Initial Testing of a Massively Parallel Ensemble Kalman Filter with the Poseidon Isopycnal Ocean General Circulation Model

CHRISTIAN L. KEPPENNE

Science Applications International Corporation, Beltsville, Maryland

MICHELE M. RIENECKER

NASA Seasonal-to-Interannual Prediction Project, Laboratory for Hydrospheric Processes, Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 24 August 2001, in final form 5 March 2002)

ABSTRACT

A multivariate ensemble Kalman filter (MvEnKF) implemented on a massively parallel computer architecture has been developed for the Poseidon ocean circulation model and tested with a Pacific basin model configuration. There are about 2 million prognostic state-vector variables. Parallelism for the data assimilation step is achieved by regionalization of the background-error covariances that are calculated from the phase–space distribution of the ensemble. Each processing element (PE) collects elements of a matrix measurement functional from nearby PEs. To avoid the introduction of spurious long-range covariances associated with finite ensemble sizes, the background-error covariances are given compact support by means of a Hadamard (element by element) product with a three-dimensional canonical correlation function.

The methodology and the MvEnKF implementation are discussed. To verify the proper functioning of the algorithms, results from an initial experiment with in situ temperature data are presented. Furthermore, it is shown that the regionalization of the background covariances has a negligible impact on the quality of the analyses.

Even though the parallel algorithm is very efficient for large numbers of observations, individual PE memory, rather than speed, dictates how large an ensemble can be used in practice on a platform with distributed memory.

1. Introduction

a. Background and motivation

Many of the early advances in ocean data assimilation have emerged from practical applications in the tropical Pacific. These applications have been driven by the need to initialize the ocean state for coupled atmosphere–ocean forecasts of the El Niño–Southern Oscillation (ENSO) phenomenon. In addition, hindcast estimates of the ocean state have been useful in diagnosing the evolution of El Niño. Over much of the world's oceans, large-scale assimilation is facilitated by the availability of satellite altimetry because of the sparsity of in situ data. However, in the tropical Pacific, the ocean observing system was vastly improved by the deployment of the Tropical Atmosphere Ocean (TAO) array of moored buoys (e.g., McPhaden et al. 1998) to support seasonal-to-interannual (SI) climate studies and prediction. One of the major successes of the Tropical Ocean Global Atmosphere program was the emergence of coupled physical models (as opposed to statistical models) with some prediction skill (e.g., Chen et al. 1995; Ji et al. 1996).

Recently, the NASA Seasonal-to-Interannual Prediction Project (NSIPP) has been established to further the utilization of satellite observations for prediction of short-term climate phenomena. NSIPP undertakes routine experimental forecasts in a research framework with global coupled ocean–atmosphere–land surface models. The initial implementation has used an ocean analysis system employing a simple assimilation methodology—a univariate optimal interpolation (UOI)—with the Poseidon isopycnal ocean general circulation model (OGCM; Schopf and Loughe 1995; Konchady et al. 1998; Yang et al. 1999). Like several other ocean data assimilation systems currently in use at other institutions (e.g., Ji and Leetma 1997), it is based on the assumption that the forecast-error covariances are approximately Gaussian and that the covariances between the temperature-field errors and the salinity-field and current-field errors are negligible.

Largely due to the high-resolution coverage and accuracy of the TAO measurements, the UOI is effective in improving surface and subsurface temperature-field

Corresponding author address: Christian L. Keppenne, Mail Code 971, NASA Goddard Space Flight Center, Greenbelt, MD 20771.
E-mail: clk@janus.gsfc.nasa.gov

estimates in the equatorial region in comparison with the estimates obtained without temperature assimilation. As a result, its introduction into the NSIPP coupled forecasting system has resulted in significant improvements in the coupled model's hindcast skill of Niño-3 temperature anomalies.

The UOI has the advantage of being inexpensive in terms of computing resources. Nevertheless, it suffers from three major shortcomings: first, it can only be used to assimilate measurements of a model prognostic variable; second, it does not use any statistical information about the expected inhomogeneous distribution of model errors; third, it is based on a steady-state error–covariance model, which gives the same weight to a unit innovation regardless of how accurate the ocean-state estimate has become as a result of previous analyses. Directly linked to this shortcoming is the failure to provide time-dependent estimates of the model errors.

In response to the first two shortcomings, a parallel multivariate OI (MvOI) system has been implemented. The MvOI uses steady-state estimates of the model-error statistics computed from ensemble runs of the OGCM in the presence of stochastic atmospheric forcing from an ensemble integration of the atmospheric general circulation model (AGCM) (Borovikov and Rienecker 2002). Yet, the MvOI cannot adjust to dynamically evolving error statistics. A parallel multivariate ensemble Kalman filter (MvEnKF) has been developed to address this shortcoming. This paper discusses its design, implementation, and initial testing.

b. Overview of the ensemble Kalman filter

Although the Kalman filter (Kalman 1960) and its generalization to nonlinear systems, the extended Kalman filter, are statistically optimal sequential estimation procedures that minimize error variance (e.g., Daley 1991; Ghil and Malanotte-Rizzoli 1991; Bennett 1992), they cannot be used in the context of a high-resolution ocean or atmospheric model because of the prohibitive cost of time stepping the model-error covariance matrix when the model has more than a few thousand state variables. Therefore, reduced-rank (e.g., Cane et al. 1996; Verlaan and Heemink 1997) and asymptotic (e.g., Fukumori and Malanotte-Rizzoli 1995) Kalman filters have been proposed. Evensen (1994) introduced the ensemble Kalman filter (EnKF) as an alternative to the traditional Kalman filter. In the EnKF, an ensemble of model trajectories is integrated and the statistics of the ensemble are used to estimate the model errors. Closely related to the EnKF are the singular evolutive extended Kalman filter (Pham et al. 1998) and the error-subspace statistical estimation algorithms described in Lermusiaux and Robinson (1999).

