

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

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

$$\begin{aligned} \mathcal{J}[\mathbf{x}(t_0)] = & (\mathbf{x}(t_0) - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}(t_0) - \mathbf{x}^b) \\ & + [\mathbf{y} - \mathcal{H}(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y} - \mathcal{H}(\mathbf{x})], \end{aligned} \quad (12.1)$$

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,

$$y_i = \mathcal{H}_i[\mathcal{M}(t_i, t_0)\mathbf{x}(t_0)] + \delta_i. \quad (12.2)$$

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

$$\begin{aligned} \mathcal{J}[\mathbf{x}(t_0), \mathbf{f}] = & (\mathbf{f} - \mathbf{f}^b)^T \mathbf{F}^{-1} (\mathbf{f} - \mathbf{f}^b) \\ & + (\mathbf{x}(t_0) - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}(t_0) - \mathbf{x}^b) \\ & + [(\mathbf{y} - \mathcal{H}(\mathbf{x}))^T \mathbf{R}^{-1} [(\mathbf{y} - \mathcal{H}(\mathbf{x}))], \end{aligned} \quad (12.3)$$

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,

$$\begin{aligned} \mathcal{J}[\mathbf{x}(t_0), \mathbf{f}] &= (\mathbf{f} - \mathbf{f}^b)^T \mathbf{F}^{-1} (\mathbf{f} - \mathbf{f}^b) \\ &+ (\mathbf{x}(t_0) - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}(t_0) - \mathbf{x}^b) \\ &+ (\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}), \end{aligned} \quad (12.4)$$

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,

$$(\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \mathbf{x} = \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y} + \mathbf{B}^{-1} \mathbf{x}^b, \quad (12.5)$$

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 $\hat{\mathbf{x}}$,

$$(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}) \hat{\mathbf{x}} = \mathbf{y} - \mathbf{H}\mathbf{x}^b. \quad (12.6)$$

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:

$$\mathbf{x}' = \begin{pmatrix} \mathbf{x}(t_0) \\ \mathbf{f} \end{pmatrix}, \quad \mathbf{B}' = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{pmatrix}, \quad \mathbf{H}' = \begin{pmatrix} \mathbf{H} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{R}' = \mathbf{R}, \quad \mathbf{y}' = \mathbf{y}. \quad (12.7)$$

Henceforth, we drop primes and simply write the objective function as

$$\begin{aligned} \mathcal{J}[\mathbf{x}] = & (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) \\ & + (\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}), \end{aligned} \quad (12.8)$$

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
     $i = i + 1$ ;
     $\mathbf{u}_i = \mathbf{r}_{i-1}$ ;
     $\mathbf{c}_i = \mathbf{A}\mathbf{u}_i$ ;
    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 $m \ll n$, 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 $K = \text{Span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{p-1}\mathbf{b}\}$, 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.

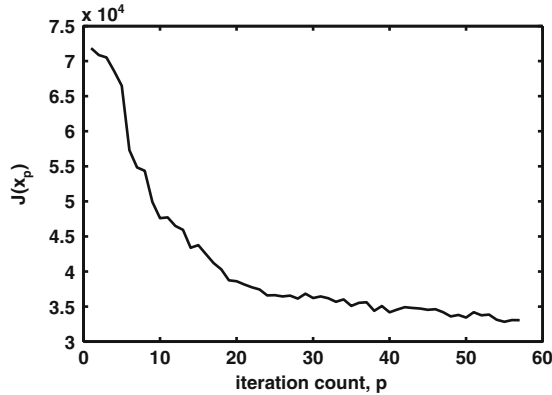


Fig. 12.2 Reduction of $\mathcal{J}(\mathbf{x})$ using GCR. The performance of the GCR solver as measured by the value of the objective function for an ocean data assimilation problem is shown. $\mathcal{J}(\mathbf{x}_p)$ is computed using (12.14) and (12.15) in the text. The application involves the 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

$$S = \sum_{i=1}^k \lambda_i(\mathcal{R}) \quad (12.9)$$

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

$$f = \left(\sum_{i=1}^p \hat{\lambda}_i(\mathcal{R}) \right) \left(\sum_{i=1}^k \hat{\lambda}_i^e(\mathcal{R}) \right)^{-1}. \quad (12.10)$$

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.

