F, Gijzen MV (1998) Singular vectors and estimates of the analysis-error covariance metric. Q J Roy. Meteorol. Soc 124:1695–1713

Berliner ML, Lu Z-Q, Snyder C (1999) Statistical design for adaptive weather observations. J Atmos Sci 56:2536–2552

Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. J Atmos Sci 56:1748–1765

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

Buizza R, Montani A (1999) Targeting observations using singular vectors. J Atmos Sci 56:2965–2985

Chen F, Dudhia J (2001) Coupling an advanced land-surface/hydrology model with the Penn State/NCAR MM5 modeling system. Part I: model description and implementation. Mon Weather Rev 129:569–585

Dudhia J (1989) Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. J Atmos Sci 46:3077–3107

Errico RM (1997) What is an adjoint model? BullAm Meteorol Soc 78:2577–2591

Errico RM (2007) Interpretations of an adjoint-derived observational impact measure. Tellus 59A:273–276

Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. Q J Roy Meteorol Soc 125:723–757

Gelaro R, Zhu Y (2009) Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. Tellus 61A:179–193

Gelaro R, Langland RH, Rohaly GD, Rosmond TE (1999) An assessment of the singular-vector approach to targeted observing using the FASTEX dataset. Q J Roy Meteorol Soc 125:3299–3327

Gelaro R, Zhu Y, Errico RM (2007) Examination of various-order adjoint-based approximations of observation impact. Met Zeitschrift 16:685–692

Gelaro R, Langland RH, Pellerin S, Todling R (2010) The THORPEX observation impact intercomparison experiment. Mon Weather Rev 138:4009–4025

Hong S-Y, Dudhia J, Chen S-H (2004) A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. Mon Weather Rev 132:103–120

Janjic ZI (1990) The step-mountain coordinate: physical package. Mon Weather Rev 118:1429–1443

Janjic ZI (1996) The surface layer in the NCEP Eta Model. In: Proceedings of the eleventh conference on numerical weather prediction, American Meteorological Society, Norfolk, 19–23 Aug 1996, pp 354–355

Janjic ZI (2002) Nonsingular implementation of the Mellor-Yamada Level 2.5 scheme in the NCEP Meso model. NCEP Office Note, No. 437, 61 p

Kain JS, Fritsch JM (1990) A one-dimensional entraining/detraining plume model and its application in convective parameterization. J Atmos Sci 47:2784–2802

Kain JS, Fritsch JM (1993) Convective parameterization for mesoscale models: the Kain-Fritsch scheme. In: Emanuel KA, Raymond DJ (eds) The Representation of Cumulus Convection in Numerical Models, American Meteorological Society, Boston, 246 p

Kalnay E (2002) Atmospheric modeling, data assimilation, and predictability. Cambridge University Press, Cambridge, 364 p

Klinker E, Rabier F, Gelaro R (1998) Estimation of key analysis errors using the adjoint technique. Q J Roy Meteorol Soc 124:1909–1933

Langland RH (2005) Issues in targeted observing. Q J Roy Meteorol Soc 131:3409–3425

Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus 56A:189–201

Langland RH, Toth Z, Gelaro R, Szunyhogh I, Shapiro MA, Majumdar SJ, Morss RE, Rohaly GD, Velden C, Bond N, Bishop CH (1999) The North Pacific Experiment (NORPEX-98): targeted observations for improved North American weather forecasts. Bull Am Meteorol Soc 180:1363–1384

LeDimet F, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus 38A:97–110

Liu J, Kalnay E (2008) Estimating observation impact without adjoint model in an ensemble Kalman filter. Q J Roy Meteorol Soc 134:1327–1335

Liu H, Zou X (2001) The impact of NORPEX targeted dropsondes on the analysis and 2–3-day forecasts of a landfalling Pacific winter storm using NCEP 3DVAR and 4DVAR systems. Mon Weather Rev 129:1987–2004

Mass CF, Dotson B (2010) Major extratropical cyclones of the northwest United States: historical review, climatology, and synoptic environment. Mon Weather Rev 138:2499–2527

McMurdie LA, Mass CF (2004) Major numerical forecast failures over the northeast Pacific. Weather Forecast 19:338–356

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA (1997) Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the longwave. J Geophys Res 102:16663–16682

Rabier F, Klinker E, Courtier P, Hollingsworth A (1996) Sensitivity of forecast errors to initial conditions. Q J Roy Meteorol Soc 122:121–150

Reynolds CA, Doyle JD, Hodur RM, Jin H (2009) Naval Research Laboratory Multiscale Targeting Guidance for T-PARC and TCS-08. Weather Forecast 25:526–543

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda M, Huang X-Y, Wang W, Powers JG (2008) A description of the advanced research WRF version 3. NCAR Technical Note TN-475

Torn RD, Hakim GJ (2008) Performance characteristics of a pseudo-operational ensemble Kalman filter. Mon Weather Rev 136:3947–3963

Torn RD, Hakim GJ, Snyder C (2006) Boundary conditions for limited-area ensemble Kalman filters. Mon Weather Rev 134:2490–2502

Tremolet Y (2008) Computation of observation sensitivity and observation impact in incremental variational data assimilation. Tellus 60A:964–978

Wedam GB, McMurdie LA, Mass CF (2009) Comparison of model forecast skill of sea-level pressure along the east and west coasts of the United States. Weather Forecast 24:843–854

Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 130:1913–1924

Chapter 24
The Advances in Targeted Observations for Tropical Cyclone Prediction Based on Conditional Nonlinear Optimal Perturbation (CNOP) Method

Feifan Zhou, Xiaohao Qin, Boyu Chen, and Mu Mu

Abstract In this chapter, we review the recent progresses in targeted observations for tropical cyclone prediction based on Conditional Nonlinear Optimal Perturbation (CNOP) method. The CNOP is a natural extension of the singular vector (SV) into the nonlinear regime and it has been used to identify the sensitive areas for tropical cyclone predictions.

The properties of the sensitive areas identified by CNOP have been first studied, including the sensitivity to the horizontal resolution, the verification area design, and the optimization period. It has been found that the CNOP sensitive areas have similarities at different horizontal resolutions, and a small variation of the verification area has minimal influence on the CNOP sensitive areas. The CNOP sensitive areas identified for special forecast times when the initial time is fixed resemble those identified for other forecast times in the linear case, while the similarities among the sensitive areas identified for different forecast times are more limited in the nonlinear case. When the forecast time is fixed, the CNOP sensitive areas are much different when they are identified at different time period ahead.

F. Zhou (✉)
Laboratory of Cloud-Precipitation Physics and Severe Storms (LACS), Institute of Atmospheric Physics, Chinese Academy of Sciences, P.O. Box 9804, Beijing, 100029, China
e-mail: zhouff04@163.com; zff@mail.iap.ac.cn

X. Qin
State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100029, China

B. Chen
National Meteorological Center of China Meteorological Administration, Beijing 100081, China

M. Mu
Key Laboratory of Ocean Circulation and Wave, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071, China

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_24,

Then the influence of the initial conditions in the sensitive areas on the targeted forecasts have been examined, and the observing system simulation experiments (OSSEs) have been performed to assess whether or not the sensitive areas can be considered as dropping sites in real time targeting. Also, the observation system experiments (OSEs) have been carried out to demonstrate the utility of the CNOP method. It is found that the impact of initial errors introduced into the CNOP sensitive areas on the forecasts is greater than that of errors fixed in the SV sensitive areas or other randomly selected areas. The OSSEs have shown that assimilating the ideal observations in the CNOP sensitive areas results in the improvements of 13–46 % in typhoon track forecasts, while the improvements of 14–25 % are obtained by assimilating the ideal observations in the SV sensitive areas. Besides, the improvements have been achieved for longer forecast times. The OSEs have shown that the DOTSTAR data in the CNOP sensitive areas has a more positive impact on the typhoon track forecast than that in the SV sensitive areas.

All the above results have demonstrated that the CNOP is a useful tool in the adaptive observations to identify the sensitive areas.

24.1 Introduction

Based on predictability studies of tropical cyclones, it is realized that the forecasts of tropical cyclone tracks and intensity could be improved when accurate initial analyses are obtained (Riehl et al. 1956; Bender et al. 1993; Zhu and Thorpe 2006; Froude et al. 2007). Consequently, it is important to supplement observations in data-sparse areas to obtain an accurate initial analysis. However, placing additional observation stations in the data-sparse areas is unnecessary; studies have shown that extensive observations obtained in the general region around the cyclone do not conclusively improve forecasts over observations obtained only in particular regions (Franklin and DeMaria 1992; Aberson 2003). Adaptive observations (also called targeted observations) are intended for this purpose: observational capabilities are intensified in areas where additional observations are expected to improve a forecast largely. These areas are considered "sensitive", in the sense that changes to the initial conditions in these areas are expected to have a larger impact on the forecast than changes in other areas. It is "adaptive" in the sense that the sensitive areas may change from day to day and case to case (Bergot 1999).

Currently, there are several strategies used for identifying the sensitive areas. One strategy is based on the adjoint technique, such as singular vectors (SVs, Palmer et al. 1998), adjoint sensitivities (Ancell and Mass 2006), and the adjoint-derived sensitivity steering vector (ADSSV) (Wu et al. 2007). Another is ensemble-based, for example, the ensemble transform (Bishop and Toth 1999), the ensemble Kalman filter (EnKF, Hamill and Snyder 2002) and the ensemble transform Kalman filter (ETKF, Bishop et al. 2001). The strategies mentioned have been tested in field experiments such as the Fronts and Atlantic Storm-Track Experiment (FASTEX; Snyder 1996; Joly et al. 1997), the North Pacific Experiment (NORPEX; Langland et al. 1999a), the Winter Storm Reconnaissance Programs (WSR; Szunyogh et al. 2000, 2002), the Dropwindsonde Observations for Typhoon Surveillance near the Taiwan region (DOTSTAR; Wu et al. 2005), the Atlantic THORPEX Regional Campaign (ATReC; Rabier et al. 2008), etc. Forecasts are generally improved by assimilation of targeted observations (Gelaro et al. 1999; Langland 2005; Wu et al. 2005; Buizza et al. 2007; Rabier et al. 2008).

The strategies mentioned above are generally linear methods. They are constrained by linear approximations. To study the effect of nonlinearity, Mu et al. (2003) proposed a novel approach of conditional nonlinear optimal perturbation (CNOP). The CNOP is an extension of the linear singular vector (SV) method in to the nonlinear regime, and it has been applied to some research fields such as El Ni∼no-Southern Oscillation (ENSO) predictability (Mu et al. 2007; Duan and Mu 2009; Duan and Luo 2010; Duan and Luo 2010; Peng et al. 2011), the nonlinear behavior of baroclinic unstable flows (Riviere et al. 2008), ensemble forecasting (Mu and Jiang 2008), and the transitions between multiple equilibria states of the ecosystem (Sun and Mu 2009).

Recently, Mu et al. (2009) suggested that the CNOP can be used to identify the sensitive areas for tropical cyclone predictions in targeted observations since the forecasts benefit more from reductions of CNOP-type initial errors than from reductions of SV-type initial errors. Then Zhou and Mu (2011, 2012a, b) used the CNOP to identify the sensitive areas, and studied the properties of the CNOP sensitive areas with respect to variations of the horizontal resolution, the verification area design and the optimization time period. Furthermore, Chen and Mu (2012) carried out sensitivity analysis by studying the impact of initial errors introduced into the CNOP sensitive areas on the forecasts. Moreover, Qin (2010a, b) and Qin and Mu (2011a, b) performed the observing system simulation experiments (OSSEs) to assess whether the sensitive areas identified by CNOP can be considered as dropping sites in realtime targeting. The observation system experiments (OSEs) using the DOTSTAR Data have also been carried out by Chen (2011) to demonstrate the utility of the CNOP method.

This chapter will summarize the above works about the approach of the CNOP to targeted observations for tropical cyclone predictions. The SV sensitive areas have also been studied correspondingly for comparison in some works. The structure of this paper is as follows. Section 24.2 provides an introduction to CNOP and SV, Sect. 24.3 introduces the properties of the CNOP sensitive areas. Section 24.4 describes the serviceability of the CNOP sensitive areas by OSSEs and OSEs. A brief summary and discussion are given in the final section.

24.2 Methodologies

24.2.1 CNOP

In this part, we briefly introduce the method of conditional nonlinear optimal perturbation. Suppose we have the following model

$$\begin{cases} \frac{\partial \mathbf{X}}{\partial t} + F(\mathbf{X}) = 0 \\ \mathbf{X}|_{t=0} = \mathbf{X}_0 \end{cases} \tag{24.1}$$

where $\mathbf{X}$ is the state vector of the model with initial value $\mathbf{X}_0$. F is a nonlinear partial differential operator. The solution of (24.1) can be expressed in discrete form:

$$\mathbf{X_t} = \boldsymbol{M}(\mathbf{X_0}) \tag{24.2}$$

where $\boldsymbol{M}$ is a nonlinear propagator, and $\mathbf{X_t}$ is the value of $\mathbf{X}$ at time t.

To measure the development of $\mathbf{X}_0$, appropriate norms must be chosen. In discrete form, this is equivalent to choosing symmetric positive definite matrices $\mathbf{C}_1$ and $\mathbf{C}_2$. An initial perturbation $\boldsymbol{\delta}\mathbf{X}_0^*$ of vector $\mathbf{X}_0$ is called CNOP if and only if

$$J(\boldsymbol{\delta}\mathbf{X}_0^*) = \max_{\boldsymbol{\delta}\mathbf{X_0^T}\mathrm{C}_1\boldsymbol{\delta}\mathbf{X_0} \le \beta} J(\boldsymbol{\delta}\mathbf{X_0}) \tag{24.3}$$

Where

$$J(\boldsymbol{\delta}\mathbf{X_0}) = [\mathbf{P}\boldsymbol{M}(\mathbf{X_0} + \boldsymbol{\delta}\mathbf{X_0}) - \mathbf{P}\boldsymbol{M}(\mathbf{X_0})]^{\mathbf{T}}\mathrm{C}_2[\mathbf{P}\boldsymbol{M}(\mathbf{X_0} + \boldsymbol{\delta}\mathbf{X_0}) - \mathbf{P}\boldsymbol{M}(\mathbf{X_0})] \tag{24.4}$$

and $\delta\mathbf{X_0^T}\mathrm{C}_1\delta\mathbf{X_0} \le \beta$ is a constraint condition of initial perturbations with the presumed positive constant β representing the magnitude of the initial uncertainty. The first guess of the initial perturbation $\delta\mathbf{X_0}$, which is usually taken as the difference between the model outputs at two times, should be adjusted to satisfy the constraint condition $\delta\mathbf{X_0^T}\mathrm{C}_1\delta\mathbf{X_0} \le \beta$. $\mathbf{P}$ is a local projection operator and takes value 1(0) within (without) the targeted region. The superscript "$^{\mathrm{T}}$" denotes the transpose of vectors or matrices. Note that the norms used in the cost function and the initial constraint condition may be the same, depending on the physical problem. It is clear that the CNOPs depend on the nonlinear model $\boldsymbol{M}$, the initial state vector $\boldsymbol{X}_0$, and the parameters β, $\boldsymbol{P}$, $\boldsymbol{C}_1$, and $\boldsymbol{C}_2$.

24.2.2 SV

Suppose that the initial perturbation $\delta\mathbf{X_0}$ is sufficiently small and the integration time interval is of moderate length, then the development of $\delta\mathbf{X_0}$ in discrete form can be approximated by

$$\boldsymbol{\delta}\mathbf{X}_t = \mathbf{L}(\boldsymbol{\delta}\mathbf{X}_\mathbf{0}) \tag{24.5}$$

where $\mathbf{L}$ is the forward tangent propagator. $\boldsymbol{\delta}\mathbf{X}_t$ is the linear development of $\delta\mathbf{X}_\mathbf{0}$ at time t. According to Barkmeijer et al. (2003), the first singular value σ_1 of $\mathbf{L}$ satisfies (with respect to the norms $\mathbf{C}_1$ and $\mathbf{C}_2$):

$$\sigma_1^2 = \max_{(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})^\mathbf{T}\mathbf{C}_1(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})\neq 0} \frac{[\mathbf{L}(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})]^\mathbf{T}\mathbf{C}_2[\mathbf{L}(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})]}{(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})^\mathbf{T}\mathbf{C}_1(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})} \tag{24.6}$$

Additionally, if $\mathbf{v}_i$ is the singular vector of $\mathbf{L}$, then

$$(\mathbf{C}_1)^{-1}(\mathbf{L}^\mathbf{T}\mathbf{C}_2\mathbf{L})\mathbf{v}_i = \sigma_i^2\mathbf{v}_i \tag{24.7}$$

where superscript "−1" denotes the inverse of the matrix. σ_i is the singular value corresponding to $\mathbf{v}_i$. The first SV (FSV) maximizes the linear development of the initial perturbations, the second SV maximizes the development under the constraint of being orthogonal to the first SV, and the third SV maximizes the development under the constraint of being orthogonal to the first two SVs, and so on (Peng and Reynolds 2006). A local projection operator P (same meaning as in (24.4)) is employed to localize the development of the perturbation in the verification region.

Actually, the FSV can be obtained by solving the following linear optimization problem (Ehrendorfer and Errico 1995):

$$J(\boldsymbol{\delta}\mathbf{X}_0^*) = \max_{\boldsymbol{\delta}\mathbf{X}_\mathbf{0}^\mathbf{T}\mathbf{C}_1\boldsymbol{\delta}\mathbf{X}_\mathbf{0}\leq\beta} J(\boldsymbol{\delta}\mathbf{X}_\mathbf{0}) \tag{24.8}$$

where

$$J(\boldsymbol{\delta}\mathbf{X}_\mathbf{0}) = [\mathbf{PL}(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})]^\mathbf{T}\mathbf{C}_2[\mathbf{PL}(\boldsymbol{\delta}\mathbf{X}_\mathbf{0})] \tag{24.9}$$

According to the linear characteristics of SV, the FSV defined by (24.8) and (24.9) equals to the FSV defined by (24.7) when it has been normalized to a unit.

24.2.3 The Model and Optimization Algorithm

The CNOP and SVs are calculated with fifth generation Pennsylvania State University–National Center for Atmospheric Research (PSU-NCAR) Mesoscale Model (MM5; Dudhia 1993) and its corresponding adjoint system (Zou et al. 1997). The following physical parameterizations are used: dry convective adjustment, grid-resolved large-scale precipitation, the high-resolution PBL scheme, and the Kuo cumulus parameterization scheme. The simulations initialized by the National Centers for Environment Predictions (NCEP) FNL (Final) Operational Global Analysis (1° × 1°) interpolated into the MM5 grids serves as the basic states for most of the calculations of the CNOPs or the SVs. The Weather Research and

Forecasting (WRF) model have been used to assimilate the operational Dropsonde Observations for Typhoon Surveillance near the Taiwan Region (DOTSTAR) in the conduction of observation system experiments (OSEs). Except for clarification, the model horizontal resolution is 60 km and there are 11 vertical levels in the following studies.

In this chapter, when the FSV is considered, it will be obtained by solving the (24.8) and (24.9). Thus both CNOP and FSV can be obtained by using the same optimization algorithm to facilitate comparison, and the optimization algorithm employed is the spectral projected gradient 2 (SPG2) (Birgin et al. 2001), which calculates the least value of a function of several variables subject to box or ball constraints. The cost function implemented for the calculation is $J_1(\delta X_0) = -J(\delta X_0)$, with the same initial constraint condition $\delta X_0^{\mathrm{T}} C_1 \delta X_0 \leq \beta$. The gradient of the cost function with respect to the initial perturbation is required for the SPG2 algorithm, and the adjoint model of MM5 is used to efficiently calculate the gradient. When other SVs have been considered, the SVs would be obtained by using the Lanczos algorithm (Ehrendorfer and Errico 1995).

For simplicity we choose $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, and $\mathbf{C}$ represents the metric of total dry energy, in a continuous expression:

$$(\boldsymbol{\delta}\mathbf{X_0})^{\mathbf{T}}\mathbf{C}(\boldsymbol{\delta}\mathbf{X_0}) = \frac{1}{D}\int_D \int_0^1 \left[\mathbf{u'^2} + \mathbf{v'^2} + \frac{c_p}{T_r}\mathbf{T'^2} + R_a T_r \left(\frac{\mathbf{p_s}'}{p_r}\right)^2\right] dz dD \quad (24.10)$$

where c_p and R_a are the specific heat at constant pressure and the gas constant of dry air respectively (with numerical values of 1005.7 J kg^{-1} K^{-1} and 287.04 J kg^{-1} K^{-1}). The reference parameters are the following: $T_r = 270$ K, $p_r = 1{,}000$ hPa. $\mathbf{u}', \mathbf{v}', \mathbf{T}', \mathbf{p_s}'$, which are components of the state vector, are the perturbed zonal and meridional wind components, temperature, and surface pressure respectively. The integration extends over the full domain D and the vertical direction z.

24.3 Properties of the CNOP Sensitive Areas

24.3.1 Definition of Sensitive Areas and Calculation of Similarity

For CNOP and FSV, the sensitive area is defined as the horizontal grid points where the function $f(i,j)$ exceeds a specified threshold value c. The function $f(i,j)$ is a vertically-integrated total dry energy function

$$f(i,j) = \int_0^1 E_{\mathrm{d}}(i,j,z)dz \quad (24.11)$$

where $E_{\mathrm{d}}(i, j, z)$ is the total dry energy (the sum of kinetic energy and available potential energy) of the CNOP or FSV at the grid point (i, j, z). The value c is specified thus the sensitive areas have a proper size.

For the composite of the SVs (short for CSV or SVs), the sensitive areas is defined as the horizontal grid points where the function $f_1(i, j)$ exceeds the threshold value c. The function $f_1(i, j)$ is expressed as follows

$$f_1(i, j) = \sum_{n=1}^{5} \frac{\sigma_n^2}{\sigma_1^2} \int_0^1 E_{\mathrm{d,n}}(i, j, z) dz \tag{24.12}$$

where $E_{\mathrm{d,n}}(i, j, z)$ is the total dry energy of the n th SV at the grid point (i, j, z). σ_n is the singular value of the n th SV. In this study, the five leading SVs have been composed.

The similarity between two vectors $\boldsymbol{X} = \{x_1, x_2, \cdots x_m\}^{\mathrm{T}}$ and $\boldsymbol{Y} = \{y_1, y_2, \cdots y_m\}^{\mathrm{T}}$ is calculated according to the following formula:

$$S_{\mathbf{xy}} = \frac{< \boldsymbol{X}, \boldsymbol{Y} >}{\sqrt{< \boldsymbol{X}, \boldsymbol{X} >}\sqrt{< \boldsymbol{Y}, \boldsymbol{Y} >}} = \frac{\sum\limits_{i=1}^{m} x_i y_i}{\sqrt{\sum\limits_{i=1}^{m} x_i^2}\sqrt{\sum\limits_{i=1}^{m} y_i^2}}. \tag{24.13}$$

24.3.2 Sensitivity of CNOP Sensitive Areas with Respect to Horizontal Resolution

From Sect. 24.2.1, it is known that the CNOPs depend on the nonlinear model $\boldsymbol{M}$, so different models due to different resolution may result in different CNOP sensitive areas. Zhou and Mu (2012a) studied this issue. In their study, a set of experiments are designed in which all the parameters are held constant except for the horizontal resolution.

Three tropical cyclones, TC Matsa (2005), TC Meari (2004), and TC Mindulle (2004), are investigated. A set of 24-h control forecasts, which served as the basic state, are integrated from 0000 UTC 5 Aug 2005 to 0000 UTC 6 Aug 2005 (TC Matsa), from 0000 UTC 26 Sep 2004 to 0000 UTC 27 Sep 2004 (TC Meari), and from 0000 UTC 28 Jun 2004 to 0000 UTC 29 Jun 2004 (TC Mindulle). For each case, the forecasts are run at 120, 60, and 30-km horizontal resolutions with 11 vertical levels. For TC Matsa, the model domain covers $28 \times 28, 55 \times 55, 109 \times 109$ (y-direction by x-direction) grids, respectively, for 120, 60, and 30-km horizontal resolutions. For TC Meari, there are $26 \times 28, 51 \times 55$, and 101×109 grids for each horizontal resolution, and for TC Mindulle the domain sizes with respect to each resolution are $21 \times 26, 41 \times 51$, and 81×101. For each case with the chosen grids, the real physical domain is the same at all resolutions, thus the verification area can be chosen the same.

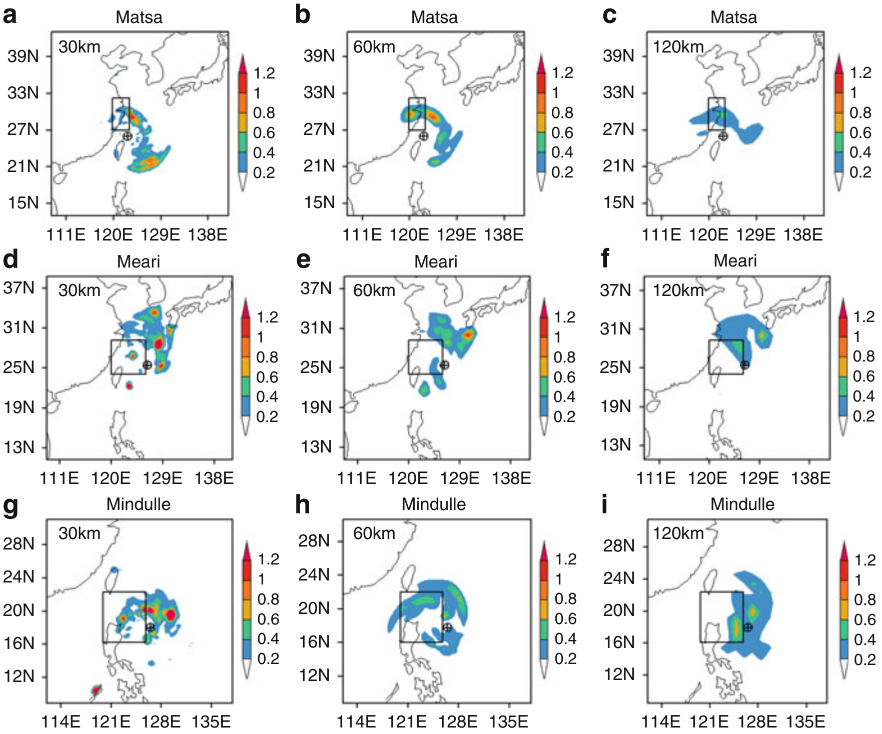


Fig. 24.1 The vertically-integrated energies of CNOP: (**a**-**c**) TC Matsa; (**d**-**f**) TC Meari; (**g**-**i**) TC Mindulle. (**a**, **d**, **g**) at a resolution of 30 km, (**b**, **e**, **h**) at a resolution of 60 km, and (**c**, **f**, **i**) at a resolution of 120 km. The *boxes* indicate the verification areas. The "⊕" symbol indicates the initial position of the cyclone (From Zhou and Mu 2012a)

For each case, the sensitive areas identified using different resolutions are different from each other (Fig. 24.1); however, common sensitive areas occur at the three resolutions, and the sizes of the common areas are different from case to case. In general, the sizes of common areas are bigger between sensitive areas at the lower resolution. This can be deduced from the similarities of the energy distributions between each resolution for the three cases (Table 24.1). Tor the three cases, the similarities between the lower resolutions (60 and 120 km) are greater than those between the finer resolutions (30 and 60 km); moreover, this illustrates that more small-scale activity would be resolved at higher resolutions.

From the analysis of the similarities, it can be induced that the sensitive areas identified at lower resolutions are also helpful for improving the forecast at finer resolution. However, to get the largest improvement at a high resolution, it is better to use the sensitive areas identified at the same resolution. A resolution at which the nonlinearity could be explored has been suggested to be used in the identification of the sensitive areas. Generally speaking, because usually the forecasts at higher resolutions are better than those at lower resolutions, and the high resolutions are

Table 24.1 The similarities among the energy distributions obtained at 30, 60, and 120-km resolutions for TC Matsa (2005), TC Meari (2004), and TC Mindulle (2004)

	30 and 60 km	60 and 120 km
TC Matsa	0.70	0.78
TC Meari	0.55	0.75
TC Mindulle	0.49	0.72

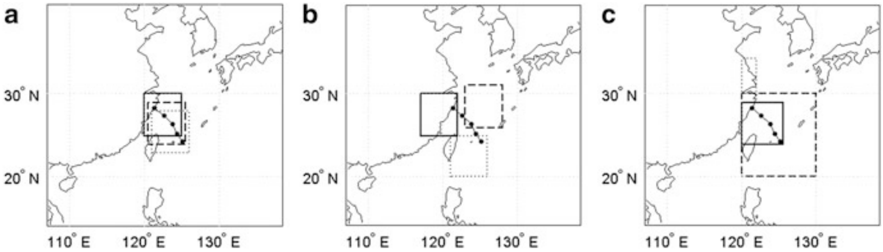


Fig. 24.2 Verification areas for different designs. The *solid* rectangles in panel (**a**), (**b**), and (**c**) are the verification areas for schemes A, D, and G respectively, while the *dashed* rectangles are for schemes B, E, and H respectively, the *dotted* rectangles are for C, F, and I respectively. The observation tracks of the cyclone are also shown in the center of the domain (From Zhou and Mu 2011)

favorable for CNOP to display the nonlinear information, which play an important role in the evolution of the initial perturbations, thus, as far as the computation condition is permitted, using the CNOP method at a high resolution to identify the sensitive areas may be more beneficial in targeted observations for tropical cyclone predictions.

24.3.3 Sensitivity of CNOP Sensitive Areas with Respect to the Verification Area Design

As indicated in Sect. 24.2.1, the CNOPs also depend on the verification area design (namely, different parameter $\boldsymbol{P}$), and this has been studied in the paper of Zhou and Mu (2011). The tropical cyclone Rananim, which occurred in the northwest Pacific Ocean in 2004, is studied.

The design of the verification area is as follows. First, they defined a control design in which the verification area includes the real cyclone tracks (the best storm tracks) during the integral time period: scheme B (Fig. 24.2a, dashed rectangle). Second, the size of the verification area is kept constant as it is moved to other places (schemes A, C, D, E, and F). Small positional variations are denoted as schemes A and C (Fig. 24.2a, solid and dotted rectangles, respectively). Large position variations are denoted as schemes D, E, and F (Fig. 24.2b, solid, dashed, and dotted rectangles, respectively). Then both the size and the position variations

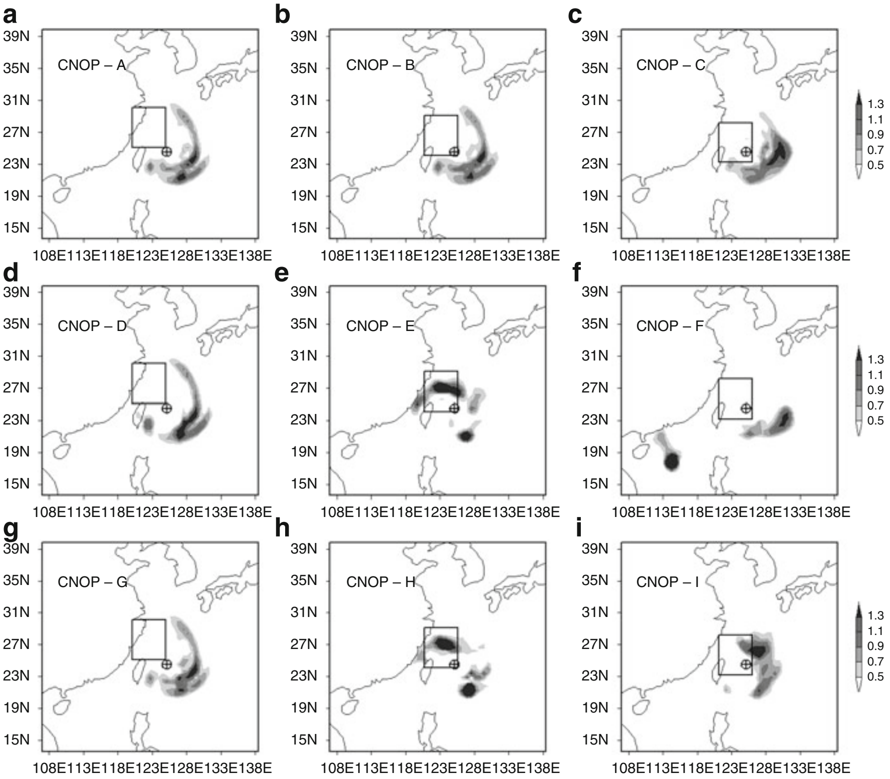


Fig. 24.3 The sensitive areas denoted by the vertically-integrated energies of CNOPs with schemes (**a**) A, (**b**) B, (**c**) C, (**d**) D, (**e**) E, (**f**) F, (**g**) G, (**h**) H, and (**i**) I respectively. The rectangles are the verification areas. The "⊕" symbol indicates the initial best position of the cyclone (From Zhou and Mu 2011)

of the verification area are considered; the designs are shown in Fig. 24.2c. Scheme G is designed with small variations (Fig. 24.2c, solid rectangle). Schemes H and I have even larger variations: Scheme H has a bigger domain, and scheme I has a smaller domain (Fig. 24.2c, dashed rectangle and dotted rectangle, respectively).

Generally, different verification area designs may result in different sensitive areas (Fig. 24.3). From the comparisons of schemes A, B, and C, it is seen that a small position change of the verification area has minimal influence on the sensitive areas. In addition, the inclusion of the best final position of the cyclone seems more important because its exclusion would result in very different sensitive areas. From the comparisons of schemes B, D, E, and F, it is found that the CNOP is sensitive to the large position changes of the verification area, which results in large differences among the identified sensitive areas. The comparisons of schemes G and B shows that the small variations in both size and position also affect the CNOP sensitive areas little, but a large variation in size or position would result in much different CNOP sensitive areas (comparing the schemes B, H, and I). Besides, although the verification area for scheme H includes almost all of the verification areas for schemes A, B, C, E, F, and G, the result of scheme H is more similar to that of scheme E than to those of schemes A, B, C, F, and G. This indicates that the areas north of the initial cyclone (scheme E) have a significant influence on the results. Therefore, verifications area should not be too large, or the results are affected by some irrelevant areas.

In general, the design of the verification area is important in tropical cyclone targeted observations. To improve a tropical cyclone forecast with targeted observations, the verification area must include the tropical cyclone tracks during the relevant time period. To do this, the ensemble forecast results must be consulted, which can introduce uncertainty into the prediction. In addition, the background field, the potential sensitive areas, as well as the economically relevant area can also provide references for the design of the verification area. The verification area cannot be too large or small, as the results would be affected for the former and the important information would be missed for the latter, also it is hard to catch the best positions of the cyclone for the latter. However, when the CNOP method is used to identify the sensitive areas, once the general position of the verification area is determined, a small variation in its size or position has minimal influence on the identification of the sensitive areas. This is a favorable characteristic for targeted observations.

24.3.4 Sensitivity of CNOP Sensitive Areas with Respect to the Optimization Period

The optimization period is a key issue in the choice of the cases. It is sure that the similarities between sensitive areas are rare for cases that occur in completely different situations; therefore, considerable attention has been paid to similarities between sensitive areas during cases occurring under similar conditions or during the temporal evolution of individual cases. The study of Zhou and Mu (2012b) focused on the latter. They have studied the following two tropical cyclones: Matsa (2005) and Meari (2004).

A set of experiments has been designed to study the time dependence of the sensitive areas in the context of targeted observations. Except for the studied time period, all parameters are held constant throughout the set of experiments. Two approaches are used. In the first approach, the initial time is fixed and forecasts are generated for 12, 24, and 36 h later. The initial times are set at 1200 UTC 4 Aug 2005 for the Matsa case (Fig. 24.4a) and 1200 UTC 25 Sep 2004 for the Meari case. In the second approach, the forecast time is fixed and forecasts are generated from initial times 12, 24, and 36 h prior to the forecast time. The forecast times are set at 0000 UTC 6 Aug 2005 for the Matsa case and 0000 UTC 27 Sep 2004 for the Meari case. The three initial times are therefore 1200 UTC 5 Aug 2005, 0000 UTC 5 Aug 2005, and 1200 UTC 4 Aug 2005 for the Matsa case (Fig. 24.4b), and 1200 UTC 26 Sep 2004, 0000 UTC 26 Sep 2004, and 1200 UTC 25 Sep 2004 for the Meari case. The optimisation time periods are the same as the forecast time periods in this study.

First, the nonlinearity of the typhoon cases is explored by comparing the linear FSVs and nonlinear CNOPs. Results suggest that for both two approaches, the

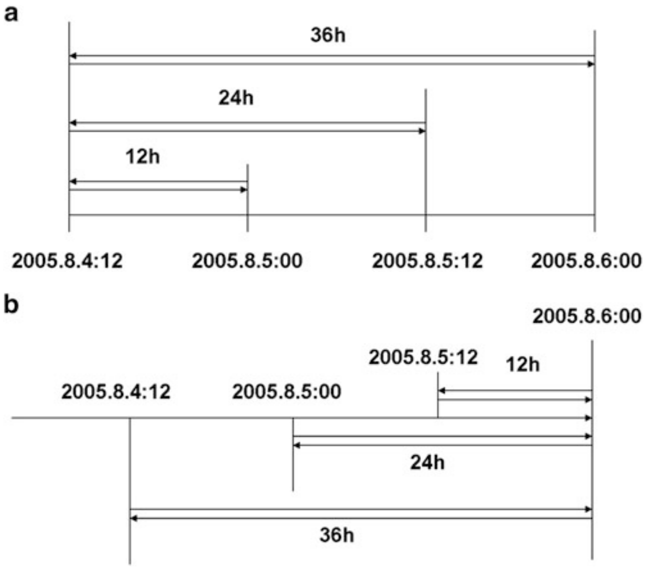


Fig. 24.4 The design of the optimization time periods for TC Matsa (2005). (**a**) for the first approach and (**b**) for the second approach (From Zhou and Mu 2012b)

nonlinearity in Matsa case is strong, especially at longer forecast integrations, since the CNOPs become progressively more different from the FSVs as the forecast time extends further from the initial time (Fig. 24.5, for the first approach, figure for the second approach not shown). While the Meari case is weak nonlinearity regardless of the optimisation time period as the patterns of the CNOPs and FSVs are similar (Fig. 24.6, for the first approach, figure for the second approach not shown). So the Meari case will be interchangeably called the "linear case" and the Matsa case the "nonlinear case" for the remainder of this part.

For the first approach, the comparison of the sensitive areas identified for the Matsa and Meari cases reveals several interesting features. In the linear case, the sensitive areas identified for a special forecast time are consistent with those identified for other forecast times when the initial time is fixed (Fig. 24.7). This result means that targeted observations deployed to improve a special time forecast would also favourably affect the forecasts at other times. In the nonlinear case, however, although there are some similarities in the sensitive areas identified for different forecast times, these similarities are limited (Fig. 24.8). This indicates that although the targeted observations deployed for a special time forecast are also beneficial for other times' forecasts, the forecast improvements for other times are limited. So it is suggested that in the nonlinear case, the deployment of targeted observations should be adaptive to obtain the largest improvement for different targeted forecasts, and should be more widespread to achieve the greatest improvement in multiple time forecasts. In addition, for both two cases, the closer the forecast times, the higher the similarities of the sensitive areas.

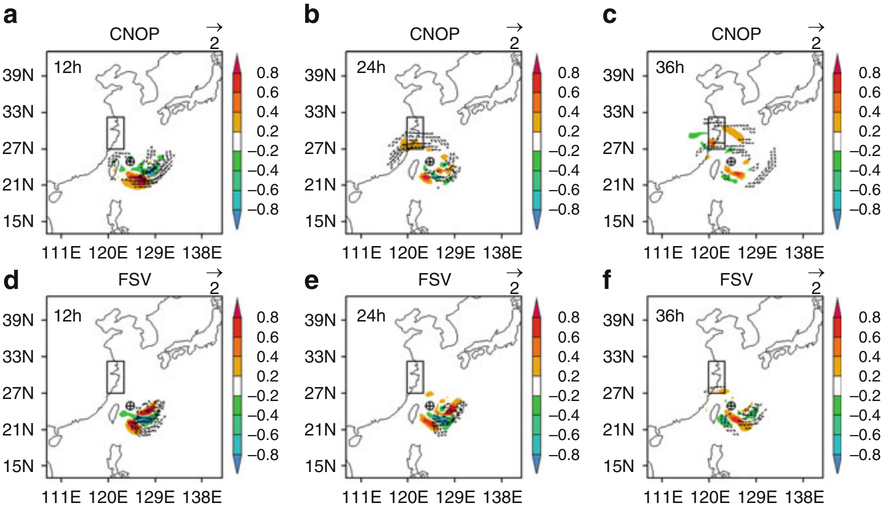


Fig. 24.5 TC Matsa (2005). The temperature (*shaded*, units: K) and wind (*vector*, units: m s^{-1}) components of (**a**), (**b**), (**c**) CNOP and (**d**), (**e**), (**f**) FSV at $\sigma = 0.7$. The *boxes* indicate the verification areas. The ⊕ indicates the position of the cyclone at 1200 UTC 04 Aug 2005. The forecasts are generated for (**a**), (**d**) 12 h, (**b**), (**e**) 24 h, and (**c**), (**f**) 36 h later (From Zhou and Mu 2012b)

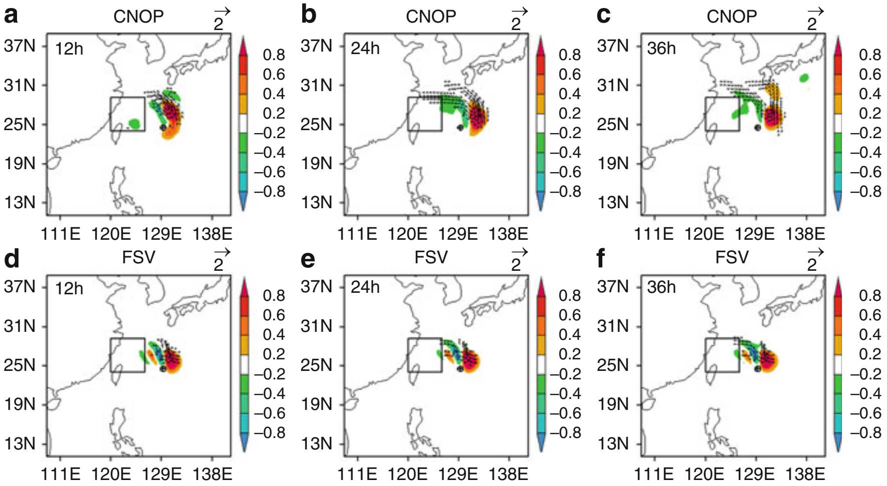


Fig. 24.6 TC Meari (2004). The temperature (*shaded*, units: K) and wind (*vector*, units: m s^{-1}) components of (**a**), (**b**), (**c**) CNOP and (**d**), (**e**), (**f**) FSV at $\sigma = 0.7$. The *boxes* indicate the verification areas. The ⊕ indicates the position of the cyclone at 1200 UTC 25 Sep 2004. The forecasts are generated for (**a**), (**d**) 12 h, (**b**), (**e**) 24 h, and (**c**), (**f**) 36 h later (From Zhou and Mu 2012b)

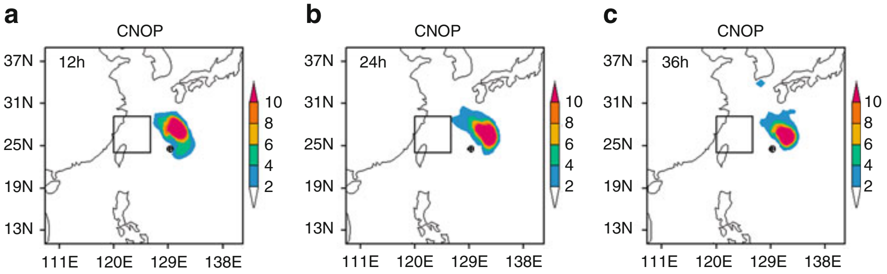


Fig. 24.7 TC Meari (2004). The vertically integrated energies of CNOP (*shaded*, units: J/kg) for the first approach. (**a**) for 12 h forecast, (**b**) for 24 h forecast, and (**c**) for 36 h forecast (From Zhou and Mu 2012b)

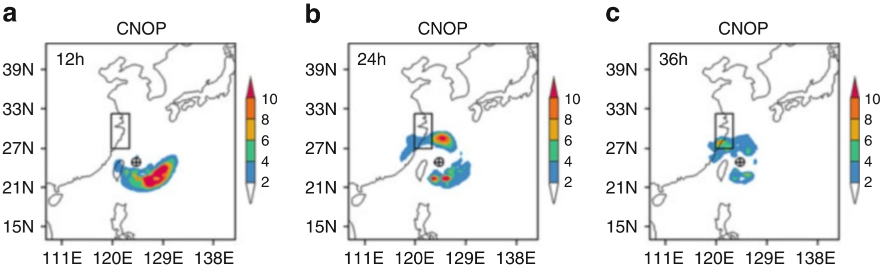


Fig. 24.8 Same as Fig. 24.7, but for TC Matsa (2005) (From Zhou and Mu 2012b)

For the second approach, the sensitive areas of the linear case move to the verification areas as the initial time is shifted closer to the forecast time (i.e., as the optimisation period is shortened; Fig. 24.9). This result is consistent with the results of previous studies that applied linear methods to cases that permitted linear approximation (Palmer et al. 1998; Kim et al. 2004; Wu et al. 2007). In such case, the background field such as the subtropical high plays an important part in the corresponding targeted forecasts. In the nonlinear case, the sensitive areas fall in disrupted-ring patterns around the initial typhoon centres, and are mainly located inside the typhoon circulation (Fig. 24.10). This indicates that the targeted forecasts in this case are affected primarily by conditions within the typhoon, while the background fields play a relatively smaller role. The results of these two cases suggest that the deployment of targeted observations intended to improve the forecast at a special time may depend strongly on the time of deployment. The time at which the targeted observations are deployed is thus of crucial importance.

Generally, the results of this study have shown that the deployment of targeted observations to improve a special forecast depends strongly on the time of deployment and it should be adaptive to achieve large improvements for different targeted forecasts.

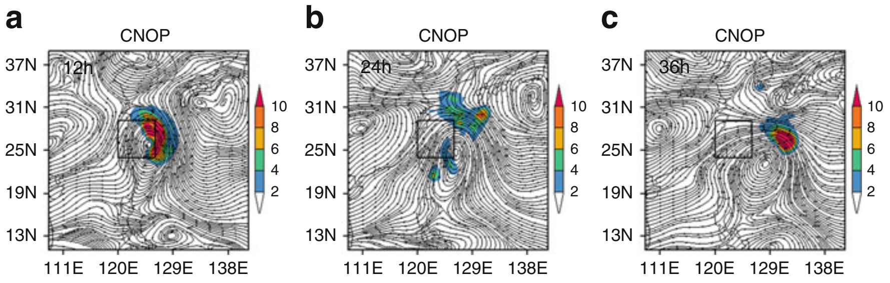


Fig. 24.9 TC Meari (2004). The vertically integrated energies of CNOP (*shaded*, units: J/kg) and the 500 hPa stream fields for the second approach. (**a**) forecast starts at 12 h ahead, (b) forecast starts at 24 h ahead, and (**c**) forecast starts at 36 h ahead (From Zhou and Mu 2012b)

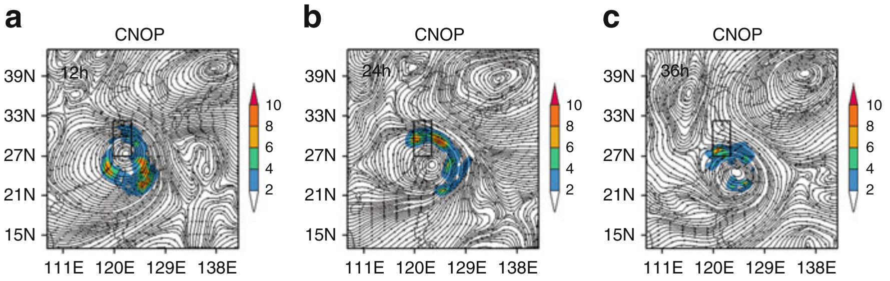


Fig. 24.10 Same as Fig. 24.9, but for TC Matsa (2005) (From Zhou and Mu 2012b)

24.4 Examination of the CNOP Sensitive Areas

In Sect. 24.3, we have studied the properties of the CNOP sensitive areas. In this section, we will examine the efficiency of the CNOP sensitive areas with a lot of cases by carrying out the sensitivity tests, OSSEs, and OSEs.