Evensen (1994) compared the EnKF to the extended Kalman filter in twin assimilation experiments involving a two-layer quasigeostrophic (QG) ocean model on a square 17×17 grid. Evensen and van Leeuwen (1996) used the EnKF to process U.S. Navy Geodesy Satellite (Geosat) altimeter data into a two-layer, regional QG model of the Agulhas current on a 51×65 grid. Houtekamer and Mitchell (1998) and Mitchell and Houtekamer (2000) used the EnKF in identical twin experiments involving a three-level, spectral QG model at triangular truncation T21 and parameterized model errors.

Keppenne (2000, hereafter K00) conducted twin experiments with a parallel MvEnKF algorithm implemented for a two-layer, spectral, T100 primitive equation model with parameterized model errors. With about 2×10^5 model variables, the state-vector size was small enough in this application to justify a parallelization scheme in which each ensemble member resides in the memory of a separate CRAY T3E processor [hereafter, processing element (PE)]. To parallelize the analysis, K00's algorithm transposes the ensemble across PEs at analysis time, so that each PE ends up processing data from a subregion of the model domain. The influence of each observation is weighted according to the distance between that observation and the center of each PE region.

To filter out noise associated with small ensemble sizes, Houtekamer and Mitchell (2001) developed a parallel EnKF analysis algorithm that applies a Hadamard (element by element) product (e.g., Horn and Johnson 1991) of a correlation function having local compact support with the background-error covariances. They tested this analysis scheme on a 128×64 Gaussian grid corresponding to a 50-level QG model using randomly generated ensembles of first-guess fields. The benefits of constraining the covariances between ensemble members using a Hadamard product with a locally supported correlation function has also been investigated by Hamill and Snyder (2000) in the context of an intermediate QG atmospheric model.

In this paper, we build upon the contributions made by each of the previously mentioned studies to implement a parallel MvEnKF for the Poseidon OGCM. Initial tests are undertaken with a 20-layer, Pacific basin configuration of the model with about 2 million state variables. The system noise is accounted for in a manner similar to that used in K00, by including a stochastic component in the forcing fields. Following Houtekamer and Mitchell (2001), an element-by-element product with an idealized three-dimensional compactly supported correlation function is used to remove spurious long-range signals from the background-error covariances.

c. Organization of the following sections

The remainder of this paper is concerned with the parallel MvEnKF design for the Poseidon OGCM. The model is briefly discussed in section 2 and the algorithms are presented in section 3. The scalability of the algorithms and the effect of distributing the analysis

calculations between PEs are discussed in section 4, where an initial test of the MvEnKF is conducted in the context of TAO temperature data assimilation. Section 5 contains a summary.

A complete description of the algorithms is available as a NASA technical report (Keppenne and Rienecker 2001, hereafter KR01). The application of the MvEnKF to the assimilation of altimeter data into Poseidon is discussed in Keppenne and Rienecker (2002a, hereafter KR02a). The impact of the assimilation of temperature on the model currents, salinity and sea surface height (SSH) is presented in Keppenne and Rienecker (2002b, hereafter KR02b).

2. The Poseidon parallel ocean model

a. Model summary

The Poseidon model (Schopf and Loughe 1995) is a finite-difference reduced-gravity ocean model that uses a generalized vertical coordinate designed to represent a turbulent, well-mixed surface layer and nearly isopycnal deeper layers. Poseidon has been documented and validated in hindcast studies of El Niño (Schopf and Loughe 1995) and has since been updated to include prognostic salinity (e.g., Yang et al. 1999). More recently, the model has been used in a numerical study of the surface heat balance along the equator (Borovikov et al. 2001) and in an examination of El Niño variations during the 1990s.

Explicit detail of the model, its vertical coordinate representation, and its discretization are provided in Schopf and Loughe (1995). The prognostic variables are layer thickness, $h(\lambda, \theta, \zeta, t)$, temperature, $T(\lambda, \theta, \zeta, t)$, salinity, $S(\lambda, \theta, \zeta, t)$, and the zonal and meridional current components, $u(\lambda, \theta, \zeta, t)$ and $v(\lambda, \theta, \zeta, t)$, where λ is longitude, θ latitude, t time, and ζ is a generalized vertical coordinate which is 0 at the surface and increments by 1 between successive layer interfaces.

Following Pacanowski and Philander (1981), vertical mixing is parameterized through a Richardson number–dependent mixing scheme implemented implicitly. An explicit mixed layer is included with a mixed layer entrainment parameterization following Niiler and Kraus (1977).