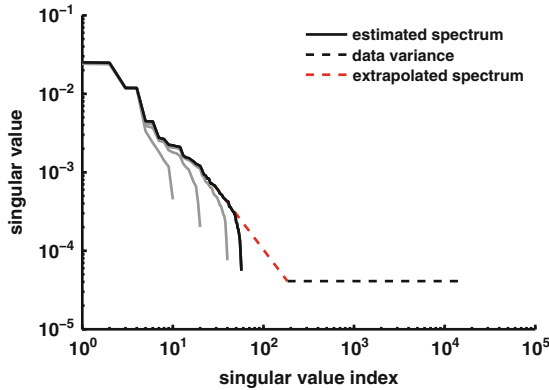


Fig. 12.3 Spectral Diagnostics from GCR. The estimated spectrum $\hat{\lambda}(\mathcal{R})$ of the representer matrix $\mathcal{R} = \mathbf{H}\mathbf{B}\mathbf{H}^T$ is shown by the *dark solid line* corresponding to the last GCR iterate ($p = 58$) in Fig. 12.2. *Solid gray lines* show $\hat{\lambda}(\mathcal{R})$ 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 $\mathbf{x}_p = \mathbf{B}\mathbf{H}^T\hat{\mathbf{x}}_p$ in (12.4), one obtains

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{x}}_p) &= \mathcal{J}^B(\hat{\mathbf{x}}_p) + \mathcal{J}^R(\hat{\mathbf{x}}_p) \\ &= \hat{\mathbf{x}}_p^T \mathbf{H}\mathbf{B}\mathbf{H}^T \hat{\mathbf{x}}_p \\ &\quad + (\mathbf{H}\mathbf{B}\mathbf{H}^T \hat{\mathbf{x}}_p - \mathbf{y})^T \mathbf{R}^{-1} (\mathbf{H}\mathbf{B}\mathbf{H}^T \hat{\mathbf{x}}_p - \mathbf{y}). \end{aligned} \quad (12.11)$$

Because the GCR solver computes the residual \mathbf{r}_p at each iterate, one has

$$(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})\hat{\mathbf{x}}_p = \mathbf{y} - \mathbf{r}_p. \quad (12.12)$$

Assuming that $\mathbf{R}\hat{\mathbf{x}}_p$ can be computed on demand, then

$$\mathbf{H}\mathbf{B}\mathbf{H}^T \hat{\mathbf{x}}_p = \mathbf{y} - \mathbf{r}_p - \mathbf{R}\hat{\mathbf{x}}_p, \quad (12.13)$$

and all terms in the expression for the objective function are computable. The contribution from the background term is

$$\mathcal{J}^B(\mathbf{x}_p) = (\hat{\mathbf{x}}_p)^T (\mathbf{y} - \mathbf{r}_p - \mathbf{R}\hat{\mathbf{x}}_p), \quad (12.14)$$

while the contribution from the observations is

$$\mathcal{J}^R(\mathbf{x}_p) = (\mathbf{r}_p + \mathbf{R}\hat{\mathbf{x}}_p)^T \mathbf{R}^{-1} (\mathbf{r}_p + \mathbf{R}\hat{\mathbf{x}}_p). \quad (12.15)$$

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 \mathbf{B} and \mathbf{R} is thus of paramount importance. This section outlines the *posterior diagnosis* strategy of Desroziers and Ivanov (2001) for validating the errors \mathbf{B} and \mathbf{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

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \quad (12.16)$$

where \mathbf{K} denotes the so-called *Kalman gain*,

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}. \quad (12.17)$$

At this optimum, the value of the objective function \mathcal{J} is given by Bennett (1992),

$$\mathcal{J}(\mathbf{x}^a) = \mathbf{d}^T \mathbf{D}^{-1} \mathbf{d}, \quad (12.18)$$