24.4.1 Sensitivity Tests

In this part, random initial errors are introduced into the CNOP sensitive areas, and their impacts on the TC forecasts are explored (Chen and Mu 2012). Two tropical cyclones, Longwang (2005) and Sinlaku (2008) are studied. For comparison, the roles of the random initial errors in the areas identified by FSV and CSV, and the randomly selected areas are considered. Based on these four areas, experiments are designed to determine which area have the greatest impact on TC forecasts.

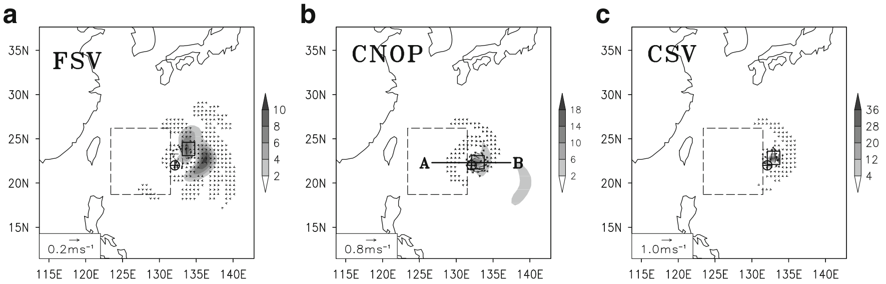


Fig. 24.11 TC Longwang (2005). VIE (*shaded*; 10 J kg^{-1}) and wind (*vector*; m s^{-1}) component of (**a**) FSV, (**b**) CNOP, (**c**) CSV over a 24-h optimization time interval initialized at 0000 UTC 30 Sep 2005. The *big boxes* (*dashed*) indicate the verification areas; the *box* (*solid*) indicates the sensitive area determined by the (**a**) FSV, (**b**) CNOP, (**c**) CSV; the symbol ⊕ indicates the initial cyclone center (From Chen and Mu 2012)

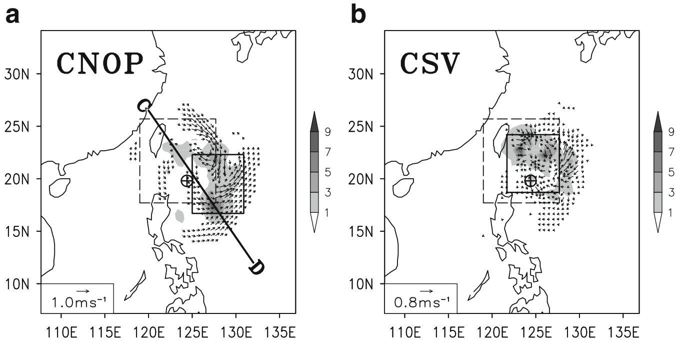


Fig. 24.12 TC Sinlaku (2008). VIE (*shaded*, 10 J kg^{-1}) and wind (vector; m s^{-1}) component of (**a**) CNOP, (**b**) CSV over a 24-h optimization time interval initialized at 0000 UTC 10 Sep 2008. The *big boxes* (*dashed*) indicate the verification areas; the *box* (*solid*) indicates the sensitive area determined by the (**a**) CNOP, (**b**) CSV; the symbol ⊕ indicates the initial cyclone center (From Chen and Mu 2012)

For simplicity, the CNOP-, CSV-, FSV-, sensitive areas and randomly selected areas are marked by CNOP_Sen, CSV_Sen, FSV_Sen, and Ran_Area, respectively. The random initial errors (Ran_Err) are added to the four areas. Notably, the random initial errors have the same size measured by the dry energy norm. The four experiments noted by RA-CN, RA-CS, RA-FS, and RA-RA, respectively represente that the initial errors are introduced into the CNOP_Sen, CSV_Sen, FSV_Sen, and Ran_Area.

First, the distributions of CNOP_Sen, CSV_Sen, and FSV_Sen are checked. Here, the maximum of the vertically integrated energy (VIE) of CNOP, CSV, and FSV are defined as the centers of CNOP_Sen, CSV_Sen, and FSV_Sen (indicated by the smaller squares in Figs. 24.11 and 24.12, differing from the above studies). It is seen that the considerable difference exists between the locations of FSV_Sen and CSV_Sen (Fig. 24.11a, c). Although the VIE maxima of CNOP and CSV are both located in the core region of TC Longwang (2005) (Fig. 24.11b, c), but the accurate locations of the two maxima are slightly different: the location of CNOP VIE maximum is one grid spacing to the south of CSV VIE maximum. Thus, the CNOP_Sen and CSV_Sen are slightly different.

The CNOP_Sen for TC Sinlaku (2008) is defined as a minimized square area, large enough to contain the two VIE maxima of CNOP, and FSV_Sen and CSV_Sen are another square area with the same size, capable of containing the two VIE maxima of CSV and located on the north side of the storm center (Fig. 24.12). In addition, the extents of CNOP_Sen, CSV_Sen, and FSV_Sen for TC Longwang (2005) are defined as a square area with 4×4 grid points, and that for TC Sinlaku (2008) is defined as a square area with 12×12 grid points.

In the four types of experiments (RA-CN, RA-CS, RA-FS, and RA-RA), 40 Ran_Errs are generated by transforming forty $1 \times 1,104$ error vectors, which are normally distributed, into forty $4 \times 4 \times 23 \times 3$ error matrices for TC Longwang (2005) and by transforming 40 $1 \times 9,936$ error vectors, which are normally distributed, into forty $12 \times 12 \times 23 \times 3$ error matrices for TC Sinlaku (2008). Specifically, the u-wind, v-wind, and temperature initial states are perturbed, and the random error vectors are zero means. Their standard deviations are 0.95 for TC Longwang (2005) and 0.32 for TC Sinlaku (2008).

Figure 24.13 show the nonlinear developments of a Ran_Err from the CNOP_Sen, CSV_Sen, and Ran_Area, respectively, for both TCs. The development of the Ran_Err from FSV_Sen is similar to that from CSV_Sen for TC Longwang (2005) since CSV_Sen and FSV_Sen for TC Sinlaku (2008) are located in the same position. Obviously, as shown in Fig. 24.13e, f, a rapid reduction in the magnitude of wind and temperature of the evolved Ran_Errs from Ran_Area occur for both TC Longwang (2005) and TC Sinlaku (2008), compared with the results of RA-CN and RA-CS. For TC Longwang (2005), the verification dry energy norms of the evolved Ran_Errs from CNOP_Sen, CSV_Sen, and Ran_Area are $63.28\,\mathrm{J\,kg^{-1}}$, $17.47\,\mathrm{J\,kg^{-1}}$ and $12\,\mathrm{J\,kg^{-1}}$, respectively, while for TC Sinlaku (2008) those values are $21.13\,\mathrm{J\,kg^{-1}}$, $14.10\,\mathrm{J\,kg^{-1}}$, and $3.08\,\mathrm{J\,kg^{-1}}$. Therefore, the locations of initial errors in CNOP_Sen may have had great impacts on the final forecasts.

Next, the impacts of 40 Ran_Errs added to CNOP_Sen, CSV_Sen, FSV_Sen, and Ran_Area are assessed according to the statistical averages of verification dry energy norms of the 40 evolved Ran_Errs from the four types of areas, respectively. Table 24.2 presents the results of these statistical averages for both TC Longwang (2005) and TC Sinlaku (2008). Notably, the errors introduced into the CNOP_Sen have the largest influences on the final forecasts. For both cases, the Ran_Errs introduced into the CSV_Sen have the next largest influences, while the Ran_Errs fixed in Ran_Area lead to the smallest changes. Thus, the growth rates of Ran_Errs introduced into sensitive areas, such as CNOP_Sen and CSV_Sen, are higher than those introduced into Ran_Area.

Additionally, another 36 randomly selected areas are considered in the RA-RA experiment for TC Sinlaku (2008) and TC Longwang (2005). Similarly, the statistical averages of nonlinear developments of the 40 Ran_Errs from the 36 local areas are obtained following the same procedure performed in

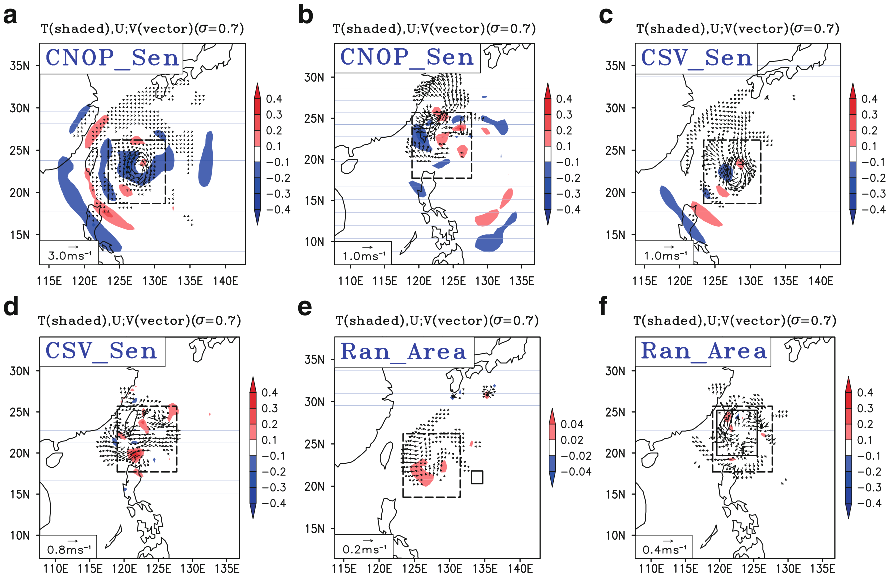

a
T(shaded),U;V(vector)(σ=0.7)
CNOP_Sen
3.0ms⁻¹
b
T(shaded),U;V(vector)(σ=0.7)
CNOP_Sen
1.0ms⁻¹
c
T(shaded),U;V(vector)(σ=0.7)
CSV_Sen
1.0ms⁻¹
d
T(shaded),U;V(vector)(σ=0.7)
CSV_Sen
0.8ms⁻¹
e
T(shaded),U;V(vector)(σ=0.7)
Ran_Area
0.2ms⁻¹
0.04
0.02
−0.02
−0.04
f
T(shaded),U;V(vector)(σ=0.7)
Ran_Area
0.4ms⁻¹
0.4
0.3
0.2
0.1
−0.1
−0.2
−0.3
−0.4
35N
30N
25N
20N
15N
10N
110E 115E 120E 125E 130E 135E 140E

Fig. 24.13 The Longwang (2005) (**a**, **c**, **e**) and Sinlaku (2008) (**b**, **d**, **f**) cases. Temperature (*shaded*; K) and wind (*vector*; m s^{-1}) components of the evolved Ran_Err from (**a** and **b**) CNOP_Sen, (**c** and **d**) CSV_Sen, and (**e** and **f**) Ran_Area on the level 0.7 at the final forecast time. The *boxes* (*dashed*) indicate the verification areas; the *boxes* (*solid*) for (**e**) and (**f**) indicate the Ran_Areas (From Chen and Mu 2012)

Table 24.2 TC Longwang (2005) and TC Sinlaku (2008). The statistical averages of the final verification dry energy norms of evolved Ran_Errs from CNOP_Sen, CSV_Sen, FSV_Sen and Ran_Area (J kg^{-1}), respectively (From Chen and Mu, Table 2)

Dry Energy (Verification Area)	CNOP_Sen	CSV_Sen	FSV_Sen	Ran_Area
Mean (40) (TC Longwang)	22.24	14.03	0.33	0.24
Mean (40) (TC Sinlaku)	27.03	19.28	19.28	7.35

the previous experiments. The geographical distribution of the 40 local areas, including CNOP_Sen, CSV_Sen, FSV_Sen, and 37 Ran_Areas, is shown in Fig. 24.14. Notably, the locations of these areas are defined around the initial storm center to the far outside, and the distance between the centers of two adjacent areas is defined as two-grid spacing for TC Sinlaku (2008) and as one-grid spacing for TC Longwang (2005). Table 24.3 shows the results of the extended experiments for both cases. Every unit of the tables (except units in the first row and column) corresponds approximately to its location in the geographical distribution of the 40 areas as shown in Fig. 24.14. For TC Longwang (2005), the statistical value

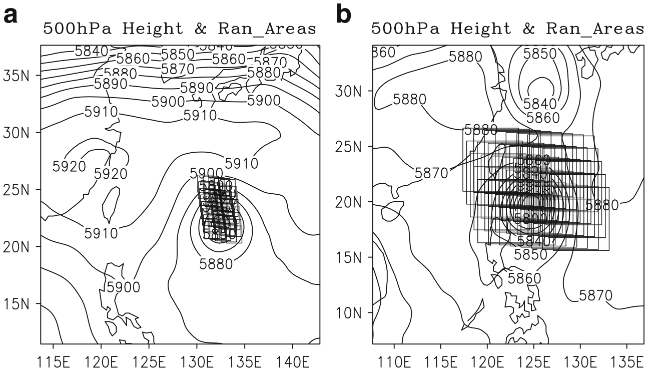

a
500hPa Height & Ran_Areas
35N
30N
25N
20N
15N
115E 120E 125E 130E 135E 140E
5840
5850
5860
5870
5880
5890
5900
5910
5920
b
500hPa Height & Ran_Areas
30N
25N
20N
15N
10N
110E 115E 120E 125E 130E 135E
5850
5840
5860
5870
5880

Fig. 24.14 Geopotential height on 500-hPa level (contour; gpm) and geographical distribution of the 40 local areas used in extended experiments (denoted by *black boxes*) for (**a**) TC Longwang (2005) and (**b**) TC Sinlaku (2008) (From Chen and Mu 2012)

Table 24.3 Statistical averages of the verification dry energy norms of the 40 evolved Ran_Errs from the 40 local areas, respectively, for (a) TC Longwang (2005) and (b) TC Sinlaku (2008) (J kg^{-1})

(a)

	J				
i	1	2	3	4	5
1	4.05	7.32	10.70	12.39	12.27
2	3.37	6.24	9.94	12.29	13.22^{a}
3	2.83	8.44	14.64	15.10	12.95
4	4.27	27.07	57.20^{b}	29.59	14.86
5	6.19	52.28	92.43^{c}	51.00	15.95
6	4.33	27.51	67.18	42.77	14.31
7	2.38	8.24	13.51	11.52	8.46
8	1.48	3.17	5.28	6.01	6.53

(b)

	J							
i	1	2	3	4	5	6	7	8
1	12.19	20.65	27.80	43.48	28.64	34.45	64.10	43.55
2	21.89	36.50	49.13^{b}	51.02	92.31	66.53	97.53	36.52
3	29.54	40.99	57.98	44.53	129.94	165.45	72.59	26.95
4	22.45	29.25	50.81	74.48	156.69	276.13^{c}	50.16	18.80
5	28.92	33.34	40.31	81.45	116.04	100.57	27.83	9.44

i represents the latitudinal direction, j indicates longitudinal direction

^{a}indicates FSV_Sen

^{b}indicates the CSV_Sen

^{c}indicates the CNOP_Sen

for CNOP_Sen is the largest, followed by those for the five Ran_Areas close to CNOP_Sen and CSV_Sen on the western and southern sides of the storm center, while the value for CSV_Sen is the seventh largest. In addition, the values for the areas on the outer ring are much lower than those for the areas near CNOP_Sen and CSV_Sen. For TC Sinlaku (2008), the value for CNOP_Sen is still the largest, and in general, the values for the areas distributed on the eastern side of the storm center are higher than those for the areas on the western side, including CSV_Sen/FSV_Sen.

These results suggested that, for a given initial error, basic state, and time interval, the growth rate of initial error strongly depended on where it is fixed, and in the CNOP sensitive areas the initial errors would grow rapidly.

24.4.2 Observing System Simulation Experiments (OSSEs)

To assess whether the sensitive regions calculated by CNOPs can be considered as dropping sites in realtime targeting, Qin (2010a) and Qin and Mu (2011a) performed observing system simulation experiments (OSSEs).

Generally, three basic components should be included in OSSEs (Hoffman et al. 1990): a four dimensional reference atmosphere, often called the nature run, the purpose of which is considered to be the 'truth'; a procedure to obtain simulated observations by sampling the nature run and adding errors; and a data assimilation system, which comprises a forecast model and the analysis procedure.

Qin (2010a) studied three typhoons Nock-Ten (2004), Matsa (2005), and Morakot (2009), and conducted nature experiments with the MM5 model but used a higher horizontal resolution of 30 km. The simulated observation data are produced by adding error to the nature run. There are two sets of data at the targeting time. Take Nock-Ten for example, one is the routine observation data that comprised the fixed stations on land and buoys in the ocean (Fig. 24.15a); the other is additional observation data obtained by dropped sondes, distributed in the CNOP sensitive areas (Fig. 24.15b) or randomly selected areas (Fig. 24.15c).

The effects (vertical mean of the variables) of utilizing additional data for Nock-Ten prediction are shown in Table 24.4. It is shown that the TPE (vertical integrated total perturbation energy) improvement produced by the CNOP is the most significant (12.74 %), while the effects caused by dropping sondes randomly have much less impact (some of them even have negative effects). This indicates that additional observations in random selected areas are useless for the forecast. By contrast, the sensitive regions identified by the CNOP prove to be better locations for additional observations. The other two cases have similar results.

In addition, Qin and Mu (2011a) performed OSSEs to evaluate the influence of additional dropsonde observation data in CNOP and SVs sensitive areas on typhoon track forecasts. In that study, the 'truth' is considered to be forecasts from 0 up to 72 h using the ERA-Interim reanalysis from ECMWF. The forecast typhoon centres at 6 h intervals are collated to represent the 'true' typhoon tracks. Forecasts during the same period, using the reanalysis from NCEP, are considered as the control run.

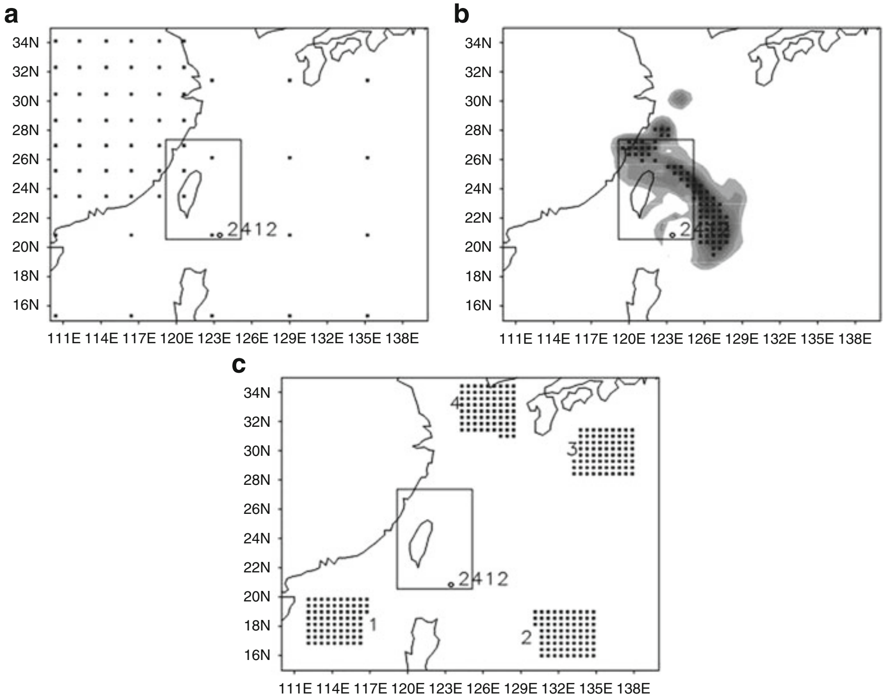


Fig. 24.15 TC Nock-Ten (2004). (**a**) the routine observations, (**b**) the CNOP sensitive areas and (**c**) the random selected areas denoted by number 1, 2, 3, 4 respectively (From Qin 2010)

Table 24.4 The RMSE of both routine and additional observation relative to only routine observation for Nock-Ten. The first column represents the kind of additional data that is utilized. The variables from the second to the eighth column stands for the vertical mean zonal wind, meridional wind, temperature, surface pressure, specific humidity, vertical wind, and vertical integrated total perturbation energy, respectively. Negative values represent that the RMSEs produced by the corresponding additional data were reduced

	U(%)	*V*(%)	*T*(%)	PP (%)	QV (%)	*W*(%)	TPE (%)
CNOP	**–2.58**	**–9.86**	**–4.01**	0.62	0.06	**–3.60**	**–12.74**
Random1	**–2.16**	**–0.96**	**–1.63**	**–0.10**	**–2.41**	**–5.23**	**–0.90**
Random2	**–0.28**	**–1.57**	0.87	**–4.78**	0.00	1.18	**–1.56**
Random3	1.60	3.93	**–0.54**	**–5.27**	0.56	**–0.59**	5.74
Random4	**–1.03**	1.85	2.62	**–4.02**	0.35	3.43	2.86

The difference between the typhoon centre positions of the control run and of the nature run is defined as the error in the typhoon track forecast without dropsonde data. After identifying the sensitive regions for the optimization period (24–48 h), simulated dropsondes are deployed at 24 h over these regions to obtain observational data, which represent the sum of the forecasts of the nature run at that time and

Table 24.5 Track forecast errors (km) without dropsondes from 24–72 h for all seven cases. The second row represents the initial time for each typhoon case (all in 2009, e.g., '082012' represents 1200UTC 20 August). The last row indicates the moving direction during this period

	Vamco	Mujigae	Koppu	Choi-Wan	Ketsana	Mirinae	Nida
Initial time (h)	082012	091000	091306	091700	092618	102918	112806
24	251.3	110	211.6	88.7	168.3	11	125.4
30	305.0	132.9	171.1	85.9	155.6	59.2	213.6
36	323.5	128.8	222.7	125.4	155.6	188.3	287.9
42	365.7	188.3	204.6	192.1	189.9	267.9	282.4
48	487.2	157.5	233.9	275.7	194.9	216.7	412.3
54	364.2	39.7	266.1	268.5	184.4	153.2	493.0
60	261.5	59.2	297.2	260.8	165.4	226.8	376.3
66	354.7	144.7	537.7	49.2	125.1		
72	199.2				101.4		
Moving direction	Northward	Westward	Northwestward	Recurved	Westward	Westward	Stagnant

randomly produced observation errors with the order of 10^{-1} of the analysis. The simulated additional dropsonde data included horizontal wind speed, horizontal wind direction, and temperature at 850, 500 and 200 hPa. The 3D-Var assimilation system of MM5 is used to assimilate the additional dropsonde data to produce an analysis at 24 h, which can be run to predict the locations of typhoon centres in the following 48 h (from 24 to 72 h). The differences between these typhoon centre positions and the nature run are defined as the typhoon track forecast errors with dropsondes. Difference between these errors and those without dropsondes are used to indicate the influence of CNOPs sensitive areas on typhoon track forecasts.

Seven typhoon events (with large track forecast errors, Table 24.5) originated in the western North Pacific during the 2009 season have been selected for analysis. It is found that CNOP sensitive areas forms (half) an annulus around the typhoon centres at targeting time for most of the typhoon cases (five of seven); and SV sensitive areas showed a maximum at the rear left quadrant with respect to the storm motion, approximately 500 km from the centre of the storms, also in five of seven cases (Fig. 24.16, typhoon Mirinae for example). Then dropping sites have been selected: the distance between adjacent sites is appropriate (about 150 km), and the total number of sites is the same in the CNOP and SV sensitive areas. See Fig. 24.16 for example. The track errors that with and without dropsondes have been compared. It is found that a varying degree of improvement in typhoon track forecasts for six of the seven cases, after assimilating simulated dropsonde data obtained for the sensitive regions (Fig. 24.17). Moreover, the improvements are not only obtained for the optimization period, for calculating CNOPs and SVs, but also for the subsequent 24 h. During the period 24–72 h, the deployment of dropsondes according to CNOPs sensitivity could reduce track forecast errors by 13–46 %, and by 14–25 % for SVs sensitivity.

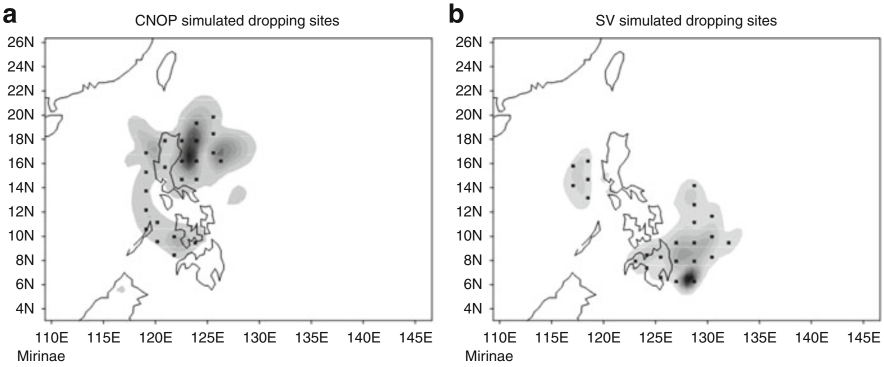


Fig. 24.16 Simulated dropping sites in sensitive regions calculated by (**a**) CNOPs and (**b**) SVs for TC Mirinae (2009). *Shaded* regions are the same as those in Fig. 24.15, *squares* represent the sites for dropping sondes (From Qin and Mu 2011a)

That is, the deployment of dropsondes in CNOP sensitive areas have an overall positive influence on typhoon track forecasts, suggesting in turn that CNOP can be utilized as an adaptive method in determining sensitive regions in adaptive observations.

24.4.3 Observation System Experiments (OSEs)

In the above section, we have demonstrated the utility of the CNOP sensitive areas in the OSSEs by using the ideal observations. In this section, we would use the real observations to show the utility of the CNOP sensitive areas (Chen 2011).

The dropsonde observations were collected under the operational Dropsonde Observations for Typhoon Surveillance near the Taiwan Region (DOTSTAR) program. Typhoon Nida occurred in 2004 has been studied. Fifteen dropwindsondes were released around Nida between 1000 and 1400 UTC 17 May 2004. The squares in Fig. 24.18 show the location of the released dropsondes. Most dropsondes were deployed every 150–200 km in a circular pattern with its center at Nida's central position. The observations include data on wind speed, wind direction, height, temperature, dewpoint temperature, and relative humidity below 196 hPa.

First, the CNOP sensitive areas and FSV sensitive areas (see Fig. 24.19) have been defined by using MM5 model. Then the 2nd, 3rd, 4th, and 5th dropsondes located near the maximal VIE area of CNOP have been chosen as the CNOP targeted observations. Similarly, the 10th $\sim$ 13th dropsondeslocated near the maximal VIE area of FSV have been chosen as the FSV targeted observations. Additionally, the 8th $\sim$ 11th dropsondes at the south side of the initial storm center are used as randomly targeted observations.

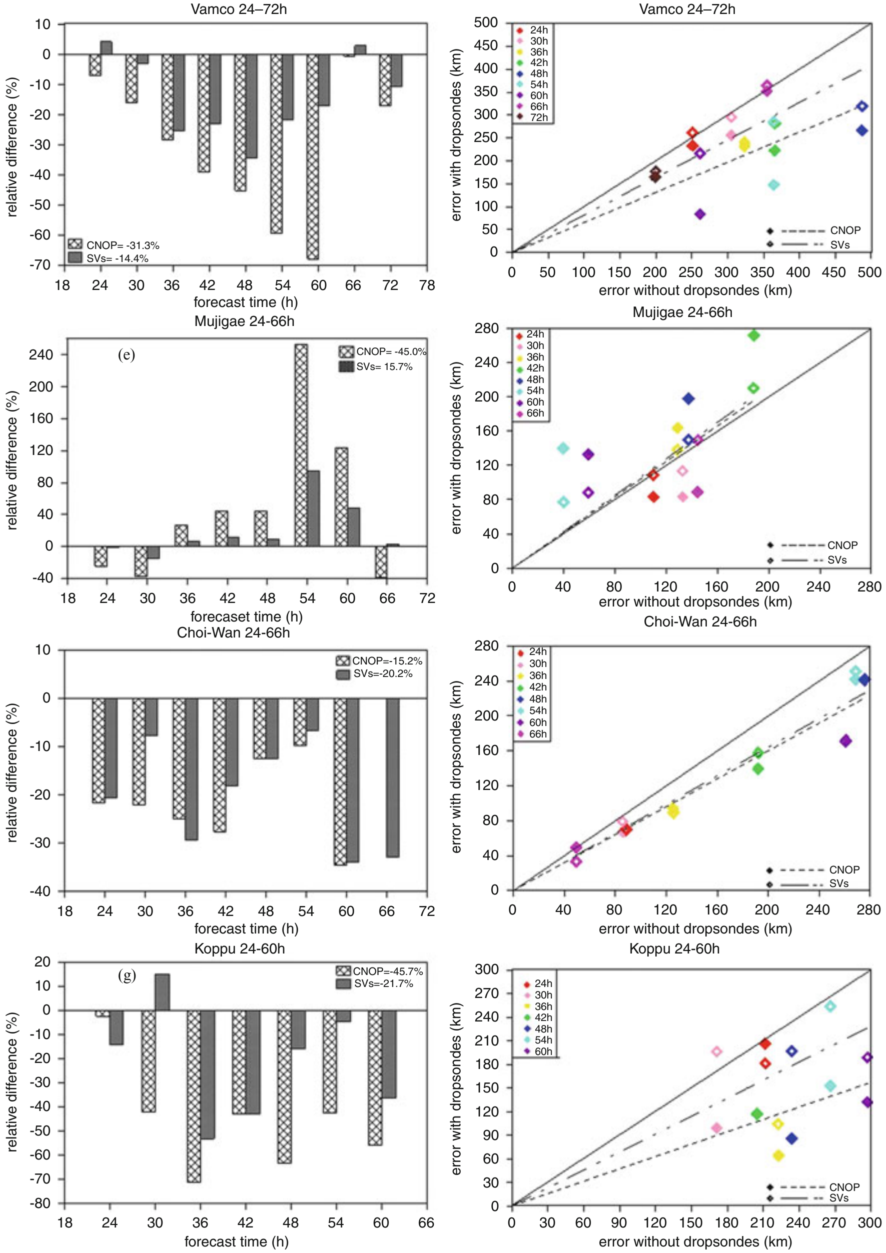


Fig. 24.17 (continued)

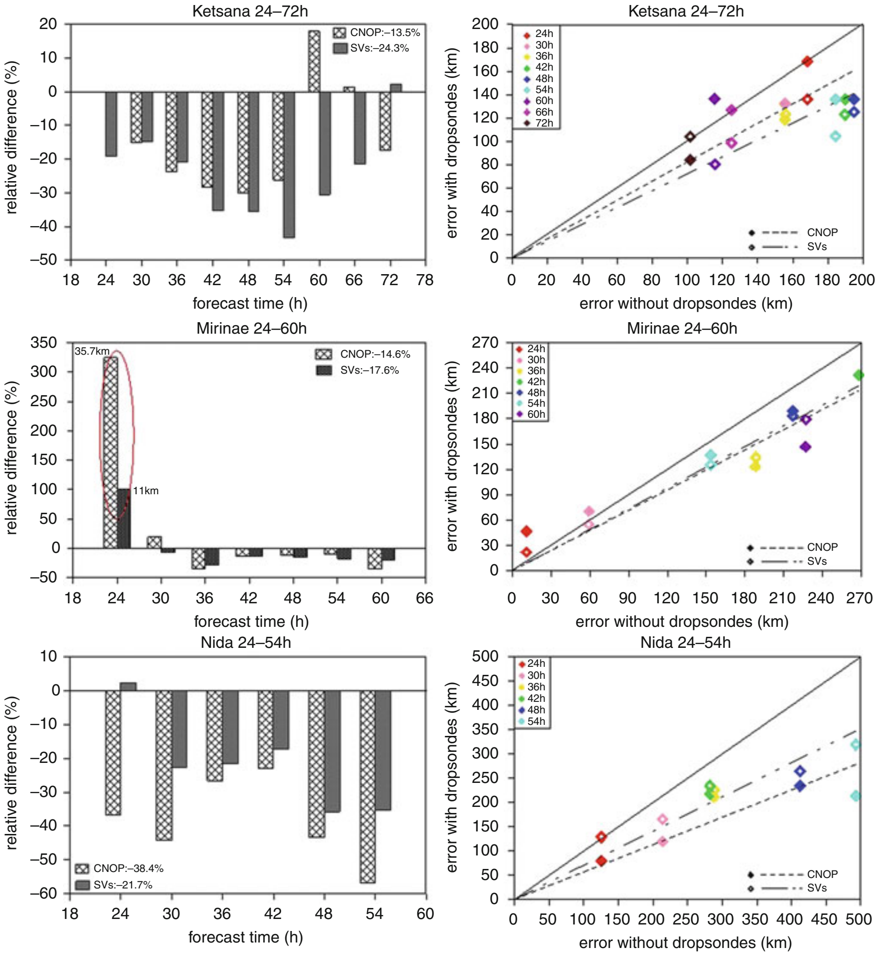


Fig. 24.17 Scatter diagrams of all track forecast errors for seven typhoon cases (*left*). The Y-axis represents the track forecast errors with dropsondes, and the x-axis represents those without dropsondes. Filled and empty diamonds denote the results of CNOPs and SVs, respectively. The colour of each diamond indicates the forecast time. Histograms on the right are relative differences corresponding to each case (From Qin and Mu 2011a)

Five kinds of experiments are designed and conducted: (1) no observations are assimilated; (2) all observations are assimilated; (3) only the CNOP targeted observations (observations in the CNOP sensitive area) are assimilated; (4) only the FSV targeted observations (observations in the FSV sensitive area) are assimilated; and (5) randomly targeted observations (observations within a randomly selected area) are assimilated. The results of the OSEs showed that the DOTSTAR data have a positive impact on the forecast of Nida's track (Table 24.6); Assimilation of the

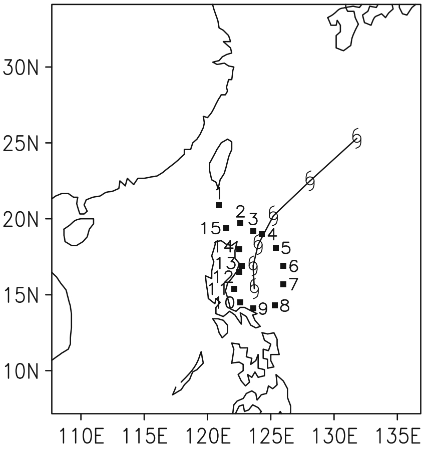


Fig. 24.18 DOTSTAR observations at 1200 UTC 17 May 2004 (*square points*) and best track of Nida from 1200 UTC 17 to 0000 UTC 20 May 2004 (*solid line*). The numbers from 1 to 15 indicate the sequence of dropsonde observations (From Chen 2011)

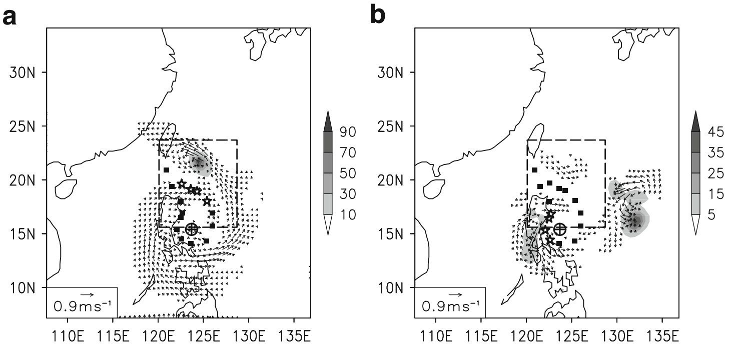


Fig. 24.19 VIE (*shaded*; J kg^{-1}) and wind (*vector*; m s^{-1}) component at the level 0.7 of (**a**) CNOP and (**b**) FSV. The box defined by *dashed lines* indicates the verification area. Pentangle points indicate the dropsonde observations used in (**a**) CNOPDROP and (**b**) FSVDROP. The symbol '⊕' indicates the position of Nida's center at 1200 UTC 17 May 2004 (From Chen 2011)

dropsondes in the CNOP sensitive area improved the track forecasts significantly, the dropsondes in the FSV sensitive areas also increased the accuracies of the 24-h and 36-h track forecasts, but the impact is comparatively small. However, the dropsondes in the randomly selected region have negative effects on the track forecasts, especially on the 24-h track forecast.

These results indicate that the CNOP method would be useful in decision making about dropsonde deployments.

Table 24.6 Verification dry energy norms (J kg^1) and 24 and 36-h forecast track errors (km) of NONDROP, ALLDROP, CNOPDROP, FSVDROP, and RANDROP as well as corresponding error reductions of ALLDROP, CNOPDROP, FSVDROP, and RANDROP

	NODROP	ALLDROP	CNOPDROP	FSVDROP	RANDROP
Dry energy (Verification area)	2288.75	2165.09	1918.48	2470.28	2622.37
Reduction (24 h forecast error)	—	5.4 %	16.2 %	–7.9 %	–14.6 %
Track error (24 h forecast)	79.83	48.34	29.02	73.36	98.81
Reduction (24 h track error)	—	39.4 %	63.6 %	8.1 %	–23.8 %
Track error (36 h forecast)	176.57	126.33	114.73	163.32	181.85
Reduction (36 h track error)	—	28.5 %	35.0 %	7.5 %	–3.0 %

24.5 Summary and Discussions

In this chapter, the recent progresses in targeted observations for tropical cyclone prediction based on Conditional Nonlinear Optimal Perturbation (CNOP) method have been reviewed. The CNOP have been used to identify the sensitive areas for tropical cyclone predictions. The first singular vector (FSV), as well as the composite singular vectors (CSV or SVs) have also been used to identify the sensitive areas for comparison.

First, the properties of the sensitive areas identified by CNOP (shorted for CNOP sensitive areas) have been studied, including the variations of the CNOP sensitive areas with respect to the changes of the horizontal resolution, the verification area design and the optimization time period. It is found that when the general position of the verification area is designed, small variations have minimal influence on targeted observations when using the CNOP method.

The CNOP sensitive areas have similarities at different horizontal resolutions, that is, the sensitive areas identified at lower resolutions could be helpful for improving the forecast at finer resolution. However, to get the largest improvement at a high resolution, it is better to use the sensitive areas identified at the same resolution. A resolution at which the nonlinearity could be explored has been suggested to be used in the identification of the sensitive areas.

The CNOP sensitive areas identified for special forecast times when the initial time is fixed resemble those identified for other forecast times in the linear case, while the similarities among the sensitive areas identified for different forecast times are more limited in the nonlinear case. This means the targeted observations deployed to improve a special time forecast would also favourably affect the forecasts at other times in the linear case, while in the nonlinear case, the targeted observations deployed for a special time forecast are limited to improve forecasts at other times. So it is suggested that in the nonlinear case, the deployment of targeted observations should be adaptive to obtain the largest improvement for different targeted forecasts, and should be more widespread to achieve the greatest improvement in multiple time forecasts. When the forecast time is fixed, the CNOP sensitive areas are much different when they are identified at different time period ahead. So the deployment of targeted observations to improve a special forecast depends strongly on the time of deployment.

Then the efficiency of the CNOP sensitive areas has been examined with a lot of events. The influence of the initial errors in various areas on the targeted forecast has been investigated based on two typhoons. Forty random initial errors have been added to the initial conditions in 40 same-size areas, and it is found that generally the initial errors in the CNOP sensitive areas would have the largest impact on the forecast.

Next, the observing system simulation experiments (OSSEs) have been performed to assess whether the CNOP sensitive areas can be considered as dropping sites in real time targeting. It is demonstrated that the energy of prediction error could be reduced by assimilating the ideal observation data in the CNOP sensitive areas for three typhoon cases. Another study of seven typhoon events originated in the western North Pacific during the 2009 season have showed that assimilating the ideal observations in the CNOP sensitive areas resulted in the improvements of 13–46 % in typhoon track forecasts, while the improvements of 14–25 % are obtained by assimilating the ideal observations in the SV sensitive areas. Besides, the improvements can be achieved for longer forecast times.

Finally, the observation system experiments (OSEs) using the DOTSTAR Data have been carried out to further examine the efficiency of the CNOP sensitive areas. Results show that the DOTSTAR data in the CNOP sensitive areas have a more positive impact on the typhoon track forecast than that in the FSV sensitive areas.

All the above results demonstrate that the CNOP is a useful tool in the adaptive observations to identify the sensitive areas.

Nevertheless, there are spaces remained to be further studied. For example, there are cases that the CNOP sensitive areas have negative impact on the typhoon forecast, analyze the reason, study more cases, and summarize the conditions when the CNOP sensitive areas would be much effective is needed. Except for the studies with horizontal resolution, the other studies have used a lower resolution of 60 km, it is necessary to use higher resolutions to study once the computational resources have been improved. The calculation of CNOP in current studies have been based on MM5 model, and it is known that the MM5 is not to be developed, so using new advanced models to calculate CNOP is urgent. Recently, Wang et al. (2011) have used the advanced model WRF to calculate CNOP, so using the CNOP based on WRF to identify sensitive areas will be an important new study. Furthermore, since in current formulation the cost function have been designed as the total dry energy in the verification area, it is expected that new cost function would be better designed thus it could directly relate to our intensions such as the track or intensity forecasts of the typhoons and could guarantee the improvement of the typhoon forecasts, and so on.

References

Aberson SD (2003) Targeted observations to improve operational tropical cyclone track forecast guidance. Mon Wea Rev 131:1613–1628

Ancell BC, Mass CF (2006) Structure, growth rates, and tangent linear accuracy of adjoint s'ensitivities with respect to horizontal and vertical resolution. Mon Wea Rev 134:2971–2988

Barkmeijer J, Iversen T, Palmer TN (2003) Forcing singular vectors and other sensitive model structures. Quart J Roy Meteor Soc 129(592):2401–2423

Bender MA, Ross RJ, Tuleya RE, Kurihara Y (1993) Improvements in tropical cyclone track and intensity forecasts using the GFDL initialization system. Mon Wea Rev 121:2046–2061

Bergot T (1999) Adaptive observations during FASTEX: a systematic survey of upstream flights. Quart J Roy Meteor Soc 125:3271–3298

Birgin EG, Martinez JE, Marcos R (2001) Algorithm 813: SPG—software for convex-constrained optimization. ACM Trans Math Softw 27:340–349

Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. J Atmos Sci 56:1748–1765

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Mon Wea Rev 129:420–436

Buizza R, Cardinali C, Kelly G, Thépaut J (2007) The value of targeted observations part II: The value of observations taken in singular vectors-based target areas. Quart J Roy Meteor Soc 133:1817–1832

Chen B-Y (2011) Observation system experiments for typhoon Nida (2004) using the CNOP method and DOTSTAR data. Atmos Ocean Sci Lett 4:118–123.

Chen B, Mu M (2012) The roles of spatial locations and patterns of initial errors in the uncertainties of tropical cyclone forecasts. Adv Atmos Sci 29:63–78

Duan WS, Luo H (2010) A new strategy for solving a class of nonlinear optimization problems related to weather and climate predictability. Adv Atmos Sci 27:741–749

Duan WS, Mu Mu (2009) Conditional nonlinear optimal perturbation: applications to stability, sensitivity, and predictability. Sci China (D) 884–906

Duan WS, Zhang R (2010) Is model parameter error related to a significant spring predictability barrier for El Ni∼no events? Results from a theoretical model. Adv Atmos Sci 27(5):1003–1013. doi:10.1007/s00376-009-9166-4

Dudhia J (1993) A nonhydrostatic version of the Penn State/NCAR mesoscale model: validation tests and simulation of an Atlantic cyclone and cold front. Mon Wea Rev 121:1493–1513

Ehrendorfer M, Errico RM (1995) Mesoscale predictability and the spectrum of optimal perturbations. J Atmos Sci 52:3475–3500

Franklin JL, DeMaria M (1992) The impact of Omega dropwindsonde observations on barotropic hurricane track forecasts. Mon Wea Rev 120:381–391

Froude LSR, Bengtsson L, Hodges KI (2007) The predictability of extratropical storm tracks and the sensitivity of their prediction to the observing system. Mon Wea Rev 135:315–333

Gelaro R, Langland RH, Rohaly GD, Rosmond TE (1999) An assessment of the singular-vector approach to targeted observations using the FASTEX dataset. Quart J Roy Meteor Soc 125:3299–3327

Hamill TM, Snyder C (2002) Using improved background-error covariance from an ensemble kalman filter for adaptive observations. Mon Wea Rev 130:1552–1572

Hoffman RN, Grassotti C, Isaacs RG, Louis JF, Nehrkorn T (1990) Assessment of the impact of simulated satellite lidar wind and retrieved 183 GHz water vapor observations on a global data assimilation system. Mon Wea Rev 118:2513–2542

Joly A, Jorgensen D, Shapiro MA, Thorpe A, Bessemoulin P, Browning KA, Cammas JP, Chalon JP, Clough SA, Emanuel KA, Eymard L, Gall R, Hildebrand PH, Langland RH, Lemaitre Y, Lynch P, Moore JA, Persson POG, Snyder C, Wakimoto RM (1997) The Fronts and Atlantic Storm-Track Experiments(FASTEX): scientific objectives and experimental design. Bull Am Meteor Soc 78:1917–1940

Kim HM, Morgan MC, Morss RE (2004) Evolution of analysis error and adjoint-based sensitivities: implications for adaptive observations. J Atmos Sci 61:795–812

Langland RH (2005) Issues in targeted observing. Quart J Roy Meteor Soc 131:3409–3425

Langland RH, Toth Z, Gelaro R, Szunyogh I, Shapiro MA, Majumdar SJ, Morss RE, Rohaly GD, Velden C, Bond N, Bishop CH (1999a) The North Pacific Experiment (NORPEX-98): targeted observations for improved North American weather forecasts. Bull Am Meteor Soc 80:1363–1384

Mu M, Jiang ZN (2008) A new method to generate the initial perturbations in ensemble forecast: conditional nonlinear optimal perturbations. Chin Sci Bull 53:2062S–2068S

Mu M, Duan WS, Wang B (2003) Conditional nonlinear optimal perturbation and its applications. Nonlin Processes Geophys 10:493–501

Mu M, Duan WS, Wang B (2007) Season-dependent dynamics of nonlinear optimal error growth and ENSO predictability in a theoretical model. J Geophys Res 112:D10113

Mu M, Zhou FF, Wang HL (2009) A method to identify the sensitive areas in targeting for tropical cyclone prediction: conditional nonlinear optimal perturbation. Mon Wea Rev 137:1623–1639

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

Peng MS, Reynolds CA (2006) Sensitivity of tropical cyclone forecasts as revealed by singular vectors. J Atmos Sci 63:2508–2528

Peng YH, Duan WS, Xiang J (2011) Effect of stochastic MJO forcing on ENSO predictability. Adv Atmos Sci 28(6):1279–1290. doi:10.1007/s00376-011-0126-4

Qin X-H (2010a) The sensitive regions identified by the CNOPs of three typhoon events. Atmos Ocean Sci Lett 3:170-175

Qin X-H (2010b) A comparison study of the contributions of additional observations in the sensitive regions identified by CNOP and FSV to reducing forecast error variance for the Typhoon Morakot. Atmos Ocean Sci Lett 3:258–262

Qin X, Mu M (2011a) Influence of conditional nonlinear optimal perturbations sensitivity on typhoon track forecasts. Quart J Roy Meteor Soc. doi:10.1002/qj.902

Qin X Mu M (2011b) A study on the reduction of forecast error variance by three adaptive observation approaches for tropical cyclone prediction. Mon Wea Rev 139:2218–2232

Rabier F, Gauthier P, Cardinali C, Langland R, Tsyrulnikov M, Lorenc A, Steinle P, Gelaro R, Koizumi K (2008) An update on THORPEX-related research in data assimilation and observing strategies. Nonlin Processes Geophys 15:81–94

Riehl H, Haggard WH, Sanborn RW (1956) On the prediction of 24-hour hurricane motion. J Meteor 13:415–420

Riviere O, Lapeyre G, Talagrand O (2008) Nonlinear generalization of singular vectors: behavior in a baroclinic unstable flow. J Atmos Sci 65:1896–1911

Sun GD, Mu M, (2009) Nonlinear feature of the abrupt transitions between multiple equilibria states of an ecosystem model. Adv Atmos Sci 26(2):293–304. doi:10.1007/s00376-009-0293-8

Szunyogh I, Toth Z, Majumdar S, Morss R, Etherton B, Bishop C (2000) The effect of targeted observations during the 1999 Winter Storm Reconnaissance program. Mon Wea Rev 128:3520–3537

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 Wea Rev 130:1144–1165

Snyder C (1996) Summary of an informal workshop on adaptive observations and FASTEX. Bull Amer Meteor Soc 77:953–961

Wang HL, Mu M, Huang XY (2011) Application of conditional non-linear optimal perturbations to tropical cyclone adaptive observation using the Weather Research Forecasting (WRF) model. Tellus 63A:939–957

Wu CC, Lin PH, Aberson S, Yeh TC, Huang WP, Chou KH, Hong JS, Lu GC, Fong CT, Hsu KC, Lin II, Lin PL, Liu CH (2005) Dropwindsonde Observations for Typhoon Surveillance near the Taiwan Region (DOSTAR): an overview. Bull Am Meteor Soc 86:787–790

Wu CC, Chen JH, Lin PH, Chou KH (2007) Targeted observations of tropical cyclone movement based on the adjoint-derived sensitivity steering vector. J Atmos Sci 64: 2611–2626

Zhang FQ et al (2002) Mesoscale predictability of the "surprise" snowstorm of 24–25 January 2000. Mon Wea Rev 130:1617–1632

Zhou FF, Mu M (2011) The impact of verification area design on tropical cyclone targeted observations based on the CNOP method. Adv Atmos Sci 28(5):997–1010. doi:10.1007/s00376-011-0120-x

Zhou FF, Mu M (2012a) The impact of horizontal resolution on the CNOP and on its identified sensitive areas for tropical cyclone predictions. Adv Atmos Sci 29:36–46. doi:10.1007/s00376-011-1003-x

Zhou FF, Mu M (2012b) The time and regime dependences of sensitive areas for tropical cyclone prediction using the CNOP method. Adv Atmos Sci 29:705–716. doi:10.1007/s00376-012-1174-0

Zhu H, Thorpe A (2006) Predictability of extratropical cyclones: The influence of initial condition and model uncertainties. J Atmos Sci 63:1483–1497

Zou X, Vandenberghe F, Pondeca M, Kuo Y-H (1997) Introduction to adjoint techniques and the MM5 adjoint modeling system. NCAR technical note NCAR/TN-435_STR

Chapter 25
GSI/WRF Regional Data Assimilation System and Its Application in the Weather Forecasts over Southwest Asia

Jianjun Xu and Alfred M. Powell, Jr.