A time-splitting integration scheme is used whereby the hydrodynamics are done with a short time step (15 min), but the vertical diffusion, convective adjustment, and filtering are done with coarser time resolution (half-daily).

b. Model setup

The version of Poseidon used here has been parallelized as in Konchady et al. (1998) using the same message-passing protocol and 2D horizontal domain decomposition used for the NSIPP-1 AGCM (Schaffer and Suarez 2000).

The experiments of section 4 use a 20-layer Pacific basin version of the parallel model with uniform 1° zonal resolution. The meridional resolution varies between 1/3° at the equator and 1° in the extratropics. Solid boundaries are imposed at 45° south and at 65° north. At lateral boundaries, a no-slip condition is used for the currents and a no-flux condition is used for mass, heat, and salinity.

There are $173 \times 164 \times 20$ grid boxes, of which 28% are situated over land, resulting in a total of 2.0422×10^6 individual prognostic variables. A 16×16 PE lattice is used as shown in Fig. 1. Although the PEs located over land do not directly take part in the ensemble integrations, they are involved in the calculation of surface heat fluxes using the atmospheric mixed-layer model of Seager et al. (1995).

Figure 2 illustrates the horizontal setup for one PE box. Locally within the box, the grid cells are numbered $1 \le i \le I$, zonally and $1 \le j \le J$, meridionally, from the box's lower-left, southwest corner. In order to minimize the communication overhead in the horizontal differencing of the model equations, the PE boxes overlap. The overlapping regions, called halo regions, have width $i_1 - 1$ to the west, $I - i_2$ to the east, $j_1 - 1$ to the south and $J - j_2$ to the north. The PE-private regions are thus defined by $i_1 \le i \le i_2$ and $j_1 \le j \le j_2$.

3. Assimilation methodology

a. Horizontal domain decomposition

Since the version of Poseidon used here is parallelized, the same domain decomposition used to run the model can be used in the analyses, provided the background-error covariance matrix, $\mathbf{P}^f$, is locally approximated. The main advantage of this parallelization strategy is that it avoids the costly ensemble transpositions across PEs that would be required at analysis time if the ensemble integration were parallelized by letting each ensemble member run on a different processor array, as in KR00. Thus, the ensemble is distributed so that the memory of each PE contains the same elements of each ensemble member's state vector. These elements correspond to every variable contained within the PE boxes, the PE-private portions of which (inner rectangles in Fig. 2) are delineated by thick lines in Fig. 1. This decomposition is used for the ensemble integrations as well as for the analyses.

b. Assimilation on geopotential surfaces

The temperature measurements from each TAO mooring are recorded at specific depths, which are fairly consistent between moorings. Since Poseidon uses an isopycnal vertical coordinate, the model fields must be interpolated to the latitude, longitude, and depth of each observation. When the UOI was implemented, the choice was made to treat the temperature observations in the usual (λ, θ, z) coordinate system in light of the

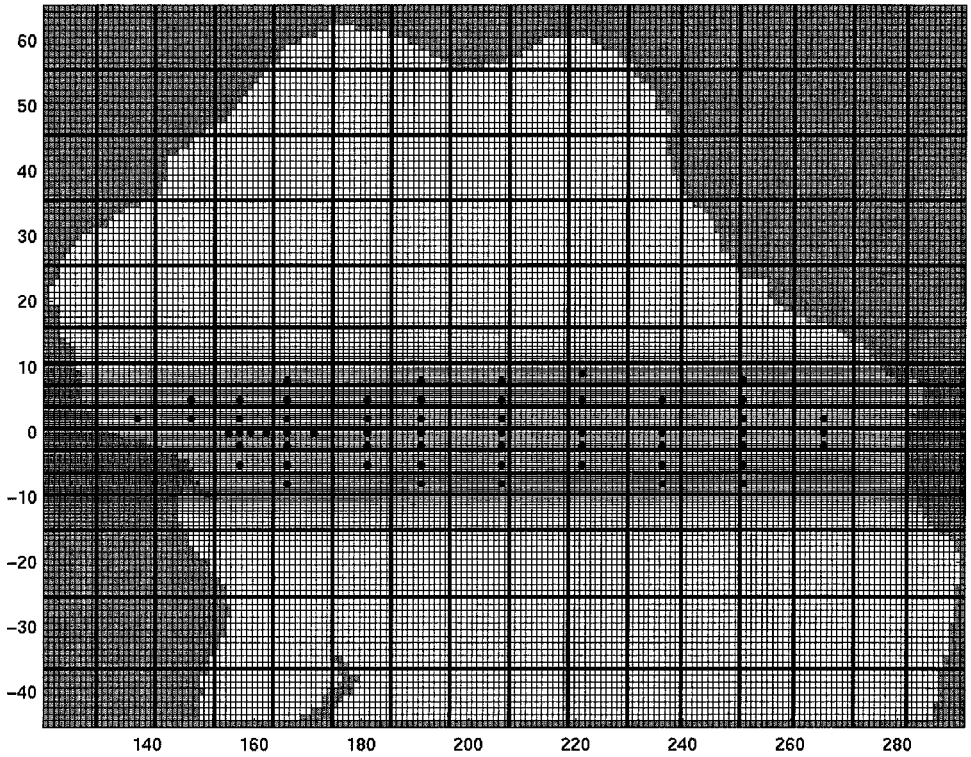

FIG. 1. Horizontal domain decomposition for the Pacific model. The thin lines delineate grid cells. The thick lines correspond to the boundaries of each PE box's PE-private area on the 16 × 16 PE lattice. Each dark circle corresponds to a TAO mooring.