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 \mathbf{B} and \mathbf{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),

$$E\{\mathcal{J}(\mathbf{x}^a)\} = E\{\chi_m^2\} = m, \quad (12.19)$$

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

$$E\{\mathcal{J}^B(\mathbf{x}^a)\} = \text{Tr}(\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{D}^{-1}), \quad (12.20)$$

and

$$E\{\mathcal{J}^R(\mathbf{x}^a)\} = \text{Tr}(\mathbf{R}\mathbf{D}^{-1}), \quad (12.21)$$

where $\text{Tr}(\mathbf{A})$ 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 \(2001\)](#)

$$E\{\mathcal{J}_l^B(\mathbf{x}^a)\} = \text{Tr}(\mathbf{\Pi}_l^B \mathbf{H}\mathbf{B}\mathbf{H}^T \mathbf{D}^{-1} \mathbf{\Pi}_l^{B T}). \quad (12.22)$$

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

$$E\{\mathcal{J}_k^R(\mathbf{x}^a)\} = \text{Tr}(\mathbf{\Pi}_k^R \mathbf{R}\mathbf{D}^{-1} \mathbf{\Pi}_k^{R T}). \quad (12.23)$$

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

$$\mathcal{J}(\mathbf{x}) = \sum_{l=1}^{\nu^B} \frac{\mathcal{J}_l^B(\mathbf{x})}{\mathbf{s}_l^B} + \sum_{k=1}^{\nu^R} \frac{\mathcal{J}_k^R(\mathbf{x})}{\mathbf{s}_k^R}, \quad (12.24)$$

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](#)),

$$\mathbf{x}^a(\mathbf{s}) = \mathbf{x}^b + \mathbf{K}(\mathbf{s})(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \quad (12.25)$$

where the tuned *Kalman gain*, $\mathbf{K}(\mathbf{s})$, takes the form

$$\mathbf{K}(\mathbf{s}) = \mathbf{B}(\mathbf{s})\mathbf{H}^T[\mathbf{H}\mathbf{B}(\mathbf{s})\mathbf{H}^T + \mathbf{R}(\mathbf{s})]^{-1} = \mathbf{B}(\mathbf{s})\mathbf{H}^T\mathbf{D}(\mathbf{s})^{-1}, \quad (12.26)$$

with $\mathbf{B}(\mathbf{s}) = \sum_{l=1}^{\nu^B} \mathbf{s}_l^B \boldsymbol{\Pi}_l^B \mathbf{B}_l \boldsymbol{\Pi}_l^{B^T}$ and $\mathbf{R}(\mathbf{s}) = \sum_{k=1}^{\nu^R} \mathbf{s}_k^R \boldsymbol{\Pi}_k^R \mathbf{R}_k \boldsymbol{\Pi}_k^R$. 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

$$\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s})) = \mathbf{d}^T \mathbf{D}^{-1} \mathbf{H} \boldsymbol{\Pi}_l^{B^T} \mathbf{B}(\mathbf{s}) \boldsymbol{\Pi}_l^B \mathbf{H}^T \mathbf{D}^{-1} \mathbf{d}, \quad (12.27)$$

with expected value

$$E\{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}))\} = \mathbf{s}_l^B \text{Tr}[\boldsymbol{\Pi}_l^B \mathbf{H} \mathbf{B}(\mathbf{s}) \mathbf{H}^T \mathbf{D}(\mathbf{s})^{-1} \boldsymbol{\Pi}_l^{B^T}], \quad (12.28)$$

and

$$\begin{aligned} \mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s})) &= [\boldsymbol{\Pi}_k^R(\mathbf{y} - \mathbf{H}\mathbf{x}^a(\mathbf{s}))]^T \mathbf{R}(\mathbf{s})^{-1} [\boldsymbol{\Pi}_k^R(\mathbf{y} - \mathbf{H}\mathbf{x}^a(\mathbf{s}))] \\ &= \mathbf{d}^T \mathbf{D}(\mathbf{s})^{-1} \boldsymbol{\Pi}_k^{R^T} \mathbf{R}(\mathbf{s}) \boldsymbol{\Pi}_k^R \mathbf{D}(\mathbf{s})^{-1} \mathbf{d}, \end{aligned} \quad (12.29)$$