Abstract In this study, the impact of directly assimilating Advanced TIROS Operational Vertical Sounder (ATOVS) radiances using the Community Radiative Transfer Model (CRTM) was evaluated to determine the impact on forecasts over Southwest Asia. The CRTM was developed by the Center for Satellite Applications and Research (STAR) and its application was promoted by the Joint Center for Satellite Data Assimilation (JCSDA). The ATOVS radiance data from the National Environmental Satellite Data and Information Service (NESDIS), the Gridpoint Statistical Interpolation (GSI) three-dimensional variational analysis (3DVAR) system from the National Centers for Environmental Prediction (NCEP), and the Advanced Research WRF (WRF-ARW) model from the National Center for Atmospheric Research (NCAR) were employed in this study.

First, this paper will describe the forecasting errors encountered from running the WRF-ARW model in the complex terrain of Southwest Asia from 1–31 May 2006. The subsequent statistical evaluation is designed to assess the model's surface and upper-air forecast accuracy. The results show that the model biases caused by inadequate parameterizations of physical processes are relatively small, except for the 2-m temperature, as compared to the nonsystematic errors resulting in part from the uncertainty in initial conditions. The total model forecast errors at the surface show a substantial spatial heterogeneity and the errors are relatively larger in higher elevation mountain areas. The performance of 2-m temperature forecasts is different from the other surface variables' forecasts; the model forecast errors

J. Xu
Environmental Science and Technological Center College of Science, George Mason University, Fairfax, 22030, VA, USA
e-mail: Jianjun.xu@noaa.gov

A.M. Powell, Jr. (✉)
NOAA/NESDIS/Center for Satellite Applications and Research (STAR) Camp Springs, WWB, 5200 Auth Road, Camp Springs, 20746, MD, USA
e-mail: Al.Powell@noaa.gov

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_25,

in 2-m temperature forecasts are closely related to the terrain configuration. The simulated diurnal variation of near-surface temperature is much smaller than the observed diurnal variation.

Second, to understand the impact of initial conditions on the accuracy of the model forecasts, the satellite radiances are assimilated into the numerical model through GSI data assimilation system. The results indicate that on average over a 30-day experiment for the 24- and 48-h (second 24-h) forecasts, the satellite data provides beneficial information for improving the initial conditions and the model errors are reduced to some degree over some of the study locations. The diurnal cycle of some forecast variables can be improved by using adequate initial conditions with satellite radiance data assimilation.

25.1 Introduction

The assimilation of satellite radiance observations into a numerical weather prediction (NWP) system is an important pathway for improving weather forecasts by providing initial conditions representative of the true state of the atmosphere. Preliminary impact studies of satellite data using satellite retrieved winds, and humidity were focused on the global system. The results show a positive impact of satellite data on numerical weather prediction forecasts, especially in the Southern Hemisphere (e.g., Tracton et al. 1980; Halem et al. 1982; Andersson et al. 1991; Mo et al. 1995; Derber and Wu 1998). The satellite data are a useful data source not only in global models but also in regional-scale models. Bouttier and Kelly (2001) demonstrated that the impact of rawinsonde data on the forecast was extremely large over regional areas, but the aircraft and satellite data seemed to have little effect.

There are two basic approaches to assimilate satellite information into a data assimilation system (DAS). The first approach is to assimilate retrieved data from radiances measured by satellite instruments. The satellite retrievals, such as humidity and wind fields, usually were provided by the satellite data provider independent of the data assimilation system. The second approach is to assimilate radiance measurements directly into a DAS. Direct radiance assimilation is theoretically superior to retrieval assimilation because the observational error statistics are more justified in direct radiance assimilation than in retrieval assimilation (Eyre et al. 1993; Derber and Wu 1998; McNally et al. 2000). This approach differs from the traditional practice of transforming the observations into analysis variables and requires an observation operator be built into the DAS to transform model variables into radiances. The linkage between forecast model state variables, such as temperature and humidity, and observed radiances is expressed mathematically by a forward radiative transfer model (RTM), which calculates radiance from model state vertical profiles.

To maximize the benefit of assimilating satellite data, it must be assimilated in the regional models in addition to the global models. Regional models have lagged global models to some extent, due to the complications from local and diabatic effects, complex nonlinear balance relationships, and the presence of lateral boundaries (Stauffer et al. 1991). The complex relationships between the different atmospheric fields and various scales of motion require a dynamical approach to data analysis and assimilation (Lorenc 1986). Regional models often contain information on structures linked to the local terrain. As a result, to obtain high-quality output from the regional models, high resolution topography is necessary.

Due to the linkage with terrain, the satellite data assimilation for regional model initialization has received the greatest attention. Therefore, the role of satellite observations for regional modeling through a month's experiments over Southwest Asia will be analyzed. Weather forecasts in Southwest Asia (SWA) are often very complex because of mesoscale variations induced by the complex terrain and diverse land use. This is a predominately a semi-arid to arid region surrounded by the Black and Caspian Seas in the north, the Mediterranean in the west, the Arabian Sea and Persian Gulf in the south, Himalayas in the east, and crossed by the impressive Tauros, Zagros, and Hindu Kush mountains. A few previous model studies (Evans and Smith 2001, 2006; Evans et al. 2004; Zaitchik et al. 2007a,b; Marcella and Eltahir 2008) provided some interesting results for the basic weather simulation in SWA using a regional climate model (RegCM2) or the MM5 model. They pointed out that the regional model had difficulty in producing an accurate simulation of precipitation in certain sub-regions, which is related to an accurate description of storm tracks, topographic interactions, and atmospheric stability.

This evaluation primarily concentrates on the forecasts of wind, temperature and precipitation since SWA is dominated by hot, dusty, windy weather (Agrawala et al., 2001). During the transitional season from winter to summer, the temperature and wind increase substantially; contrastingly, the precipitation decreases significantly. During this seasonal transition, the occurrence of blowing sand/dust and unstable local-scale weather events increases as well, and the prediction accuracy of these events is highly dependent upon the accuracy of the temperature, precipitation and wind forecasts from the model.

Some recent studies have evaluated the WRF-ARW model based on objective error statistics for precipitation forecasts. Cheng and Steenburgh (2005) produced surface sensible weather forecasts with WRF-ARW and Eta models over the western United States. Their results suggest that improvements in initialization are more important than improvements in the physics for land surface processes. Gallus and Bresch (2006) compared the impacts of the WRF dynamics core physics package, and initial conditions on warm season rainfall forecasts over central United States. They found that the sensitivity of rainfall forecasts to the physics, dynamics, and initial conditions are dependent on the rainfall events. For heavier rainfall, sensitivity to initial conditions is generally less substantial than the sensitivity to changes in the dynamic core or physics. For light rainfall, the WRF model using NCAR physics is much more sensitive to a change in the dynamic core than the WRF model using NCEP physics. Wan and Xu (2011) pointed out, in a case study of the flash-flood that occurred in the central Guangdong Province of Southeast China during June 20–21 2005, that the model simulation largely depends on three factors: model resolution, physical process schemes and the initial conditions.

The studyshowed that the initial conditions are improved by using the satellite data assimilation and result in a reduced forecast error for heavy rainfall location. Therefore, it is not surprising that considerable effort has focused on improving the estimates of the model initial states through data assimilation.

This paper is organized as follows. Section 25.2 describes the real-time configuration of WRF-ARW and data assimilation system. The observational datasets used in the verification are given in Sect. 25.3. Section 25.4 explains the methodology used in the evaluation. The results of the forecast error for the May 2006 case are presented in Sect. 25.5. Section 25.6 investigates the impact of data assimilation on the forecasts. Finally, a summary and discussion are given in Sect. 25.7.

25.2 Model and Data Assimilation System

25.2.1 ARW WRF Regional Model

The numerical weather prediction model used in this study is the WRF model (Michalakes et al. 2001; Skamarock et al. 2005), which is a nonhydrostatic, fully compressible, primitive equation model. Lead institutions involved in the effort to develop this model include the National Center for Atmospheric Research (NCAR), Air Force Weather Agency (AFWA), National Centers for Environmental Prediction (NCEP), National Oceanic and Atmospheric Administration (NOAA), and other government agencies and universities. WRF is built around a software architectural framework in which different dynamical cores and model physics packages are presented within the same code. With the WRF model, it is possible to mix and match the dynamical cores and physics packages of different models to optimize performance since each model has strengths and weaknesses in different areas and weather events. It uses a terrain-following hydrostatic pressure coordinate and the Arakawa C grid staggering.

25.2.2 GSI 3DVAR Data Assimilation System for ARW WRF Regional Model

The Gridpoint Statistical Interpolation (GSI) analysis system (Kleist et al. 2009a,b) is developed based on NCEP's current three-dimensional variational analysis (3DVAR) system known as the Spectral Statistical Interpolation (SSI) (Parrish and Derber 1992; Derber et al. 1991). The SSI has the advantage that the statistics of the background error, both structure and amplitude, can be easily obtained and applied in the analysis procedure. It is simpler to apply a diagonal background error covariance in spectral space than to convolve the corresponding smoothing kernel with the innovations in physical space. However, with only a diagonal covariance in spectral space, the structure function is limited to being geographically homogeneous and isotropic about its center (Parrish and Derber 1992; Courtier et al. 1998). One has little control over the spatial variation of the error statistics when a simplified diagonal background error covariance in spectral space is used. With some computational cost associated with the extra transforms in and out of the physical space in each iteration of the optimization solver, spatially inhomogeneous, for example, latitude-dependent, variances can be applied, but it is not as easy to construct inhomogeneous and/or anisotropic shapes for the covariance profiles in spectral space. The GSI helped overcome this shortcoming.

The current GSI regional analysis system employs NCEP's Nonhydrostatic Mesocale Model (NMM) WRF and NCAR's ARW WRF mass core (Liu and Weng 2006; Xu et al. 2009; Xu and Powell 2011; Wan and Xu 2011), and the input data can be either binary or netcdf format datasets. DAS/forecast model interface has been adapted separately for the WRF NMM core and the WRF mass core. For the ARW WRF mass core, the inputs/outputs are made on a C-grid, no interpolation is needed for the mass variables (T,Q), but the wind variables (u & v) are interpolated in x and y to mass points respectively

All interpolations are linear in each direction; the projection information is not required. The code automatically determines the local scale information needed for transforming from global coordinates to local coordinates, properly rotating winds to the model frame, and dx, dy are needed for local derivatives. All of these procedures can be determined from the two dimension fields available in both NMM and ARW mass core files given the earth latitude and longitude and dx, dy for every grid point.

Eventually, GSI can be connected to other models in a systematic way. Part of this has already been accomplished by eliminating the need to specify map projections for the horizontal domain definition.

The Assimilation system produces an analysis through the minimization of an objective function given by

$$\mathrm{J} = \mathbf{x}^{\mathrm{T}}\mathrm{B}^{-1}\mathbf{x} + (\mathrm{H}\mathbf{x} - \mathbf{y})^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{H}\mathbf{x} - \mathbf{y})$$

$$\mathrm{J} = 1/2(\mathbf{x}^{\mathrm{T}}\mathrm{B}^{-1}\mathbf{x}) + (\mathrm{H}\mathbf{x} - \mathbf{y})^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{H}\mathbf{x} - \mathbf{y})$$

where $\mathbf{x}$ is a vector of analysis increment, B is the background error covariance matrix, $\mathbf{y}$ is innovation vector, $\mathbf{y} = \mathbf{y}_{\mathrm{obs}} - \mathrm{H}\mathbf{x}_{\mathrm{guess}}$, R is the observational and representativeness error covariance matrix, and H is the transformation operator from the analysis variable to the form of the observations.

For the SSI which is tied to isotropic and homogeneous background error covariance matrix (B), the spectral model can conveniently and easily handle the pole. In contrast, the GSI allows for non-homogeneous and anisotropic B formulation (Wu et al. 2002), distinguishes between land and sea, the tropics, and midlatitudes, and is easy to use in both global and regional applications. Currently background error cannot change in outer iteration (due to preconditioning in the inner iteration). In regard to this problem, Derber suggests that the two outer iterations appear to work reasonably well except for precipitation (personal communication). Often, we run three outer iterations so we can see the fit to the observations at the end of the second outer iteration. The background error variances, which vary by wavenumber and vertical mode, are fixed in time and estimated from scaled differences between 24- and 48-h forecasts valid at the same time (see Parrish and Derber 1992). For the regional system, the background error statistics use the same vertical grid structure as the first guess. The background error covariance matrix is extracted through the interpolation of NCEP's Global Forecast System (GFS) counterpart.

The observation error covariance matrix (R) should not only contain information on the observational error but also errors in representativeness (Lorenc 1986). Thus, this matrix should include the error in the radiative transfer modeling, but the specification of this matrix is difficult. It is clear that the errors are probably correlated spatially because of the errors in the radiative transfer, instrument errors and errors arising from imperfect cloud clearing, emissivity correction, and other components. However, these correlations are probably quite different from the spatial correlations found in the temperature and moisture retrievals and are currently not well known. For this reason, the GSI system has chosen these errors to be spatially uncorrelated. In addition, because the microwave inter-channel error correlations are not known, they have been set equal to zero.

For the radiance data, the transformation is more complicated. The temperature, moisture and pressure on the Gaussian grid are bilinearly interpolated in the horizontal to the observation region to create a temperature and moisture profile.

25.3 Observed and Analyzed Datasets

Observed precipitation. The observed precipitation data are taken from the Climate Prediction Center (CPC) Famine Early Warning System (FEWS) program, which is derived from geostationary satellite retrieved precipitation merged with rain gauge and model analysis. The merging technique has been shown to significantly reduce bias and random error compared to individual precipitation data sources, thus increasing the accuracy of rainfall estimates (Xie and Arkin 1996). Geostationary satellite data is utilized for the determination of cloud top temperature. METEOSAT 5 thermal Infrared (IR) digital data at 5 km pixel resolution is accessed every 30 min and then reformatted and converted to a geographic grid with a 0.1° resolution. The grid is 751 × 501 points, which begins with point (1, 1) at 20 °E, 10 °N and ends at point (751, 501) at 95 °E, 60 °N. A horizontal resolution of 0.1° was chosen for the estimated computations to correspond with the absolute positioning error for the satellites of approximately 10 km. Arrays are used to accumulate the occurrences of cloud top temperatures below 235 °K and 275 °K. Rain gauge reports transmitted via the Global Telecommunications System (GTS) are received every 6 h and are utilized in the CPC Climate Assessment Data Base (CADB) for monitoring of climate anomalies. Automated quality control of these GTS observations within the CADB is done prior to the processing of precipitation estimates.

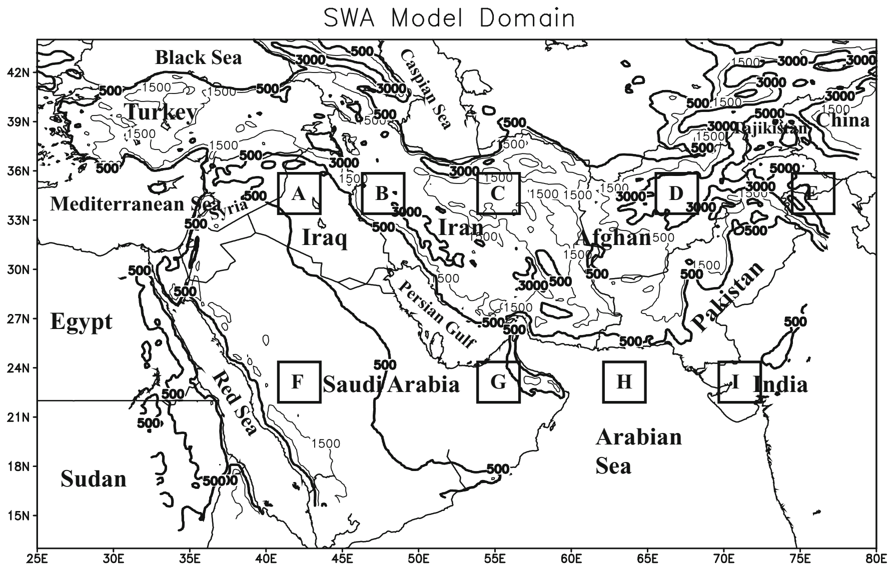


Fig. 25.1 Domain of model and subregion definition. Shaded indicates the elevation of terrain (unit: m). The sub-regions are defined as north Iraq (*A*; 34°–36 °N, 41°–43 °E); northwest Iran (*B*; 34°–36 °N, 46°–48 °E); north central Iran (*C*; 34°–36 °N, 54°–56 °E); central Afghanistan (*D*; 34°–36 °N, 66°–68 °E); west Himalaya Mountain (*E*; 34°–36 °N, 74°–76 °E); west Saudi Arabia (*F*; 22°–24 °N, 41°–43 °E); east Saudi Arabia (*G*; 22°–24 °N, 51°–53 °E); Arabian Sea (*H*; 22°–24 °N, 63°–65 °E) and Northwest India (*I*; 22°–24 °N, 70°–72 °E)

Observed temperature. The maximum and minimum temperature at the 2-m level with 0.5° × 0.5° gridded datasets are created by the NOAA's CPC, which is taken from observational stations of the WMO GTS datasets. The interpolation method is based on the previous rainfall estimation algorithm (Xie et al. 1996).

Analyzed temperature and wind field. The temperature and wind fields are taken from the NCEP Global Forecasting System (GFS) analysis data (GFS_ANL), which is gridded to a horizontal resolution of 1° × 1°.

25.4 Topography and Evaluation Method

To investigate the spatial heterogeneity of complex terrain in SWA region, nine representative sub-regions are depicted in Fig. 25.1. They are defined as north Iraq (A; 34°–36 °N, 41°–43 °E); northwest Iran (B; 34°–36 °N, 46°–48 °E); north central Iran (C; 34°–36 °N, 54°–56 °E); central Afghanistan (D; 34°–36 °N, 66°–68 °E); west Himalaya mountains (E; 34°–36 °N, 74°–76 °E); west Saudi Arabia (F; 22°–24 °N, 41°–43 °E); east Saudi Arabia (G; 22°–24 °N, 51°–53 °E); Arabian Sea (H; 22°–24 °N, 63°–65 °E); and west India (I; 22°–24 °N, 70°–72 °E).

Table 25.1 The averaged height of topography (Hgt: meter), vegetation type (Veg) and soil type (Soil) in the nine sub-regions (defined as Fig. 25.1) over SWA

	A	B	C	D	E	F	G	H	I
Hgt	328	2,557	737	3,833	4,839	958	67	0	75
Veg	Barren	Grass	Barren	Shrub land	Wooded tundra	Barren	Barren	Water	Mixed Dryland/ Cropland
Soil	Loam	Loam	Clay loam	Loam	Loam	Sandy loam	Loam	Water	Loam

The nine sub-regions effectively represent the heterogeneity of complex terrain in SWA region. Table 25.1 displays the average elevation of the topography (Hgt), vegetation type (Veg) and soil type (Soil) over these nine regions. Except for the water type in the Arabian Sea (marked H), the soil types in all other eight regions are loam; and the vegetation types include barren, grass, shrub land, wooded land, mixed dry/crop land and water. Three regions (B, D, E) with terrain above 2,500 m are covered by short plants with grass (B), shrubland (D) and wooded tundra (E). Three regions with terrain under 1,000 m (A, C and F) and the two plains regions (G and I) are practically free of any plants.

This evaluation is designed to present the model errors of surface temperatures, precipitation, wind speeds and upper atmospheric variables for both 24-h (hour) and 48-h (hour) (e.g. the second 24-h) forecasts. The statistical measures used to quantify model forecast performance are bias (forecast – observation), mean-square error (MSE), and standard deviation (SD) error. The MSE represents the total model forecast error including contributions from both systematic and nonsystematic/random errors. Systematic error may be caused by a consistent misrepresentation of physical parameters such as radiation or model convection. Nonsystematic errors are caused by uncertainties in the model initial conditions or unresolvable differences in scales between the forecasts and observations (Nutter and Manobianco 1999).

If X represents any of the parameters under consideration for a given time and vertical level, then the forecast error is defined as $X' = X_f - X_a$, where the subscripts f and a denote forecast and analyzed/observed quantities, forecasts and analyses, the bias is computed as

$$Bias = \overline{X'} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} X'_i \tag{25.1}$$

the mean-square error is computed as

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(X'_i\right)^2 \tag{25.2}$$

the SD error is computed as

$$SD = \left[\frac{1}{N} \sum_{i=1}^{N} \left(X_i' - \overline{X'} \right)^2 \right]^{1/2} \tag{25.3}$$

In (25.1, 25.2, and 25.3), N is used rather than N-1 so that a decomposition following Murphy (1988, eq. (9)) could be applied to the MSE:

$$MSE = (\overline{X'})^2 + (SD)^2 \tag{25.4}$$

Therefore, the total model forecast error (e.g. MSE) consists of contributions from model squared biases $(\overline{X'})^2$ (i.e., systematic error) and squared standard deviation $(SD)^2$ error (i.e., nonsystematic error) in the forecast and observed data. A fraction is defined to indicate the ratio of systematic error in the total model forecast error as follows:

$$E_r = \left((\overline{X'})^2 / MSE \right) \times 100\% \tag{25.5}$$

In (25.5), note that if the model bias is less than 50 %, most of the MSE is due to random, nonsystematic type variability in the errors.

25.5 Forecast Error

Similar to Air Force Weather Agency's (AFWA) operational setup, a 15-km grid spacing centered over the Southwest Asia (SWA) region (Fig. 25.1) is used to encompass the regions complex topography and associated spatial variability in surface characteristics. To assess model predictive skill, 48-h (hour) forecasts are made for each day starting at 00Z for the period of May 1 through May 31, 2006. Forecasts without data assimilation are labeled CTRL in order to distinguish them from the forecasts with data assimilation found in Sect. 25.6. The initial atmospheric and lateral boundary conditions, including soil moisture and temperature, are taken from the NCEP Global Forecast System (GFS) real time forecasts with 3-h intervals, which is gridded to a horizontal resolution of 1° × 1°. Through the WRF Preprocessing System (WPS), the global soil categories, land use categories, terrain height, annual mean deep soil temperature, monthly vegetation fraction, monthly albedo, maximum snow albedo, and slope category are interpolated into the model grids of the study domain. The physics packages used in the CTRL forecasts are the WRF Single Moment 5-class (WSM5) microphysics scheme, Yonsei University planetary boundary layer (YSUPBL) scheme, Noah land surface scheme, Grell-Devenyi ensemble cumulus scheme, Rapid Radiative Transfer Model (RRTM) longwave radiation, and the Dudhia shortwave radiation scheme.

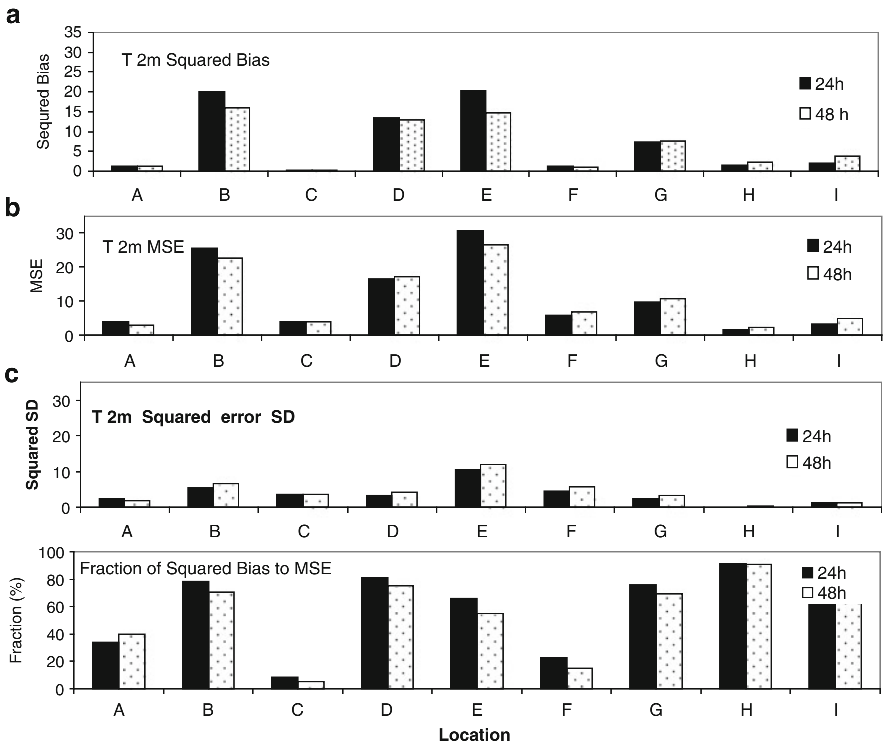


Fig. 25.2 *Squared bias*, mean square error (*MSE*), squared standard deviation (*SD*) error and fraction of Squared bias to MSE (*Er*) for 2-m temperature (°C) forecasts from 1 to 31 May 2006. Results are plotted for averaged 24- and 48-h forecasts with respect to the regions (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*) defined in Fig. 25.1. Unit: square of temperature (°C) in (**a**), (**b**) and (**c**) (Adapted from Xu et al. (2009))

In the following section, WRF-ARW model forecast error characteristics for 24- and 48-h (e.g. second 24-h) forecasts and diurnal variation are described. Forcasts results are produced during the 30-day period starting from May 1 through May 30, 2006.

25.5.1 24-h and 48-h Forecasts

25.5.1.1 2-Meter Temperature ($T_{2\text{-}M}$)

Squared biases (WRF forecasts – GFS analysis) in 2-m temperature forecasts vary with terrain elevation (Fig. 25.2a). Biases are larger over high terrain areas (E, B, D) for the 24- and 48-h forecasts. While, the biases are significantly smaller in low terrain regions (A, C, F, G, I) or water areas (H).

Even though the magnitude of the squared SD error (Fig. 25.2c) in the highest terrain region (E: west Himalayas mountains) is near equivalent to that of the forecast bias, it is very small in the other areas. However, the biases and corresponding MSEs are comparable in magnitude over most of the other mountain areas (Fig. 25.2b). The fraction of squared biases to the MSEs (Fig. 25.2d) is greater than 50 % in most of the areas, which showed clearly that a large contribution to the total model forecast errors in these regions are derived from a systematic model error. The result indicates an apparent model deficiency in the description of surface temperature in high terrain areas.

To illustrate the above point, squared biases, MSEs, and squared SD error in the whole SWA region are depicted in Fig. 25.3. For the 24-h forecasts, the total model forecast errors are dominated by the model systematic errors (Fig. 25.3a–c). The fraction of squared biases to the MSEs (Fig. 25.3d) exceeds 50 %; the distribution of total model forecast errors is also dependent on the topography of the model domain (Fig. 25.3a, b vs. Fig. 25.1). The 48-h forecast errors are a little higher than the 24-h forecast errors (Fig. 25.3e–h).

25.5.1.2 Precipitation

In contrast, the precipitation MSEs in the 24-h forecasts are dominated by squared SD error (Fig. 25.4) over all nine selected sub-regions. The biases are not correlated to the height of terrain. The maximum of the squared bias (Fig. 25.4a) over the highest terrain region is much smaller than the squared SD error. The fraction of squared biases to the MSE (Fig. 25.4d) is far less than 50 % in all selected sub-regions, which showed clearly that a larger contribution to the total model forecast error is from a nonsystematic model error. These results indicate an apparent model problem in the description of the initial conditions or the model resolution. The 48-h forecast errors are much higher than the 24-h forecast errors in most of areas.

For the whole study domain, the MSEs in the 24-h forecasts are obviously dominated by the model nonsystematic errors (Fig. 25.5a–c). The fraction of squared biases to the MSEs (Fig. 25.5d) is under 50 % except for some Himalaya mountain areas. The distribution of total model forecast errors has nothing to do with the structure of higher terrain. The areas of 48-h forecast errors greater than 20 mm^2 clearly extend over wider areas in the model domain (Fig. 25.5e–h).

25.5.1.3 Wind Speed at 10-M

Similar to precipitation, the MSEs in 10-m wind speed are largely associated with the nonsystematic errors in most of the sub-regions (Fig. 25.6a–c). The largest model bias occurs over northwestern Iran (B) (Fig. 25.6a). The biases over the west Himalaya mountain region (E), east Saudi Arabia (G) and west India (I) are almost zero. The fractions of squared biases to the MSEs (Fig. 25.6d) are under 50 % over all selected areas.

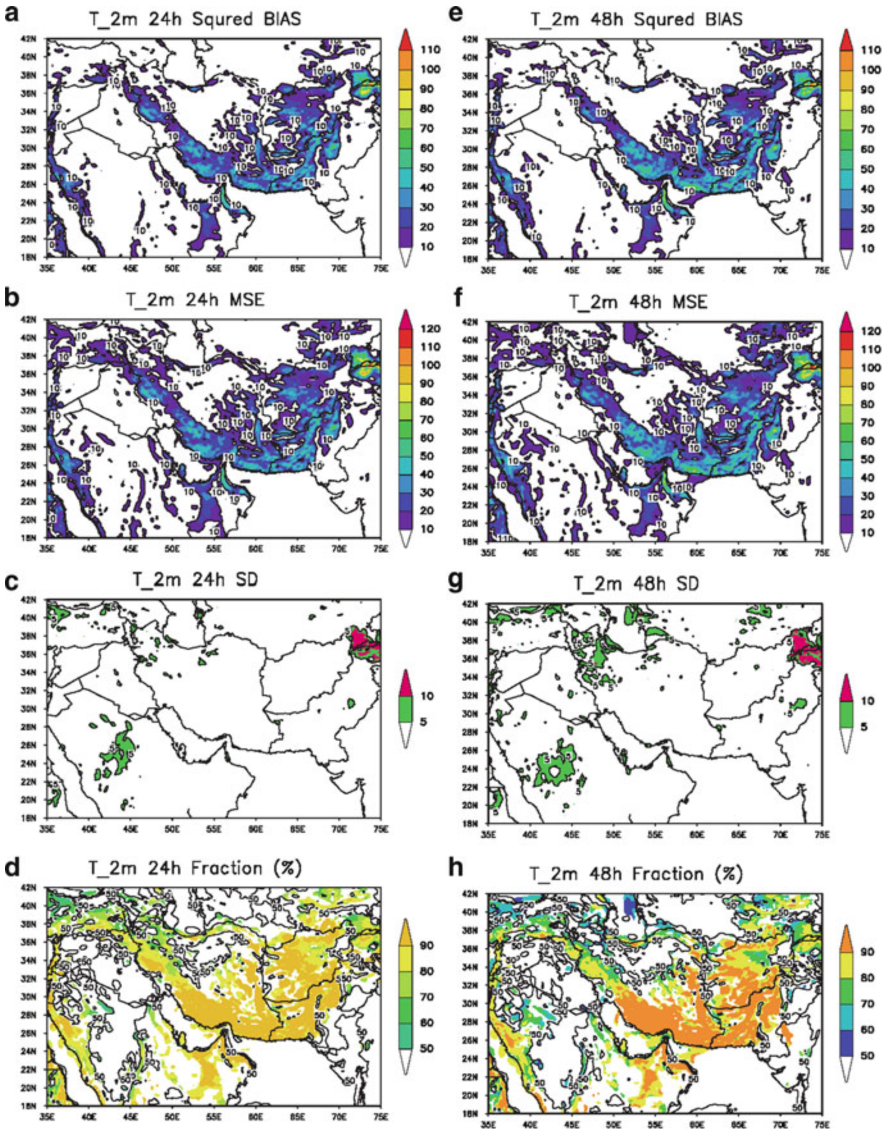


Fig. 25.3 *Squared Bias*, mean square error (*MSE*), Squared standard deviation (*SD*) error and fraction of squared bias to MSE (*Er*) of 2-m temperature for 24-h (**a–d**) and 48-h (**e–h**) forecasts for 30-day average from 1 to 31 May 2006. Unit: square of temperature (°C) in (**a**), (**b**), (**c**) and (**e**), (**f**), (**g**) (Adapted from Xu et al. (2009))

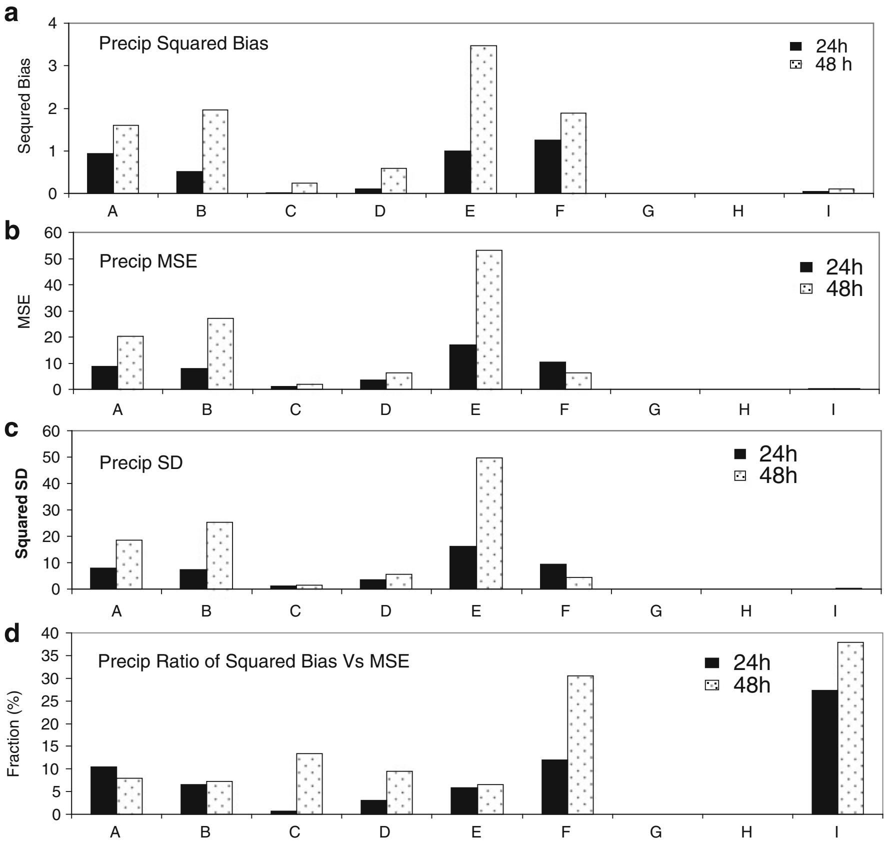


Fig. 25.4 Same as Fig. 25.2 but for precipitation (mm/day) forecasts. Unit: square of precipitation amounts (mm/day) in (**a**), (**b**), (**c**) (Adapted from Xu et al. (2009))

Over SWA domain, the MSEs of 10-m wind speed in 24-h forecasts correspond fairly well to the low SD errors (Fig. 25.7b, c). The forecast errors from systematic error in the western mountains of Iran are relatively large values (Fig. 25.7a). The 10-m wind speed statistical fields are quite different from the 2-m temperature fields, and the nonsystematic model errors compose a much larger portion of the total forecast errors for 2-m temperature forecasts (Fig. 25.7d). The 48-h forecast errors are similar to the 24-h forecast errors in most of the areas (not shown).

The above results suggest that the MSEs near the surface contain a substantial spatial heterogeneity, as seen by the relatively larger errors in higher mountainous areas. However, the source of the errors indicates a significant difference among temperature, precipitation, and wind speed. The inaccuracies in 2-m temperature forecasts are mainly from systematic errors, which are controlled largely by the physical representation in the model. In contrast, the inaccuracies in precipitation and 10-m wind speed forecasts are dominated more by nonsystematic errors,

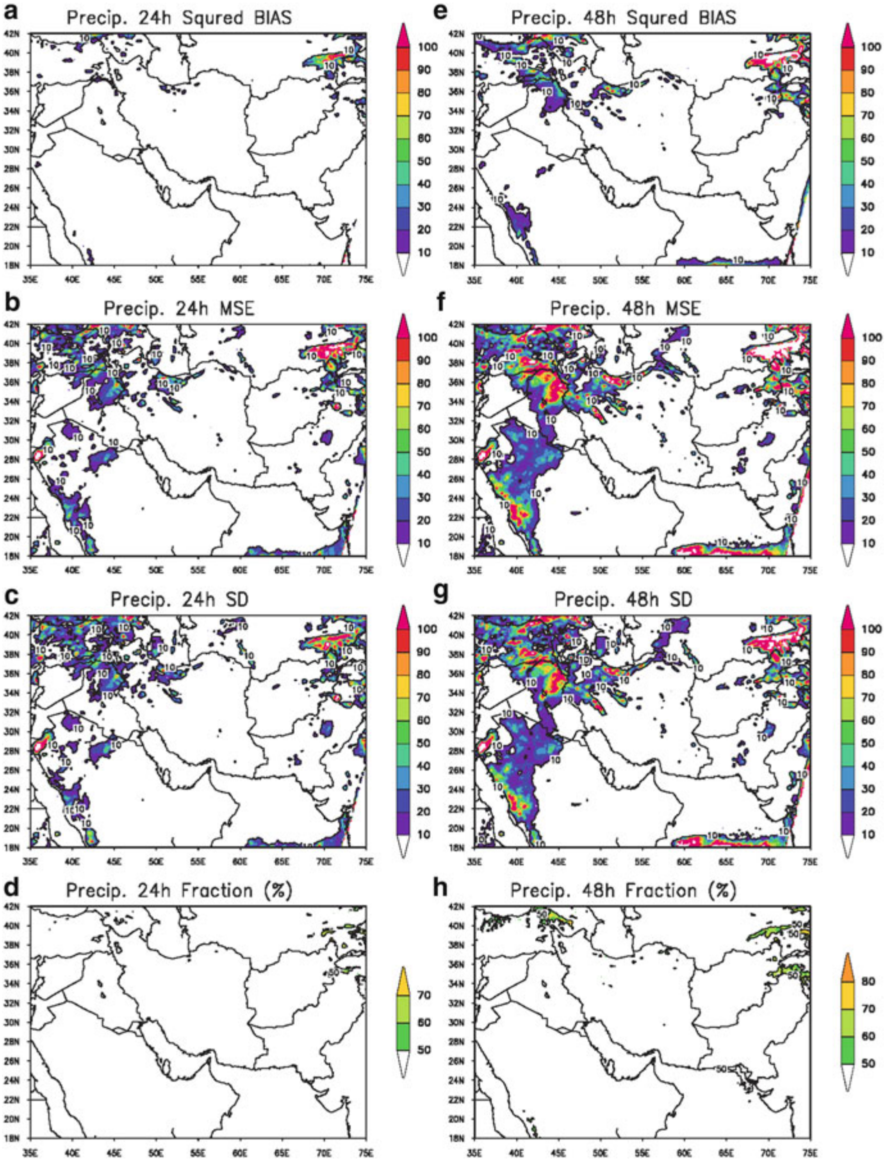


Fig. 25.5 Same as Fig. 25.3 but for precipitation (mm/day) forecasts. Unit: square of precipitation amounts (mm/day) in (**a**), (**b**), (**c**) and (**e**), (**f**), (**g**) (Updated from Xu et al. (2009))

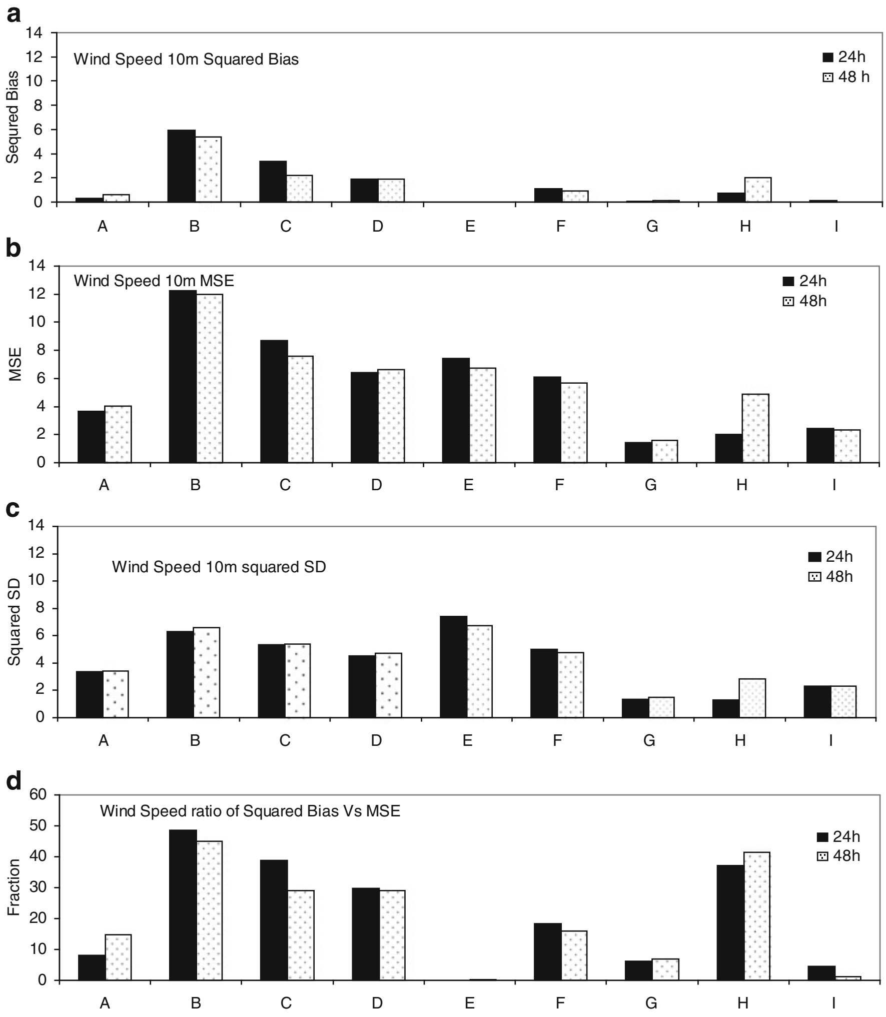


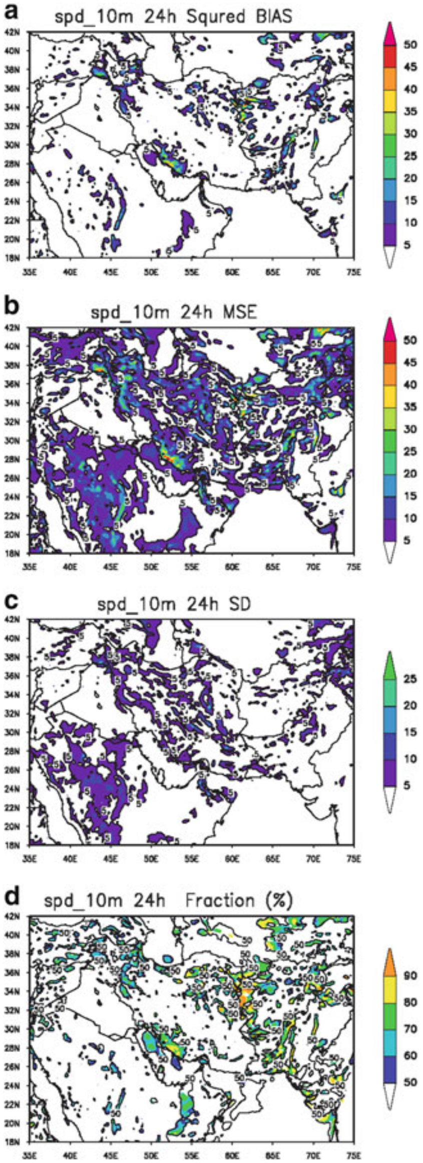
Fig. 25.6 Same as Fig. 25.2 but for 10-m wind speed (ms^{-1}) and 24-h forecasts only. Unit: square of wind speed (ms^{-1}) in (**a**), (**b**), (**c**) (Updated from Xu et al. (2009))

which we postulated to be errors derived from the random inadequacies of initial conditions.