absence of corresponding salinity observations. To maintain compatibility with the UOI, which interpolates model fields vertically to a series of prespecified depths (hereafter, levels) prior to each analysis, the same approach is used here and the background covariances are calculated on levels rather than on layers. Therefore, the T, S, u, and v fields are converted from isopycnals to levels and the analysis increments are calculated on the levels before being mapped back to the isopycnals. The mapping is made as though the vertical variations of the field were piecewise linear, with the discontinuities in the slope occurring in the middle of the layers. This is illustrated in Fig. 3 for the temperature field. Sixteen levels in the depth range from 0 to 500 meters are used in section 4. To minimize vertical interpolation errors, the levels include the depths of the TAO temperature measurements.

This scheme results in only T, S, u, and v being updated. The layer thicknesses, h, are left unchanged by the assimilation. The procedure allows the model to dynamically recalculate h from the new density distribution and the target interface buoyancies, as it does at every time step (see Schopf and Loughe 1995).

c. Ensemble size

With the MvEnKF, PE memory imposes constraints on both the domain decomposition and the ensemble size. The platform used in this work is a 1024-PE CRAY T3E-600 with 128MB local RAM per PE. The Pacific basin version of Poseidon is typically run on 64 PEs. The goal is for the MvEnKF runs to be done on a few times as many PEs. In this study, 256 PEs are used and the memory available on these PEs imposes a limit of about 40 ensemble members on this platform. Encouraging results have been obtained with comparably sized ensembles by Mitchell and Houtekamer (2000) with a three-level QG model and by K00 with a two-layer shallow water model.

d. Decomposition of analysis between PEs

The small ensemble size introduces the need to filter out spurious long-range correlations when the background covariances are computed. Following Houtekamer and Mitchell (2001) and a suggestion by Gaspari and Cohn (1999), this filtering is achieved through a Hadamard product (i.e., $\mathbf{A} \cdot \mathbf{B}$ such that $\{\mathbf{A} \cdot \mathbf{B}\}_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$) of the error covariance matrices with a local compactly supported correlation function. This function is the product of a horizontal correlation function, $C_h(r_h^{(12)})$, $r_h^{(12)} = [(\lambda_2 - \lambda_1)^2/l_\lambda + (\theta_2 - \theta_1)^2/l_\theta]^{0.5}$, and a vertical correlation function, $C_v(r_v^{(12)})$, $r_v^{(12)} = |z_2 - z_1|/l_z$, where $(\lambda_i, \theta_i, z_i)$ are the coordinates of point i. In this study, $C_h = C_v = C_0$, where C_0 is defined by Gaspari and Cohn (1999, their section 4.10). The nor-

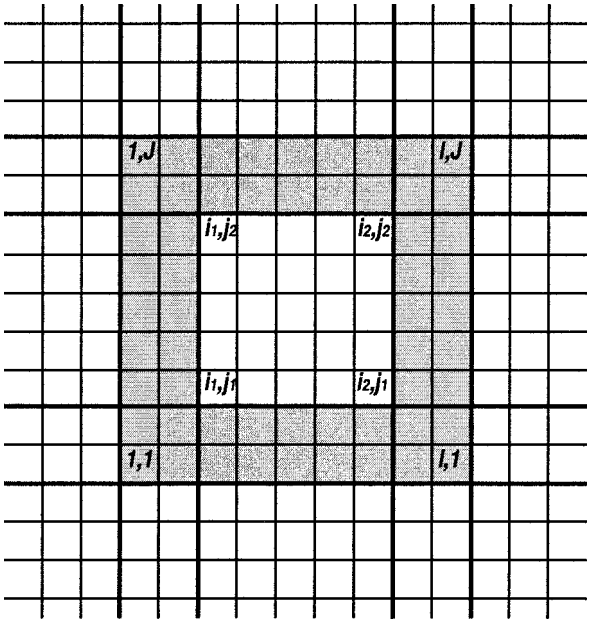

FIG. 2. Schematic setup for one PE. The halo regions are colored gray. The thin lines delineate grid cells. The thick lines delimit the halo regions and PE boundaries. In this example, $I = J = 9$, $i_1 = j_1 = 3$, and $i_2 = j_2 = 7$.

malization is such that $C_0(r) = 0$, $r \geq 2$. The correlation scales used in section 4 for the assimilation of TAO temperature data are $l_\lambda = 30°$, $l_\theta = 15°$, and $l_z = 500$ m. Shorter correlation scales give better results when gridded altimeter data are assimilated (KR02a).

Although the TAO temperature data assimilated in section 4 are sufficiently few (about 600 at each analysis) for each PE to process them all, an approach whereby each PE processes data from a subregion of the model domain is used. When more numerous data are assimilated, such as in KR02a, the regionalization becomes a necessity.