with expected value

$$E\{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}))\} = \mathbf{s}_k^R \text{Tr}[\boldsymbol{\Pi}_k^R \mathbf{R}(\mathbf{s}) \mathbf{D}(\mathbf{s})^{-1} \boldsymbol{\Pi}_k^{R^T}]. \quad (12.30)$$

The criterion for the tuning parameters is that the relations

$$\mathbf{s}_l^B = \frac{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}))}{\text{Tr}[\boldsymbol{\Pi}_l^B \mathbf{H} \mathbf{B}(\mathbf{s}) \mathbf{H}^T \mathbf{D}(\mathbf{s})^{-1} \boldsymbol{\Pi}_l^{B^T}]} \quad (12.31)$$

and

$$\mathbf{s}_k^R = \frac{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}))}{\text{Tr}[\boldsymbol{\Pi}_k^R \mathbf{R}(\mathbf{s}) \mathbf{D}(\mathbf{s})^{-1} \boldsymbol{\Pi}_k^{R^T}]} \quad (12.32)$$

are exactly satisfied. [Desroziers and Ivanov \(2001\)](#) proposed an iterative approach (*fixed-point algorithm*) to solve (12.31) and (12.32), namely,

$$\mathbf{s}_{l \ i+1}^B = \frac{\mathcal{J}_l^B(\mathbf{x}^a(\mathbf{s}_i))}{\text{Tr}[\boldsymbol{\Pi}_l^B \mathbf{H} \mathbf{B}(\mathbf{s}_i) \mathbf{H}^T \mathbf{D}(\mathbf{s}_i)^{-1} \boldsymbol{\Pi}_l^{B^T}]} \quad (12.33)$$

$$\mathbf{s}_{k \ i+1}^R = \frac{\mathcal{J}_k^R(\mathbf{x}^a(\mathbf{s}_i))}{\text{Tr}[\boldsymbol{\Pi}_k^R \mathbf{R}(\mathbf{s}_i) \mathbf{D}(\mathbf{s}_i)^{-1} \boldsymbol{\Pi}_k^{R^T}]}, \quad (12.34)$$

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, $Tr[\Pi_l^B \mathbf{H}\mathbf{B}(\mathbf{s})\mathbf{H}^T \mathbf{D}(\mathbf{s})^{-1} \Pi_l^{B^T}]$ and $Tr[\Pi_k^R \mathbf{R}(\mathbf{s})\mathbf{D}(\mathbf{s})^{-1} \Pi_k^{R^T}]$. Because the matrices $\mathbf{H}\mathbf{B}\mathbf{H}^T$ 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

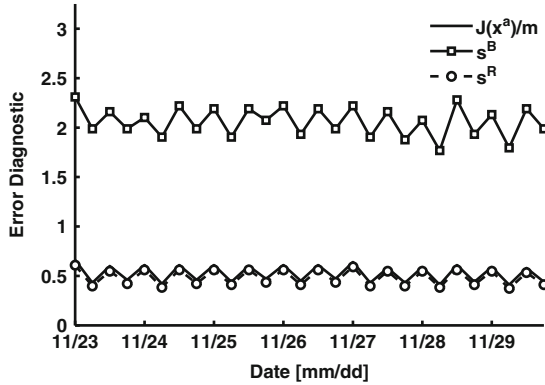


Fig. 12.4 NAVDAS-AR posterior error diagnostics. The reduced value of the objective function divided by the number of observations is consistently smaller than unity ($\mathcal{J}(\mathbf{x}^a)/m < 1$; *solid line*), its expected value if both background and observation errors are correctly scaled (12.19). 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
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
- Girard 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>