25.5.1.4 Temperature at 500 hPa

The squared bias is very small except that the Himalaya mountain region (E) increases up to 17 °C for 24-h forecast (Fig. 25.8a) and there is a larger value for

Fig. 25.7 Same as Fig. 25.3, but for 10-m wind speed and 24-h forecasts only. Unit: square of wind speed (ms^{-1}) in (**a**), (**b**), (**c**) (Updated from Xu et al. (2009))



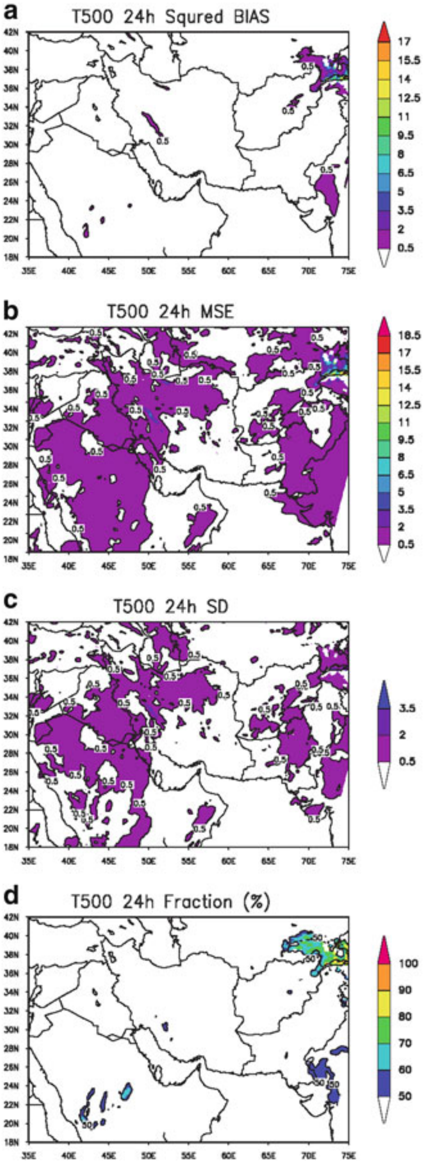


Fig. 25.8 Same as Fig. 25.3 but for 500 hPa temperature and 24-h forecasts only. Unit: square of temperature (°C) in (**a**), (**b**), (**c**) (Updated from Xu et al. (2009))

48-h forecasts (not shown). The larger magnitude of MSEs is randomly distributed over the central Saudi Arabia, southeast Iraq, northwest Iran and west Himalaya mountain region (Fig. 25.8b). The corresponding SD error (Fig. 25.8c) reveals that nonsystematic errors compose a substantial portion of the total error. The fraction of squared biases to the MSE (Fig. 25.8d) is far less than 50 % except for west Himalaya mountain region (E), which showed clearly that a larger contribution to the MSEs is from a nonsystematic total model forecast error. Compared to 24-h forecasts, the 48-h forecasts' bias is higher over most of study areas (not shown).

25.5.1.5 Winds at 200 hPa

Similar to the upper level temperature forecasts, the wind forecasts (the first 24-h forecasts shown only) at 200 hPa (Fig. 25.9) indicate that the MSEs are dominated by nonsystematic errors in either the zonal or meridional wind component or both. For the zonal wind forecasts, the large MSE over Himalaya mountain region is consistent with nonsystematic error, as well as, the Arabian Sea also has a strong nonsystematic error signature. For the meridional wind component, the larger forecast errors occur over a different place relative to the zonal wind forecasts. The larger MSEs for the zonal wind forecasts in the Himalaya mountain region disappear in the meridional wind field.

In summary, the 2-m temperature forecast error is typically caused by systematic error and is most associated with the elevated terrain; by contrast, precipitation, 10-m wind speed, and upper level forecast errors are dominated by the nonsystematic errors, which do not appear correlated with terrain.

25.5.2 Diurnal Variation

Based on model forecasts, the Southwest Asian domain-wide mean of the 2-m temperature exhibits a minimum near 0000 UTC followed by a sharp increase to a maximum near 1200 UTC (not shown). The difference of variables (temperature and wind speed) at maximum (1200 UTC) and minimum (0000 UTC) time is defined as the diurnal cycle variation in this study.

The Southwest Asia region's mean diurnal cycle of 2-m temperature during the 30-day study period (Fig. 25.10a) shows that the amplitude of temperature diurnal cycle for model forecasts is considerably lower than the value in the WMO GTS observations. Note a slight deepening in the diurnal temperature cycle on May 2, 7 and 17 in GTS observations that is not reproduced in the model forecasts. These two points indicate the near surface diurnal temperature cycle in model forecasts has a serious problem.

The 10-m model forecast wind speeds exhibit a different behavior from that of temperature. Similar to the NCEP GFS analysis data (GFS_ANL), the amplitude of the wind speed diurnal cycle in the model (Fig. 25.10b) shows a strong diurnal

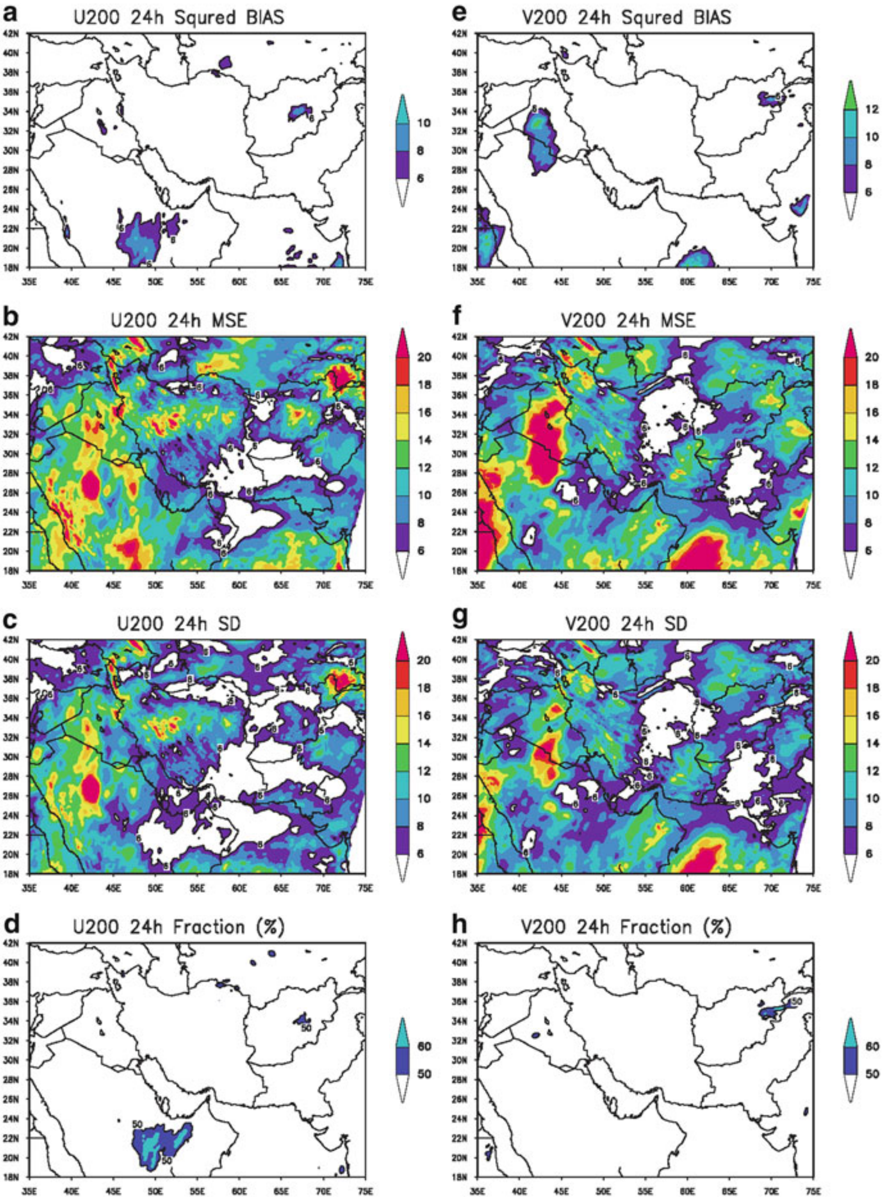


Fig. 25.9 Same as Fig. 25.3 but for 200 hPa zonal wind (**a–d**), meridional wind (**e–h**) and 24-h forecasts only (Updated from Xu et al. (2009))

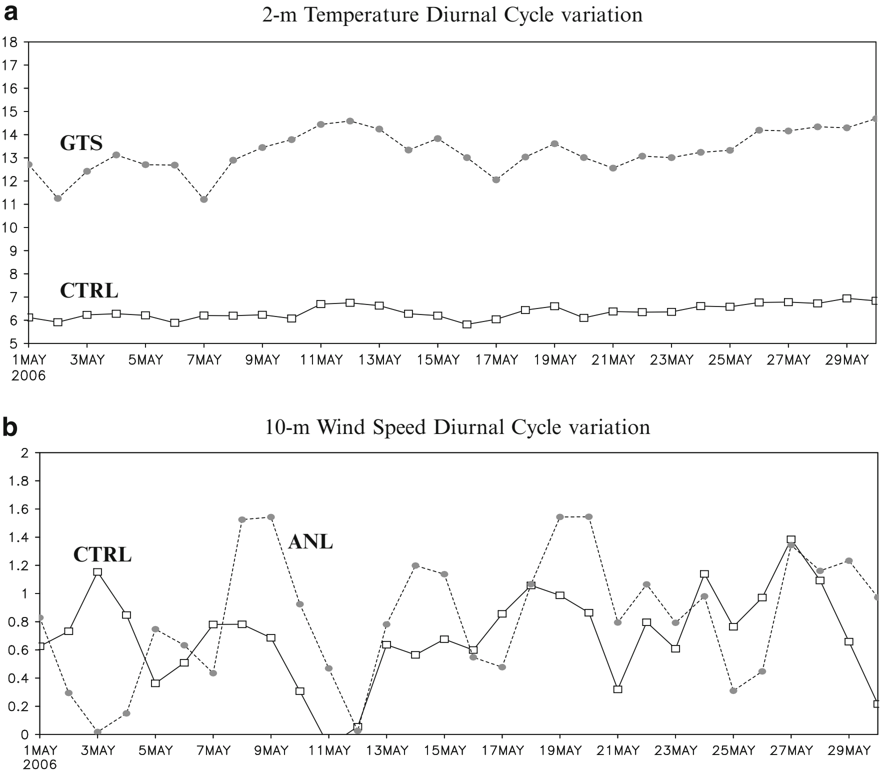


Fig. 25.10 Diurnal cycle variation, (**a**) 2-m temperature (°C), (**b**) 10-m wind speed (ms^{-1}). Twenty-four-hours forecasts only (Updated from Xu et al. (2009))

variation. The magnitude of model forecasts values are fairly consistent with the analysis values, except for the large difference on May 3. There is no evidence of a sharp gap between the model forecasts and the analysis data.

However, the spatial domain of the SWA areas appears to cover about four time zones. The whole domain average did not reflect significantly the diurnal cycle of the different regions. The 30-day period mean of diurnal cycle (Fig. 25.11) displayed the variation by region. The results show that over regions in the western and northeastern part of Southwest Asia, including the Saudi Arabian desert and northern border of Afghanistan, the model forecasts of 2-m temperature (Fig. 25.11b) are in much better agreement with the GTS observations (Fig. 25.11a) than in the Zagros mountains of western Iran, and Indian northwest deserts. Note the amplitude of the diurnal cycle in the model is much smaller than the GTS observations.

For the diurnal cycle variation in 10-m wind speed, the model forecasts (Fig. 25.11d) over Southwest Asia has a similar amplitude and distribution to the NCEP GFS analysis data (Fig. 25.11c) except for the clear mesoscale features in the

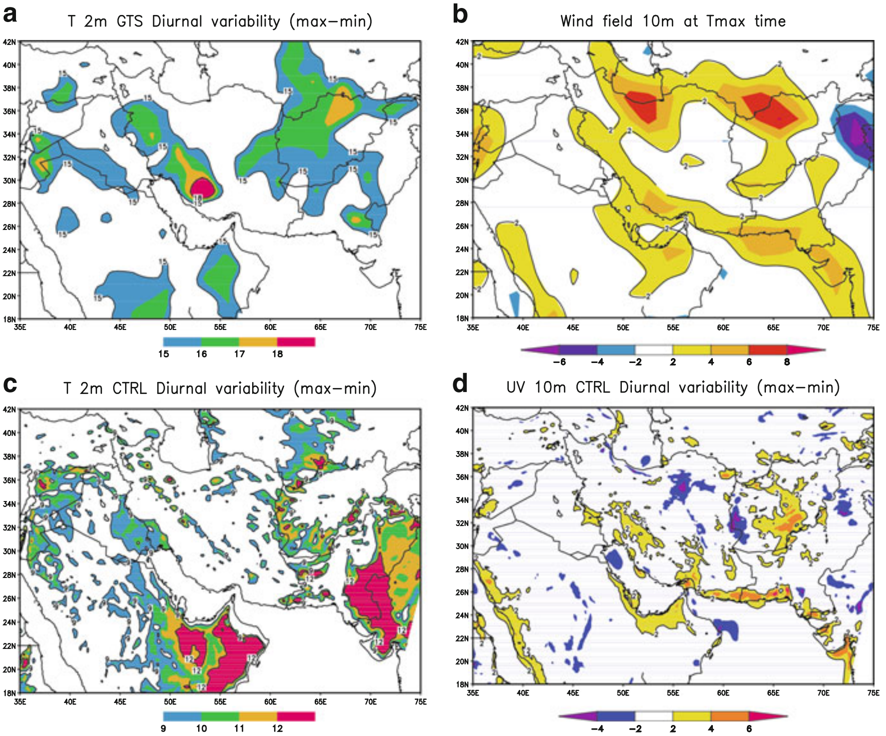

a
T 2m GTS Diurnal variability (max−min)
15
16
17
18
b
Wind field 10m at Tmax time
−6
−4
−2
2
4
6
8
c
T 2m CTRL Diurnal variability (max−min)
9
10
11
12
d
UV 10m CTRL Diurnal variability (max−min)
−4
−2
2
4
6
42N
40N
38N
36N
34N
32N
30N
28N
26N
24N
22N
20N
18N
35E
40E
45E
50E
55E
60E
65E
70E
75E

Fig. 25.11 Diurnal cycle variation of 2-m temperature (°C) in (**a**) GTS observation, (**b**) CTRL model forecasts, and 10-m wind speed (ms^{-1}) in (**c**) NCAR GFS analysis, (**d**) CTRL model forecasts. Twenty-four-hours forecasts only (Updated from Xu et al. (2009))

model forecasts. Note that the analysis data suggests a strong diurnal cycle variation over northwest Iran and north Afghanistan. Overall, however, the model forecasts of the diurnal cycle are consistent with the analysis.

25.6 Impact of Satellite Data Assimilation

Results from the previous section suggest that, aside from the 2-m temperature, errors in forecast variables are dominated by nonsystematic errors, which are caused by uncertainties in the model initial conditions or unresolvable differences in scales between the forecasts and observations (Nutter and Manobianco 1999). The model initial conditions are very important factors affecting model forecasts. For the purpose of understanding the role of initial conditions in the accuracy of forecasts, we will now consider satellite observation data assimilation in this section.

25.6.1 Experiment Design

In this study, the GSI analysis system is integrated with the WRF-ARW mesoscale system, and the Advanced TIROS-N (Television and Infrared Observation Satellite) Operational Vertical Sounder (ATOVS) radiance observations are employed. The ATOVS datasets supplied by National Environmental Satellite, Data, and Information Service (NESDIS) are composed of radiances from the Advanced Microwave Sounding Unit (AMSU) and the High-Resolution Infrared Sounder (HIRS)/3. Two separate radiometers (AMSU-A and AMSU-B) compose the AMSU platform. The AMSU-A is a cross-track, stepped-line scanning total power radiometer. The instrument has an instantaneous field-of-view of 3.3° at the half-power points providing a nominal spatial resolution at nadir of 48 km. The AMSU-B is a cross-track, continuous line scanning, total power radiometer and has an instantaneous field-of-view of 1.1° (at the half-power points). Spatial resolution at nadir is nominally 16 km. The antenna provides a cross-track scan, scanning ±48.95° from nadir with a total of 90 earth fields-of-view per scan line.

The AMSU-A and AMSU-B radiance data used here have undergone substantial preprocessing by NESDIS to remove various biases before being made available. The data have been statistically limb corrected (adjusted to nadir) and surface emissivity corrected in the microwave channels. Figure 25.12 shows an example of the scan position of the two microwave sensors of NOAA-15 and -16 during the study period. It is clear that NOAA-16 data covers most Southwest Asia and AMSU-B has a higher density of observations than AMSU-A.

Derber and Wu (1998) pointed out that the presence of a single data point containing large errors can result in substantial degradation of the analysis and subsequent forecast. For this reason, a simple quality control has been developed. To achieve similar radiances between instruments, the observed brightness temperature data have been modified empirically with different adjustment procedures for each instrument. In the GSI analysis system, this check includes two steps. First, a location check (including removal of observations outside the domain) and thinning procedure (excluding location/time duplicates and incomplete observations) is performed to ensure vertical consistency of upper-air profiles. Secondly, numerous quality control (QC) checks are redone based on various quality parameters after the model brightness temperatures are obtained from the radiative transfer model. The quality parameters are formulated in terms of the expected observational error variance as a function of the channels and have been adjusted for their position across the track of the scan, whether it is over land, sea, snow, sea ice, or a transition region, for elevation, the difference between the model and the real topography, and the latitude. In Fig. 25.13, the statistics show that the number of observations used in the GSI regional data assimilation system is quite different. AMSU-B has many more observations than the two AMSU-A platforms. For NOAA-15 (Fig. 25.13), the maximum number of AMSU-B observations for all 30 days range from 50,000 to 150,000 pixels, and for AMSU-A, the number is only around 40,000 pixels. For NOAA-16, the number of AMSU-B observations exceeded 150,000 pixels, while

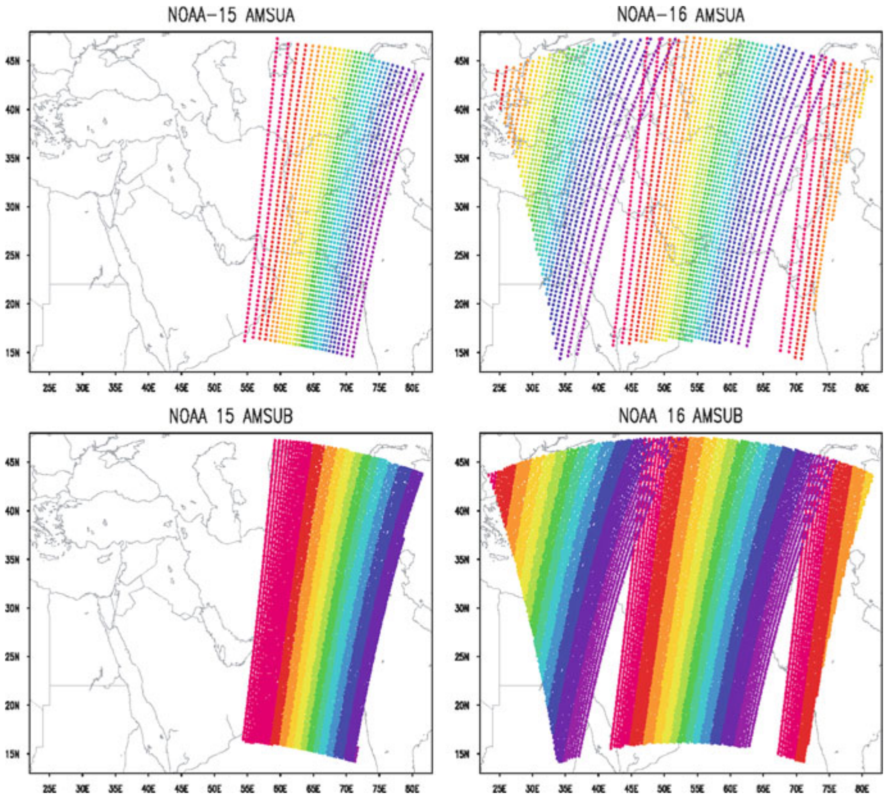


Fig. 25.12 Scan coverage of ATOVS (AMSU-A, AMSU-B) radiance being used in current data assimilation system at 00Z during May 2006 (Updated from Xu et al. (2009))

the AMSU-A was under 60,000 pixels. On average for the 30 days, the evidence shows through this two-step checking procedure, the amount of radiance data going into the model is reduced substantially. The percent usage of AMSU-A radiance data was over 40 %, but for AMSU-B it was only 16 %.

It is obvious that bias correction and quality control toss out non-useful data. This is less taxing on the minimization procedure within variational data assimilation systems However, because of the imperfections inherent in bias correction and quality control schemes, a lot of valuable observations are tossed out. Future studies should continue to refine good bias correction and quality control schemes.

For the control experiments described in Sect 25.2 (referred to as CTRL), the initial conditions generated for the GFS forecasts were assimilated using several different satellites, such as AMSU-A/B, High Resolution Infrared Radiation Sounder (HIRS), Microwave Sounding Unit (MSU) and so on. For the purpose of eliminating the effect of the radiance assimilation in the first guess field from GFS global analysis data, we first generated a spin-up run for 6 h from 18Z on previous day to 00Z on the forecast day in the data assimilation (referred to as DA) experiments; then the AMSU-A, AMSU-B radiance data are assimilated in the

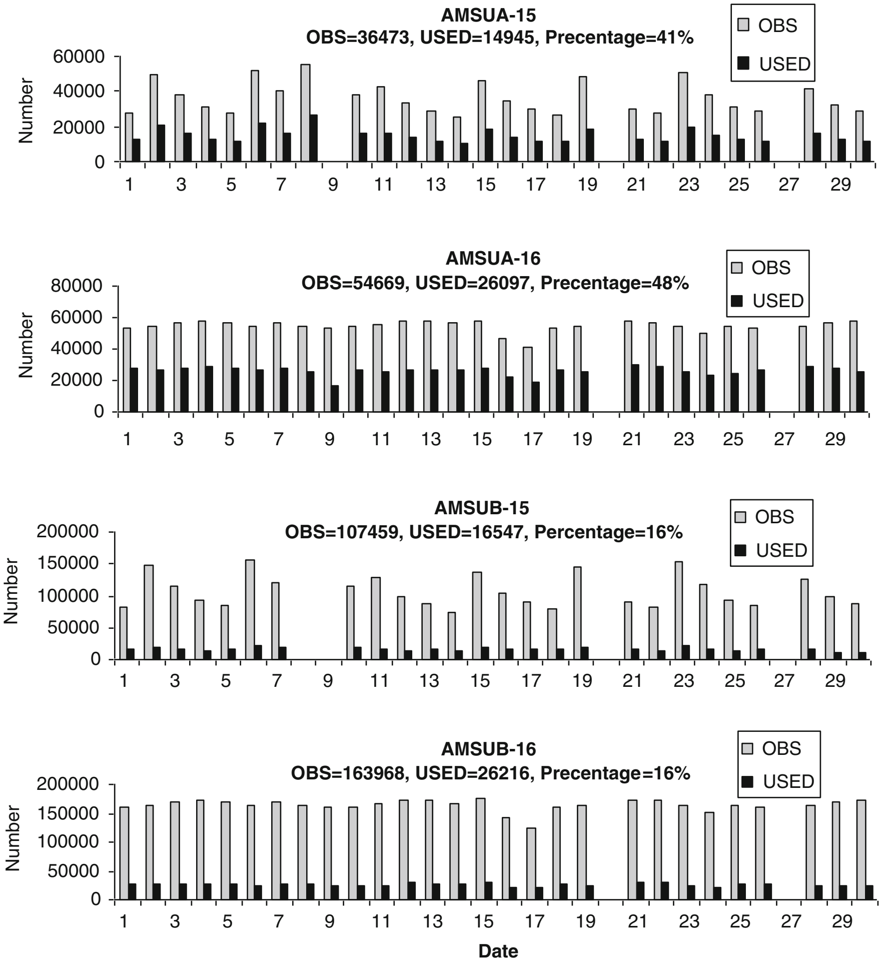


Fig. 25.13 Total number of radiance and used percentage in the forecast experiments as a function of date for AMSU-A and AMSU-B in NOAA 15 and NOAA 16, respectively (Updated from Xu et al. (2009))

ARW WRF forecast model to modify the initial condition at 00Z on each day, and then integrated using the same forecast lengths as in the CTRL experiment.

25.6.2 Results

To understand clearly the effect of ATOVS radiance data assimilation on the forecasts over Southwest Asia, three statistical parameters – bias, correlation, and mean square error skill scores – are calculated against the observation data.

25.6.2.1 Bias

The absolute bias difference between the DA and the CTRL experiment is defined as $|Bias|_{DA} - |Bias|_{CTRL}$. The 30-day's mean will be investigated first. For the 24-h forecasts, the absolute bias difference in 2-m temperature forecast (Fig. 25.14a) shows that the bias is reduced in DA over most of the Southwest Asia region. The bias in Iran, Afghanistan and Pakistan is on average 0.3–1.8 °C less than the CTRL forecasts, with the largest impact occurring on the south or southwest slope of the Afghanistan Hindu Kush mountain areas (see Fig. 25.1).

The absolute bias difference in 10-m wind speed for 24-h forecasts (Fig. 25.14b) reveals that the largest impacts to DA are over the Arabian Sea, Persian Gulf and the border between Pakistan and Afghanistan, where there are minimal high terrain effects. Whereas the impact of the satellite data assimilation on 2-m temperature is observed near the mountain areas, while the impact on the 10-m wind speed is apparentin places far away from these mountain regions, and especially over water areas. However, the evidence shows that the bias increased in many areas including southeast Iran, northwest India and the other areas.

Compared to the 24-h forecasts in the CTRL experiment, the precipitation forecast bias with DA (Fig. 25.14c) decreased slightly over the Mediterranean Sea, Black Sea coast, Saudi Arabian desert, and the Iranian Zagros mountain areas. However, the absolute bias of the precipitation forecast for the DA experiment increased over the Himalaya mountain areas.

For the upper levels, the absolute bias difference in 500 hPa temperature, geopotential height and wind field forecasts are presented in Fig. 25.15. The radiance data assimilation reduces the forecast bias of the geopotential height (Fig. 25.15b) and wind field (Fig. 25.15c, d) over most of Southwest Asia areas. These results show the impact of satellite radiance data assimilation on the upper level geopotential height and wind field forecasts are not associated with the configuration of terrain. Meanwhile, the 500 hPa temperature forecasts are modulated by the radiance assimilation differently (Fig. 25.15a). Here, the satellite data assimilation does not improve the temperature forecasts over the central Southwest Asia areas including Saudi Arabia, Iranian Zagros mountains and Afghanistan Hindu Kush mountains.

25.6.2.2 Mean-Square-Error Skill Scores

Murphy (1988) found forecasting skill scores are generally defined as measures of the relative accuracy of two forecasts, where one of the two forecasts is defined as a "reference system". For the following experiments, the CTRL forecasts are considered as the reference system. Based on the mean-square-error, the skill score (SS) can be expressed as follows:

$$SS(d, r, a) = 1 - [MSE\ (d, a)/MSE\ (r, a)] \tag{25.6}$$

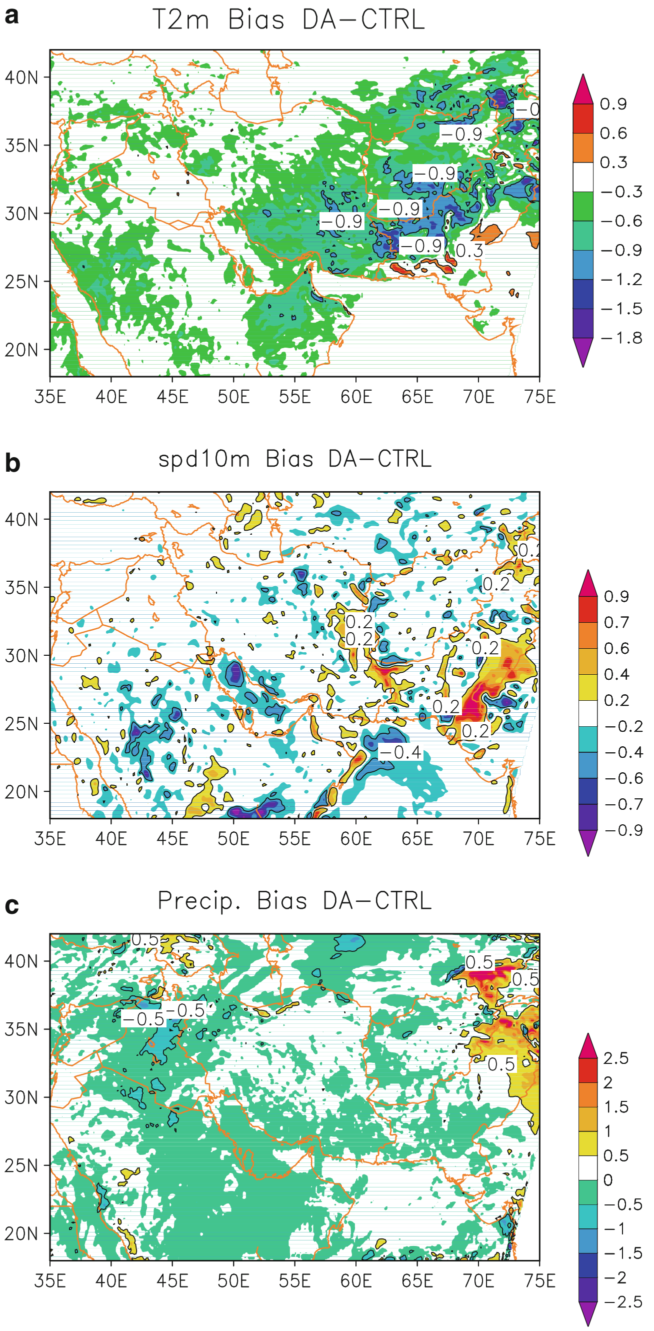


Fig. 25.14 Bias (model – observation) of 2-m temperature (**a**: °C), 10-m wind speed (**b**: ms^{-1}) and precipitation (**c**: mm/day) for 24-h forecasts averaged over the 1-month period of May, 2006 (Updated from Xu et al. (2009))

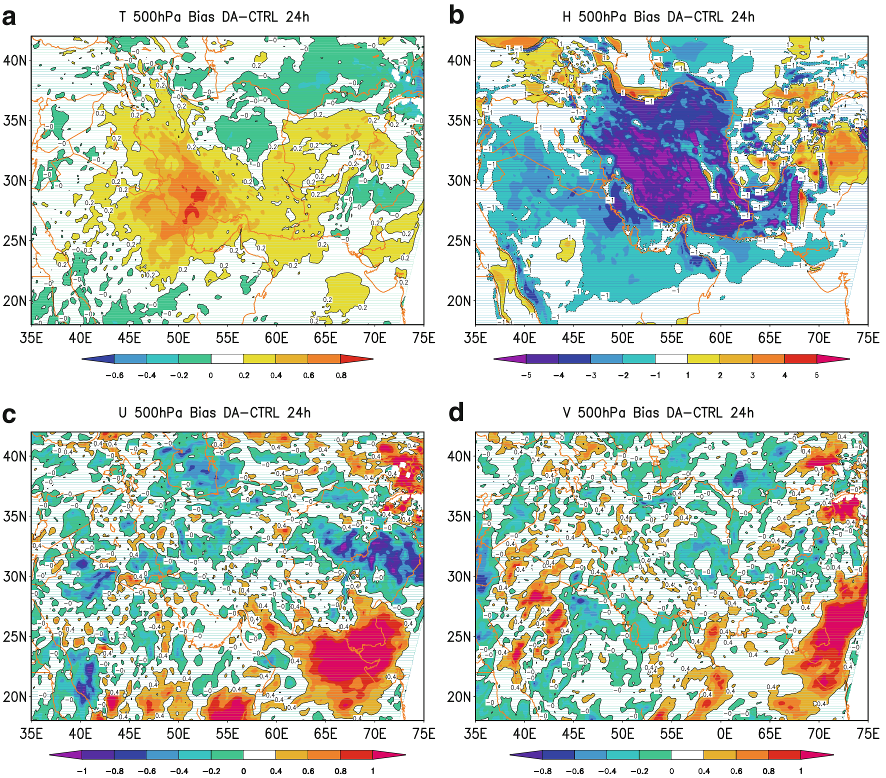


Fig. 25.15 Bias (model – observation) of temperature (**a**: °C), geopotential height (**b**: gpm), zonal wind (**c**: ms^{-1}) and meridional wind (**d**: ms^{-1}) at 500 hPa for 24-h forecasts averaged over the 1-month period of May, 2006 (Updated from Xu et al. (2009))

Note that SS in (25.6) is a function of the DA forecasts (d), the CTRL reference forecasts (r), and the analyzed quantity (a). The *MSE (d, a) and MSE (r, a)* are as defined in (25.2) indicating the mean-square-error of DA and CTRL forecasts relative to the analysis, respectively. Therefore, the greater positive SS values reflect increasing positive skill over the performance of the reference forecasts.

Figure 25.16 depicted the results for the 2-m temperature, 10-m wind speed and precipitation forecasts over the nine locations defined in Fig. 25.1. With regards to the 2-m temperature forecasts, the statistical analysis (Fig. 25.16a) indicates that all SS in the different locations are positive for the 24- and 48-h forecasts, but the SS for 48-h forecasts in most regions is greatly diminished in relation to that of the 24-h forecasts. The SS in the north Iranian Zagros Mountains (B) and west Himalaya Mountains (E) is about 10–20 % less than that in the lower mountains or plain areas. Compared with the results shown in Fig. 25.2, we find that the forecast errors in the high mountain areas are mainly from the model systematic errors and the nonsystematic errors make a relatively smaller contribution to the total forecast

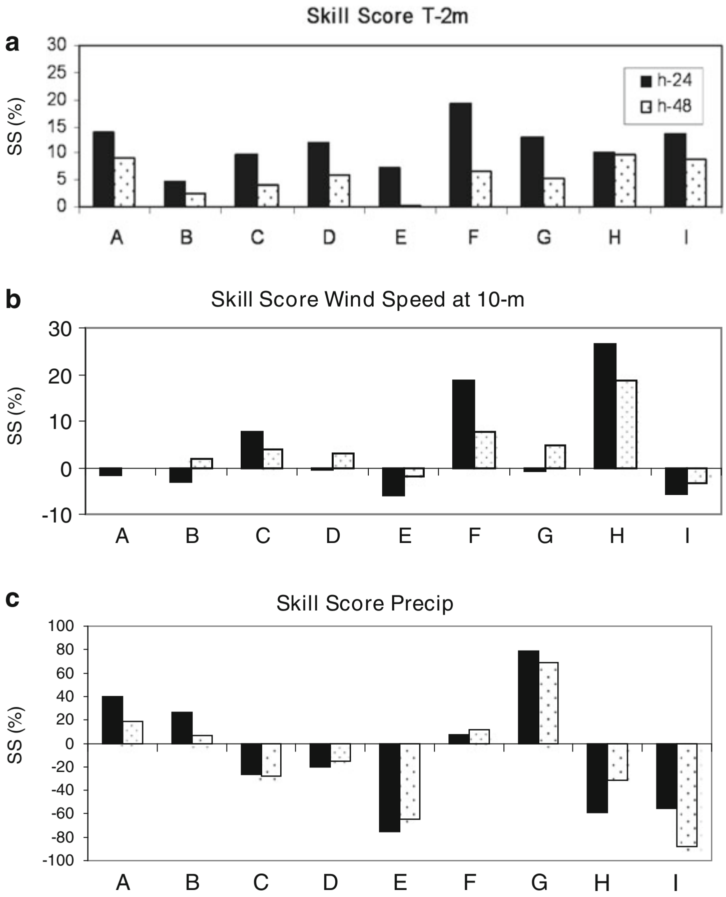


Fig. 25.16 Mean square error skill scores (*SS*) for 2-m temperature (**a**), 10-m wind speed (**b**) and 24 h accumulated precipitation (**c**). Results are plotted for averaged 24- and 48-h forecasts as a function of defined locations (Updated from Xu et al. (2009))

error. Satellite data assimilation, at least for the AMSU-A and AMSU-B radiances, seems not to make a significant contribution to the accuracy of surface temperature forecasts in the higher mountain areas.

In contrast, the 10-m wind speed in Fig. 25.16b shows a reverse SS value from the surface temperature. Six of nine locations including all high mountain areas (B, D, E) show a negative skill score, which means the satellite data assimilation produced a negative impact, but the SS in the Arabian Sea increases by 25 % and 20 % for 24- and 48-h forecasts, respectively. For the precipitation forecasts, the results suggest (Fig. 25.16c) that the satellite data assimilation only has a positive impacts on improvement of forecast biases over Iraq (A), North of Iran (B) and Saudi Arabia desert (F, G). The other five sub-regions become worse.

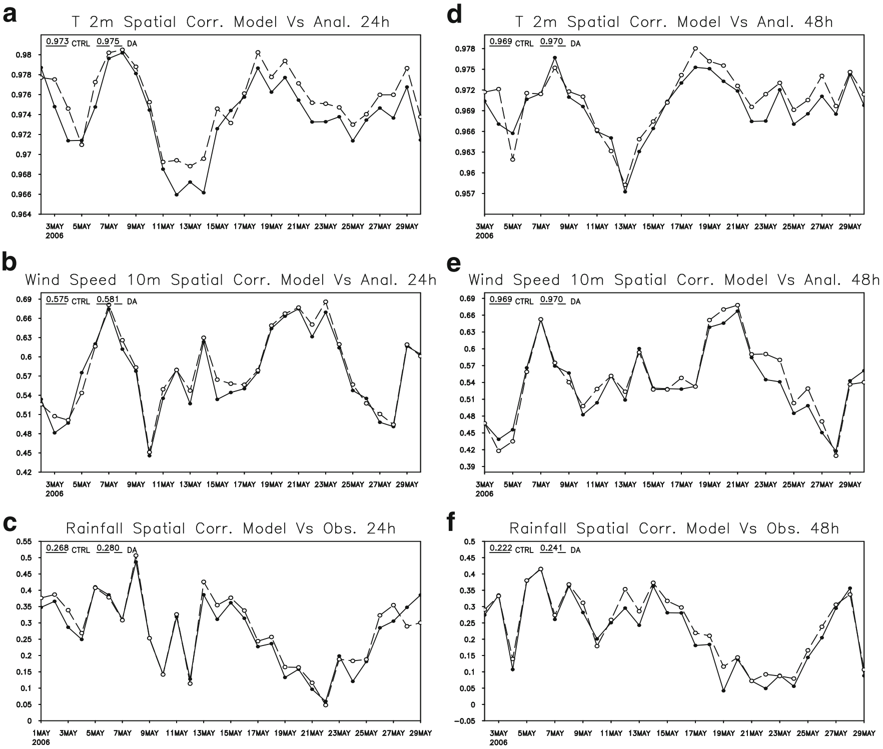


Fig. 25.17 Pattern correlation of model forecasts and observation of 2-m temperature, 10-m wind speed and rainfall for 24-h (**a–c**) and 48-h (**d–f**) forecasts (Updated from Xu et al. (2009))

25.6.2.3 Pattern Correlation

In order to evaluate the spatial agreement between the model and the observations quantitatively, pattern correlations (Walsh and McGregor 1997) were calculated between the model simulated and observed fields. The pattern correlation ρ_p of two spatial fields is simply the correlation of a series of points (i) from one field with corresponding values from the other field:

$$\rho_p = \frac{\sum (X_{oi} - \bar{X}_o)(X_{fi} - \bar{X}_f)}{\sqrt{\sum (X_{oi} - \bar{X}_o)^2}\sqrt{\sum (X_{fi} - \bar{X}_f)^2}} \tag{25.7}$$

where $\bar{X}_o$ and $\bar{X}_f$ are the means of the observational field (X_o) and model simulated field (X_f) fields respectively.

Figure 25.17 shows the pattern correlation of observational and model forecasted values for 2-m temperature, 10-m wind speed and precipitation over the whole prediction domain. The pattern correlation coefficient between observations and

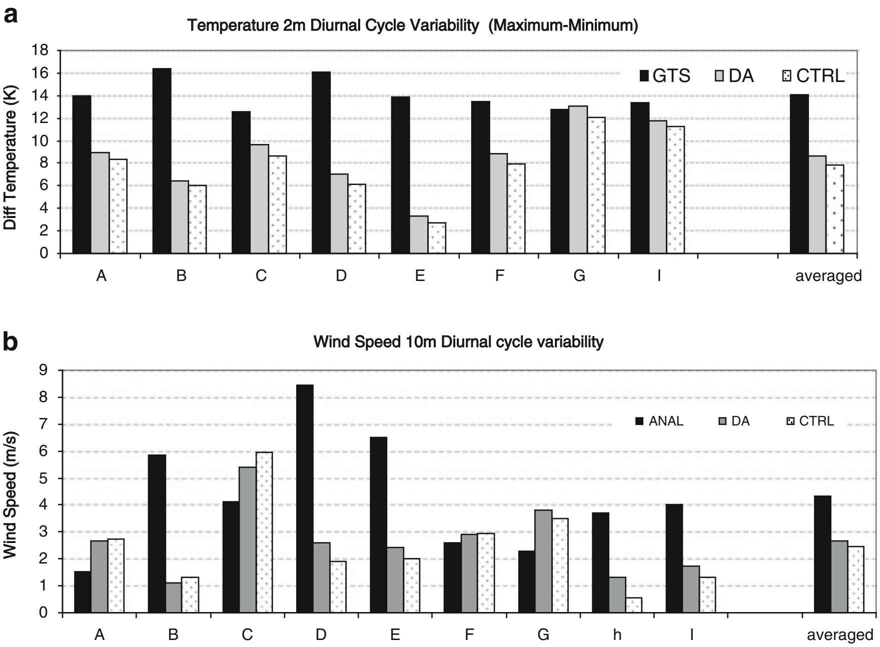


Fig. 25.18 Diurnal variation over the sub-regions of Southwest Asia shown in Fig. 25.1 for (**a**) 2-m temperature (°C) and (**b**) 10-m wind speed (ms^{-1}) (Updated from Xu et al. (2009))

model forecasts of these three surface variables increases slightly after the satellite data assimilation for 24- and 48-h forecast. For a 30-day average in 24-h forecasts (Fig. 25.17a–c), the correlation coefficient in the CTRL gets to 0.973, 0.268 and 0.575 for 2-m temperature, 10-m wind speed and precipitation, respectively. The corresponding values for the DA experiment are 0.975, 0.280 and 0.581. The 48-h forecasts have show results (Fig. 25.17d–f). The results indicate that the forecast pattern improvement is very limited although the correlation coefficient increases in the DA experiment.

25.6.2.4 Diurnal Variation of Near Surface Temperature and Wind Field

The analysis of near surface temperature and wind field variability is based on the eight selected sub-regions (the Arabian Sea (H) was omitted due to the lack of GTS temperature data there). The diurnal variation of the 30-day mean 2-m temperature is presented in Fig. 25.18. It is apparent that the amplitude of the diurnal cycle in model forecasts of temperature in the CTRL and DA are relatively lower than in the GTS observations over seven of eight selected sub-regions (Fig. 25.18a). Note that the amplitude of the diurnal cycle in the DA experiment is closer to the GTS observations than in the CTRL experiment. These results demonstrate that the diurnal cycle of surface air temperature can be improved slightly by the assimilation of satellite radiance data.

For the analysis of the 10-m wind fields, the reference data used is still the NCEP GFS analysis data. In contrast to surface temperature, it is not readily apparent that the amplitude of the diurnal cycle has been improved in DA (Fig. 25.18b). The performance is quite different in these selected sub-regions. The diurnal cycle of the wind speed (Fig. 25.18b) in the analysis data is considerably larger than in the model forecasts over the five sub-regions B, D, E, H, and I, where B, D and E are three high mountain sub-regions. But it is clear that the amplitude of the diurnal cycle in the DA experiment has been modified closer to the analysis data.

25.7 Summary and Discussion

25.7.1 Summary

This paper presented an objective verification and impact of radiance data assimilation on weather forecasts over the complex terrain areas of Southwest Asia using the National Center for Atmospheric Research (NCAR) mesoscale model (WRF-ARW) and Joint Center for Satellite Data Assimilation (JCSDA) GSI analysis system. The numerical experiments are conducted for a one month period May 2006. The results are summarized as follows:

The model biases caused by inadequate parameterization of physical processes, except for the 2-m temperature, are relatively small compared to the nonsystematic errors resulting, in part, from the uncertainty in the initial conditions. The total forecast errors at the surface show a substantial spatial heterogeneity; there is a relatively larger error in the higher mountain areas. However, the sources of the error indicate a unique difference between temperature, precipitation and wind speed. While the error in 2-m temperature is mainly from systematic error, which is largely controlled by the physical representation of terrain (i.e., the errors are positively correlated with terrain elevation); the errors in 10-m wind speed and precipitation have a greater contribution from nonsystematic error, which is more likely related to uncertainty in the initial conditions.

The amplitude of the diurnal cycle of the model 2-m temperature is much smaller than the GTS observations. However, the model forecasts of the diurnal cycle are consistent with the NCEP GFS analysis data. There is no evidence of a sharp gap between the model forecasts and the analysis data.

The ATOVS satellite data provides useful information for improving the initial conditions, and the model error was reduced to some degree. The bias and mean square error skill score (SS) shows that satellite data assimilation produces a better forecast over some areas; however, it seems not to make a significant contribution to the accuracy of forecasts in the higher mountain areas. Although the improvement in correlation coefficient growth is very small, the forecast patterns are improved in the DA experiment,.

25.7.2 Discussion

In this study, the weather forecasts using the WRF-ARW system were evaluated over the mountain areas of Southwest Asia. Due to the complexity of the high terrain and lack of knowledge in the estimation of physical processes in this area, forecasters should have greater awareness of these limitations of the model forecasts in this region.

First of all, the parameterization of physical processes plays a significant role in the forecasting of surface temperature. For the 2-m temperature forecasts, the systematic error component is larger than the random errors, and it is related to the elevation of terrain. It should be noted that the areas of high bias shown in Fig. 25.3a correspond with the areas of rapid elevation change. These are the areas where a difference in terrain height between the datasets would have the largest effect. They are also the areas where the difference between the observational station elevation and mean grid point elevation has the largest value. The lapse rate effects due to these terrain height differences is probably another reason for the 2 m temperature bias. In contrast with the temperature fields, random errors play a much bigger role in the forecasting of the upper level precipitation and 10-m wind fields. The random errors constrain forecasters from presenting high quality forecast guidance and are caused by a combination of uncertainty in the initial conditions and unreasonable model scales. The detailed statistical results presented in Sect. 25.4 are specific to the surface and the upper levels at nine locations. The basic error characteristics for one forecasting variable change by the selected region, and may not be representative of errors of other forecast variables. For example, in the preliminary investigation of temperature errors, the results demonstrated that the maximum 2-m temperature biases occurred over the high mountain areas while the temperature biases at 500 hPa were found over most of Southwest Asia and it was not related to the terrain configuration.

Note that the results presented here are for only one month of experimental model runs; the accuracy of the forecast performance needs to be further verified and investigated with more real-time forecasts. As expressed by Manning and Davis (1997), "These statistics would provide additional information to model users and alert model developers to those research areas that need more attention." The additional and complementary need for verification strategies in the WRF-ARW model is elucidated in reference papers (Skamarock et al. 2005).

Second, random errors are very complicated. It is only partially attributed to the uncertainty in initial conditions. An accurate representation of initial conditions would help users to compare the latest forecast guidance with current observations and make appropriate adjustments in real time. The assimilation of satellite radiance observations into a numerical weather prediction (NWP) model provides initial conditions more closely representative of the true state of the atmosphere. The results shown here demonstrate the positive impact of satellite data on weather prediction in most of the Southwest Asia areas, but the impacts are not as obvious in the high terrain areas, such as the Himalaya Mountain and Iranian Mountain regions.

This feature implies that the random errors arise not only from the uncertainty in initial conditions, but also from another parameter like the resolution of the model horizontal scale. This issue will be investigated in future work.

Acknowledgements The authors thank Dr. Liu from NCAR and Dr. Russ Treadon from JCSDA for answering many questions regarding the WRF-ARW and GSI data assimilation system.

This work was supported by the National Oceanic and Atmospheric Administration (NOAA), National Environmental Satellite, Data, and Information Service (NESDIS), Center for Satellite Applications and Research (STAR). The views, opinions, and findings contained in this publication are those of the authors and should not be considered an official NOAA or U.S. Government position, policy, or decision.