The regionalization of the analysis is justified by the Hadamard product: only data within an ellipse with semiaxes $2l_\lambda$ and $2l_\theta$ exhibit nonzero correlations with the state variables within each grid cell. Taking advantage of this fact, the region from which the observations assimilated on each PE are collected is chosen to be the smallest rectangle, with sides $\lambda_{i2,j1} - \lambda_{i1,j1} + 4l_\lambda$ and $\theta_{i1,j2} - \theta_{i1,j1} + 4l_\theta$, containing all the ellipses that correspond to the PE-private grid cells of this PE. This is illustrated in Fig. 4.

e. Analysis procedure

Without the Hadamard product of the background-error covariances with the compactly supported correlation function, the EnKF analysis can be written as

$$\mathbf{y}_i = \Xi(\mathbf{x}^f - \langle \mathbf{x} \rangle^f), \tag{1a}$$

$$\mathbf{l}_i = \mathcal{L}(\mathbf{y}_i + \langle \mathbf{x} \rangle^f) - \mathcal{L}(\langle \mathbf{x} \rangle^f), \tag{1b}$$

$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m), \quad \mathbf{L} = (\mathbf{l}_1, \ldots, \mathbf{l}_m),$$

$$(\mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{W})\mathbf{b}_i = \mathbf{d} - \mathcal{L}(\mathbf{y} + \langle \mathbf{x} \rangle^f) + \mathbf{e}_i, \tag{1c}$$

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{Y}\mathbf{L}^{\mathrm{T}}\mathbf{b}. \tag{1d}$$

In (1) and throughout this discussion, uppercase boldface symbols represent matrices, lowercase boldface symbols represent vectors and lowercase regular (i.e., not bold) symbols denote scalar variables. The vector, $\mathbf{d}$ ($n_d \times 1$) contains n_d observations, $\mathbf{x}_i$ ($n_x \times 1$), $1 \leq i \leq m$, is the ith ensemble-state vector of length n_x, and m stands for the ensemble size. The superscripts a and f refer to the analyzed state and the forecast, respectively, Ξ is an optional smoothing operator and $\langle\ \rangle$ denotes an ensemble average. The vectors $\mathbf{y}_i$ ($n_x \times 1$) and

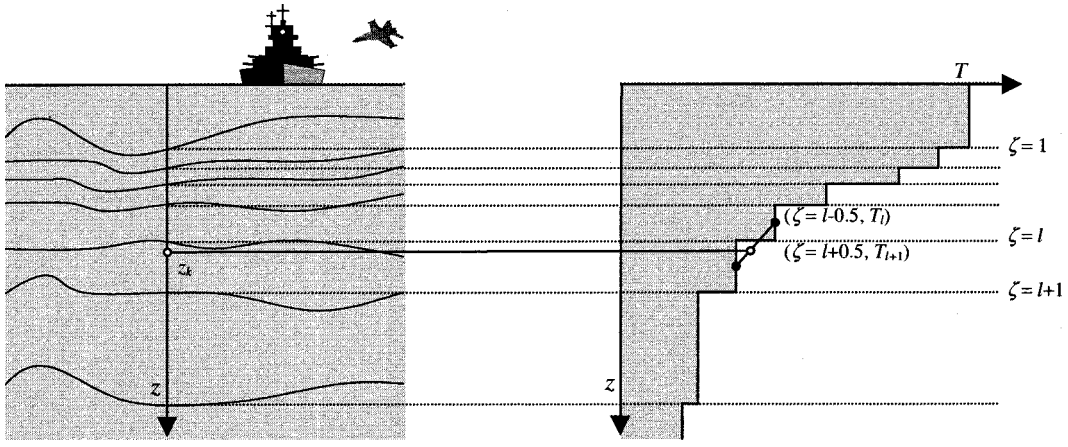

FIG. 3. Mapping of the model temperature field to a specified level, $z = z_k$. Within the current grid cell, z_k is contained between the layer interfaces $\zeta = l$ and $\zeta = l + 1$. In the model discretization, only the layer-average temperature matters. Yet, to avoid ambiguities when more than one specified level passes through the same layer in the grid cell, the field is interpolated linearly as shown.

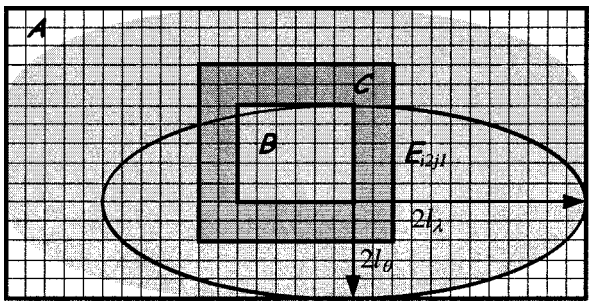

FIG. 4. Domain decomposition for the analysis. The outer rectangle delimits the area, $\mathcal{A}$, from which the data assimilated on one PE are collected. The innermost rectangle depicts the boundary of the PE-private area, $\mathcal{B}$. The ellipse delimits the influence region of the PE-private area's southeastern corner cell, (i_2, j_1). The shaded area contains the ellipses for all grid cells, (i, j), contained in $\mathcal{B}$. The region $\mathcal{C}$ contains all the PE's grid cells including the halo regions.

$\mathbf{l}_i$ ($n_d \times 1$) are columns of the matrices $\mathbf{Y}$ ($n_x \times m$) and $\mathbf{L}$ ($n_d \times m$), respectively, and $\mathcal{L}(\mathbf{x})$ is a measurement operator that relates the state vector to the observations. Matrix $\mathbf{W}$ ($n_d \times n_d$) is the observation-error covariance matrix. It includes measurement errors as well as representation errors. The representer matrix, $\mathbf{R} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, maps the background-error covariance matrix, $\mathbf{P}^f$ ($n_x \times n_x$), to the error subspace of the measurements. The elements of $\mathbf{b}_i$ are the representer-function amplitudes used to update $\mathbf{x}_i$.