References

Agrawala S, Barlow M, Cullen H, Lyon B (2001) The drought and humanitarian crisis in central and Southwest Asia: a climate perspective. Published by International Research Institute for climate prediction (IRI), Lamont-Doherty Earth Observatory of Columbia University, Palisades, IRI special report 01-11

Andersson E, Hollingsworth A, Kelly G, Lönnberg P, Pailleux J, Zhang Z (1991) Global observing system experiments on operational statistical retrievals of satellite sounding data. Mon Wea Rev 119:1851–1864

Bouttier F, Kelly G (2001) Observing-system experiments in the ECMWF 4D-Var data assimilation system. Quart J. Roy Meteor Soc 127(Part B):1469–1488

Cheng YY, James Steenburgh W (2005) Evaluation of surface sensible weather forecasts by the WRF and the Eta models over the western United States. Wea Forecast 20:812–821

Courtier P et al (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: formulation. Quart J Roy Meteor Soc 124:1783–1807

Derber JC, Parrish DF, Lord SJ (1991) The new global operational analysis system at the National Meteorological Center. Wea Forecast 6:538–547

Derber JC, Wu W-S (1998) The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. Mon Wea Rev 126:2287–2299

Evans JP, Smith R (2001) Modelling the climate of South West Asia. In: Ghassemi F et al (eds) Proceedings international congress on modelling and simulation, MODSIM01, Australian National University, Canberra, Dec 10–13

Evans JP, Smith R (2006) Water vapor transport and the production of precipitation in the Eastern Fertile Crescent. J Hydrometeorol 7:1295–1307

Evans JP, Smith RB, Oglesby RJ (2004) Middle East climate simulation and dominant precipitation processes. Int J Clima 24:1671–1694

Eyre JR, Kelly G, McNally AP, Andersson E, Persson A (1993) Assimilation of TOVS radiances through one dimensional variational analysis. Quart J Roy Meteor Soc 119:1427–1463

Gallus WA, Bresch James F (2006) Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. Mon Wea Rev 134:2632–2641

Halem M, Kalnay E, Baker WE, Atlas R (1982) An assessment of the FGGE satellite observing system during SOP-1. Bull Am Meteor Soc 63:407–429

Kleist DT, Parrish DF, Derber JC, Treadon R, Errico RM, Yang R (2009a) Improving incremental balance in the GSI 3DVAR analysis system. Mon Wea Rev 137:1046–1060. http://dx.doi.org/10.1175/2008MWR2623.1

Kleist DT, Parrish DF, Derber JC, Treadon R, Wan-Shu W, Lord S (2009b) Introduction of the GSI into the NCEP Global Data Assimilation System. Wea Forecast 24:1691–1705. http://dx.doi.org/10.1175/2009WAF2222201.1

Liu Q, Weng F (2006) Detecting the warm core of a hurricane from the special sensor microwave imager sounder. Geophys Res Lett 33:L06817. doi:10.1029/2005GL025246

Lorenc AC (1986) Analysis methods for numerical weather prediction. Quart J Roy Meteor Soc 112:1177–1194

Manning KW, Davis CA (1997) Verification and sensitivity experiments for the WISP95 MM5 forecasts. Wea Forecast 12:719–735

Marcella MP, Eltahir EA (2008) Modeling the hydroclimatology of Kuwait: the role of subcloud evaporation in semiarid climates. J Clim 21(12):2976–2989

McNally AP, Derber JC, Wu W, Katz BB (2000) The use of TOVS level-1b radiances in the NCEP SSI analysis system. Quart J Roy Meteor Soc 126:689–724

Michalakes J, Chen S, Dudhia J, Hart L, Klemp J, Middlecoff J, Skamarock W (2001) Development of a next generation regional weather research and forecast model. In: Zwieflhofer W, Kreitz N (eds) Developments in teracomputing: proceedings of the ninth ECMWF workshop on the use of high performance computing in meteorology, World Scientific, pp 269–276

Mo KC, Wang XL, Kistler R, Kanamitsu M, Kalnay E (1995) Impact of satellite data on the CDAS–reanalysis system. Mon Wea Rev 123:124–139

Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon Wea Rev 116:2417–2424

Nutter PA, Manobianco J (1999) Evaluation of the 29-km Eta model. Part I: objective verification at three selected stations. Wea Forecast 14:5–17

Parrish DF, Derber JC (1992) The National Meteorological Center's spectral statistical interpolation analysis system. Mon Wea Rev 120:1747–1763

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG (2005) A description of the advanced research WRF version 2. NCAR technical note NCAR/TN-468+STR, 94

Stauffer DR, Seaman NL, Binkowski FS (1991) Use of four-dimensional data assimilation in a limited-area mesoscale model part II: effects of data assimilation within the planetary boundary layer. Mon Wea Rev 119:734–754

Tracton MS, Desmarais AJ, van Haaren RJ, McPherson RD (1980) The impact of satellite soundings on the National Meteorological Center's analysis and forecast system—the Data Systems Test results. Mon Wea Rev 108:543–586

Walsh K, McGregor J (1997) An assessment of simulations of climate variability over Australia with a limited area model. Int J Clim 17:201–223

Wan Q, Xu J (2011) A numerical study of the rainstorm characteristics of the June 2005 flash flood with WRF/GSI data assimilation system over south-east China. Hydrol Process 25:1327–1341. doi:10.1002/hyp.7882

Wu WS, Purser RJ, Parrish DF (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. Mon Weather Rev 130:2905–2916

Xie P, and Arkin PA, (1996) Analysis of global monthly precipitation using gauge observations, satellite estimates, and numerical model prediction. J Clim 9:840–858

Xie P, Rudolf B, Schneider U, Arkin P (1996) Gauge-based monthly analysis of global land precipitation from 1971 to 1994. J Geophys Res 101(D14):19023–19034

Xu J, Rugg S, Horner M, Byerle L (2009) Application of ATOVS radiance with ARW WRF/GSI data assimilation system in the prediction of hurricane Katrina. Open Atmos Sci J 2:288–303

Xu J, Powell A (2012) Dynamical downscaling precipitation over the Southwest Asian: impacts of radiance data assimilation on the hindcasts of the WRF-ARW model. Atmos Res 111:90–103

Zaitchik BF, Evans JP, Smith RB (2007a) Regional impact of an elevated heat source: the Zagros plateau of Iran. J Clim 20:4133–4146

Zaitchik BF, Evans JP, Geerken RA, Smith RB (2007b) Climate and vegetation in the Middle East: inter-annual variability and drought feedbacks. J Clim 20:3924–3941

Chapter 26
Studies on the Impacts of 3D-VAR Assimilation of Satellite Observations on the Simulation of Monsoon Depressions over India

A. Chandrasekar and M. Govindan Kutty

Abstract Variational data assimilation provides a convenient means of optimally combining the "first-guess" or "background" meteorological fields with the observations. The background fields are typically obtained from the numerical weather prediction output of a model while the observations can be either the meteorological model variables or even non-model variables. In the three-dimensional variational (3D-VAR) method the analysis state is obtained by optimally combining the "first-guess" and the "observations" at the same analysis time.

The present article begins with a brief overview of the characteristics of the monsoon disturbances that form over the Indian region during the summer monsoon season. Subsequently, the 3D-VAR method is briefly introduced together with details of the mesoscale model employed in this study. The next section outlines the results of the impact of the 3D-VAR assimilation of satellite observations in the simulation of a few monsoon disturbances over India using the Weather Research and Forecast (WRF) model. The satellite observations utilized in the 3D-VAR assimilation study presented in this article include (1) temperature and humidity profiles from Moderate Resolution Imaging Spectroradiometer (MODIS), (2) temperature and humidity profiles from Advanced TIROS Vertical Sounder (ATOVS), and (3) total precipitable water from Special Sensor microwave imager (SSMI), respectively. In order to discern the impact of 3D-VAR assimilation of satellite observations a (base or control) numerical experiment called "control run" is performed, which is identical to the assimilated run (called "3D-VAR run") except

A. Chandrasekar (✉)
Department of Earth and Space Sciences, Indian Institute of Space Science and Technology,
Valiamala, Thiruvananthapuram, 695547, India
e-mail: chandra@iist.ac.in

M.G. Kutty
Center for Analysis & Prediction of Storms, National Weather Centre,
University of Oklahoma, 73072, USA
e-mail: kutty@ou.edu

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_26,

that no observations are assimilated in the control run. The results of the simulation between the assimilated run and the control run are compared with one another as well as with global analysis and Tropical Rainfall Measurement Mission (TRMM) and Quick Scatterometer (QuikSCAT) observations.

The results of the study indicate that the assimilation of satellite observations, in general, does improve the simulation of the various monsoon disturbances over India, although the improvements are not uniformly very marked for all the monsoon disturbances and for all the satellite observations. Assimilating MODIS temperature and humidity profiles have yielded better results for two of the depressions as compared to the ATOVS and SSM/I assimilations. Also the results of the study indicate that assimilating total precipitable water from SSM/I has lower impact as compared to assimilating temperature and humidity profiles from ATOVS and MODIS.

26.1 Introduction

26.1.1 Southwest Indian Monsoon

The planetary monsoon circulation is one of the most important components of the large-scale circulation over the tropical regions (B. Wang 2006). The Indian summer monsoon circulation is one of the most spectacular manifestations of the global planetary monsoon circulations. In addition to the planetary scale nature of the Indian summer monsoon, there are several weather systems (also called '*monsoon disturbances*'), such as *monsoon depressions* and *low-pressure systems* which are embedded within the overall planetary scale monsoon circulation. The monsoon disturbances that form over India during the June to September summer monsoon season, includes systems such as monsoon depression, and low-pressure systems such as the *onset vortex*, the *mid-tropospheric cyclone* and *offshore trough/vortex* respectively. The above-mentioned monsoon disturbances not only provide for copious rainfall over several regions over India, but also contribute significantly to the seasonal Indian monsoon rainfall. The monsoon disturbances are usually associated with the strengthening of the *monsoon trough* over India and in most situations, herald the "*active phase*" of the Indian summer monsoon rainfall (Krishnamurti et al. 1977; Krishnamurti and Ardanuy 1980; Saha et al. 1981). The monsoon trough is a trough in the surface pressure chart over India oriented along the northwest-southeast direction from the heat low over Pakistan to the north Bay of Bengal and is seen during the summer monsoon season. When the axis of the monsoon trough is south of its normal position with its eastern end dipping into the Bay of Bengal, active summer monsoon conditions prevail over India. This active phase causes heavy rainfall over the plains of north India, central parts of India as well as along the Indian west coast. Furthermore, during such active phases, monsoon depressions are known to form over the north Bay of Bengal.

26.1.2 Monsoon Depression

The monsoon depressions (Sikka 1977) are systems that are intermediate in terms of intensity between the relatively weak low-pressure systems that have wind speeds less than 8.5 m s^{-1} and the tropical cyclones that have associated wind speeds exceeding 17 m s^{-1}. The preferred region of formation of these monsoon depressions over India is between 20° and 30°N and 80° and 90°E. Typically the average number of monsoon depressions that can form over India during the summer monsoon months of June to September is about 6 with the month of August accounting for about two monsoon depressions. The maximum number of monsoon depressions, that forms over India is however higher. While some of the monsoon depressions owe their origin to weak easterly waves traveling from the east, the other depressions can develop in situ over the North Bay of Bengal. The monsoon depressions that form over North Bay of Bengal have a horizontal radial extent of 1,000 km. Monsoon depressions are systems known to be typically cold core below 700 hPa and warm core aloft. Because of the above fact, the strongest winds associated with the monsoon depressions are observed near 700 hPa. At higher levels, the cyclonic circulation associated with the monsoon depression weakens and is absent at and above 300 hPa. The location of the center of the monsoon depression slopes south-westward with height. The winds associated with the monsoon depressions are asymmetric with stronger winds south of the depression center at low levels. The maximum horizontal convergence of moist air together with the associated maximum spatial precipitation pattern is found in the south-west sector of the monsoon depression. The principal zone of heaviest precipitation associated with a monsoon depression occurs at about 200–400 km away from the center while a secondary zone of relatively lower rainfall is seen at about 800 km to the west of the depression center.

Typically the monsoon depression moves in the west-to-west to north-west direction (Mooley and Shukla 1989). During the months of June and September, the movements of the monsoon depression can follow either the northerly direction or they can recurve over the Bay of Bengal. However, during the other 2 months of July and August, most of the monsoon depressions move in the west-north-westerly direction over India. During the month of July, the average speed of a monsoon depression is between 1.2 and 2.4 ms^{-1} to the east of 85°E while the average speed of the depression is between 4.8 and 9.6 ms^{-1}, to the west of 85°E. The average life period of a monsoon depression is about 5 days for a depression which has formed over the Bay of Bengal while the same is about 3 days for a depression that has formed over the Arabian Sea and over land.

The monsoon depressions (Shukla 1978) which form during the Indian summer monsoon season do not generally intensify into a tropical cyclone. The existence of low level westerly winds together with strong upper level easterlies during the Indian summer monsoon season is responsible for the existence of strong wind shear in the vertical. Such strong vertical wind shears do not aid in the manifestation of penetrative convection, the latter essential for the formation of large scale organization of cumulus convection associated with a tropical cyclone. Monsoon depressions generally weaken in intensity after reaching the central parts of India. These then weaken to low pressure systems and move in a west-northwest direction and merge with the seasonal low over northwest India (Sikka 1980).

26.2 Data Assimilation

Geophysical (atmospheric/oceanic) information is essentially utilized to test hypothesis (i.e. testing our understanding of the system), attribute cause and effect (i.e. to understand the cause of geophysical events) and to make forecasts, i.e. to predict future geophysical events (Lahoz et al. 2010). Broadly, the geophysical information is available through two broad sources, namely, (1) "observation" which are nothing but measurements of the geophysical system, and (2) "models", which have been built based on the earlier "measurements" gathered of the system as well as our understanding of the evolution of the geophysical system. It is true that both observations and models have errors. The model errors arise due to the fact that the models are imperfect in the sense that our understanding of the physical processes associated with the geophysical system is somewhat "incomplete". Furthermore, model errors also appear due to the need to limit resolution of the digitization of the continuous governing equations due to computational costs. The observational errors are characterized as random, systematic and also due to representativeness (Lahoz et al. 2010). Furthermore, the "observations" have gaps since the measurements of the system are in general discrete in space and time. It is logical to fill the gaps in the "observation" by using the information based on the behaviour of the system, namely the "models". A methodology of objectively combining "observations" and "model" information to yield an "optimum" or "best estimate" of the geophysical system is called "data assimilation".

The basic premise in the data assimilation methodology is that combining "observations" and "model" information together with knowledge of their respective errors will yield combined information that is more valuable than the individual information, provided the process of combining both the information is robust. In data assimilation, the model takes the information from the observation and propagates this information to unobserved regions successfully filling in the so called "gaps" in the observation. Data assimilation can also provide estimation of unobserved quantities. While Panofsky (1949) utilized polynomial functions to fit to the observation values, in the early development of objective analysis of meteorological data, Gilchrist and Cressman (1954) improved the above method by introducing the concept of "region of influence" for each observation. Gilchrist and Cressman also proposed the use of a background field from a previous forecast. Bergthorsson and Doos (1955) optimized the weights given to each observation based on the accuracy of the various types of observations while Cressman (1959) proposed variation of above method involving multiple iterations of the analysis. This was followed by data assimilation method based on "optimal interpolation"

(OI) in which the weights given to observations were related to observation errors (Gandin 1963).

Furthermore, the OI method considered and utilized the importance of the background field information and its error characteristics as useful source of information. The OI method, when first implemented in operational centers worldwide in late 1970s and early 1980s, had to invoke major approximations in order to meet the calculations feasible. The advent of variational methods for data assimilation in the mid 1980s saw the emergence of an important breakthrough in data assimilation research.

26.2.1 Variational Data Assimilation

The underlying physical principle of variational data assimilation schemes is that the analysis $\mathbf{x}^{\mathrm{a}}$ is the optimum state vector that minimizes a global "cost function" J, the latter providing a measure of the mismatch between a model state vector $\mathbf{x}$ and the background state $\mathbf{x}^{\mathrm{b}}$ and observation $\mathbf{y}$. This cost function as utilized in three-dimensional variational (3D-VAR) method is given as

$$\mathbf{J} = 0.5\,[\mathbf{x} - \mathbf{x}^{\mathrm{b}}]^{\mathrm{T}}\mathbf{B}^{-1}[\mathbf{x} - \mathbf{x}^{\mathrm{b}}] + 0.5\,[\mathbf{y} - \mathrm{H}(\mathbf{x})]^{\mathrm{T}}\,\mathbf{R}^{-1}\,[\mathbf{y} - \mathrm{H}(\mathbf{x})] \tag{26.1}$$

where $\mathbf{B}$ is error covariance of background state, $\mathbf{R}$ is error covariance of observation (including the representativeness errors) and T indicates transpose. The 3D-VAR method was implemented operationally in mid 1990s at the National Centre for Environmental Prediction (NCEP) first and later at European Centre for Medium Range Weather Forecast (ECMWF). The minimization of cost function J in 3D-VAR is usually performed in "control space". The error covariance of the background state $\mathbf{B}$ is usually estimated from the difference between pairs of forecasts that verify at the same time (Parrish and Derber 1992), the so called "NMC" method. The observations are assumed to have no bias and no serious errors associated with the malfunctioning of instruments. The observation errors are assumed to be Gaussian. The 3D-VAR method assumes that all observations are valid at the same time and further assumes that the background errors and observation errors are not correlated. An important advantage of the variational method is that one can utilize observation variables which are different from the model state variables.

An important extension of the 3D-VAR method is called the "four-dimensional variational method" (4D-VAR) in which the cost function minimization is performed over a time window, the latter accounting for observations spread over time and lasting typically 6 h or 12 h for operational weather forecasts. The cost function in the 4D-VAR method is as follows

$$\mathbf{J} = 0.5\,[\mathbf{x}_{\mathbf{o}} - \mathbf{x}_{\mathbf{o}}^{\mathrm{b}}]^{\mathrm{T}}\mathbf{B}_{\mathbf{o}}^{-1}[\mathbf{x}_{\mathbf{o}} - \mathbf{x}_{\mathbf{o}}^{\mathrm{b}}] + 0.5\sum_{i=1}^{N}\,[\mathbf{y}_{\mathbf{i}} - \mathrm{H}(\mathbf{x}_{\mathbf{i}})]^{\mathrm{T}}\,\mathbf{R}^{-1}\,[\mathbf{y}_{\mathbf{i}} - \mathrm{H}(\mathbf{x}_{\mathbf{i}})] \tag{26.2}$$

In the 4D-VAR method it is necessary to estimate the time evolution of the perturbation using a linear model and to calculate the adjoint of the above linear model. Furthermore the 4D-VAR method is considered to be computationally intensive since this involves forward integration of the model and backward integration of the adjoint model, over many iterations to obtain the minimization of cost function. In this chapter, results of our studies on the impacts of 3D-VAR assimilation of satellite observations on the simulation of monsoon depressions over India are provided. Although 4D-VAR is superior, our study has been restricted to 3D-VAR for the following reasons, (1) we felt that it is more appropriate and prudent to take up 3D-VAR studies first rather than go for 4D-VAR, and (2) the computational costs associated with the study of 4D-VAR.

26.2.2 Assimilation of Satellite Observation

The basic problem in numerical weather prediction (NWP) is that the observations are at least two orders smaller than the number of degrees of freedom of the model. Most meteorological systems form over the sea which is a data sparse region. Satellites provide an excellent platform to obtain observations of the atmosphere over the sea. Unlike the conventional observations such as radiosondes/rawinsondes, the quantities measured by satellites do not directly relate to the atmospheric quantities such as temperature, humidity, wind direction, wind speed, etc. What essentially satellite measures are the radiation that reaches the top of the atmosphere at given frequencies in the case of passive radiometers and the back scattered radiation emitted by a surface (say a sea surface) in the case of active scatterometer.

The most common of satellite observations to be assimilated in a "data assimilation" methodology is the satellite derived vertical air temperature and humidity profiles. Other important meteorological observations obtained from satellite are the sea surface temperature (SST), surface wind speed and wind direction over the sea, rainfall rate, total precipitable water, cloud motion wind vector (CMV) at different levels of the atmosphere. While the satellite derived temperature and humidity profiles can be directly assimilated in a NWP model, the variational method allows for the direct assimilation of satellite radiance observation. For direct satellite radiance assimilation, the observational operator H incorporates a radiative transfer model that maps the atmospheric profile to radiance space. The above procedure of directly assimilating satellite radiance is better since radiance errors are more easily characterized than the retrieval errors. The following subsections provide brief information of the various satellite sensors (QuikSCAT, SSMI, ATOVS and MODIS) which are normally utilized in 3D-VAR assimilation impact studies.

26.2.3 Various Satellite Observations

26.2.3.1 QuikSCAT

The National Aeronautical and Space Administration (NASA) launched the Quick Scatterometer (QuikSCAT) in June 1999 at an altitude of 800 km with a swath width of 1,800 km and having an orbital period of 101 min in a polar orbiting configuration. The sensor (Seawinds) aboard the QuikSCAT is a 13.4 GHz Ku-band conical-scanning microwave radiometer which measures the ocean surface wind vector from the relationship between the sea surface roughness and the back scattered radar signal. The accuracy of the retrieved ocean surface wind from QuikSCAT is about 2 m s^{-1} in wind speed and 20° in wind direction for winds of magnitude 3–20 m s^{-1} (Shirtliffe 1999). It is known that rainfall can affect the accuracy of the scatterometer sea surface wind measurements (Weissman et al. 2002; Hoffman and Leidner 2005). While in the scatterometer, light winds can get overestimated by excess back scatter from the rain, strong winds can be underestimated due to rainfall attenuation. The QuikSCAT Operational Standard Data Products (L2B) are being processed and distributed by NASA Jet Propulsion Laboratory (JPL) Physical Oceanography Distributed Active Archive Center (PO DAAC). The sea surface wind vectors retrieved from QuikSCAT have been validated with wind data from ocean buoys and were found to be in good agreement with the buoy data (Ebuchi et al. 2002) with the root mean square (rms) differences of about 1.01 m s^{-1} and 23° for the wind speed and wind direction, respectively. The horizontal spatial resolution of the QuikSCAT data wind vector data is 25 km while the reference height of the surface wind vector from QuikSCAT is 10 m.

26.2.3.2 Spectral Sensor Microwave Imager (SSM/I)

The SSM/I sensor was first launched aboard the polar orbiting Defense Meteorological Satellite Program (DMSP) of the Unites States Navy in June 1987. The SSM/I sensor is a conical scanning, linearly polarized, four-frequency and seven channel passive microwave radiometer. The SSM/I sensor aboard the DMSP satellite has an orbital period of 102 min. The SSM/I sensor also has an incidence angle of 53°, a swath width of 1,400 km, a mean altitude of 830 km and a horizontal spatial resolution of 25 km. Detailed information about the SSM/I sensor is given in Hollinger (1989). The retrieved total precipitable water (TPW) observation from the SSM/I satellite is used as observations in our recent study whose results will be presented in the following section. Also, version–5 multistage regression algorithm is used for the retrieval of SSM/I data products.

26.2.3.3 Advanced TIROS Operational Vertical Sounder (ATOVS)

The Advanced (Television and Infrared Observational Satellite) TIROS Operational Vertical Sounder (ATOVS) is a sounding instrument package, first flown on the

National Oceanic and Atmospheric Administration (NOAA)-KLM satellite series. The ATOVS sensor comprises of the Advanced Microwave Sounding Units A and B (AMSU-A, AMSU-B), and the High Resolution Infrared Radiation Sounder (HIRS/3). The NOAA-TOVS has three infrared channels at 8.3, 7.3, and 6.7 μm and can provide moisture information for the following three layers: 1,000–700 hPa, 700–500 hPa, and 500–300 hPa. The variable scale-height algorithm is used to derive the vertical humidity profile at standard levels 1,000, 850, 700, 500, 400 and 300 hPa from the precipitable water vapor in the three layers. The method and its validation are outlined in Rajan et al. (2002). Rajan et al. (2002) found that the associated root mean square error (RMSE) when validated against near radiosonde observations was less than 10 %. The temperature sounding data used in this study are from the HIRS/2 instrument onboard TOVS. TOVS satellite makes a morning and an evening pass (around 7:30 A.M. and 7:30 P.M. Indian Standard Time) over the Indian subcontinent. Hence, the TOVS temperature and humidity profiles are ingested at the nearest analysis time, i.e., at 00 and 12 UTC.

ATOVS data archived in the NOAA Comprehensive Large Array Stewardship system (CLASS) is available in the raw Level 1b format. The ATOVS data has been quality controlled, and assembled into discrete data sets, and to which the calibration information as well as the information of location on Earth are appended. This Sounding and Imager Data from the High Resolution Picture Transmission (HRPT) direct read out stream of NOAA-ATOVS satellite is processed end-to-end using the ATOVS and AVHRR (Advanced Very High Resolution Radiometer) Processing Package (AAPP). The output of AAPP is called the Level 1d data, in which factors such as instrument reflectance and/or brightness temperatures are mapped on a common instrument grid (HIRS in this study) with navigation, calibration and contamination information appended. This Level 1d data is then fed to the International ATOVS Processing Package (IAPP); to retrieve bias corrected parameters for assimilation. The data has a horizontal resolution of about 42 km and gives temperature and humidity sounding in 42 vertical levels, up to 10 hPa level. Satellite data observed and pre-processed in the time window of ±1.5 to ±2 h from the assimilation time are ingested. English et al. (2000) investigated the impact of assimilating ATOVS data in a numerical weather prediction (NWP) model, and concluded that the information provided by the radiance observations reduced the forecast errors by about 20 % in the southern hemisphere and by about 5 % in the northern hemisphere.

26.2.3.4 Moderate Resolution Imaging Spectroradiometer (MODIS)

The “Moderate Resolution Imaging Spectroradiometer” (MODIS) is a key instrument onboard the Terra and Aqua satellites launched in 1999 and 2002, respectively. Both Terra MODIS and Aqua MODIS are viewing the entire Earth’s surface every 1–2 days, acquiring data over several wavelengths in two important regions of the electromagnetic spectrum, namely the near-infrared (NIR) and the infrared (IR) to monitor atmospheric temperature and moisture, respectively. The MODIS data is extremely valuable as it can be utilized to improve our understanding of the processes in the Earth system encompassing the land, oceans, and in the lower atmosphere of the Planet Earth. The MODIS observations are indeed performing a vital role in the development of globally, interactive Earth system models which are utilized to predict the global changes accurately. The predicted global changes can assist policy framers in arriving at sound and pragmatic decisions concerning the protection of the Earth's environment. A brief technical description of the MODIS instrument is as follows. The MODIS instrument provides high radiometric sensitivity in 36 spectral bands ranging in wavelength from 0.4 to 14.4 μm. Two bands are imaged at a nominal resolution of 250 m at nadir, with five bands at 500 m, and the remaining 29 bands are at 1 km resolution. The MODIS temperature and water vapour profiles consists of 30 gridded variables related to atmospheric stability, atmospheric temperature, moisture profiles, total atmospheric water vapour and total ozone. All the above-mentioned variables are available during both day time and night time conditions at 5 km pixel resolution whenever nine field of view (FOV) pixels or more are cloud free. The atmospheric temperature and humidity profiles available at high spatial resolution from MODIS provides an extensive source of information on the atmospheric structure in clear skies and hence can be utilized most fruitfully to improve the initial state of the atmosphere. A validation of temperature and humidity profiles with concurrent Arabian Sea Monsoon experiment (ARMEX) Global Positioning System (GPS) radiosonde data was performed during July 2002. The root mean square error (RMSE) of temperature below 500 hPa was 1–2.5° K while the RMSE for the specific humidity profile was less than 2 gkg^{-1} (Simon and Rahman 2003). The temperature and humidity profiles available on 14 vertical levels were utilized in this study.

26.2.3.5 Tropical Rainfall Measuring Mission (TRMM)

The Tropical Rainfall Measuring Mission (TRMM) sensor provides for a broad sampling footprint between 35°N and 35°S, and is responsible for the detailed and comprehensive dataset on the space and time distribution of rainfall and latent heating over the oceanic and tropical continental regions. The TRMM algorithm 3B42 provides adjusted 24 h cumulative estimates of the rain using merged microwave and infrared precipitation information (Adler et al. 2000). The TRMM adjusted Geostationary Observational Environmental satellite precipitation index (AGPI) is produced by using cases of coincident TRMM combined instruments with TRMM Microwave Imager (TMI) and Precipitation Radar (PR) algorithm (Haddad et al. 1997). The 3B42 algorithm provides for a three hourly rain rate at 0.25° × 0.25° horizontal resolution. In this study the TRMM data obtained by using the 3B42V6 algorithm was utilized for the validation of model predicted rainfall.

26.3 Literature Study of Earlier Assimilation Studies

Numerical Weather Prediction, especially in the short-range requires accurate initial conditions. A large number of studies have shown that assimilating various observations such as satellite data, Doppler Weather Data have improved the initial conditions and have resulted in better model performance (Gal-Chen et al. 1986; Lipton and Vonder Haar 1990; Lipton et al. 1995; Ruggiero et al. 1999; Zou and Xiao 2000; Pu et al. 2002; Fan and Tilley 2005; Chou et al. 2006; Chen 2007; Zhang et al. 2007; Zapotocny et al. 2007; Govindankutty et al. 2008; Kelly et al. 2008; Singh et al. 2008a; Singh et al. 2008b; Brennan et al. 2009; Rakesh et al. 2009a; Sinha and Chandrasekar 2010; Singh et al. 2010; Govindankutty et al. 2010; Singh et al. 2011a,b; Kumar et al. 2011; Singh et al. 2011c). Chen (2007) investigated and compared the impact of assimilating SSM/I, and the QuikSCAT satellite surface winds, on the simulations of Hurricane Isidore. The results of the above study indicated that the increment of the QuikSCAT wind analysis was higher than that from the SSM/I analysis. Furthermore, the results also showed that the increase in low-level wind speeds enhanced the air-sea interaction processes and improved the simulated intensity for the hurricane in the assimilated QuikSCAT run. Also, the non-availability of the surface wind direction information from the SSM/I data resulted in less improved simulation as compared to the QuikSCAT assimilated run. The above study showed that the position of the center of hurricane over the ocean which is usually misrepresented at the model initial time can be improved due to assimilation of high-resolution surface wind information.

Rakesh et al. (2009b) investigated the impact of assimilating QuikSCAT surface wind vectors, SSM/I wind speed and the Total Precipitable Water (TPW) for forecasts of wind, temperature, and humidity from 1 month long assimilation experiments during July 2006. In the above study, the control (without assimilation of satellite data) as well as 3D-Var sensitivity experiments (with assimilation of satellite data) using MM5/WRF were made for 48 h starting daily at 0000 UTC July 2006. Rakesh et al. (2009b) utilized the control run results as a baseline for assessing the impact of MM5/WRF 3D-Var satellite data sensitivity experiments. Rakesh et al. (2009b) found from their results that the forecast errors in predicted wind, temperature and humidity at different levels are lower in WRF model as compared to the MM5 model, except for the temperature prediction at lower level. Also, their results indicated that the rainfall pattern and prediction skill from day one and day two forecasts by WRF model is superior to the MM5 model. Furthermore, Rakesh et al. (2009b) found that the spatial distribution of forecast impact for wind, temperature, and humidity fields showed that on an average, for 24 and 48-h forecasts, the assimilation of satellite data did improve the MM5/WRF initial conditions and resulted in reduced errors of the predicted meteorological fields. Among the assimilation experiments, MM5/WRF wind speed prediction was found most beneficial due to ingestion of QuikSCAT surface wind and SSM/I TPW data while for the temperature and humidity prediction the improvement in assimilation is mostly due to ingestion of SSM/I TPW data (Rakesh et al. 2009b). The results of the study also indicated that the largest improvement in the MM5/WRF rainfall prediction was associated with the SSM/I TPW assimilation. The assimilation of only the SSM/I wind speed in MM5/WRF model, however resulted in some degradation in the simulation of the humidity and rainfall fields (Rakesh et al. 2009b).

Singh et al. (2011a) have recently investigated the impact of ATOVS radiance on the analysis and forecasts of WRF model over the Indian region during the 2008 summer monsoon and found a positive impact of the assimilated ATOVS radiance on both the analysis as well as the short-range forecasts. The above study, in addition to the control (no assimilation) run, also utilized satellite radiances from AMSU-A, AMSU-B and HIRS sensors. Singh et al. (2011b) have also compared the performances of Kalpana and HIRS water vapor radiances in the WRF 3D-Var assimilation system for the period 10–20 July 2008 over the Indian region and found that the assimilation of Kalpana radiances provided significant improvements to the results as compared to the assimilation of HIRS radiances. Singh et al. (2011c) investigated the impact of assimilating Oceansat2 surface wind vectors for the month of July 2010 over the Indian region. The results of the above study indicated that the assimilation of Oceansat2 surface wind vectors led to small, but positive, impact on the forecast (particularly later hours of forecast) of mid-tropospheric moisture, temperature, and upper tropospheric winds. Also, the assimilation of Oceansat2 surface wind vectors improved the precipitation forecast (as compared to the control run) for moderate to heavy rainfall thresholds when validated against TRMM rainfall.

26.4 Case Studies of the Impact of Satellite Data Assimilation Using WRF-3D-VAR System on Three Monsoon Depressions That Formed Over India

This section investigates the impact of assimilation of the vertical profiles of air temperature, and humidity from the MODIS and the ATOVS, and total precipitable water (TPW) from the SSM/I sensors using the WRF model. Three cases of monsoon depressions have been considered for the above investigation in the present study: (1) a monsoon depression that formed over the Bay of Bengal, between September 19–22, 2006, (2) a monsoon depression that formed over the Bay of Bengal, between September 02–05, 2006, and (3) a monsoon depression that formed over the Bay of Bengal, between June 18–22, 2007. The following sections represent descriptions and features of the monsoon depressions investigated, the experimental design, together with the results and discussions as well as conclusions of the study.

26.4.1 Monsoon Depression That Formed During 19 to 22 September 2006

A low pressure area formed over the north-east Bay off Arakan coast and the adjoining east-central Bay in the evening of 18 September 2006. The system was over the north-east Bay on 19 September 2006 and became well marked in the same evening. The low pressure area moved over Gangetic West Bengal and adjoining Bay on 20 September 2006 and subsequently intensified into a depression on 21 September 2006 03 UTC, close to Jamshedpur, India. The depression remained stationary over Jamshedpur till 21 September 2006 12 UTC and then moved slightly north-westwards and was centered in Jharkhand, about 50 km east of Ranchi, at 03 UTC and 12 UTC of 22 September 2006. The depression moved in the north-eastward direction and lay centered close to Dhanbad on 23 September 2006 03 and 12 UTC. The land depression caused heavy rainfall over the north-eastern and the central parts of India during 21–24 September 2006.

26.4.1.1 Numerical Experiment

The WRF model is configured with 24 vertical levels and with two domains of 36 and 12 km grid spacing, using a two way nesting option. The number of grid cells in the east–west (EW)-north–south (NS) direction being, 118 × 130 and 271 × 271 for the coarser and finer resolutions, respectively. The physics options employed in the WRF model utilized in this study include the WRF Single Moment (WSM) class-3 simple ice scheme for micro physics, Rapid Radiative Transfer Model (RRTM) scheme for long wave radiation, NOAH land surface model for land surface, Kain-Fritsch scheme for cumulus parameterization and the Yonsei University (YSU) scheme for the Planetary boundary layer parameterization scheme. The NCEP-GFS forecast data available at a horizontal resolution of $1^\circ \times 1^\circ$ and a time resolution of 6 h have been used to develop the initial and lateral boundary conditions. Four numerical experiments are performed to study the impact of temperature, humidity and total precipitable water separately in the simulated structure of monsoon depression. All the simulations are started with the same initial conditions on 18 September 2006 18 UTC, but the observations are assimilated at different times depending on the availability of the satellite data over the domain. For a given analysis/forecast time the satellite observations that fall in ±90 min window is assimilated. The first experiment, called the control (CTRL) run, has utilized the NCEP-GFS data for creating initial and lateral boundary conditions. The model integrations are performed from 18 September 2006 18 UTC to 22 September 2006 12 UTC without any assimilation of satellite observations. The second experiment, called the "ATOVS run" which started at 18 UTC on 18 September 2006 and has assimilated the ATOVS temperature and humidity profiles into the model using

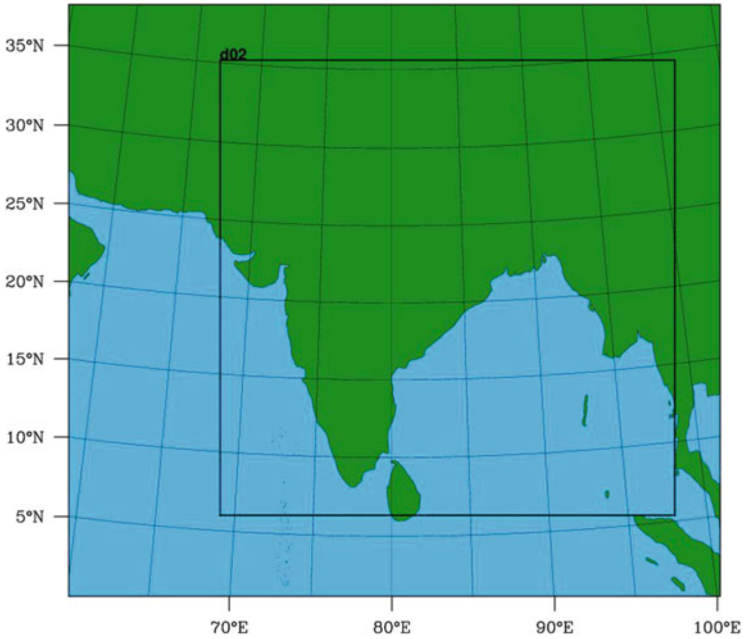


Fig. 26.1 The 36 km (outer domain) and 12 km (inner domain) utilized in this study

3D-VAR after 6 h of forecast; i.e., from 19 September 2006 00 UTC and has ingested the ATOVS observations in a 12 h interval up to 00 UTC 20 September 2006. The model is subsequently integrated for the next 60 h in a free forecast mode without any further assimilation of the satellite data. The third experiment is named as the "MODIS run" in which the vertical profiles of temperature and humidity observations from the MODIS satellite have been assimilated in 12 h interval from 18 September 2006 18 UTC to 19 September 2006 18 UTC and the model was subsequently run in a free forecast mode up to 22 September 2006 12 UTC. The high resolution MODIS data is subjected to "thinning" before ingesting the data to the observational pre-processor. The fourth experiment, named as the "SSM/I run" is similar to the ATOVS experiment except that it incorporates SSM/I total precipitable water (TPW) instead of ATOVS temperature and humidity observations. All the model runs are subjected to a 6 h interval data cycling to maintain the dynamical consistency of the model simulation The CTRL run is subjected to data cycling without any assimilation of observations (Chen et al. 2008). Figure 26.1 shows the model domains with the outer domain depicting horizontal grid resolution of 36 km while the inner domain shows the 12 km horizontal resolution. All the results shown in this article are from the finer 12 km resolution only.

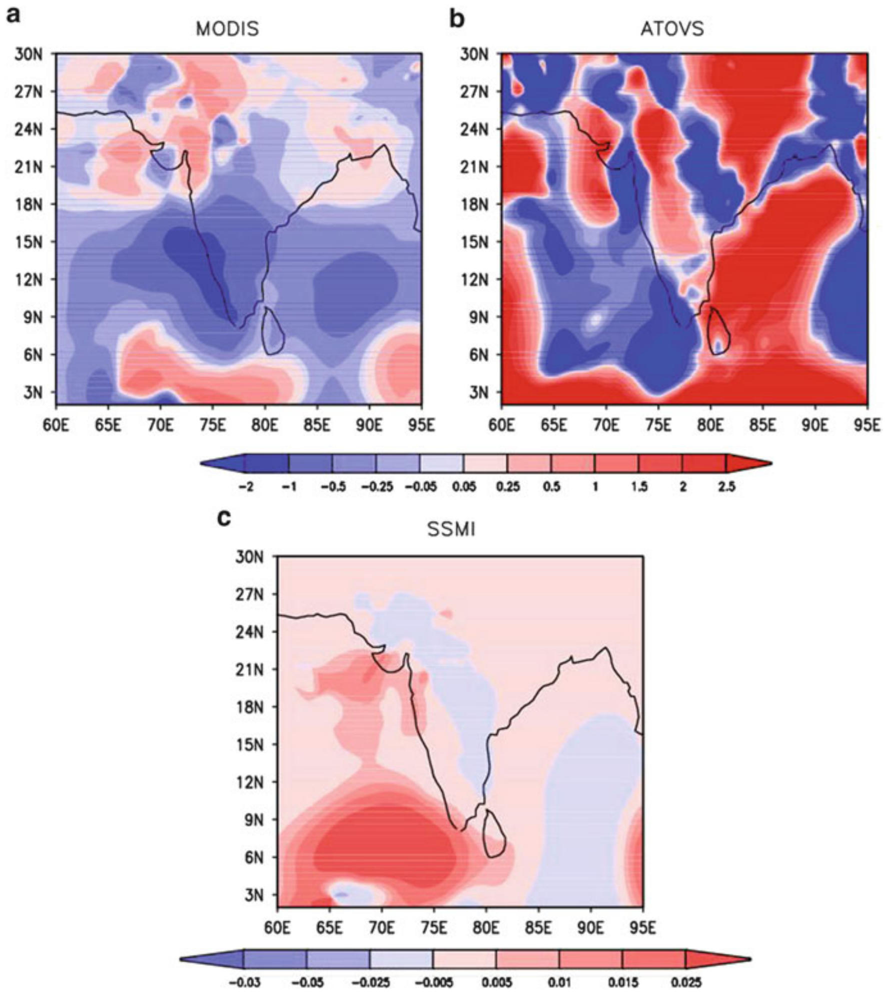


Fig. 26.2 Analysis increment at 850 hPa in the wind speed of (**a**) MODIS, (**b**) ATOVS and (**c**) SSM/I for 19–22 September 2006 depression

26.4.1.2 Results and Discussion

Initial Conditions

The impact of assimilation of satellite observations on the initial conditions can be obtained by calculating the analysis increment. Analysis increment can be defined as the difference in the model variables before and after the assimilation of the observations. It does not give a qualitative or quantitative verification whether the initial condition from satellite data assimilation is better or worse, but provides a measure of how much the observation impacted the initial analysis. Figure 26.2

depicts the analysis increments due to assimilation of the satellite observations such as MODIS, ATOVS and SSM/I on the 850 hPa wind speed; the figure shows clear differences in the wind speed at 850 hPa. In contrast to the increment due to the assimilation of SSM/I TPW, which are negligibly small, the assimilation of MODIS and ATOVS temperature and humidity profiles have shown significant and marked analysis increment values over the domain of study. Among the three 3D-VAR experiments, the ATOVS 3D-VAR assimilation has resulted in the largest spatial extent of positive value of increments up to 2.5 ms^{-1} and the largest negative values of increments up to $-2\,\mathrm{ms}^{-1}$ when compared to the MODIS and the SSM/I 3D-VAR experiments.

Mean Sea Level Pressure (MSLP) fields

Figure 26.3 shows the MSLP (a–d) and the lower tropospheric wind at 850 hPa (e–h) from NCEP-FNL (Final Analysis), and 24 h accumulated precipitation from TRMM satellite (i and j) for the 19–22 September 2006 depression. The depression is initially shown as a low pressure area with a minimum central MSLP of 1,000 hPa, which later intensified into a depression with a central minimum pressure of 996 hPa as evident from Fig. 26.3a–d. The maximum accumulated precipitation of TRMM satellite observations are seen over the west coast of the Bay of Bengal during 20–22 September 2006.

Figure 26.4a–p depicts the simulated MSLP patterns on 20 September 2006 00 UTC and subsequent predicted values at 24, 48 and 60 h of forecast for the CTRL experiment (Fig. 26.4a–d), the ATOVS experiment (Fig. 26.4e–h), the SSM/I (Fig. 26.4i–l) experiment and the MODIS (Fig. 26.4m–p) experiment, respectively. The simulated MSLP fields are then compared with the NCEP FNL analysis. The CTRL run simulates an intense depression with a minimum central pressure of 990 hPa. The system as simulated in the CTRL run intensifies in the 24, 48 and 60 h of forecast. The ATOVS experiment simulates the weakest depression of all the four numerical experiments, with a lowest central minimum pressure of 998 hPa. The simulated MSLP pattern of the SSM/I experiment is similar to that of the CTRL experiment, clearly indicating that the ingestion of the SSM/I total precipitable water has not produced significant impact in the numerical simulation. The MODIS experiment however simulates a well organised monsoon depression with the well known west-north-westward movement of the depression (Godbole 1977). Also, the results of the MODIS experiment simulate the depression well in terms of the location of the depression center and not with respect to the intensity. Among the four experiments, the MODIS run shows the maximum inland penetration, though there are some inadequacies in accurately simulating the location of the depression centre vis-a-vis the NCEP-FNL analysis. In terms of intensity of MSLP, the ATOVS experiment results are quite close to the NCEP-FNL values than the other three experiments.

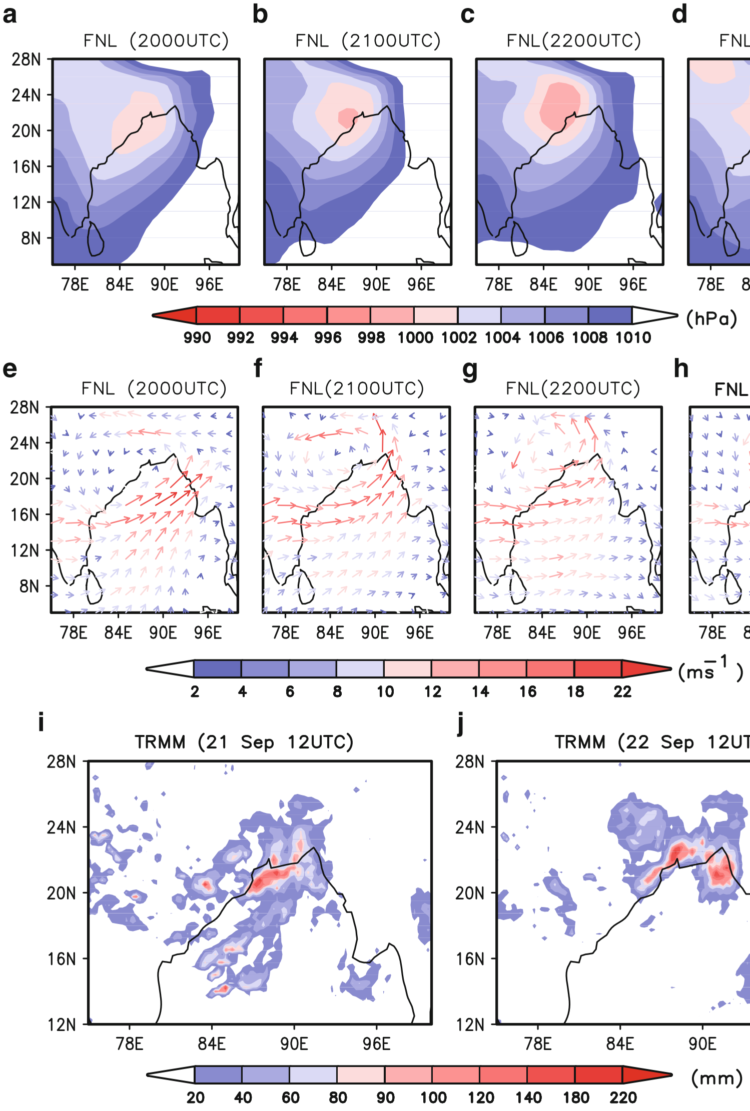


Fig. 26.3 NCEP-FNL analysis sea level pressure fields at 00 UTC on 20–22 September 2006 (**a**–**d**), NCEP-FNL lower tropospheric wind speed (850 hPa) (**e**–**h**) and 24 h accumulated TRMM rainfall (**i**–**j**) for 19–22 September 2006 depression

Wind Speed

The lower tropospheric wind vectors simulated by all the four experiments at 850 hPa are shown in Fig. 26.5 which includes the CTRL run (Fig. 26.5a–d), ATOVS experiment (Fig. 26.5e–h), SSM/I experiment (Fig. 26.5i–l) and MODIS experiment

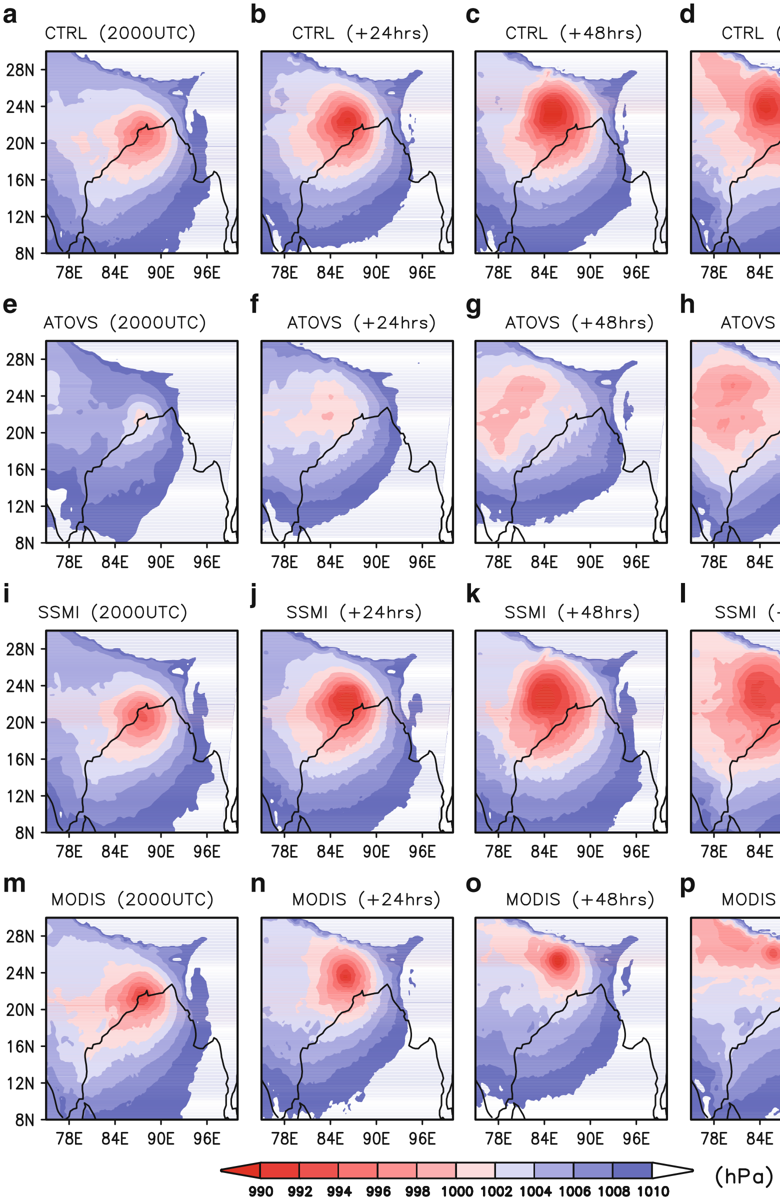


Fig. 26.4 Sea level pressure simulated by CTRL (**a**–**d**), ATOVS (**e**–**h**), SSM/I (**i**–**l**) and MODIS (**m**–**p**) runs on 20 September 2006 00 UTC and at 24, 48 and 60 h of forecast for 19–22 September 2006 depression

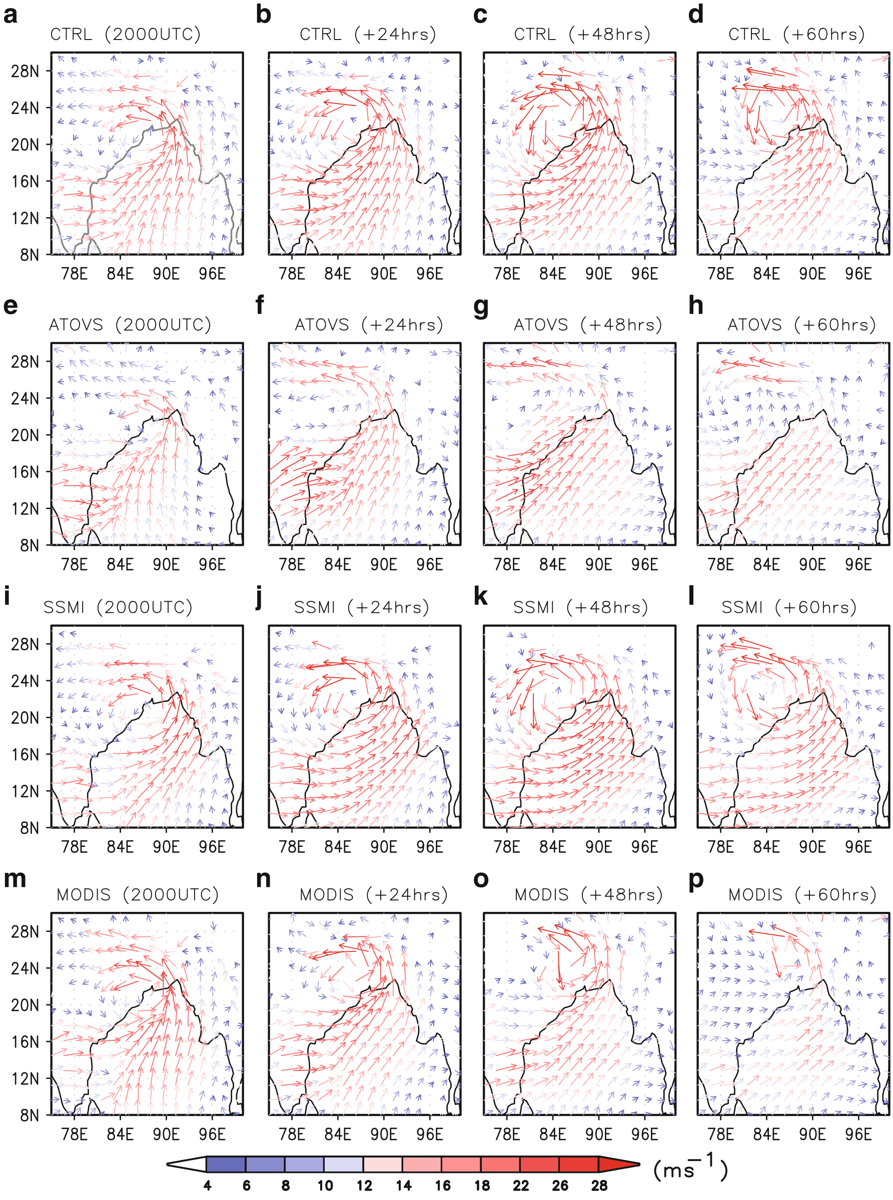


Fig. 26.5 Wind vector simulated by CTRL (**a**–**d**), SSM/I (**e**–**h**), ATOVS (**i**–**l**) and MODIS (**m**–**p**) runs on 20 September 2006 00 UTC and at 24, 48 and 60 h of forecast for 19–22 September 2006 depression

(Fig. 26.5m–p) for the 19–22 September 2006 depression. All the three experiments other than the ATOVS run have simulated higher wind speed values when compared to the NCEP-FNL analysis. The ATOVS run has simulated the weakest wind patterns among all the four experiments. The spatial distribution of the 850 hPa wind vectors depict an intense cyclonic circulation for the CTRL and the SSM/I experiments as compared to the ATOVS and the MODIS runs. From Fig. 26.5p, it is evident that the cyclonic circulation of the lower tropospheric wind started weakening during the 60 h of forecast of the MODIS experiment. On 20 September 2006 at 00 UTC, the simulated model results indicate that the monsoon current has strengthened over peninsular India and the central Bay of Bengal region, a feature associated with fresh surge of cross-equatorial air that usually precedes the formation of a monsoon depression (Sikka 1977).

Forecast Impact Parameter

In order to investigate further on the impact of assimilation on the simulated wind speed, "forecast impact" parameter (FI) of each assimilation experiment has been calculated with reference to the QuikSCAT wind observations for the 19–22 September 2006 depression. Following Wilks (2006), the forecast impact (FI) parameter for any variable based on the ratio of root mean square error (rmse) in the model forecasts for the control and the assimilation experiments can be defined as,

$$\mathrm{FI} = \left[1 - \frac{rmse_E}{rmse_C}\right] \times 100\% \tag{26.3}$$

where rmse_E and rmse_C are the rmse of the assimilation and control experiments with both the rmses being calculated with respect to the observations. The spatial distribution of the FI parameter for all the three assimilation experiments is shown in Fig. 26.6. A positive value of the FI implies the improvement in the predicted wind speed due to the assimilation of observations. All the three assimilation experiments exhibit positive "FI" parameter over most of the Bay of Bengal region. Among the three assimilation experiments, the ATOVS run shows the highest spatial extent of maximum positive value of the "FI" parameter, which is above 50 %. Though the SSM/I experiment also exhibits positive "FI" parameter values over the domain, its magnitude of the "FI" parameter is lower as compared to the MODIS and the ATOVS experimental "FI" values.

Root Mean Square Error of Wind Speed Profiles

The domain averaged rmse for the wind speed profiles at different levels and times are presented in Table 26.1 for the 19–22 September 2006 depression. RMSE values are calculated with reference to the IMD rawinsonde observations. The model values corresponding to each observation are taken from the nearest model grid

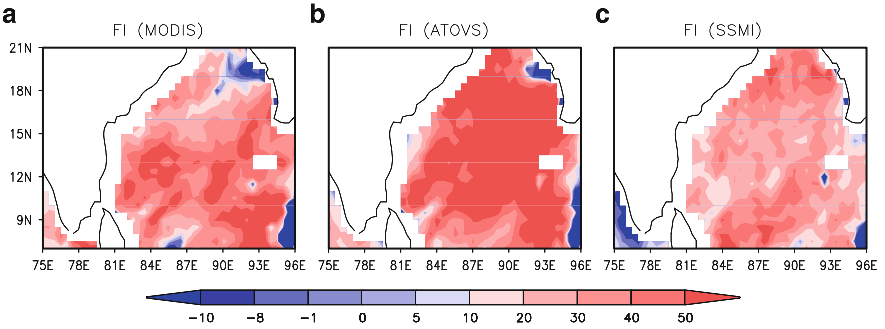


Fig. 26.6 Spatial distribution of forecast impact (FI) for (**a**) MODIS, (**b**) ATOVS, (**c**) SSM/I calculated against quikscat observations. for 19–22 September 2006 depression

corresponding to the observation location. The improvements if any, in the results of the assimilation experiments are established by comparing its RMSE values with that of the CTRL run. At 950 hPa level, the RMSE value of the wind speed is found to be higher for all the assimilation experiment when compared to the CTRL run for the forecast valid on 20 September 2006 12 UTC. The ATOVS experiment gives lower RMSE values for wind speed at 920, 850, 780, 700, 500 and 300 hPa levels while the MODIS experiment gives lower RMSE at 850, 780, 700, 620 and 300 hPa levels when compared to CTRL run. The RMSE values of the SSM/I experiment is found to be higher than that of the CTRL run and the other two experiments in almost all the levels for 20 September 2006 12 UTC. On 21 September 2006 12 UTC, it can be seen that the RMSE of the wind speed of the assimilated experiments have reduced in almost all the levels when compared to the CTRL run. On the third day of the forecast (22 September 2006 12 UTC), the wind speed simulated by the MODIS experiment has exhibited lower RMSE values up to 700 hPa level from the surface. However, above 700 hPa level, the RMSE of the wind speed rapidly increases and persists till the upper levels of the troposphere when compared to the CTRL experiment. The rms error values of the wind speed are lower for the ATOVS experiment when compared to the CTRL run in the lower and upper levels of the troposphere for all the three mentioned times of the forecast. For all the three times the SSM/I experiment shows higher RMSE values of wind speed as compared to the CTRL run at almost all the levels. Though the results of the MODIS and ATOVS runs did not yield lower RMSE values for all the times and at all the levels, they do show a consistent reduction in rmse of the wind speed for most of the times, thus showing some positive impact due to assimilation.

Rainfall

Due to its convective nature and complex interactions with the terrain features and vegetations it is difficult to forecast the precipitation fields very precisely.

Table 26.1 Root mean square error of wind speed profiles averaged over the domain for 19–22 September 2006 depression

	20 September 2006 12 UTC				21 September 2006 12 UTC				22 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	4.31	4.62	5.51	4.44	6.97	5.42	6.19	6.31	5.54	5.44	4.40	6.11
920	4.76	4.65	5.33	4.74	7.41	5.73	6.78	7.02	5.32	5.11	4.07	5.87
850	6.95	6.12	6.40	7.05	5.71	6.01	5.08	5.75	3.78	5.26	2.28	4.86
780	5.78	5.04	5.35	6.35	4.86	6.72	4.23	5.78	5.18	5.91	4.10	6.15
700	4.83	5.06	4.12	4.94	4.70	6.38	3.98	4.74	4.68	5.58	5.72	5.51
500	4.19	3.85	4.57	4.24	6.29	3.01	6.14	6.28	20.98	17.52	22.14	19.71
400	3.83	4.63	4.40	5.37	5.59	4.68	5.50	5.43	4.09	5.43	4.90	5.04
300	5.49	4.24	4.54	5.93	5.35	4.85	4.55	5.73	4.28	3.71	5.82	4.21
200	4.57	6.98	7.16	5.47	5.06	3.15	5.83	4.73	4.77	3.34	6.21	4.96

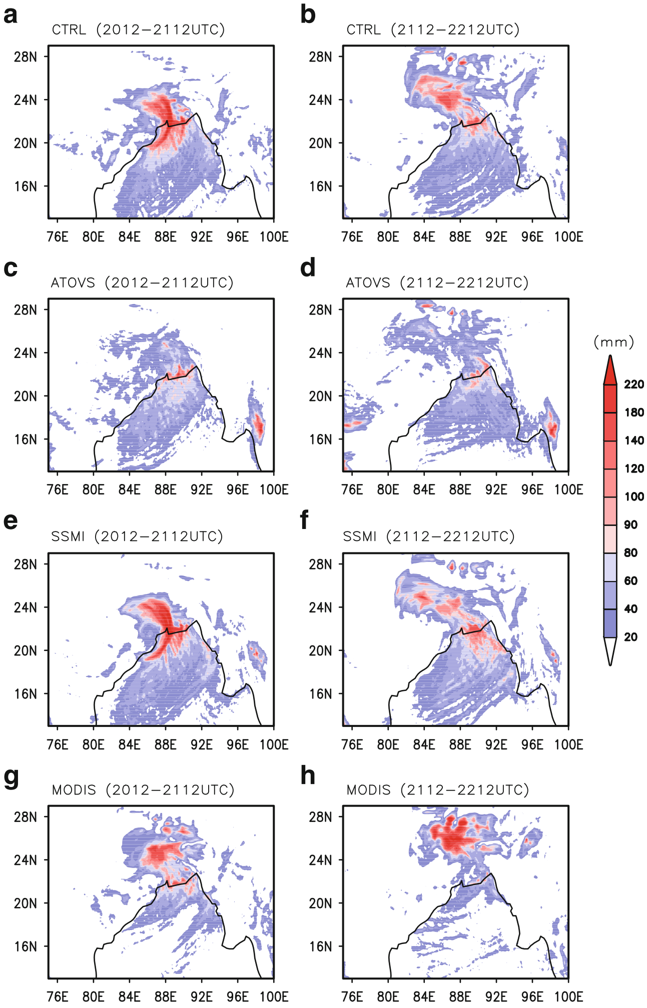


Fig. 26.7 Spatial distribution of 24-h accumulated precipitation for CTRL (**a**, **b**), ATOVS (**c**, **d**), SSM/I (**e**, **f**) and MODIS (**g**, **h**) for 19–22 September 2006 depression

Hence it is important to investigate carefully the skill of the model in simulating the intensity and spatial distribution of the rainfall associated with the monsoon depression. Figure 26.7 shows the spatial distribution of 24-h accumulated precipitation simulated by the four numerical experiments on day-one and day-two of the forecast for the 19–22 September 2006 depression. These are then compared with the TRMM satellite observations for validation which are depicted in Fig. 26.3i, j respectively. It can be readily seen that the simulated precipitation by the CTRL, the SSM/I and the MODIS experiments are over predicting the intensity of the rainfall when compared to the TRMM observations. The ATOVS experiment simulates less rainfall than that of the TRMM observations and the other three experiments, for both days of the forecast. Furthermore, there are also positional errors in the location of maximum precipitation as simulated by all the models. Even though the MODIS run overestimates the precipitation over land, the results of the above run are in better agreement with observations over the sea. Therefore, the MODIS run is in better agreement with TRMM for the day-two of the forecast, since it simulates much less rain over the sea as compared to the CTRL, the SSM/I and the ATOVS runs.

Equitable Threat Score and Bias Score

Further quantitative analysis of the simulated rainfall is performed by calculating the statistical skill scores namely "equitable threat score" (ETS) and "bias score" (BS) using the contingency table (Wilks 2006; Colle et al. 1999)**.** The Bias Score (BS) is a measure of the ratio of the frequency of forecast events to the frequency of observed events. The bias score indicates whether the forecast system has a tendency to underpredict (Bias < 1) or overpredict (Bias > 1) events. The bias score does not however, measure how well the forecast corresponds to the observations. The Equitable Threat Score (ETS) measures the fraction of observed and/or forecast events that are correctly predicted, with a provision for hits associated purely with "random chance". The ETS is often used in the verification of rainfall in NWP models since its "equitability" allows scores to be compared more fairly across different rainfall thresholds. The ETS penalizes both misses and false alarms in the same manner and also it does not concern itself with the source of forecast error. While higher values of the threat score represent enhanced skill of precipitation forecast, the maximum value of ETS is one. The ETS and BS are calculated for all the four numerical experiments based on the 48 h accumulated precipitation for various threshold values. The threshold values used are 40, 50 60, 70, 80, 90 and 100 mm. The results are presented in Figs. 26.8 and 26.9. From the Fig. 26.8, it is seen that the MODIS experiment exhibits highest skill of precipitation forecast for all the various threshold values when compared to the CTRL run, while the ATOVS and the SSM/I experiments, do show some precipitation predictability skill of the model although lower than the MODIS run. The ATOVS experiment shows higher skill scores at the lower threshold values which declines with increase of threshold values. Among the four numerical experiments, the SSM/I experiment shows the least skill score. However, all the experiments show a decrease in ETS values with increase in the threshold, indicating the difficulty in the prediction of high intense rainfall events. More quantitative verification of precipitation forecast is carried out using the "BS". Bias score of value of 1.0 implies that the model precipitation forecast has the same frequency (areal coverage) as that of the observations. The BS value greater than one for any model run indicates that the above run is over estimating the precipitation while a BS value less than one signifies the under estimation of precipitation when compared to the observation. Figure 26.9 gives

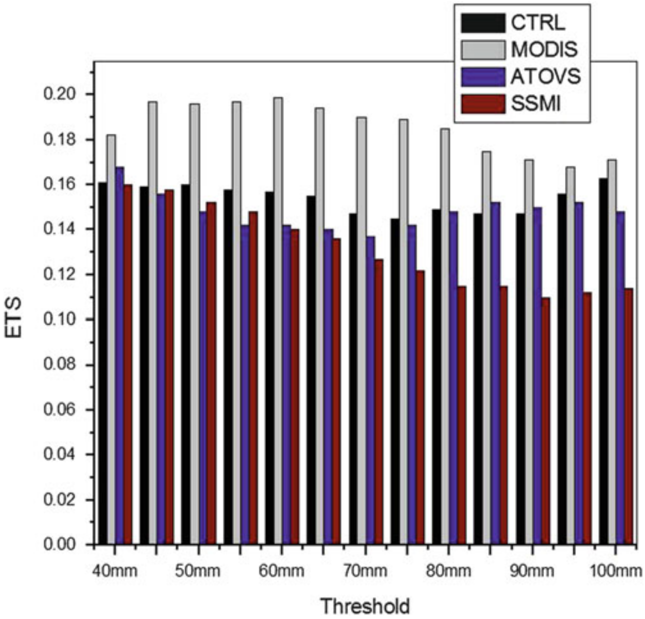


Fig. 26.8 Equitable threat score for 48-h accumulated precipitation for CTRL, ATOVS, SSM/I and MODIS for 19–22 September 2006 depression

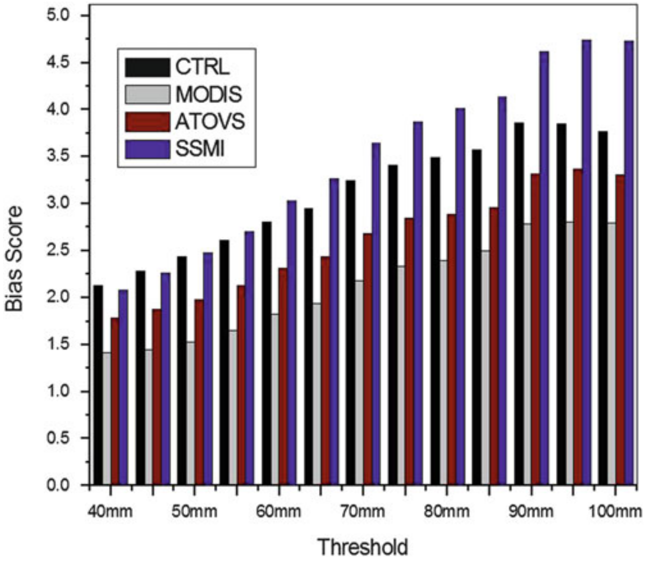


Fig. 26.9 Bias score for 48-h accumulated precipitation for CTRL, ATOVS, SSM/I and MODIS for 19–22 September 2006 depression

the BS values exhibited by all the four experiments for various threshold values with respect to the TRMM observations. All the four experiments reveal a BS value greater than one for all the thresholds, indicating that all the four model experiments are over estimating the precipitation features at all the thresholds. As expected, the MODIS experiment has the lowest BS values above one (i.e., lowest overestimation) at all the thresholds which indicates that the results of the MODIS experiment is relatively closer to the observations. The CTRL and the SSM/I run show higher BS values which indicate that both the above experiments have the highest degree of overestimation of the intensity of the rainfall.

Improvement Parameter ($\acute{\eta}$)

"Improvement parameter" for the 24 h accumulated precipitation valid at 12 UTC of 21 September 2006 and 22 September 2006 are calculated. Improvement parameter for any variable can be defined as,

$$\acute{\eta} = |(\text{observation} - \text{control})| - |(\text{observation} - \text{experiment})| \tag{26.4}$$

A positive value of "improvement parameter" is a clear indication of the positive impact of assimilation of observation and vice versa. The results of the MODIS experiment shows both positive and negative values of "improvement parameter" over the domain of study (Fig. 26.10). The maximum positive of η can be seen on the day-one of the forecast having values up to about 150. Negative impact of assimilation can be seen over the north of the domain. For the second day of forecast, the positive impact is more prominent over the Bay of Bengal region than over the land for the MODIS experiment. The ATOVS experiment also shows the same patterns of spatial distribution of "improvement parameter" as that of the MODIS experiment with highest positive values over the land. However, the negative impact is less over the land for the ATOVS experiment when compared to the MODIS experiment on day one of the forecast. In contrast to the first day, the second day exhibits more negative $\acute{\eta}$ over the land with values reaching up to –20. Also the positive impact over the Bay of Bengal region is not as prominent as that of the MODIS run. Hence, it can be concluded that despite the improvements in the ATOVS experiment improvement parameter on the first day of the forecast, the same decreased on the second day of the forecast. Undesirably large analysis increment values generated due to the assimilation of the ATOVS temperature and humidity profiles may be one of the reasons for the disappointing results. For the SSM/I run, the day one forecast shows smaller extent of positive values of "improvement parameter" over the domain. Furthermore, even though the second day forecast of the SSM/I run has produced positive $\acute{\eta}$ values over the oceans, the positive impact is found to be lower over the land regions.

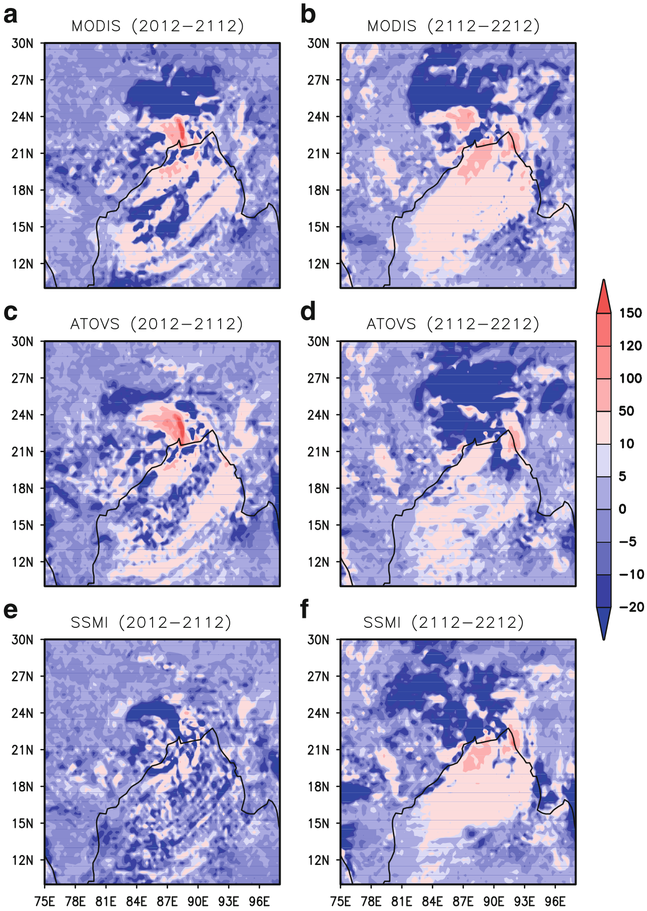


Fig. 26.10 Improvement parameter for 24-h accumulated precipitation for MODIS (**a**, **b**), ATOVS (**c**, **d**) and SSM/I (**e**, **f**) for 19–22 September 2006 depression

RMSE of Temperature Profiles

The model predicted temperature and humidity profiles are verified with India Meteorological Department (IMD) radiosonde observations for the 19–22 September 2006 depression. The RMSE values for the temperature and the dew-point temperature are calculated for 10 vertical levels at three different times of forecast. As mentioned before, the model values corresponding to each observation are taken from the nearest model grid corresponding to the observation location. The domain averaged RMSE for temperature profiles are presented in Table 26.2. As part of the verification, the RMSE of the temperature of the assimilation experiments are compared with the RMSE of the temperature of the CTRL run. At 950 hPa level on 20 September 2006 12 UTC, the RMSE of temperature is found to be higher for the

Table 26.2 Root mean square error of temperature profiles averaged over the domain for 19–22 September 2006 depression

	20 September 2006 12 UTC				21 September 2006 12 UTC				22 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	2.20	2.66	1.84	2.57	2.06	2.43	2.03	2.29	3.84	2.40	3.97	3.36
920	1.71	2.31	1.50	2.07	2.00	2.19	2.06	2.24	3.11	2.26	3.46	3.01
850	3.16	3.28	3.07	3.25	2.09	2.27	2.12	2.21	2.92	2.60	3.31	2.66
780	3.04	2.93	2.91	2.85	1.85	2.14	2.02	2.03	2.54	2.35	2.88	2.54
700	7.95	7.31	7.99	7.51	2.29	2.46	2.55	2.52	2.60	2.64	2.99	2.71
620	5.80	5.15	5.75	5.57	2.24	2.64	2.76	2.42	3.56	3.50	3.70	3.25
500	4.03	3.82	3.98	3.93	2.75	2.56	2.88	2.82	3.32	3.05	3.35	3.12
400	3.34	3.40	3.38	3.27	2.24	2.11	2.31	2.20	3.55	3.43	3.51	3.61
300	4.40	4.21	4.48	4.48	2.65	2.58	2.75	2.78	4.61	3.86	4.29	4.77
200	4.85	4.56	5.04	5.02	5.00	4.32	4.59	5.00	5.33	4.52	4.86	5.25

assimilation experiments except for the MODIS run. The ATOVS run shows higher error values of temperature up to 850 hPa level. From 780 to 200 hPa levels the RMSE values of temperature for the ATOVS run are found to be lower indicating an improved forecast for temperature in the mid-troposphere and upper troposphere for the ATOVS experiment on 20 September 2006 12 UTC. The RMSE values of temperature decreases with height for the SSM/I run on 20 September 2006 12 UTC; indicating a positive impact of assimilation of SSM/I total precipitable water at this time. The ATOVS experiment shows a decrease in the RMSE values of temperature with height for the forecast valid on 21 September 2006 12 UTC. Furthermore, significant decrease in the error is seen above the mid-troposphere (500 hPa) level. In contrast with the results of the ATOVS experiment, the MODIS run shows lower RMSE values of temperature at the lower levels on the second day which then increases with height up to 300 hPa levels, while the SSM/I run depicts higher RMSE of temperature at all the levels, on the second day, indicating the positive impact of assimilation of total precipitable water has somewhat decreased on the second day of the forecast. On 22 September 2006 12 UTC, the ATOVS and the SSM/I experiments have performed well with lower RMSE value at almost all the levels, while the MODIS experiment shows higher RMSE values for temperature at the lower and mid-troposphere.

RMSE of Dew-Point Temperature Profiles

Table 26.3 gives the domain averaged RMSE of dew-point temperature with respect to the IMD radiosonde observation. Previous studies (Cox et al. 1998; Sandeep et al. 2006) have noted that the mesoscale models have difficulty in accurately simulating the upper level moisture content. The lack of upper level moisture observations for assimilation can be regarded as one of the reasons for the above difficulty. Hence it is important to validate the simulated humidity profiles of all the experiments using the available observations. As expected, the assimilation of temperature and humidity profiles from the ATOVS and the MODIS have reduced the RMSE values of the moisture content at the lower levels of troposphere for the forecast valid on 20 September 2006 12 UTC when compared to the CTRL run. However, the RMSE of dew point temperature increases with height in the middle and upper troposphere, indicating the inability in accurately simulating the humidity profiles even after the assimilation of observation. In contrast with the other two assimilation experiments, the SSM/I experiment has shown significant improvement in the simulation of humidity profiles by having lower RMSE values of the dew point temperature for all the levels, on the day one of the forecast. The second day of the forecast shows higher values of RMSE of the dew point temperature at all the levels for the MODIS experiment while the ATOVS run depicts lower RMSE values of the dew point temperature at the lower levels of troposphere. On 22 September 2006 12 UTC, both the MODIS and the ATOVS provide results having higher RMSE values of dew-point temperature at almost all the levels while the SSM/I run has significantly reduced the error values. Hence, it can be concluded

Table 26.3 Root mean square error of dew point temperature profiles averaged over the domain for 19–22 September 2006 depression

	20 September 2006 12 UTC				21 September 2006 12 UTC				22 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	2.53	2.32	2.33	2.30	2.96	2.38	3.40	2.92	1.81	2.41	2.00	2.44
920	3.98	3.55	4.01	3.57	2.54	2.13	2.82	2.62	2.25	3.55	2.29	2.81
850	3.30	2.94	3.24	2.84	2.58	2.78	2.91	2.88	4.50	5.14	4.65	4.44
780	5.67	5.10	5.60	5.12	2.88	2.98	3.26	3.34	4.98	5.19	5.01	4.87
700	13.59	11.67	12.70	11.85	4.05	3.59	4.13	3.94	5.99	6.89	5.28	5.78
620	10.84	9.67	11.20	10.44	7.30	8.07	7.64	6.98	7.80	8.37	7.99	7.70
500	9.03	9.28	11.65	7.22	7.40	8.79	8.00	6.90	8.29	7.47	7.87	8.51
400	9.87	11.20	12.76	9.91	7.88	7.77	6.92	8.74	11.58	10.47	11.37	11.49
300	8.62	10.58	9.87	7.80	5.31	7.67	3.98	7.36	8.81	7.29	6.82	11.31
200	11.11	12.56	17.70	10.95	6.57	11.09	5.68	8.98	11.55	7.94	7.92	3.98

that the assimilation of SSM/I total precipitable water has resulted in significant positive impact on the simulation of the humidity profiles in a numerical mesoscale model. The results obtained in this study indicate that the mesoscale model predicts the moisture fields in the lower levels better than that of the upper levels. The above result is due to the availability of more moisture information at the lower levels (Cox et al. 1998).

Apparent Heat Source and Moisture Sink

The main driving energy associated with the tropical disturbances is the latent heat release due to cumulus convection. Many studies have investigated the formation and evolution of the tropical cloud clusters, which manifest due to intense convection (Manabe et al. 1970; Wallace 1971; Nitta 1970). The study of monsoon depression, one of the prominent tropical monsoonal disturbances, associated with deep cumulus convection and large cloud clusters, provides an excellent case for a detailed diagnostic study with a special emphasis on convective processes. The main purpose of this section is to diagnose the role of the vertical distribution of heating and cooling associated with convection in the monsoonal environment. Following Yanai et al. (1973), heat and moisture budget analysis has been performed for the model experiments in this section. To be consistent with the model simulated heat and moisture fields, the computations for the heat and moisture budget analysis are performed in σ coordinates. The budget analysis for heat and moisture has also been performed for NCEP-FNL analysis in σ coordinates for comparing the model diagnostics with the analysis. Apart from the heat and moisture budget analysis, the time averaged and area averaged temperature anomaly and relative vorticity of the simulated depression in the vertical have also been validated with the respective fields of NCEP-FNL analysis. All the three cases of monsoon depressions, which are part of our investigation, are utilized for this analysis.

Figure 26.11a, b depict the 60 h time averaged profiles of the area averaged apparent heat source (Q_1) and apparent moisture sink (Q_2) respectively for the 19–22 September 2006 monsoon depression that formed over the Bay of Bengal. The area averages were computed for 3° × 3° region around the centre of the depression. As can be seen from Fig. 26.11a, the CTRL experiment is showing the maximum heating rate at the model height of $\sigma = 0.55$. The MODIS run also simulates a similar heating profile as that of the CTRL run, but with a lower magnitude. Considering the NCEP-FNL analysis as the true estimate of heating, it is clear that both the CTRL and the MODIS experiments overestimate the rate of heating in the vertical. The magnitude of heating rate in the ATOVS and the SSMI experiments are lower when compared to the other two experiments and the NCEP-FNL analysis. The maximum heating rates due to moisture effects is seen in the CTRL run among all the four model experiments, reaching a maximum of about 16 K day^{-1} over the mid-troposphere. The vertical heating rate simulated by the MODIS experiment closely follows the NCEP-FNL analysis which implies that the MODIS experiment better simulates the apparent moisture sink in the vertical.

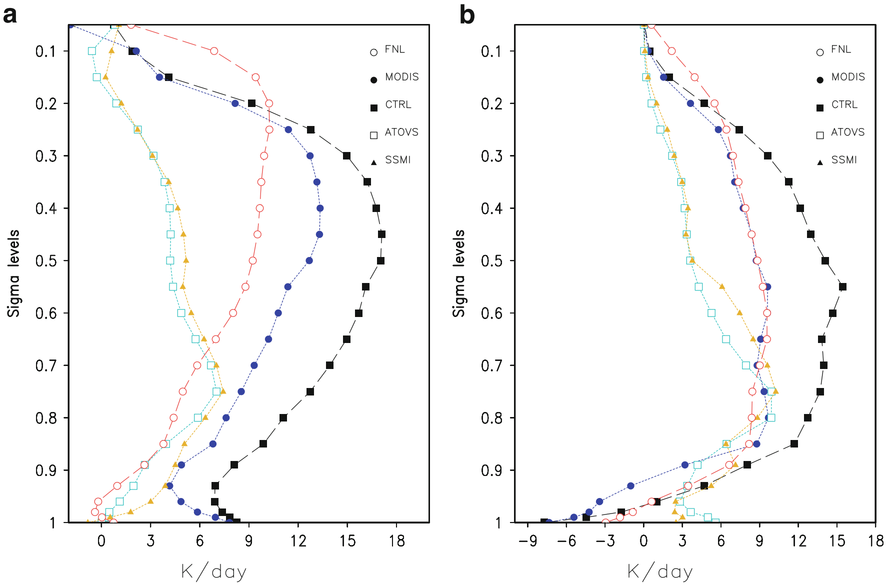


Fig. 26.11 Vertical profiles of the 60 h time averaged and area averaged over a 3° × 3° region around the centre of the depression for the (**a**) apparent heat source (Q_1) (**b**) apparent moisture sink (Q_2) for the 19–22 September 2006 depression

Earlier studies (Schlesinger 1994) showed that the heating due to apparent moisture sink has a minimum at about 9 km ($\sigma = 0.225$) with the maximum peak observed at about 3 km ($\sigma = 0.725$). The relative absence of a pronounced maximum peak associated with the apparent moisture sink is well captured in all the four model experiments.

26.4.2 Monsoon Depression That Formed During 02 to 05 September 2006

The depression which formed over the north Bay of Bengal during the first week of September 2006 was first seen as a low pressure area over the same region in the morning. The system became well marked in the forenoon and subsequently intensified into a depression and lay centered about 180 km southeast of Balasore, Orissa at 12 UTC on 03 September 2006. Moving in a north-westerly direction, the depression crossed the Orissa coast close to Chandbali at 12 UTC on 04 September 2006. Furthermore, the system moved north-westwards and weakened into a well marked low pressure area over north Chhattisgarh and adjoining east Madhya Pradesh on 5 September 2006.

26.4.2.1 Numerical Experiment

The model settings and the domain used are the same as in Sect. 26.4.1. The NCEP-GFS forecast data available at a horizontal resolution of 1° × 1° and a time resolution of 6 h have been used to develop the initial and lateral boundary conditions. Four numerical experiments are performed as in Sect. 26.4.1 to study the impact of temperature, humidity and total precipitable water in the simulated structure of monsoon depression. All the simulations are started with the same initial conditions on 01 September 2006 18 UTC; however, the observations are assimilated at different times depending on the availability of the satellite data over the domain. For the CTRL experiment, the model integrations are performed till 05 September 2006 12 UTC. In the ATOVS run, the ATOVS temperature and humidity profiles are ingested into the model using 3D-VAR, initially, after 6 h of forecast, i.e., on 02 September 2006 00 UTC and subsequently ingested up to 03 September 2006, 00 UTC in a 12 hourly interval. The WRF model is then subsequently integrated for the next 60 h in free forecast mode without any further assimilation of the satellite data. In the "MODIS run", the MODIS temperature and humidity observations have been assimilated in 12 h intervals from 01 September 2006 18 UTC to 02 September 2006 18 UTC. Subsequently, the model is run in a free forecast mode up to 05 September 2006 12 UTC. As in Sect. 26.4.1, the MODIS data is subjected to "thinning". The "SSM/I run" is similar to the ATOVS experiment except that it incorporates SSM/I total precipitable water instead of ATOVS temperature and humidity observations. The CTRL run is subjected to data cycling without any assimilation of observations. The domain of the study is same as that of Sect. 26.4.1. The results presented are from 12 km resolution domain only. All the satellite observations have shown impact on the initial condition, with the ATOVS observation having maximum significance while the SSM/I observations have the lowest significance.

26.4.2.2 Results and Discussion

Initial Conditions

The analysis increment of wind speed at 850 hPa level for the MODIS, ATOVS and SSM/I experiments are presented in Fig. 26.12 for the 2–5 September 2006 depression. It can be seen that the assimilation of the satellite observations has introduced clear increments in the 850 hPa wind speed. The assimilation of MODIS temperature and humidity observations has introduced negative increments over the Arabian Sea together with a small pocket of positive increments to the south of the peninsular India. However, the increment values of 850 hPa wind speed in the MODIS run is almost zero over the Bay of Bengal region. The assimilation of ATOVS temperature and humidity observations has introduced larger spatial distribution of positive values over the Arabian Sea, the land surface and the Bay

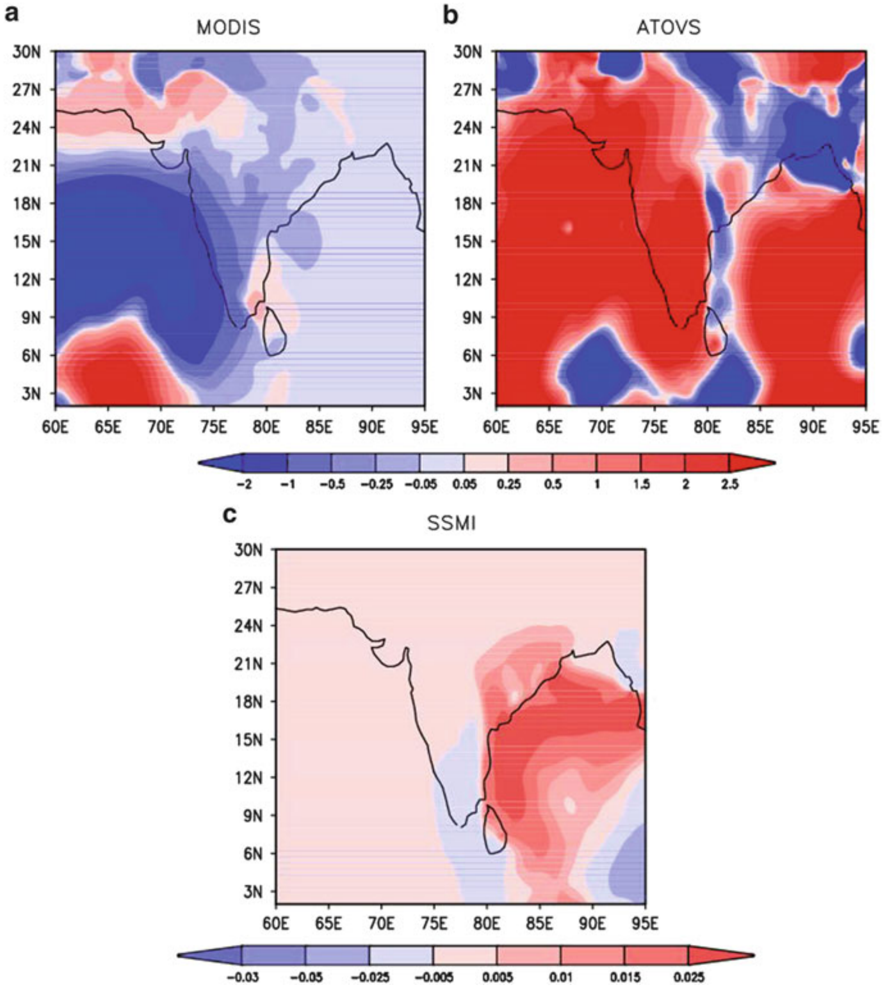


Fig. 26.12 Analysis increment at 850 hPa in the wind speed of (**a**) MODIS, (**b**) ATOVS and (**c**) SSM/I for 2–5 September 2006 depression

of Bengal region. Even though the negative increments of 850 hPa wind speed are seen over the domain for the ATOVS run, its spatial extent is found to be less. Though the ingestion of total precipitable water using SSM/I has introduced little increment values of 850 hPa wind speed as compared to the ATOVS and the MODIS experiments, the SSM/I experiment has shown positive values of increment of 850 hPa wind speed with a maximum spatial distribution over the Bay of Bengal region.

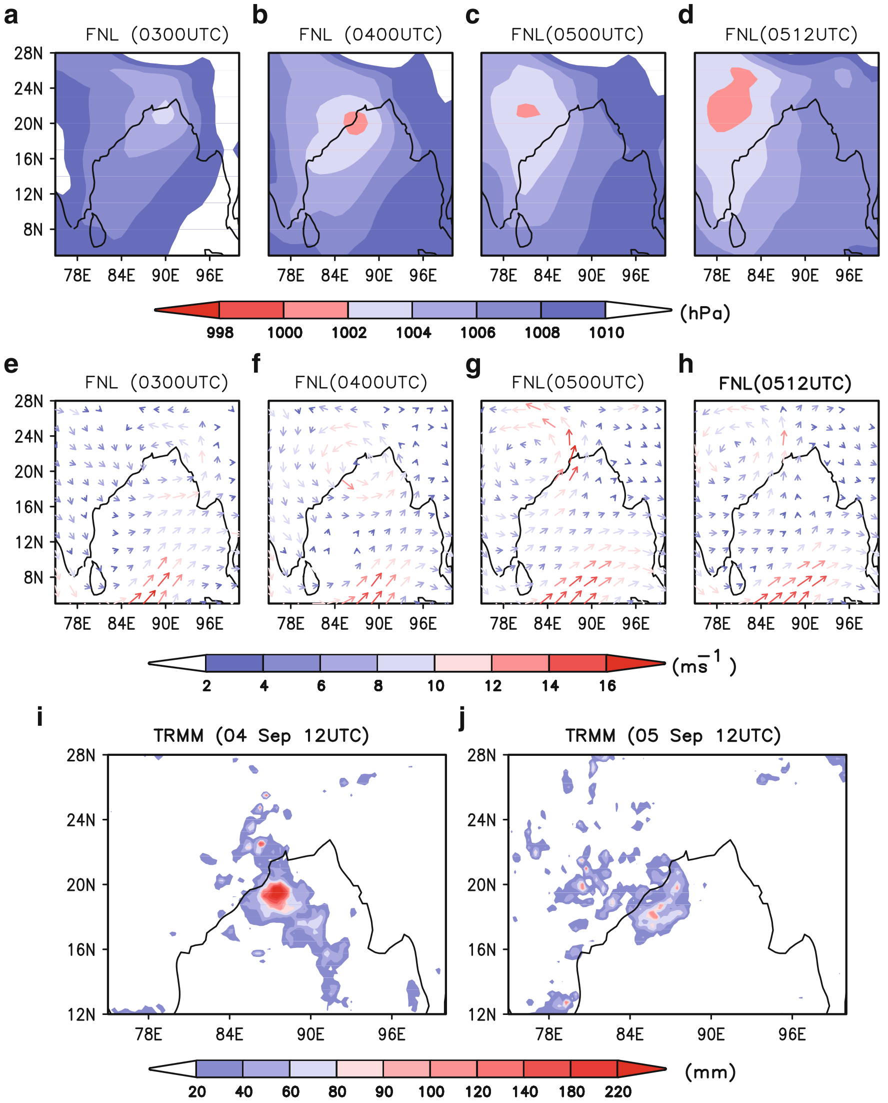


Fig. 26.13 NCEP-FNL analysis sea level pressure fields at 03–05 September 2006 00 UTC (**a**–**d**), NCEP-FNL lower tropospheric wind speed (850 hPa) (**e**–**h**) and 24 h accumulated TRMM rainfall (**i**–**j**)

MSLP

The spatial distribution of MSLP patterns of NCEP-FNL are depicted in Fig. 26.13a–d. From the figure it can be seen that the depression started as a low pressure area over the Head Bay region. The low pressure area subsequently intensified into a depression with a minimum central pressure of 998 hPa. The system experienced landfall before 5 September 2006 and started weakening subsequently after landfall. The MSLP plots of the NCEP-FNL therefore are consistent with that of the observations. Figure 26.14a–p depicts the MSLP patterns on 03 September 2006 00 UTC and subsequent predicted values at 24, 48 and 60 h of forecast for the CTRL experiment (Fig. 26.14a–d), the ATOVS experiment (Fig. 26.14e–h), the SSM/I (Fig. 26.14i–l) experiment and the MODIS (Fig. 26.14m–p) experiment for the 2–5 September 2006 depression. From the spatial distribution of MSLP patterns it can be seen that the CTRL and the ATOVS experiments have failed to simulate accurately the evolution of the depression over the Bay of Bengal. After 48 h of forecast the CTRL run simulates a low pressure region over the south east of the Bay of Bengal, a feature absent in the observation records. A pattern similar to the CTRL run with a lower intensity of MSLP is seen in the results of the ATOVS experiment too. The above-mentioned erroneous low pressure region is not seen in the SSM/I and the MODIS experiments. However, the SSM/I run, simulates the location of the depression centre at the 60 h of forecast near Hyderabad which is much south of the observed position of the monsoon depression at that time. The experiment which assimilated MODIS observation has, however, stimulated the exact location of the depression with slightly higher intensity of MSLP than that of the observed.

Wind Speed

The lower tropospheric wind vector simulated by all the four experiments at 850 hPa are shown in Fig. 26.15 for the 2–5 September 2006 depression, which includes the CTRL run (Fig. 26.15a–d), ATOVS experiment (Fig. 26.15e–h), SSM/I experiment (Fig. 26.15i–l) and MODIS experiment (Fig. 26.15m–p).