The $n_d \times 1$ vector, $\mathbf{z}_i = \mathbf{d} - \mathcal{L}(\mathbf{y}_i + \langle\mathbf{x}\rangle^f) + \mathbf{e}_i$ in (1c), contains the innovations with respect to the ith ensemble member. The rationale for applying Ξ to presmooth $\mathbf{x}_i$ prior to their calculation is discussed in section 3h. Following Burgers et al. (1998), $\mathbf{e}_i$ is a random perturbation chosen such that $\langle\mathbf{e}_i\rangle = 0$ and $\langle\mathbf{e}_i\mathbf{e}_i^{\mathrm{T}}\rangle = \mathbf{W}$. Its role is to maintain the influence of observation uncertainty in the error covariances estimated directly from the ensemble so that these covariances are consistent with the theoretical estimates. Its inclusion helps prevent the ensemble from collapsing, resulting in a systematic error underestimation.

When $\mathcal{L}(\mathbf{x})$ is a linear operator and Ξ is an identity mapping, (1) simplifies to the usual Kalman filter analysis equations (e.g., Gelb 1974) applied to update each ensemble member in turn.

When the Hadamard products with the compactly supported correlation function are introduced and when the subscript ranges are explicitly written down, (1c) and (1d) are replaced by

$$c_{pq} = c_{qp} = C_h(r_h^{(pq)})C_v(r_v^{(pq)}), \qquad 1 \le p \le n_d, \quad 1 \le q \le n_d, \tag{2a}$$

$$(\mathbf{C} \cdot \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{W})\mathbf{b}_i = \mathbf{d} - \mathcal{L}(\mathbf{y}_i + \langle\mathbf{x}\rangle^f) + \mathbf{e}_i, \qquad 1 \le i \le m, \tag{2b}$$

$$\boldsymbol{\eta}_{kp} = C_h(r_h^{(kp)})C_v(r_v^{(kp)}), \qquad 1 \le k \le n_{\mathrm{box}}, \quad 1 \le p \le n_d, \tag{2c}$$

$$\left.\begin{aligned} \boldsymbol{\gamma}_{ik} &= \mathbf{L}^{\mathrm{T}}\mathbf{b}_i \cdot \boldsymbol{\eta}_k, \\ x_{ik}^a &= x_{ik}^f + \mathbf{y}_k \circ \boldsymbol{\gamma}_{ik}, \end{aligned}\right\} \qquad 1 \le i \le m, \tag{2d}$$

$$1 \le k \le n_{\mathrm{box}}, \tag{2e}$$

where $\circ$ and $\cdot$ refer, respectively, to the inner product of two vectors and to the Hadamard product of two matrices, and $\mathbf{C}$ ($n_d \times n_d$) is a compactly supported correlation matrix whose elements are defined by (2a), where the indices p and q refer to the data w_p and w_q. The components of the $n_d \times 1$ vector $\boldsymbol{\eta}_k$ defined by (2c) contain idealized correlations between the (λ, θ, z) coordinates, of grid box k and the coordinates of each measurement. To simplify the notation, only one subscript is used to identify the grid box. The index, $1 \le k \le n_{\mathrm{box}}$, thus loops over the three dimensions of the (λ, θ, z) coordinate system. The $m \times 1$ vector, $\mathbf{y}_k = \{y_{1k}, \ldots, y_{mk}\}$, contains optionally smoothed deviations from the ensemble mean of the m ensemble state vectors in the kth grid box. It is thus a single row of matrix $\mathbf{Y}$. With the MvEnKF, y_{ik} actually has four components, that is,

$$y_{ik} = \Xi[(T, S, u, v)_{ik} - (\langle T\rangle, \langle S\rangle, \langle u\rangle, \langle v\rangle)_k].$$

The $m \times 1$ vector, $\boldsymbol{\gamma}_{ik}$, contains the weights with which the elements of $\mathbf{y}_k$ in the kth grid box, $\{T, S, u, v\}_{ik}$, are combined to update the ith ensemble member. In each grid box, the analysis update, (2c–e), involves m matrix-vector multiplications of $\mathbf{L}^{\mathrm{T}}$ by $\mathbf{b}_i \cdot \boldsymbol{\eta}_k$ (2d). If the analysis calculations were not distributed between PEs, or if the observations allowed to influence the variables of each grid box were not limited to a subregion of the entire domain as a result of imposing compactly supported background covariances, these multiplications would be very costly. For the Poseidon model distributed across 256 PEs, they correspond to a tolerable fraction of the total cost of the MvEnKF. For example, when TAO temperature data are assimilated every 5 days as in section 4, the ensemble integration takes about 1100 s per analysis cycle while the analysis takes about 380 s. Of these, about 270 s are spent in the matrix-vector products of (2d).