The CTRL, ATOVS and the SSM/I experiments simulate almost the same patterns of wind vector field with a strong south-westerly flow over the extreme south of the Indian peninsula. The south-westerly flow is found to be slightly intense for the SSM/I run. However, the 850 hPa cyclonic circulation usually associated with the monsoon depression, is found to be absent in the case of the CTRL, ATOVS and SSM/I runs. The MODIS experiment simulates a weak cyclonic vortex over the Head Bay region and over the inland regions at 24, 48 and 60 h of forecast. The notable feature about the MODIS experiment is the absence of the south-westerly flow as seen in the other experiments. Hence it can be concluded that the CTRL ATOVS, SSM/I experiments failed to simulate adequately the cyclonic circulation associated with the monsoon depression over the Head Bay region. The reason for the above failure may be partly due to the erroneous simulation of a low pressure system over south-east of Bay of Bengal and the presence of a stronger south-westerly flow over the extreme south of the Indian peninsula.

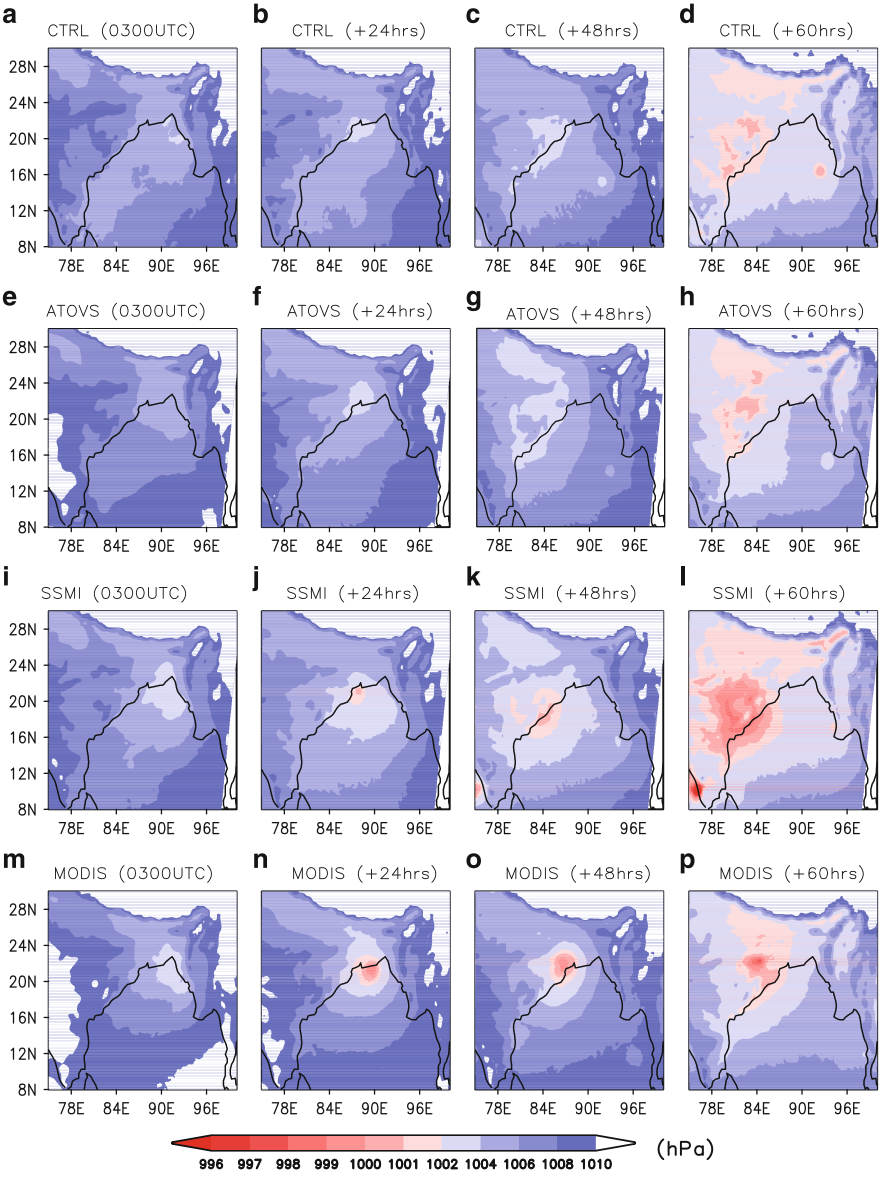


Fig. 26.14 Sea level pressure simulated by CTRL (**a–d**), SSM/I (**e–h**), ATOVS (**i–l**) and MODIS (**m–p**) runs on 03 September 2006 00 UTC and at 24, 48 and 60 h of forecast

Forecast Impact Parameter

The "Forecast Impact" parameter has been calculated to validate the impact of assimilation of the satellite observations on the wind speed for the 2–5 September

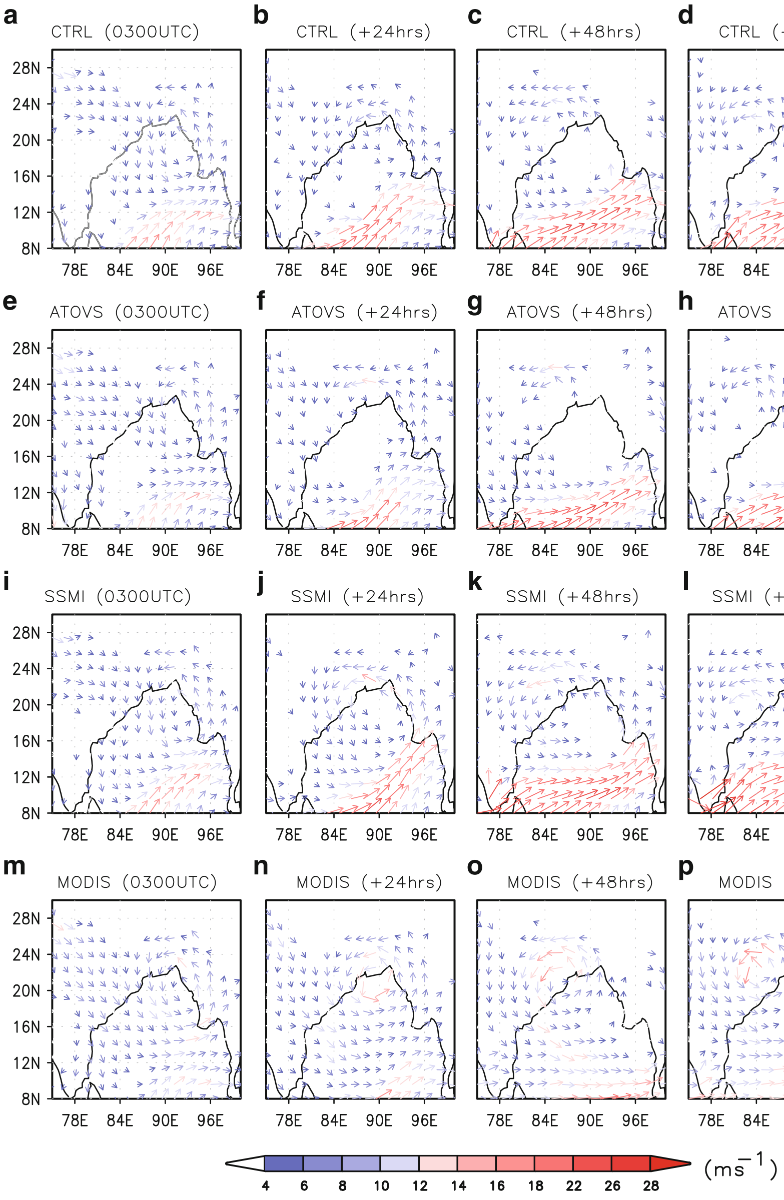


Fig. 26.15 Wind vector simulated by CTRL (**a–d**), SSM/I (**e–h**), ATOVS (**i–l**) and MODIS (**m–p**) runs on 03 September 2006 00 UTC and at 24, 48 and 60 h of forecast

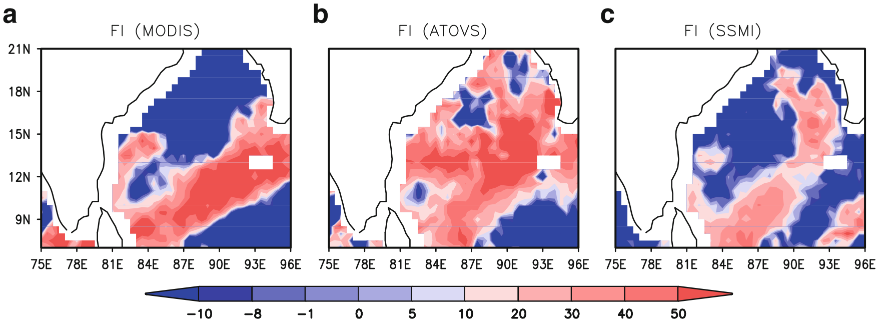


Fig. 26.16 Spatial distribution of forecast impact (FI) for (**a**) MODIS, (**b**) ATOVS, (**c**) SSM/I calculated against QuikSCAT observations for 2–5 September 2006 depression

2006 depression. The verification of the model results are made with respect to QuikSCAT wind observations. The spatial distribution of FI is presented in Fig. 26.16. For the MODIS experiment the spatial distribution of FI parameter reveals a blend of positive and negative values over the Bay of Bengal region. The assimilation of MODIS observations has produced negative impact on the 10 m wind speed over the north Bay of Bengal and over the south-east regions of the domain, while the central part depicts higher positive of FI for the wind speed. The ATOVS run exhibits a more positive impact on the wind speed with the larger spatial extent of positive values over the domain while for the SSM/I run the spatial extent of negative values are more prominent. The magnitude of positive values for FI is less for the SSM/I run as compared to the other two experiments.

RMSE of Wind Speed Profiles

The domain averaged values of RMSE for wind speed is depicted in Table 26.4. The MODIS and the ATOVS experiment have reduced the RMSE values of wind speed for 2–5 September 2006 depression at almost all the levels except 620, 500 and 200 hPa levels when compared to the CTRL run during the forecast time valid on 03 September 2006 12 UTC. The assimilation of SSM/I total precipitable water did reduce the RMSE value of wind speed for all the levels as compared to CTRL run: for the ATOVS run on 3rd September 2006, the error increased up to 400 hPa level and then the error started decreasing at higher levels. The MODIS experiment has however given better results with low RMSE of wind speed on the second day of the forecast. The SSM/I run has shown higher RMSE of wind speed on the lower and mid-troposphere together with lower RMSE of wind speed at the upper levels during 04 September 2006 12 UTC. The MODIS experiment shows lower RMSE values of wind speed on the third day of forecast while the RMSE of wind speed is higher for the ATOVS run and is lower for the SSM/I experiment at the upper levels of the troposphere.

Table 26.4 Root mean square error of wind speed profiles averaged over the domain for 2–5 September 2006 depression

	03 September 2006 12 UTC				04 September 2006 12 UTC				05 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	3.72	2.75	2.13	2.81	3.00	2.51	2.75	3.68	4.05	3.03	3.30	4.77
920	4.58	3.67	3.19	3.73	4.28	3.90	3.70	4.81	3.77	2.82	2.61	4.32
850	3.49	2.97	2.37	2.95	3.87	4.17	2.35	4.41	4.17	3.56	3.44	2.25
780	4.09	3.51	3.11	3.29	3.47	4.57	2.73	4.62	8.23	7.75	7.24	5.48
700	4.51	3.95	3.92	3.93	4.48	4.87	3.37	5.02	4.41	4.87	3.61	3.53
620	5.31	5.78	5.46	5.47	3.53	4.06	2.66	3.35	2.70	3.00	2.63	3.17
500	5.06	5.25	4.66	4.38	3.16	3.48	2.07	2.02	4.76	3.95	4.51	4.10
400	7.29	6.33	5.03	4.67	2.38	2.75	2.42	2.44	4.19	4.43	4.51	3.78
300	4.77	3.13	4.86	3.92	2.72	2.34	3.76	3.00	3.66	4.07	6.01	4.57
200	4.62	5.31	5.07	4.43	8.88	8.84	8.77	7.39	4.38	4.14	6.73	4.07

Rainfall

Figure 26.13i, j show the spatial distribution of 24 h accumulated precipitation valid on 4 September 2006 12 UTC and 5 September 2006 12 UTC respectively for the 2–5 September 2006 depression.

The TRMM observation shows that the maximum precipitation is over the east coast of India, and also over the Bay of Bengal. On the second day, the rainfall intensity has decreased considerably indicating the weakening of the monsoon depression. The rainfall patterns simulated by the CTRL (Fig. 26.17a, b), ATOVS (Fig. 26.17c, d) and SSM/I (Fig. 26.17e, f) runs show an unrealistically intense and extreme precipitation over the south of the Bay of Bengal. Also, in the above three runs, the simulated precipitation intensity is weak over the Head Bay and over the coastal regions of Orissa at variance with the TRMM observations. The amount of precipitation simulated by these three experiments over the Orissa coast and the adjacent land area are less than 40 mm. However, the MODIS experiment (Fig. 26.17g, h) provides the best results out of all the model runs, accurately simulating the location and intensity on the first day of the forecast. The MODIS run, however, is at variance with TRMM observations simulating less rain over the sea on the second day of the forecast. Moreover, the unrealistic heavy rainfall pattern over the south Bay of Bengal seen in the other three model runs is absent in the MODIS experiment.

Improvement Parameter (ή)

The spatial distribution of the "improvement parameter" for precipitation is shown in Fig. 26.18 for the 2–5 September 2006 depression. Positive improvement in the precipitation features can be seen over the south of the Bay of Bengal for all the experiments. For the MODIS experiment most of the negative values are seen over the sea near the Head Bay region while the same is over the land for the ATOVS experiment on 4 September 2006 12 UTC. On the second day of forecast, the precipitation feature shows more positive improvement over the land for both the MODIS and the ATOVS experiment. The distribution of negative ή values are prominent over the south east of the Bay of the Bengal on the first day of forecast, while the second day witnessed a better precipitation forecast with larger distribution of positive ή values over the land for the SSM/I run.

RMSE of Temperature Profiles

The assimilation of temperature and humidity profiles from the ATOVS and the MODIS sensors have reduced the RMSE values of temperature at 950, 920, 700, 500, 400 and 200 hPa levels when compared to the CTRL run for the forecast time valid on 03 September 2006 12 UTC which is shown in Table 26.5 for the 2–5 September 2006 depression. The SSM/I run depicts lower RMSE at the upper levels

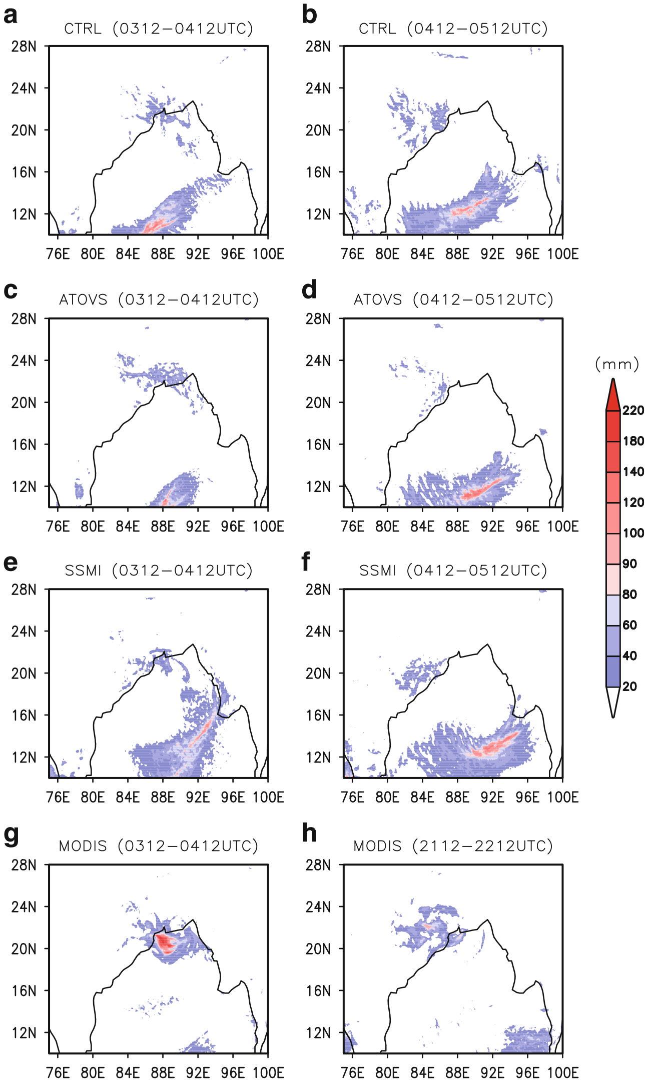


Fig. 26.17 Spatial distribution of 24-h accumulated precipitation for CTRL (**a**, **b**), ATOVS (**c**, **d**), SSM/I (**e**, **f**) and MODIS (**g**, **h**) for 2–5 September 2006 depression

of the troposphere on the day one of the forecast period. The MODIS experiment gives lower RMSE values on the second day of forecast except for the 500 and 300 hPa levels while the ATOVS run shows lower RMSE values at 780, 700, 620, 400 and 200 hPa levels. The SSM/I run however exhibits an inconclusive error pattern for temperature on 4 September 2006 12 UTC. On the third day of the

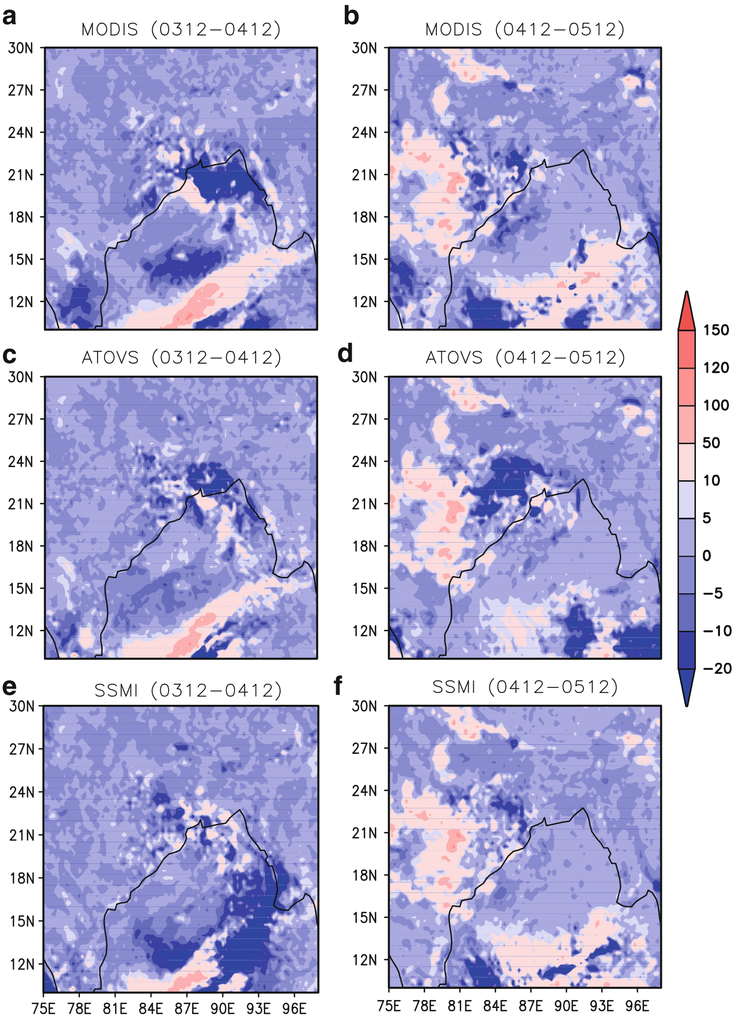


Fig. 26.18 Improvement parameter for 24-h accumulated precipitation for MODIS (**a**, **b**), ATOVS (**c**, **d**) and SSM/I (**e**, **f**) for 2–5 September 2006 depression

forecast the SSM/I run however shows higher RMSE of temperature at almost all the levels when compared to the CTRL run.

RMSE of Dew-Point Temperature Profiles

The RMSE of dew-point temperature in presented in Table 26.6 for the 2–5 September 2006 depression. As can be seen from the table, the ATOVS experiment shows lower RMSE values of dew point temperature than that of the

Table 26.5 Root mean square error of temperature profiles averaged over the domain for 2–5 September 2006 depression

	03 September 2006 12 UTC				04 September 2006 12 UTC				05 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	3.08	2.93	2.86	2.98	3.25	3.32	2.78	3.54	2.91	3.13	3.57	3.19
920	3.05	3.03	2.75	3.26	2.73	2.96	2.19	3.11	3.17	3.24	3.50	3.42
850	1.84	2.02	2.57	2.21	2.28	2.30	1.92	2.46	2.32	2.45	2.57	2.61
780	1.42	1.51	2.01	1.59	1.79	1.75	1.74	1.71	2.73	2.53	2.61	2.70
700	2.05	1.88	2.03	2.11	2.29	2.08	1.98	2.25	2.60	2.62	2.65	2.60
620	1.76	1.60	1.90	1.94	2.51	2.21	2.41	2.52	2.04	1.91	2.07	2.00
500	1.96	1.72	1.78	1.66	1.81	1.87	1.96	1.79	2.01	2.24	2.03	2.05
400	2.69	2.32	2.55	2.44	2.00	1.73	1.94	1.89	2.46	2.52	2.28	2.33
300	3.03	3.00	3.41	2.88	2.02	2.19	2.08	2.17	3.13	3.07	2.66	3.36
200	3.38	3.67	3.30	3.51	4.31	4.26	4.21	4.57	5.28	5.06	3.42	5.51

Table 26.6 Root mean square error of dew point temperature profiles averaged over the domain for 2–5 September 2006 depression

	03 September 2006 12 UTC				04 September 2006 12 UTC				05 September 2006 12 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	3.06	2.74	3.37	2.99	2.42	2.80	3.16	2.68	4.25	4.32	4.52	3.84
920	2.94	2.58	2.61	2.87	1.79	2.21	2.47	2.23	3.68	3.72	3.75	3.17
850	3.82	3.20	3.52	3.50	3.21	3.02	2.67	3.40	3.92	3.80	3.44	3.17
780	4.36	3.91	4.09	4.16	5.37	5.34	5.43	6.34	3.59	4.12	3.92	3.72
700	9.54	8.51	8.49	9.21	7.81	6.03	7.83	7.94	3.53	5.89	6.61	6.34
620	8.30	6.82	7.77	8.76	6.21	4.59	7.12	6.18	10.00	10.72	9.37	10.65
500	11.68	10.11	10.80	12.69	8.28	6.33	8.02	9.71	10.53	11.11	10.07	11.70
400	9.46	8.43	8.78	9.31	9.18	7.63	7.27	7.89	9.91	10.75	11.07	11.01
300	7.63	6.00	8.05	7.07	3.83	6.27	6.57	3.92	8.76	9.41	9.80	9.73
200	8.94	10.74	8.90	9.58	15.37	15.51	16.31	16.59	7.56	9.27	6.71	6.55

CTRL experiment for all the levels except 200 hPa. The MODIS run also shows improvements in the humidity profiles with lower RMSE of dew point temperature at almost all the levels except 950 and 850 hPa due to assimilation. The SSM/I run shows lower error values of dew point temperature at 950, 920, 850, 780, 700, 400 and 300 hPa levels against the CTRL run on 03 September 2006 12 UTC. On the second day of forecast, the ATOVS run has produced significant improvements in the dew-point temperature profile by reducing the RMSE of the dew point temperature at all the levels except 950 and 300 hPa. The MODIS and the SSM/I run have shown higher RMSE of dew-point temperature at almost all the levels. However, on the second day of the forecast, the assimilation experiments have failed to show marked improvements in the humidity profiles on 5 September 2006 12 UTC as compared to the CTRL experiment.

26.4.3 Depression That Formed During 18 to 22 June 2007

The depression that formed over the Bay of Bengal during the third week of June 2007 was initially seen as a low pressure area over the east central Bay and neighbourhood on 20 June 2007. The low pressure area intensified to a depression and lay centred at 15.5°N, 86.0°E on 21 June 2007 03 UTC. The depression further intensified into a deep depression and lay centred over 16°N, 84.0°E at 12 UTC on the same day. At 03 UTC of 22 June 2007, the system was over coastal Andhra Pradesh. The system subsequently moved westwards.

26.4.3.1 Numerical Experiments

The model settings and the domain used are the same as that for the two previous cases. The NCEP-GFS forecast data available at a horizontal resolution of 1° × 1° and a time resolution of 6 h are used to develop the initial and lateral boundary conditions. Four numerical experiments are performed to study the impact of temperature, humidity and total precipitable water in the simulated structure of monsoon depression. All the simulations are started with the same initial conditions on 18 June 2007 18 UTC, but the observations are assimilated at different times depending on the availability of the satellite data over the domain. The CTRL experiment, has utilized the NCEP-GFS data for creating initial and lateral boundary conditions. The model integrations are performed from 18 June 2007 18 UTC to 22 June 2007 06 UTC without any assimilation of satellite observations. The second experiment, called the ATOVS run is started from 18 June 2007 18 UTC. For the ATOVS run, the ATOVS temperature and humidity profiles have been ingested initially into the model using 3D-VAR after 6 h of forecast, i.e., on 19 June 2007 00 UTC and again up to 20 June 2007, 00 UTC in a 12 hourly interval. The WRF model is then subsequently integrated for the next 54 h in a free forecast mode without any further assimilation of the satellite data. The third experiment, named the MODIS run, ingests the temperature and humidity observations from MODIS in a 12 hourly interval from 18 June 2007 18 UTC to 19 June 2007 18 UTC and the model is subsequently run in a free forecast mode up to 22 June 2007 06 UTC. The high resolution MODIS data is subjected to thinning before ingesting the data to the observational pre-processor. The fourth experiment (SSM/I) is similar to the ATOVS experiment except that it incorporates SSM/I total precipitable water instead of ATOVS temperature and humidity observations. All the model runs include 6 h interval data cycling to maintain the dynamical consistency of the model simulation. The CTRL run is subjected to data cycling without any assimilation of observations. The results presented are from 12 km resolution domain only.

26.4.3.2 Results and Discussion

Initial Conditions

As in previous cases, the ATOVS experiment depicts larger spatial distribution of positive analysis increment values of 850 hPa wind speed with negative increment seen towards the south of the domain. For the MODIS run, most of the positive increment values of 850 hPa wind speed are seen towards north of the domain with negative values concentrated towards the southern tip of the Indian peninsula and the adjacent Bay of Bengal. As seen in the previous cases, the assimilation of SSM/I total precipitable water shows very small increment values of 850 hPa wind speed over the domain. The differences in the initial conditions are presented in Fig. 26.19 for the 18–22 June 2007 depression.

MSLP

Figure 26.20a–d show the spatial distribution of MSLP patterns of the depression from the NCEP-FNL analysis. An intense monsoon depression with a minimum central pressure of 990 hPa is shown in the figure. As per the NCEP analysis, the system made the landfall after 22 June 2007 00 UTC which is consistent with the observation records. The results of all the numerical experiments from Fig. 26.21 show that the intensity of the depression is over estimated in all the experiments except the ATOVS run. The ATOVS experiment did predict the intensity of the depression with almost the same magnitude as that of the NCEP-FNL analysis. None of the four experiments can accurately predict the landfall of the depression. Moreover, a slow movement of depression is seen prominently in the MODIS run.

Wind Speed

The spatial pattern of the 850 hPa wind vector simulated by the numerical experiments is shown in Fig. 26.22 for the 18–22 June 2007 depression. The NCEP-FNL wind pattern showed landfall of the cyclonic system on 22 June 2007 00 UTC

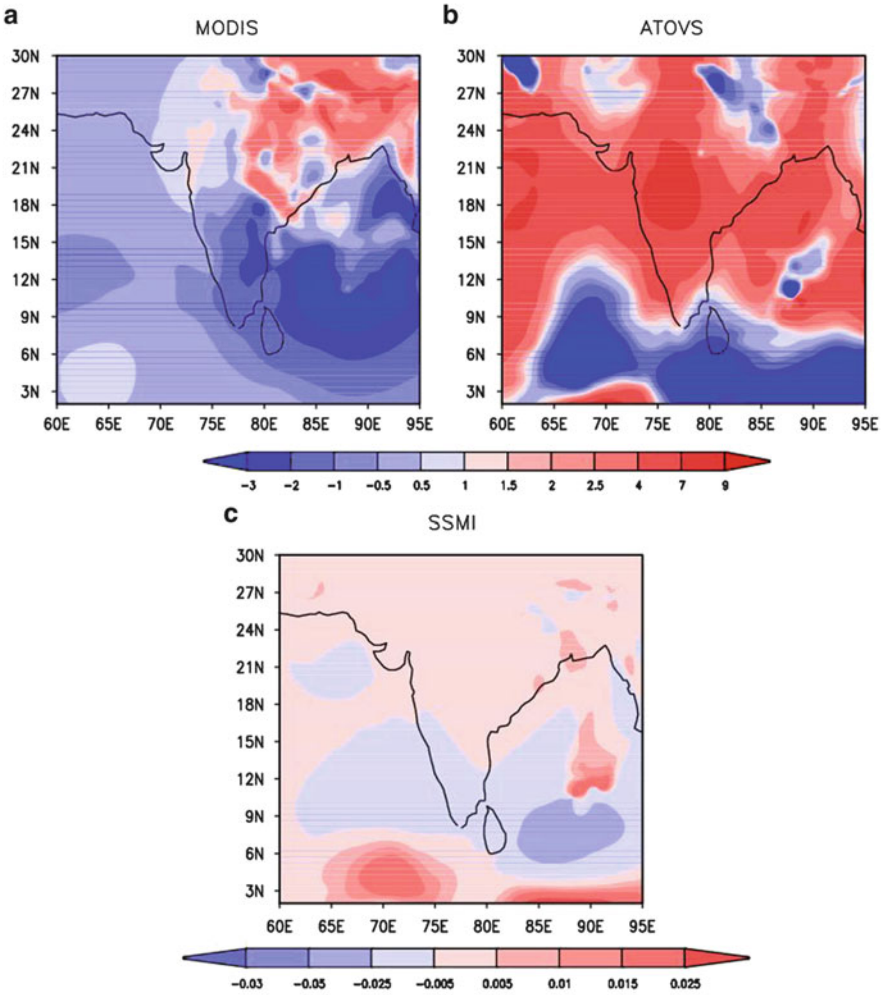


Fig. 26.19 Analysis increment at 850 hPa in the wind speed of (**a**) MODIS, (**b**) ATOVS and (**c**) SSM/I for 18–22 June 2007 depression

which is evident from the Fig. 26.20e–h. It is seen from Fig. 26.22 that the low level south-westerly monsoon flow is prominent in all the four experiments. The maximum wind speed in the CTRL, the SSM/I and the MODIS runs exceed 28 m s^{-1} for most of the times. The cyclonic circulation associated with the depression moves in north-westward direction in all the experiments. However, the initial cyclonic circulation feature simulated by the MODIS experiment is much east of the observed location.

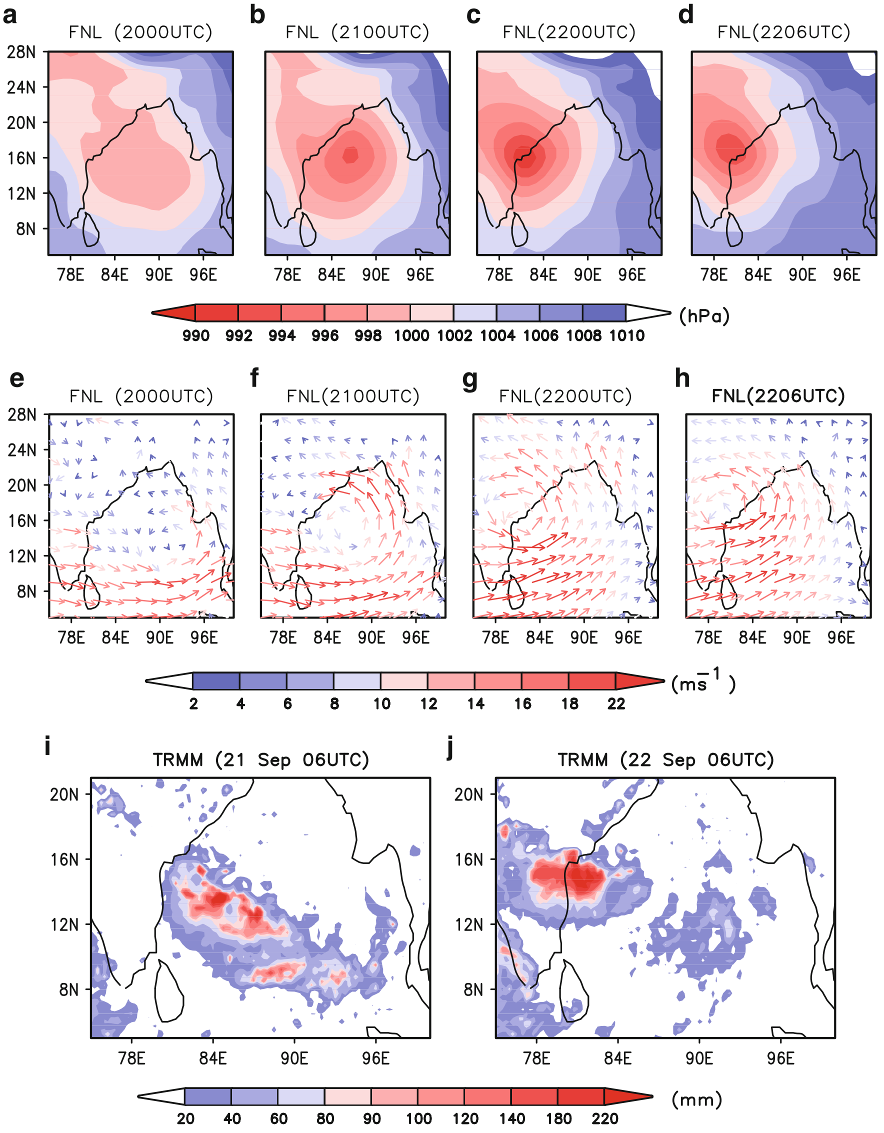


Fig. 26.20 NCEP-FNL analysis sea level pressure fields at 20 June 2007 00 UTC – 22 June 2007 06 UTC (**a–d**), NCEP-FNL lower tropospheric wind speed (850 hPa) (**e–h**) and 24 h accumulated TRMM rainfall (**i–j**)

Forecast Impact Parameter

As in the previous case, the FI parameter has been calculated against QuikSCAT observations to study the impact of assimilation of the satellite observations on the surface wind speed over the sea. From Fig. 26.23, the MODIS run shows negative FI

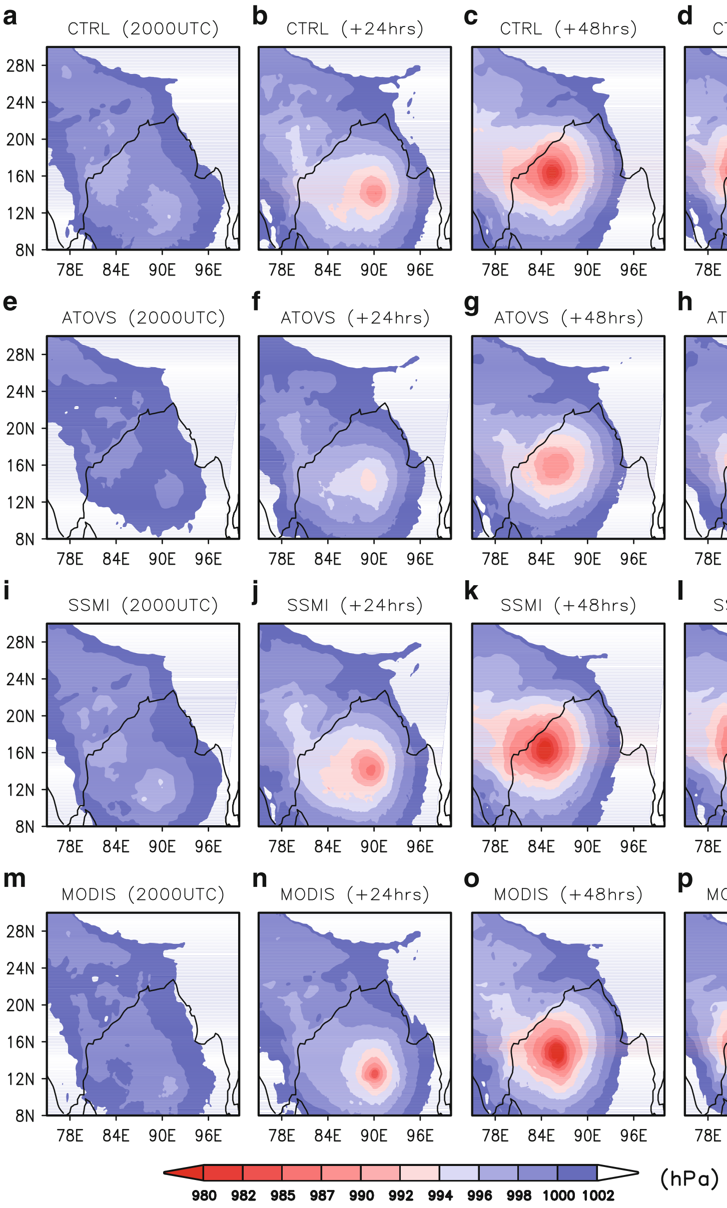


Fig. 26.21 Sea level pressure simulated by CTRL (**a–d**), SSM/I (**e–h**), ATOVS (**i–l**) and MODIS (**m–p**) runs on 20 June 2007 00 UTC and at 24, 48 and 54 h of forecast

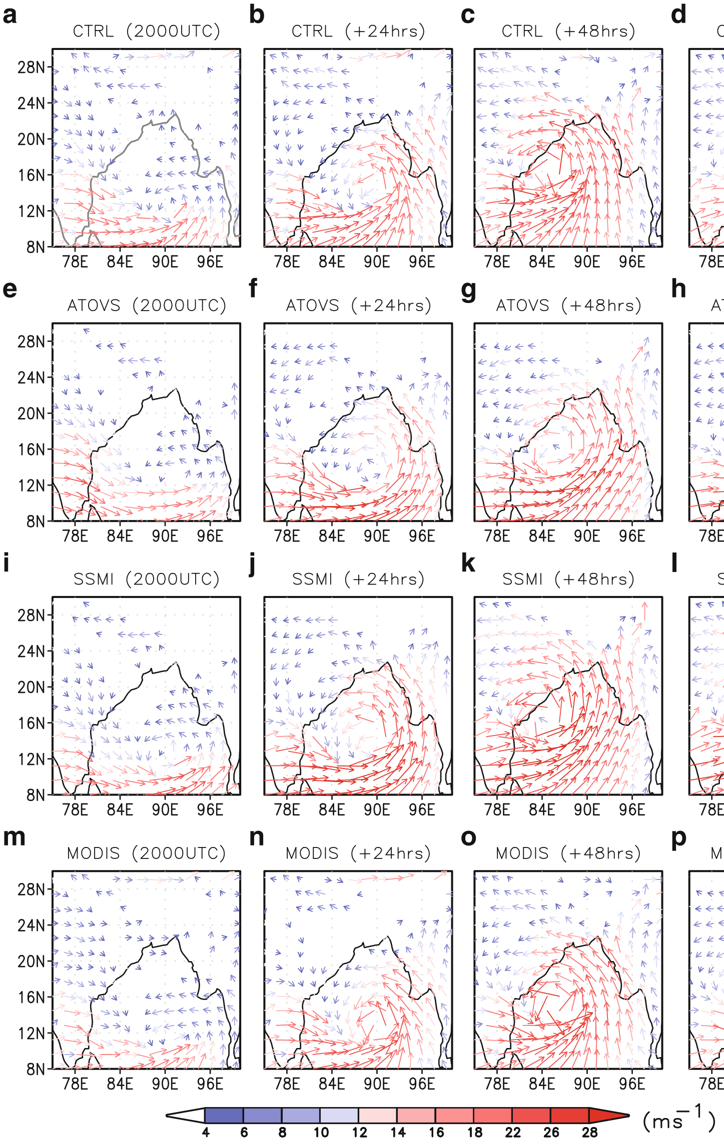


Fig. 26.22 Wind vector simulated by CTRL (**a–d**), SSM/I (**e–h**), ATOVS (**i–l**) and MODIS (**m–p**) runs on 20 June 2007 00 UTC and at 24, 48 and 54 h of forecast

values over most of the sea. However, the assimilation of temperature and humidity profiles from ATOVS and total precipitable water from SSM/I have resulted in significant positive impact.

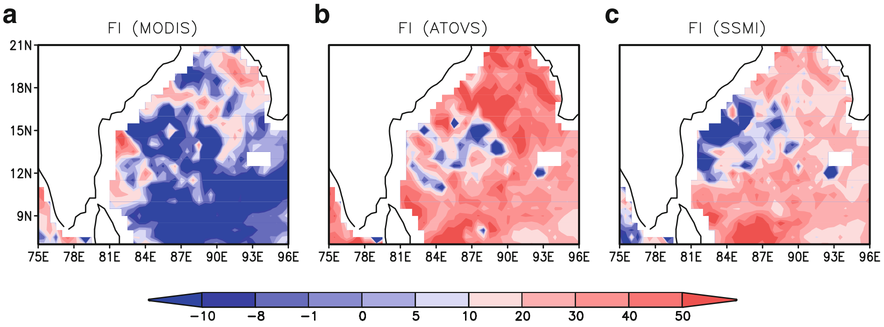


Fig. 26.23 Spatial distribution of forecast impact (FI) for (**a**) MODIS, (**b**) ATOVS, (**c**) SSM/I calculated against QuikSCAT observations for 18–22 June 2007 depression

RMSE of Wind Speed Profiles

The RMSE of wind speed profiles are shown in Table 26.7 for the 18–22 June 2007 depression. The RMSE of wind speed simulated by the ATOVS experiment shows higher error values at most of the levels except 780, 500 and 300 hPa when compared to the CTRL run on 20 June 2007 12 UTC. The MODIS experiment however, shows lower RMSE of wind speed at the first two levels from the surface, after which the errors increase while the SSM/I experiment shows higher error values of wind speed at all the levels in the vertical as compared to the CTRL run. On 21 June 2007 12 UTC, the ATOVS run shows lower RMSE of wind speed at 950, 920 700, 620, 400, 300 and 200 hPa levels while the MODIS run gives lower RMSE of wind speed at the four following levels, viz., 700, 400, 300 and 200 hPa only. The SSM/I run shows better results by simulating lower error values of wind speed at this forecast time at 920, 850, 700, 400, 300 and 200 hPa levels. On 22 June 2007 00 UTC, the ATOVS run shows higher RMSE of wind speed for most of the levels except the first two levels from the surface. The MODIS run, at this time, however, does not provide a conclusive error pattern of wind speed while the SSM/I run shows higher values of the forecast error of wind speed at all the levels in vertical.

Rainfall

TRMM observations of 24 h accumulated precipitation are shown in Fig. 26.20i, j valid for 21 and 22 June 2007 06 UTC. The above figure shows intense rainfall over the Bay of Bengal. Moreover, the location of maximum precipitation is observed over the land (landfall) on 22 June 2007 06 UTC. The rainfall features simulated by all the numerical experiments are shown in Fig. 26.24 for the 18–22 June 2007 depression. Though all the experiments simulate the rainfall features over the Bay of Bengal, none among them are able reproduce the landfall as seen in the TRMM. Also, the precipitation features are overestimated in all the

Table 26.7 Root mean square error of wind speed profiles averaged over the domain for 18–22 June 2007 depression

	20 June 2007 12 UTC				21 June 2007 12 UTC				22 June 2007 00 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	2.78	3.47	1.82	3.72	3.47	2.53	3.95	3.87	6.27	4.52	5.17	7.60
920	3.16	4.06	2.56	4.06	4.23	3.85	4.98	4.06	5.91	5.17	5.67	7.65
850	2.38	2.96	2.48	2.50	6.87	7.49	7.37	6.07	5.80	5.98	5.49	7.33
780	3.56	3.46	3.98	3.60	5.73	6.08	6.46	5.80	5.71	6.00	5.43	7.32
700	4.83	5.05	6.19	4.93	4.81	4.63	4.72	3.44	4.75	6.10	5.45	6.17
620	5.02	5.46	6.41	5.34	3.89	3.65	4.14	3.34	5.23	6.05	5.43	5.73
500	5.32	5.06	6.45	6.72	3.32	4.42	3.77	5.64	5.16	5.23	4.73	5.72
400	2.43	2.65	3.26	3.99	5.71	5.08	5.11	3.90	3.68	3.98	3.90	4.36
300	3.52	3.25	3.58	3.64	4.91	4.88	4.54	3.78	4.12	4.51	3.51	5.74
200	4.16	4.16	7.69	4.02	5.70	4.72	4.06	5.24	17.80	17.51	17.79	17.09

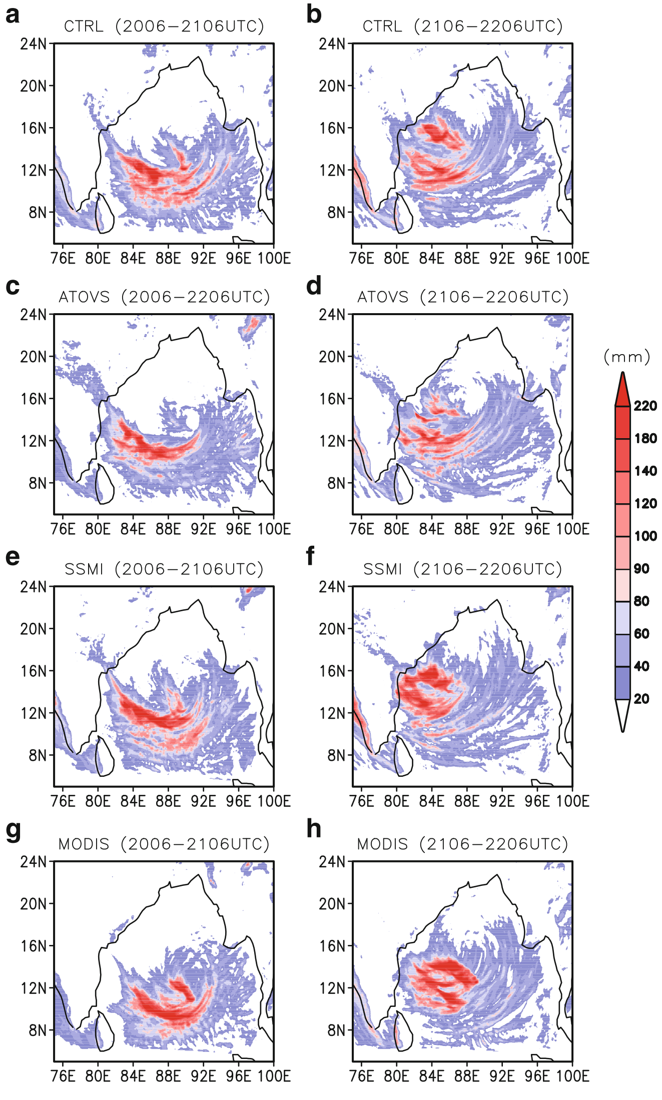


Fig. 26.24 Spatial distribution of 24-h accumulated precipitation for CTRL (**a**, **b**), ATOVS (**c**, **d**), SSM/I (**e**, **f**) and MODIS (**g**, **h**) for 18–22 June 2007 depression

four experiments. However, the SSM/I experiment has simulated the maximum precipitation band closer to that of the observation when compared to the other numerical experiments on the second day of the forecast. It is indeed surprising that the SSM/I experiment contributed to improved forecast of rainfall pattern despite the absence of profiles of vertical temperature and vertical humidity observations. It is to be noted that the observations of "total precipitable water" (TPW) were assimilated

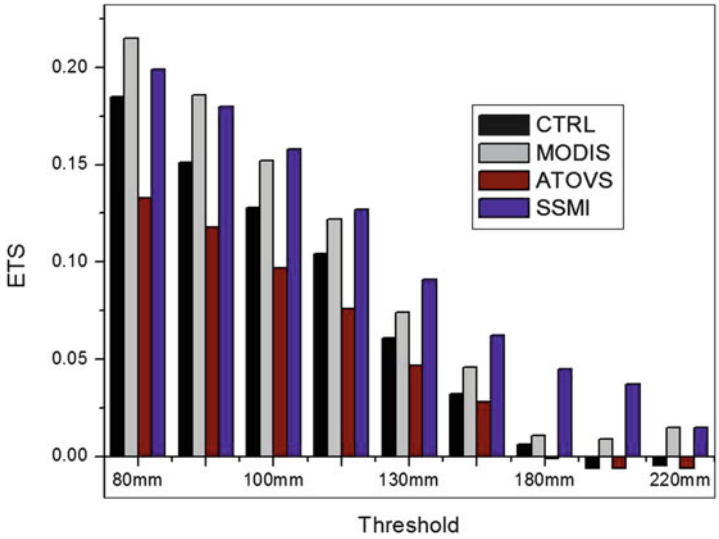


Fig. 26.25 Equitable threat score for 48-h accumulated precipitation for CTRL, ATOVS, SSM/I and MODIS for 18–22 June 2007 depression

in the SSM/I experiment. The retrieved TPW observations contain valuable moisture information and one would expect that the analyzed and forecasted vertical structure of moisture would get improved by assimilating SSM/I TPW. Such assimilation of TPW observations are also known to improve the initial temperature field and the initial geopotential height field (Lu et al. 2011).