f. System-noise representation

The theory of the Kalman filter (e.g., Gelb 1974) assumes that the first- and second-order statistics of the errors in the model and external forcing are known. Higher-order statistics are neglected. Let the evolution of the true state be represented by

$$\frac{\partial \mathbf{x}}{\partial t} = \mathbf{F}(\mathbf{x}, t) + \boldsymbol{\xi}(\mathbf{x}, t), \tag{3}$$

where $\boldsymbol{\xi}$ combines forcing errors and errors due to imperfections in the model, and is commonly known as system noise or process noise. The vector $\mathbf{F}$ ($n_x \times 1$)

contains the right-hand sides of the model partial differential equations, which include the model hydrodynamics, physics, and forcing. It is assumed that the model and forcing are unbiased, that is, $[\langle \boldsymbol{\xi}(\mathbf{x}, t)\rangle] = 0$, and that the $\boldsymbol{\xi}$ vectors are uncorrelated in time:

$$\langle \boldsymbol{\xi}(\mathbf{x}_k, t_k)\boldsymbol{\xi}(\mathbf{x}_l, t_l)\rangle = \boldsymbol{\Gamma}(\mathbf{x}_k, \mathbf{x}_l)\delta(t_k - t_l), \qquad (4)$$

where the system-noise covariance matrix, $\boldsymbol{\Gamma}$, is assumed known. Of course, the unbiased assumption is rarely correct in practice. This is especially true with ocean models in which the thermocline layer is usually too diffuse.

In an effort to separately account for the bias and unbiased forecast-error components, an algorithm, derived from Dee and Da Silva (1998), to estimate and correct systematic model errors has been implemented into the MvEnKF code. However, it is not used in this study. Rather, as in most atmospheric and oceanographic data assimilation applications, no attempt is made here to dissociate the bias from the unbiased error component.

In meteorological and oceanographic data assimilation, the statistics of $\boldsymbol{\xi}$ are generally unknown and are the object of parameterization. Adaptive Kalman filters that simultaneously estimate the state and system-noise statistics have been developed. Blanchet et al. (1997) summarize and compare several adaptive filtering algorithms. In practice, the prohibitive cost of the adaptive filters has limited their application in meteorology and oceanography.

Motivated by the current lack of information about the model-error statistics, the system noise is represented solely by modeling the errors in the surface wind stress and heat flux forcing. A system-noise representation in which not only the forcing errors but also the model errors are parameterized is in development.

Because of the focus on seasonal variability, the forcing errors (uncertainties) are modeled on SI timescales, with each ensemble member being forced by a different monthly mean perturbation of the monthly mean basic state. The basic state is the superposition of the climatological seasonal cycle with interannual anomalies. The climatology is provided by Special Sensor Microwave Imager (Atlas et al. 1996) winds and Earth Radiation Budget Experiment heat flux data. The interannual anomalies are obtained by integrating the AGCM over observed sea surface temperature (SST) data (Reynolds and Smith 1994) starting from different initial conditions. Thus, the perturbations applied are due entirely to internal atmospheric chaos. By using the same SST, each member of the atmospheric ensemble used to force the ocean ensemble has the same SI phase. The spread of the atmospheric ensemble is meant to be representative of the uncertainty of the forcing products used to force the model in nonensemble runs.

g. Parallel algorithm

All information exchanges between PEs use message-passing functions from the Goddard Earth Modeling System (GEMS) library. The GEMS functions provide a high-level, object-oriented interface to the CRAY native SHMEM (shared memory) communication library (see Schaffer and Suarez 2000).

The analysis relies principally on two GEMS functions, which are mentioned here in template form in order to simplify the discussion. The first function, pe_collect (. . .), is used to collect data from either the entire two-dimensional PE array or from a row or column of this array. The second function, halo (. . .), updates its array argument in the halo regions of each PE (gray areas in Fig. 2), after each PE has modified its PE-private elements of this array corresponding to the inner rectangle in Fig. 2.

Seen from the point of view of one PE referred to as the current PE, the parallel MvEnKF analysis involves the following steps. This overview starts after the current PE has obtained the observations, $\mathbf{d}^{\mathbf{b}}$, made within its PE-private region ($\mathcal{B}$ in Fig. 4). For a more detailed discussion, the reader is referred to KR01.

- Step 1: Vertical interpolation of the *T*, *S*, *u*, and *v* fields from the model layers to the analysis levels as explained in section 3b.
- Step 2: Calculation of the anomalies with respect to the ensemble mean over the entire domain of the current PE (area $\mathcal{C}$ in Fig. 4), $\mathbf{x}_i^{\mathbf{c}f} - \langle \mathbf{x}^{\mathbf{c}}\rangle^f$, $1 \le i \le m$.
- Step 3: Calculation of $\mathbf{y}^{\mathbf{c}}$, the current PE's portion of $\mathbf{y}_i$ in (1a). Prior to this step, calls to pe_collect ( ) are used to collect the state elements required to apply the optional smoothing operator, Ξ (section 3h).
- Step 4: Identification of the PE-private data required by the other PEs. For each other PE, the current PE calculates which elements of its $\mathbf{d}^{\mathbf{b}}$ fall inside that other PE's rectangle, $\mathcal{A}$, which is the region from which the other PE will need to collect data (Fig. 4).
- Step 5: Evaluation of the measurement operator. The current PE calculates a $n_d^{\mathbf{b}} \times m$ matrix, $\mathbf{L}^{\mathbf{b}}$, where $n_d^{\mathbf{b}}$ is the number of observations contained in its PE-private region. The element at the intersection of the *p*th row and *i*th column of $\mathbf{L}^{\mathbf{b}}$ is