ETS and BS

The ETS and BS values with respect to the observed 48 h accumulated precipitation from TRMM are shown in Figs. 26.25 and 26.26 respectively for the 18–22 June 2007 depression. The threshold values are set above 80 mm since the maximum precipitation values exceed 250 mm over the Bay of Bengal. All the experiments show higher skill score at the lower threshold values; however the skill score decreases with the increase in the threshold values. The MODIS run shows the highest skill score of all the four experiment at 80 and 90 mm threshold values. The SSM/I run shows somewhat improved forecast skill at higher threshold values while all the other experiments have failed to do so. The SSM/I and the ATOVS experiments are clearly over-predicting precipitation which is evident from the bias score values. The MODIS run shows almost the same areal frequency of precipitation with respect to the TRMM observations at lower threshold values (refer Fig. 26.26); however the BS value, starts increasing at higher threshold values.

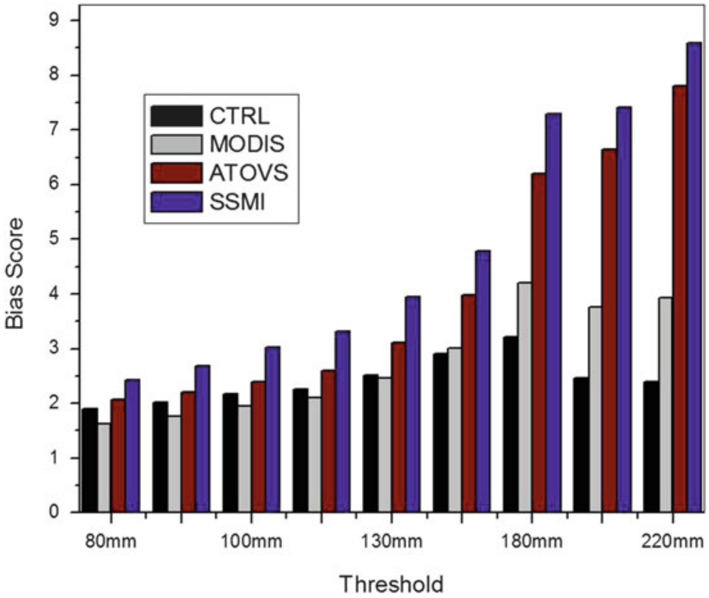


Fig. 26.26 Bias score for 48-h accumulated precipitation for CTRL, ATOVS, SSM/I and MODIS for 18–22 June 2007 depression

Improvement Parameter (ή)

The spatial distribution of η shows almost identical features in all the three assimilation experiment which is evident from Fig. 26.27. The first day of forecast witnessed negligible improvement in the rainfall features simulated by all the assimilation experiments. The second day did show some positive improvement over the peninsular region of the Indian subcontinent.

RMSE of Temperature Profiles

The RMSE values of temperature in vertical are shown in Table 26.8 for the 18–22 June 2007 depression. Lower RMSE values of temperature are seen in 950, 920, 700, 300 and 200 hPa levels for the ATOVS experiment when compared to the CTRL run, while the MODIS run shows lower RMSE of temperature at 780, 620 and 300 hPa levels. The SSM/I run shows improvement in the temperature forecast at 950, 920, 850, 500, 400, 300 and 200 hPa levels on 20 June 2007 12 UTC when compared to the CTRL run. On 21 June 2007 12 UTC, the ATOVS and the MODIS experiments show improved temperature forecast for 6 levels in the vertical, while the SSM/I run shows lower RMSE of temperature for 3 levels only. On 22 June 2007

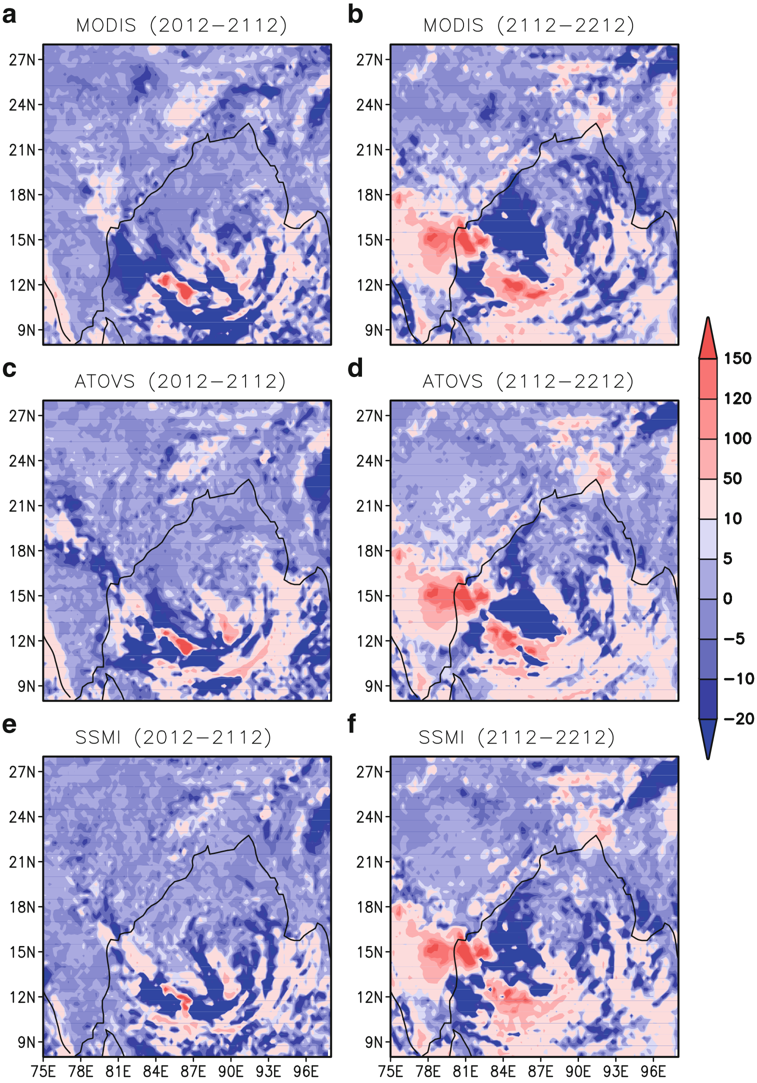


Fig. 26.27 Improvement parameter for 24-h accumulated precipitation for MODIS (**a**, **b**), ATOVS (**c**, **d**) and SSM/I (**e**, **f**) for 18–22 June 2007 depression

00 UTC, the RMSE of temperature are found to be lower in 6 levels for the ATOVS run, lower in 4 levels for MODIS run and lower in 5 levels for the SSM/I run, when compared to the CTRL experiment.

Table 26.8 Root mean square error of temperature profiles averaged over the domain for 18–22 June 2007 depression

	20 June 2007 12 UTC				21 June 2007 12 UTC				22 June 2007 00 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	3.40	2.97	3.75	2.89	3.31	3.09	3.71	2.70	3.73	3.24	3.45	3.12
920	3.54	3.24	3.90	3.12	3.48	3.34	3.80	3.02	2.63	2.15	2.47	2.08
850	2.47	2.14	2.83	2.20	2.85	3.03	3.12	2.91	2.11	2.05	2.18	2.08
780	2.02	2.18	2.01	2.39	2.75	3.16	2.97	3.04	2.23	2.56	2.29	2.37
700	2.11	2.00	2.29	2.23	2.40	2.49	2.37	2.48	2.91	2.82	3.10	2.99
620	1.88	1.88	1.72	1.93	2.39	2.29	2.36	2.45	3.13	2.96	3.09	3.72
500	2.93	3.08	2.94	2.89	2.81	2.54	2.72	2.69	3.01	2.91	2.76	3.15
400	2.32	2.36	2.48	2.25	2.12	2.12	2.09	2.18	2.83	2.88	2.80	2.76
300	2.53	2.40	2.35	2.45	3.30	3.17	3.09	14.71	2.87	2.95	3.07	2.97
200	1.79	1.78	1.94	1.71	6.28	6.15	6.06	12.61	3.84	3.90	3.71	3.73

RMSE of Dew-Point Temperature Profiles

Root mean square error values for dew-point temperature are presented in Table 26.9 for the 18–22 June 2007 depression. Though the ATOVS run shows higher RMSE of dew point temperature at the lower levels, interestingly, the upper levels show an improved forecast of dew-point temperature especially at 200 and 300 hPa levels on the day one of the forecast. The MODIS run shows large RMSE values of dew point temperature for almost all the levels. The assimilation of total precipitable water from SSM/I however does reduce the error values of dew point temperature at 950, 920, 780, 400, 300 and 200 hPa levels on 20 June 2007 12 UTC. On the second day of the forecast, the ATOVS, MODIS and the SSM/I runs do not show significant improvement in the forecast of the dew-point temperature while on 22 June 2007 00 UTC, the SSM/I experiment did show improved better result with lower RMSE of dew point temperature than the other experiments.

26.4.4 Conclusions

The impact of assimilating satellite observations of temperature, humidity and total precipitable water on the prediction of three monsoon depressions which formed over the Bay of Bengal are investigated. Out of the three depression cases, the first one is a land depression, the second one is a weak depression and third one is a strong monsoon depression event over the Bay of Bengal. Four simulations are undertaken for each of the three monsoon depression cases. The simulation without assimilation of any observations is called the ‘CTRL run’, while simulation that assimilated ATOVS, MODIS, SSM/I observations are called ATOVS run, MODIS run, SSM/I run respectively. The results of the model simulations are compared with one another and also with the TRMM and QuikSCAT observations as well as with NCEP-FNL analysis. The general conclusions based on the above investigation are as follows.

The results of the study provide direct and good evidence of the impact of assimilation of temperature and humidity profiles and the total precipitable water to a certain extent. From the analysis increment, it can be seen that the assimilation of the ATOVS temperature and humidity profiles results in larger analysis increment of 850 hPa wind speed as compared to the other assimilation experiments. The simulated sea level pressure field of MODIS experiment is found to be relatively in better agreement with that of the NCEP-FNL analysis in the first two cases. Also the assimilation of the MODIS temperature and humidity profiles does produce significant improvement in the precipitation patterns when compared to the TRMM observation. Equitable threat score and bias score calculated to quantitatively validate the precipitation forecast shows that the MODIS experiment yields better results for the land depression and the weak depression events. The high-resolution dense observations associated with MODIS data can possibly be the reason for the better forecast produced by the MODIS experiment. The SSM/I run, however, shows

Table 26.9 Root mean square error of dew point temperature profiles averaged over the domain for 18–22 June 2007 depression

	20 June 2007 12 UTC				21 June 2007 12 UTC				22 June 2007 00 UTC			
Pressure levels (hPa)	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I	CTRL	ATOVS	MODIS	SSM/I
950	3.07	3.01	3.34	2.77	4.28	3.38	4.31	3.20	2.85	2.93	3.00	2.21
920	2.67	2.72	2.85	2.51	4.67	3.67	4.61	3.65	3.14	2.94	3.43	2.23
850	3.04	3.36	2.71	3.13	3.25	2.89	3.00	2.99	2.68	2.48	2.74	2.64
780	3.85	3.31	3.85	3.79	3.75	3.88	3.73	3.61	2.36	2.62	2.81	2.74
700	3.55	3.61	3.32	3.85	5.17	5.50	5.57	4.00	3.45	6.22	3.86	4.27
620	3.45	4.18	4.66	4.28	5.79	6.04	4.45	3.90	6.72	9.16	8.82	5.88
500	6.57	6.86	7.33	6.88	5.89	7.45	6.76	6.54	6.13	7.55	6.12	4.96
400	7.38	5.32	8.06	7.68	10.08	7.11	9.11	9.00	5.26	6.10	5.70	6.22
300	10.74	8.39	15.51	9.45	19.59	20.26	22.26	20.43	7.72	6.37	6.85	8.00
200	10.79	5.70	8.57	7.31	4.48	5.82	4.38	3.01	9.14	4.46	9.34	7.52

improved forecast of rainfall patterns in the third case study. The ATOVS experiment shows significant forecast impact on the surface wind speed which is validated against the QuikSCAT observations. There is some discernable reduction in the RMSE of wind speed, temperature and dew-point temperature in some levels for all the assimilation experiments. The assimilation experiments do show some difficulty in accurately simulating the upper level features of the troposphere especially for the dew-point temperature, which is evident from the higher RMSE of dew point temperature at these heights. However, the assimilation of total precipitable water from SSM/I does provide for some positive improvements in the dew-point temperature simulation. The absence of vertical temperature and humidity profiles from the SSM/I sensor has contributed to the lower impact of assimilation for the SSM/I experiment. Despite the positive impact on the simulation and forecast of the monsoon depressions with 3D-VAR assimilation technique, the results of the study also reveals marked positioning error of the depression centre in the forecast in most of the numerical experiments. The above marked positioning errors at large forecast lead time clearly indicate that issues other than data assimilation are also important and hence a solution of overcoming the marked positioning error problem requires further detailed investigations, addressing issues, which are beyond the scope of this study.

Acknowledgements The authors acknowledge NASA, NOAA and NCEP for providing access to the various satellite (MODIS, ATOVS, SSM/I, QuikSCAT and TRMM) observations as well as the NCEP GFS and FNL fields. The first author acknowledges funding support for this work from Space Application Centre, Ahmadabad, India. The second author acknowledges the CSIR and IIT Kharagpur, India for providing research fellowship and all research facilities to undertake this work. The above work includes a small portion of the PhD work of the second author.

References

Adler RF et al (2000) Tropical rainfall distributions determined using TRMM combined with other satellite and rain gauge information. J Appl Meteor 39:2007–2023

Bergthorsson P, Doos B (1955) Numerical weather map analysis. Tellus 7(3):329–340

Brennan MJ, Hennon CC, Knabb RD (2009) The operational use of QSCAT ocean surface vector winds at the national hurricane center. Weather Forecast 24:621–645

Chen SH (2007) The impact of assimilating SSM/I and QuikSCAT satellite winds on Hurricane Isidore simulation. Mon Wea Rev 135:549–566

Chen SH et al (2008) A study of the characteristics and assimilation of retrieved MODIS total precipitable water data in severe weather simulations. Mon Wea Rev 136:3608–3638

Chou SH et al (2006) Assimilation of atmospheric infrared sounder (AIRS) data in a regional model. In: Preprints, 14th conference on satellite meteorology and oceanography. American Meteorological Society, Atlanta, CD_ROM, P5.12

Colle BA, Westrick KJ, Mass CF (1999) Evaluation of MM5 and Eta 10 precipitation forecast over the Pacific Northwest during the cool season. Weather Forecast 14:137–154

Cox R, Bauer BL, Smith T (1998) A mesoscale model intercomparison. Bull Am Meteorol Soc 79:265–283

Cressman GP (1959) An operational objective analysis system. Mon Wea Rev 87:367–374

Ebuchi N, Graber HC, Caruso MJ (2002) Evaluation of wind vectors observed by QuikSCAT/SeaWinds using ocean buoy data. J Atmos Oceanic Technol 19:2049–2062

English SJ et al (2000) A comparison of the impact of TOVS and ATOVS satellite sounding data on the accuracy of numerical weather forecasts. Q J R Meteorol Soc 126:2911–2931

Fan X, Tilley JS (2005) Dynamic assimilation of MODIS-retrieved humidity profiles within a regional model for high-latitude forecast applications. Mon Wea Rev 133:3450–3480

Gal-Chen T, Schmidt BD, Uccellini LW (1986) Simulation experiments for testing the assimilation of geostationary satellite temperature retrievals into a numerical prediction model. Mon Wea Rev 114:1213–1230

Gandin LS (1963) Objective analysis of meteorological fields. Gridromet (in Russian) English translation by the Israel program for scientific translations, Leningrad, 242 pp

Gilchrist B, Cressman GP (1954) An experiment in objective analysis. Tellus 6:309–318

Godbole RV (1977) The composite structure of the monsoon depression. Tellus 29:25–40

Govindankutty M et al (2008) The impact of assimilation of MODIS observations using WRF-VAR for the prediction of a monsoon depression during September 2006. Open Atmos Sci J 2008; 2:68–78

Govindankutty M et al (2010) Impact of 3D-VAR assimilation of Doppler weather radar wind data and IMD observation for the prediction of a tropical cyclone. Int J Remote Sens 31:6327–6345

Haddad ZS et al (1997) The TRMM (Day-1) radar/radiometer combined rain-profiling algorithm. J Meteor Soc Jpn 75:799–809

Hoffman RN, Leidner SM (2005) An introduction to the near-realtime QuikSCAT data. Weather Forecast 20:476–493

Hollinger J (1989) DMSP special sensor microwave/imager calibration/validation. Naval Research Laboratory Final Report 153, Naval Research Laboratory, Washington, DC

Kelly G et al (2008) Impact of SSM/I observations related to moisture, clouds, and precipitation on global NWP forecast skill. Mon Weather Rev 136:2713–2716

Krishnamurti TN, Ardanuy P (1980) The 10 to 20 day westward propagating mode and breaks in the monsoon. Tellus 32:15–26

Krishnamurti TN, Molinari J, Pant HL, Wong V (1977) Downstream amplification and formation of monsoon disturbances. Mon Wea Rev 105:1281–1297

Kumar P, Singh R, Joshi PC, Pal PK (2011) Impact of additional surface observation network on short range weather forecast during summer monsoon 2008 over Indian subcontinent. J Earth Sys Sci 120:53–64

Lahoz W, Khattatov B, Menard R (eds) (2010) Data assimilation: making sense of observations. Springer, Heidelberg

Lipton AE, Vonder Haar TH (1990) Mesoscale analysis by numerical modeling coupled with sounding retrieval from satellites. MonWea Rev 118:1308–1329

Lipton AE et al (1995) Satellite-model coupled analysis of convective potential in Florida with VAS water vapor and surface temperature data. Mon Wea Rev 123:3292–3304

Lu Q, Yang X, Wu C, Zheng J, Qin D, Yang H, Zhang P (2011) An initial study on assimilating satellite-derived total precipitable water in a variational assimilation system. In: PIERS proceedings, Suzhov, 12–16 Sept, pp 576–580

Manabe S, Holloway L, Stone HM (1970) Tropical circulation in a time integration of a globe model of the atmosphere. J Atmos Sci 27:580–613

Mooley DA, Shukla J (1989) Main features of the westward moving low pressure systems which form over the Indian region during the monsoon season and their relationship with the monsoon rainfall. Mausam 40: 137–152

Nitta T (1970) A study of generation and conversion of eddy available potential energy in the tropics. J Meteorol Soc Jpn 48:524–528

Panofsky HA (1949) Objective weather-map analysis. J Appl Meteor 6:386–392

Parrish DF, Derber JC (1992) The national-meteorological-centers' spectral statistical interpolation analysis system. Mon Wea Rev 120(8):1747–1763

Pu Z et al (2002) The impact of TRMM data on mesoscale numerical simulation of super typhoon Paka. Mon Wea Rev 130:2248–2258

Rajan D et al (2002) Impact of NOAA/TOVS derived moisture profile over the ocean on global data assimilation and medium range weather forecasting. Atmosphera 15:223–236

Rakesh V et al (2009a) Intercomparison of the performance of MM5/WRF with and without satellite data assimilation in short-range forecast applications over the Indian region. Meteorol Atmos Phys105 133–155

Rakesh V, Singh R, Pal PK, Joshi PC (2009b) Impact of satellite observed surface wind and total precipitable water on WRF short-range forecasts over Indian region during monsoon 2006. Weather Forecast 24:1706–1731

Ruggiero FH et al (1999) Coupled assimilation of geostationary satellite sounder data into a mesoscale model using the Bratseth analysis approach. Mon Wea Rev 127:802–821

Saha KR, Sanders F, Shukla J (1981) Westward propagating predecessors of monsoon depressions. Mon Wea Rev 109:330–343

Sandeep S, Chandrasekar A, Singh D (2006) The impact of assimilation of AMSU data for the prediction of a tropical cyclone over India using a mesoscale model. Int J Remote Sens 27:4621–4653

Schlesinger RE (1994) Heat, moisture and momentum budgets of isolated deep midlatitude and tropical convective clouds as diagnosed from three-dimensional model output part-I: control experiments. J Atmos Sci 51:3649–3673

Shirtliffe G (1999) QuikSCAT science data products user's manual. Jet Propulsion Laboratory Publication, Pasadena, CA, p 90

Shukla J (1978) CISK, barotropic and baroclinic instability and the growth of monsoon depressions. J Atmos Sci 35:495–500

Sikka DR (1977) Some aspects of life history, structure and movement of monsoon depression. Pure Appl Geophys Basel 115:1501–1529

Sikka DR (1980) Some aspects of large scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in planetary and regional scale circulation parameters. Proc Indian Acad Sci Earth Planet Sci 89:179–195

Simon B, Rahman SH (2003) Application of MODIS data for mesoscale processes. In: Proceedings scale interaction and monsoon variability, Munnar

Singh R et al (2008a) The impact of variational assimilation of SSM/I and QuikSCAT satellite observations on the numerical simulation of Indian Ocean tropical cyclone. Wea. Forecasting 23:460–476

Singh R et al (2008b) Impact of atmospheric infrared sounder data on the numerical simulation of a historical mumbai rain event Wea. Forecasting 23:892–913

Singh R, Pal PK, Joshi PC (2010) Assimilation of Kalpana very high resolution radiometer water vapor channel radiances into a mesoscale model. J Geophys Res 115:D18124

Singh R et al (2011a) Assimilation of the multisatellite data into the WRF model for track and intensity simulation of the Indian Ocean tropical cyclones. Meteorol Atmos Phys 111:103–119

Singh R, Kishtawal CM, Pal PK (2011b) A comparison of the performance of Kalpana and HIRS water vapor radiances in the WRF 3D-Var assimilation system for mesoscale weather predictions. J Geophys Res 116:D08113

Singh R, Kumar P, Pal PK (2011c) Assimilation of Oceansat-2-Scatterometer-derived surface winds in the weather research and forecasting model. IEEE Trans Geosci Remote Sens 99:1–7

Sinha PK, Chandrasekar A (2010) Improvement of mesoscale forecasts of monsoon depressions through assimilation of QuikSCAT wind data: two case studies over India. Open Atmos Sci J 4:160–177

Wallace JM (1971) Spectral studies of tropospheric wave disturbances in the tropical western pacific. Rev/Geophys Space Phys 9:557–612

Weissman DE, Bourassa MA, Tongue J (2002) Effects of rain rate and wind magnitude on SeaWinds scatterometer wind speed errors. J Atmos Ocean Technol 19:738–746

Wang B (ed) (2006) The Asian monsoon. Springer, Heidelberg, 787 pp

Wilks D (2006) Statistical methods in the atmospheric sciences: an introduction, 2nd edn. Academic , San Diego, 627 pp

Yanai M, Esbensen S, Chu JH (1973) Determination of bulk properties of tropical cloud clusters from large-scale heat and moisture budgets. J Atmos Sci 30:611–627

Zapotocny TH et al (2007) A two season impact study of satellite and in situ data in the NCEP global data assimilation system. Weather Forecasting 22:887–909

Zhang X et al (2007) The impact of multisatellite data on the initialization and simulation of Hurricane Lili's (2002) rapid weakening phase. Mon Wea Rev 135:526–548

Zou X, Xiao Q (2000) Satellites on variational bogus data assimilation scheme initialization and simulation of a mature hurricane. J Atmos Sci 57:836–860

Chapter 27
Parameter Estimation Using an Evolutionary Algorithm for QPF in a Tropical Cyclone

Xing Yu, Seon Ki Park, and Yong Hee Lee

Abstract In this study the quantitative precipitation forecast (QPF) related to a tropical cyclone is performed using a high-resolution mesoscale model and an evolutionary algorithm. For this purpose two parameters of the Kain-Fritsch convective parameterization scheme, in the Weather Research and Forecasting (WRF) model, are optimized using the micro-genetic algorithm (GA). The autoconversion rate (c) and the convective time scale (T_c) are target parameters. The fitness function is based on a QPF skill score. Typhoon Rusa (2002) is simulated in a grid spacing of 25 km. The default value of c is $0.03\,\mathrm{s}^{-1}$ while that of T_c is limited to a range between 1800 s and 3600 s as a function of grid resolution. To produce the best QPF skill, at least for this tropical cyclone case, c is optimized to $0.0004\,\mathrm{s}^{-1}$ and T_c to 1922s. Our results indicate that parameters of subgrid-scale physical processes need to be adjusted to produce better QPF in a tropical cyclone, sometimes to values far different from the default values in a numerical model. Such adjustment may be dependent on the characteristics of weather systems as well as geographical locations.

X. Yu
Tropical Marine Science Institute, National University of Singapore, 18 Kent Ridge Road, Singapore 119227, Singapore
e-mail: tmsyx@nus.edu.sg

S.K. Park (✉)
Department of Atmospheric Science and Engineering, Ewha Womans University, Seoul 120-750, Republic of Korea
e-mail: spark@ewha.ac.kr

Y.H. Lee
Forecast Research Laboratory, National Institute of Meteorological Research,
Korea Meteorological Administration, Seoul 156-720, Republic of Korea

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7_27,

27.1 Introduction

Numerical weather prediction (NWP) models have uncertainties involved in the subgrid-scale physical processes, most of which have to be parameterized (Navon 2009). In the formation of cloud and precipitation, the convection and microphysical processes are important, and interactions among hydrologic, boundary layer and land surface processes are conducted mostly by cumulus convection (Arakawa 2004). Parameterizations, including convective parameterization (CP), contain numerous parameters whose values are not known precisely. Parameters in CP scheme can affect the model performance and hence the forecast accuracy. Therefore optimization of parameters can potentially improve the accuracy of numerical forecasts.

Genetic algorithms (GAs) have been used for global search by combining the use of random number generation and information from precious iterations to evaluate and improve a population of points at a time (Goldberg 1989). They are based on natural genetic and selection mechanism, and ideas of constructing an optimization procedure are borrowed from Genetics (Holland 1975, 1992). With capability of achieving global optimal solution, GAs have been applied to various atmospheric problems (Fang et al. 2009; Jackson et al. 2004; Kishtawal et al. 2003; Singh et al. 2005a, b).

A standard GA had been applied to a heavy rainfall case in Korea by Lee et al. (2006) to improve the quantitative precipitation forecasting (QPF) through optimization of both a physical and a computational parameter. This study focuses on optimal parameter estimation in a CP scheme to improve the QPF in the Weather Research and Forecasting (WRF) model using a micro-GA. Section 27.2 describes the parameters to be optimized, the typhoon case and the experiments design, and Sect. 27.3 describes the computational procedures of the micro-GA. Results are presented in Sect. 27.4, and conclusions are provided in Sect. 27.5

27.2 Methods

27.2.1 Description on Parameters

Deep convection is generally parameterized when horizontal grid spacing is greater than about 10 km but CP is typically avoided at higher resolution (Kain et al. 2008). In this study, we will investigate the resolution dependency of performance of the Kain-Fritsch (KF) CP scheme (Kain 2004; Kain and Fritsch 1993), in the process of GA optimization, by comparing three simulation runs – one with no CP scheme, another with the default KF scheme, and the other with the improved KF scheme. Here, two parameters are optimized by GA.

One parameter to be optimized is a convective time scale (T_c). As illustrated by Fritsch and Chappell (1980), the CP problem is mainly to determine T_c and the grid element temperature (water vapor) after convection. The KF scheme assumes that at least 90 % of the environmental convective available potential energy (CAPE) is consumed over T_c, which is limited between 1800 s and 3600 s (Kain 2004; Kain and Fritsch 1993). T_c is proportional to the model grid spacing and inversely proportional to averaged winds between 500 hPa and the lifting condensation level.

The KF scheme is originally designed for a mesoscale model with grid spacing of 20–50 km, thus T_c matches well with the lifetime of a convective cell. This scheme is dictated by the time it takes the cloud to grow to the point that precipitation forms and the time it then takes the precipitation to fall to low levels (Emanuel 1994). However, given typical mean horizontal wind speeds of 10 m s^{-1}, T_c will be fixed to 1800 s if grid spacing is less than 18 km. Narita and Ohmori (2007) and Saito et al. (2007) suggested that a shorter T_c of 900 s can improve the QPF with a 10 km grid spacing in the Japan Meteorological Agency's operational mesoscale model.

The other parameter to be optimized, c, controls microphysical feedback from the parameterized convection to its environment (Correia et al. 2008), and its mathematical formulation follows Ogura and Cho (1973) as function of the amount of condensate at the layer bottom, the amount of condensate lost by the parameterized updraft, the layer depth and updraft velocity.

The value of constant c is 0.01 s^{-1} in the old KF scheme (Kain and Fritsch 1990, 1993), and is set to 0.03 s^{-1} in the new KF scheme of the WRF model. 2004). Correia et al. (2008) found a smaller value of 0.005 s^{-1} directly increases the hydrometeor feedback at the expense of the convective precipitation. Many studies have shown the auto-conversion processes are responsible for determining the organization and structure of convective systems (Correia et al. 2008; Tao et al. 1995; Zhu and Zhang 2004).

27.2.2 Case Description and Experiments Design

Typhoon Rusa (2002) landed over the southwestern part of the Korean Peninsula (KP) at 0600 UTC 31 August, 2002. Its central sea-level pressure stayed between 950 and 960 hPa. It moved across the Korea Peninsula and produced a 100-year record-breaking heavy precipitation amount of 870 mm d^{-1} at Kangnung, located at the central-eastern coast of the KP (Gu et al. 2005; Lee and Choi 2010; Park and Lee 2007).

In this study, the numerical model WRF Version 3.2 is used. The computational domain has a size of $1,800\,\text{km} \times 2,100\,\text{km}$ centered at (127°E, 34°N), with a grid spacing of 25 km. The initial and boundary conditions are supplied by the NCEP Final Analysis (FNL) data on $1° \times 1°$ with 6 h interval and the 3DVAR is used to assimilate conventional surface and sounding observations. The simulations are initiated at 00 UTC 30 August 2002, and ended at 12 UTC 31 August 2002. Schemes for physical processes include: the YSU PBL, the WSM3 simple ice microphysics, the Dudhia radiation and RRTM, and the Noah land surface model. For the CP scheme, three groups of experiments are set up: (1) the KF scheme with default parameters values (KFEXs), (2) the KF scheme with the parameter values optimized by GA (OPTMs) and (3) no CP scheme (NOCPs).

27.3 Micro-Genetic Algorithm (Micro-GA)

27.3.1 The Optimization Process

The micro-GA, suggested by Goldberg (1989) and Krishnakumar (1989) is a small-population genetic algorithm with reinitialization while standard GAs mostly use large populations to achieve diversity upon "convergence". It requires less computational time than standard GAs (Krishnakumar, 1989; Lee et al. 2005; Wang et al. 2010). The procedural details of micro-GA were described by Carroll (1996), Liong et al. (2005), and Wang et al. (2010). With a small population there will be rapid convergence to a possible suboptimal solution, by generating new population members as soon as a convergence has been achieved in a GA cycle.

The micro-GA has been demonstrated to yield marked improvement over conventional large-population GAs. Although the range of application of micro-GA is becoming extensive, its applicability has yet to be explored fully, and is certainly needed for atmospheric design problems where computational requirement is enormous.

For experiments in this study, the micro-GA is initialized with a random sample of individual solutions, with the population size set to five following Lee et al. (2005) and Wang et al. (2010). The chromosomes are generated based on a tournament selection method in order to select parent genes on which the uniform crossover operation is applied. The micro-GA does not have mutation operations, and the algorithm stops when the prescribed number of generations (100 in this study) is reached. The ranges of T_c and c to search are set to $600\,\mathrm{s} \leq T_c \leq 3600\,\mathrm{s}$ and $0.0001\,\mathrm{s}^{-1} \leq c \leq 0.1\,\mathrm{s}^{-1}$, respectively. The chromosome length is set to 10; thus, the precision for T_c and c is about 3 s (i.e., $3000/2^{10}$) and $0.0001\,\mathrm{s}^{-1}$ (i.e., $0.0999/2^{10}$), respectively.

27.3.2 Fitness Function

The fitness function to be optimized is defined by using a QPF skill score – the equitable threat score (*ETS*) (see Hamill 1999; Lee et al. 2006; Yang and Tung 2003). The forecast amounts are computed from the 24-h accumulated total (grid resolved + parameterized) precipitation at a forecast period of 12–36 h (from 1200 UTC 30 to 1200 UTC 31 August 2002). The 24-h observations of total precipitation are based on the hourly precipitation data from 615 Automatic Weather Stations (AWSs) of the Korean Meteorological Administration (KMA).

However, some AWSs were destroyed during the 24 h period due to intensive rainfall. Therefore, we interpolated the irregular (also with different amount of stations) hourly precipitation to the corresponding model grid spacing (i.e., 25 km), and then added the above gridded precipitation in 24 h to a total precipitation observation. We only consider the model grids falling into the territory of South Korea, thus the total number of points N to be verified is 129 for a 25 km grid spacing.

27.4 Results

The micro-GA is performed for optimal estimation of two parameters in the KF scheme – the convective time scale (T_c) and the auto-conversion rate (c), for the case of Typhoon Rusa (2002) using the WRF. The optimized values are obtained in 100 generations (with population size of 5) using the micro-GA. For the parameter T_c, the optimized value is about 1922 s which locates in the default range. However, for the parameter c, the optimized value is $0.0004\,\mathrm{s}^{-1}$ that is two orders smaller than the default value ($\mathrm{c} = 0.03\,\mathrm{s}^{-1}$). A smaller c in the KF scheme implies that the condensed cloud water is more detrained to the grid-resolved environment rather than converted to convective precipitation falling down (Correia et al. 2008).

Figure 27.1 compares the ETSs of three groups of experiments (NOCP, KFEX and OPTM). It turns out that the ETSs with OPTM are much higher than those with NOCP and KFEX in all thresholds, indicating a significant improvement in quantitative precipitation forecasts. The sum of ETS with OPTM, KFEX and NOCP is 7.84, 2.78 and 2.33, respectively. Here, both KFEX and NOCP have almost no forecast skill for light precipitations ($< 60\,\mathrm{mm\,d}^{-1}$), while NOCP generally performs better than KFEX at light to moderate precipitation rate ($< 200\,\mathrm{mm\,d}^{-1}$). For heavy precipitate rate ($> 200\,\mathrm{mm\,d}^{-1}$), NOCP loses forecast skill abruptly but KFEX has considerable forecast skill – higher than that for moderate precipitation rate and even closer to OPTM. This indicates that the KF scheme is useful for forecasting high precipitation rate at the grid resolution of 25 km, relevant to the forecast with optimized parameters.

Figure 27.2a depicts horizontal distribution of the observed 24-h precipitation amount from 1200 UTC 30 to 1200 UTC 31 August 2002. Heavy precipitations ($> 100\,\mathrm{mm\,d}^{-1}$) are observed at the northeastern to southwestern parts of South Korea with three regions of local maximum ($> 400\,\mathrm{mm\,d}^{-1}$) located at mountainous areas – one at the north-eastern coast of South Korea near the eastern side of the Taebaek Mountain Range and two at the southern side of the Sobaek Mountain Range. Correspondingly, light to moderate precipitation (5–100 $\mathrm{mm\,d}^{-1}$) occurs at the northwestern part of South Korea. More details of the precipitation processes in this typhoon have been analyzed by Park and Lee (2007) and Lee and Choi (2010).

Figure 27.2b represents the 24-h accumulated precipitation from 12 UTC 30 August to 12 UTC 31 August 2002 from three experiments (NOCP, KFEX and OPTM) with a 25 km horizontal resolution. OPTM shows the best agreement with

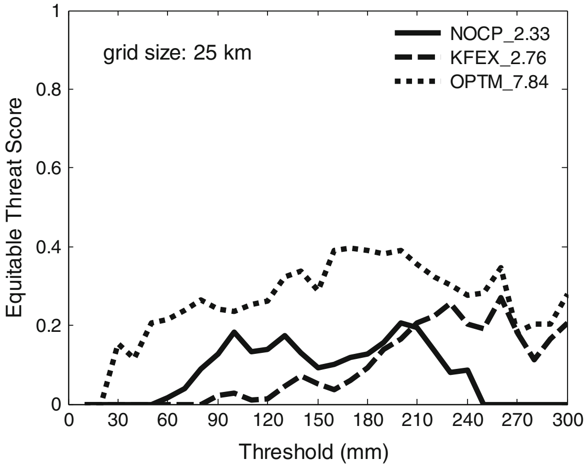


Fig. 27.1 The ETSs of precipitation forecast in terms of various thresholds (in mm) of NOCP (*solid*), KFEX (*dashed*) and OPTM (*dash-dotted*) for a horizontal resolution of 25 km. Numbers beside each experiment name are the sum of ETSs. The precipitation is accumulated from 1200 UTC 30 to 1200 UTC 31 August 2002

observation in terms of both location and amount of precipitation (Fig. 27.2d). Both NOCP (Fig. 27.2b) and KFEX (Fig. 27.2c) have spurious local precipitation maxima ($>$ 200 mm d^{-1}) at central-western coast of South Korea, where the observation records only 50–100 mm d^{-1} (see Fig. 27.2a). The spurious maximum of KFEX is larger than that of NOCP, resulting in lower forecast skills of KFEX at thresholds less than 210 mm (see Fig. 27.1). Meanwhile, those false local maxima did not appear in experiments with the optimized parameters.

27.5 Conclusions

In this study, optimal estimation of two parameters in the Kain-Fritsch convective parameterization scheme is performed to improve the quantitative precipitation forecast (QPF) for Typhoon Rusa (2002), which brought heavy rainfall in the Korean Peninsula. The micro-GA is applied to find the best parameter values with a QPF skill score as a fitness function, using the WRF model at a grid spacing of 25 km. Among the two parameters, the auto-conversion rate c has a default value of 0.03 s^{-1} while the convective time scale T_c has a default range between 1800 s and 3600 s.

It turns out that, in order to produce the highest QPF skill at least for this tropical cyclone case, c should be optimized to 0.0004 s^{-1}; thus the auto-conversion is considered to be effectively turned off. The optimized T_c value is 1922 s. By applying a set of two optimized parameters, the performance of WRF with a 25 km resolution has been maximized in terms of the QPF skill for Typhoon Rusa (2002).

In this study, we have applied the micro-GA only to improve the QPF skill of a tropical cyclone. However, other forecast aspects of tropical cyclones, such as track and intensity, can be also improved via optimal parameter estimation by defining different fitness function (e.g., squared error of track distance, mean sea

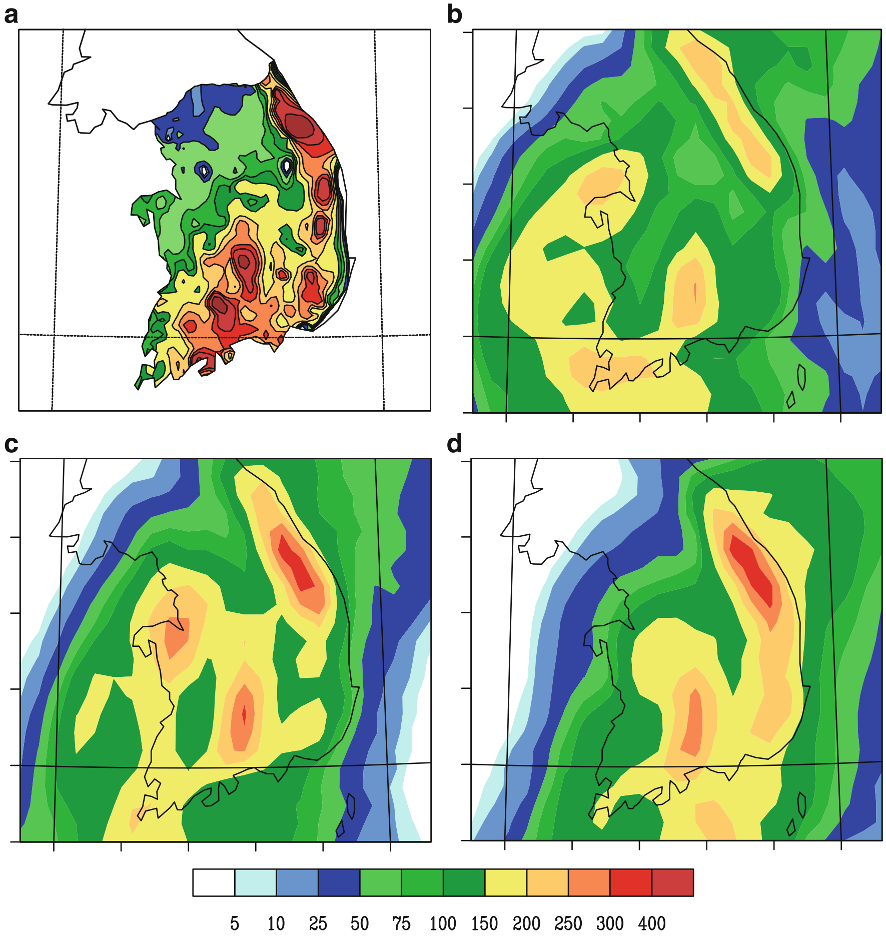


Fig. 27.2 Horizontal distribution of accumulated total precipitation from 1200 UTC 30 to 1200 UTC 31 August 2002 (in mm). (**a**) observation, (**b**) NOCP (**c**) KFEX and (**d**) OPTM with a horizontal resolution of 25 km

level pressure, or maximum surface wind speed between the observations and model solutions).

Our results indicate that such phenomena can be forecasted more accurately, at least for heavy rainfall, via optimal parameter estimation of operational numerical weather prediction models.

Acknowledgements This research is supported by the Ministry of Education, Science and Technology, Korea, through the National Research Foundation grant (No. 2009-0093464).

References

Arakawa A (2004) The cumulus parameterization problem: past, present, and future. J Climate 17:2493–2525

Carroll DL (1996) Genetic algorithms and optimizing chemical oxygen-iodine lasers. Developments in theoretical and applied mechanics vol XVIII School of Engineering, The University of Alabama, Birmingham, pp 411–424

Correia J, Arritt RW, Anderson CJ (2008) Idealized mesoscale convective system structure and propagation using convective parameterization. Mon Wea Rev 136:2422–2442

Emanuel KA (1994) Atmospheric convection. Oxford University Press, New York, 580 pp

Fang CL, Zheng Q, Wu WH, Dai Y (2009) Intelligent optimization algorithms to VDA of models with on/off parameterizations. Adv Atmos Sci 26:1181–1197

Fritsch JM, Chappell CF (1980) Numerical prediction of convectively driven mesoscale pressure systems Part I: convective parameterization. J Atmos Sci 37:1722–1733

Goldberg DE (1989) Genetic algorithm in search, optimization and machine learning. Addison-Wesley, Reading, 432 pp

Gu JF et al (2005) Assimilation and simulation of typhoon Rusa (2002) using the WRF system. Adv Atmos Sci 22:415–427

Hamill TM (1999) Hypothesis tests for evaluating numerical precipitation forecasts. Wea Forecast 14:155–167

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, 228 pp

Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, Cambridge, MA, 211 pp

Jackson C, Sen MK, Stoffa PL (2004) An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. J Climate 17:2828–2841

Kain JS, Fritsch JM (1990) A one-dimensional entraining/detraining plume model and its application in convective parameterization. J Atmos Sci 47:2784–2802

Kain JS, Fritsch JM (1993) Convective parameterization for mesoscale models: The Kain-Fritsch scheme. Paper presented at the representation of cumulus convection in numerical models, Meteorological Monographs American Meteorological Society, No. 46, pp 165–170

Kain JS (2004) The Kain-Fritsch convective parameterization: an update. J Appl Meteorol 43:170–181

Kain JS et al (2008) Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. Wea Forecast 23:931–952

Kishtawal CM, Basu S, Patadia F, Thapliyal PK (2003) Forecasting summer rainfall over India using genetic algorithm. Geophys Res Lett 30:2203

Krishnakumar K (1989) Micro-genetic algorithms for stationary and non-stationary function optimization Paper presented at intelligent control and adaptive systems, Philadelphia pp 289–296

Lee DK, Choi SJ (2010) Observation and numerical prediction of torrential rainfall over Korea caused by Typhoon Rusa (2002). J Geophys Res 115:D12105

Lee J, Kim SM, Park HS, Woo BH (2005) Optimum design of cold-formed steel channel beams using Micro Genetic Algorithm. Eng Struct 27:17–24

Lee YH, Park SK, Chang DE (2006) Parameter estimation using the genetic algorithm and its impact on quantitative precipitation forecast. Ann Geophys 24:3185–3189

Liong SY, Phoon KK, Pasha MFK, Doan CD (2005) Efficient implementation of inverse approach for forecasting hydrological time series using micro GA. J Hydroinformatics 7:151–163

Narita M, Ohmori S (2007) Improving precipitation forecasts by the operational nonhydrostatic mesoscale model with the Kain-Fritsch convective parameterization and cloud microphysics. In: Preprints, 12th conference on mesoscale process, edited, Watervillle Valley, CD-ROM, 3.7

Navon IM (2009) Data assimilation for numerical weather prediction: a review. In: Park SK Xu L (eds) Data assimilation for atmospheric oceanic and hydrologica applications. Springer, Berlin, pp 21–65

Ogura Y, Cho HR (1973) Diagnostic determination of cumulus cloud populations from observed large-scale variables. J Atmos Sci 30:1276–1286

Park SK, Lee E (2007) Synoptic features of orographically enhanced heavy rainfall on the east coast of Korea associated with Typhoon Rusa (2002). Geophys Res Lett 34:L02803

Saito K et al (2007) Nonhydrostatic atmospheric models and operational development at JMA. J Meteor Soc Japan 85B:271–304

Singh R, Joshi PC, Kishtawal CM (2005a) A new technique for estimation of surface latent heat fluxes using satellite-based observations. Mon Wea Rev 133:2692–2710

Singh R, Kishtawal CM, Joshi PC (2005b) Estimation of monthly mean airsea temperature difference from satellite observations using genetic algorithm. Geophys Res Lett 32:L02807

Tao WK, Scala J, Ferrier B, Simpson J (1995) The effects of melting processes on the development of a tropical and a midlatitude squall line. J Atmos Sci 52:1934–1948

Wang Q, Fang HB, Zou XK (2010) Application of Micro-GA for optimal cost base isolation design of bridges subject to transient earthquake loads. Struct Multidiscip Opitimiz 41:765–777

Yang MJ, Tung QC (2003) Evaluation of rainfall forecasts over Taiwan by four cumulus parameterization schemes. J Meteor Soc Jpn 81:1163–1183

Zhu T, Zhang DL (2004) Numerical simulation of Hurricane Bonnie (1998). Part II: sensitivity to varying cloud microphysical processes. J Atmos Sci 63:109–126

Index

S.K. Park and L. Xu (eds.), *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, DOI 10.1007/978-3-642-35088-7,