  $$\mathbf{L}^{\mathbf{b}}_{pi} = \mathcal{L}^p(\mathbf{y}_i + \langle \mathbf{x}\rangle^f) - \mathcal{L}^p(\langle \mathbf{x}\rangle^f),$$

  where $\mathcal{L}^p$, discussed in KR01, is an interpolation operator that maps its argument to the location of $d_p^{\mathbf{b}}$, the *p*th PE-private observation on the current PE.
- Step 6: Calculation of $\mathbf{z}$, the innovations with respect to the ensemble mean for the current PE's private region. The innovation corresponding to $d_p^{\mathbf{b}}$ is

  $$z_p^{\mathbf{b}} = d_p^{\mathbf{b}} - \mathcal{L}^p(\langle \mathbf{x}^{\mathbf{c}}\rangle^f).$$

- Step 7: Gathering of $\mathbf{L}^{\mathbf{a}}$, a $n_d^{\mathbf{a}} \times m$ matrix analogous to $\mathbf{L}^{\mathbf{b}}$, but corresponding to the $n_d^{\mathbf{a}}$ measurements made within area $\mathcal{A}$ (Fig. 4). The function pe_collect ( ) is called N_{PE} times, where N_{PE} is the number of PEs.

Each call results in a different PE completing the collection of its version of $\mathbf{L}^a$.

- Step 8: Collection of the innovations, $\mathbf{z}^a$, required by each PE. As for gathering $\mathbf{L}^a$, pe_collect ( ) is called N_{PE} times. Each PE passes to pe_collect ( ) the elements of its $\mathbf{z}^b$ innovation vector required by the other PEs.
- Step 9: Calculation of the representer amplitudes by solving a local equation system corresponding to the restriction to area $\mathcal{A}$ of the global equations (2). A local representer matrix, $\mathbf{R}^a = \mathbf{L}^a(\mathbf{L}^a)^T$, and its Hadamard product with a local compactly supported correlation matrix, $\mathcal{C}^a \cdot \mathbf{R}^a$, are computed. Then, local versions of the m right-hand sides of (2b) are calculated as $\mathbf{z}^a - \mathbf{L}_i + \mathbf{e}_i$, $1 \le i \le m$. Finally, the local equivalent of (2b) is solved m times, yielding the $\mathbf{b}_i$ vectors for the current PE.
- Step 10: Computation of the portions of the analysis increments corresponding to each PE-private grid box. A local version of (2c–e) is used. Then, calls to halo ( ) are used to fill the elements of $\mathbf{x}^a - \mathbf{x}^f$ in the current PE's halo regions.
- Step 11: Transformation of the T, S, u, and v increments from the analysis levels to averages on the model layers. This step is the reciprocal of step 1.

h. Prefiltering and other miscellaneous features

The resolution of the model is much higher than that of the observational TAO network and includes spatial scales much shorter than those of the SI variability of interest in this application. Hence, it was found that it is best to apply the smoothing operator, Ξ in (1a), in the application discussed in section 4. In this respect, Ξ serves to remove from the representer matrix, $\mathbf{R}$, *short-range* covariance structure that is not resolved by the observing system. Spurious *long-range* covariances are filtered out by imposing that the covariance functions be compactly supported.

In the current implementation, the Ξ operator relies on successive applications of a simple one-dimensional recursive (infinite impulse response) filter that is applied horizontally in each layer to damp small-scale variability prior to calculating $\mathbf{L}$ and after subtraction of the ensemble-mean from each ensemble member's state vector, as indicated in (1a). The filter equations and its response function are discussed in KR01.

Incremental analysis updating (IAU; e.g., Bloom et al. 1996) is used to insert the analysis increments into each ensemble member's state vector in a gradual manner. The IAU is used here for two reasons. First, it lessens the unwanted effects of intermittent data assimilation, specifically initialization shocks resulting from imbalances between the model fields when the analysis increments are inserted directly. Second, it allows the model to gradually adjust the h field in response to the T, S, u, and v increments without violating the constraints imposed by the continuity equation.

As is common when several measurements are made at the same location between successive analyses and to reduce the overall computational burden, the observations are smoothed temporally. A discussion of this operation, sometimes referred to as superobing and introduced by Lorenc (1981), can be found in KR01.

4. Verification

a. Effect of parallel decomposition on analysis

The parallel algorithm relies on the assumptions that the analysis calculations can be partitioned, resulting in each processor assimilating local data, and that the partitioning does not have a deleterious effect on the analysis results. To verify these assumptions, the impact of performing a different local inversion on each processor rather than inverting the global system matrix, $\mathbf{S} = \mathbf{C} \cdot \mathbf{L}\mathbf{L}^T + \mathbf{W}$ in (2b), is examined in this section.

The squares in Fig. 5a show the evolution of the root-mean-square (rms) innovation (rmsi) in an experiment in which TAO temperature data are assimilated into Poseidon every 5 days for January 1993 to March 1993 using a 40-member ensemble distributed on 256 PEs (the details of the experiment are given in section 4c). In this case, sufficiently few data are involved that (2b) can also be solved on each PE without distributing the observations (i.e., without partitioning $\boldsymbol{S}$ between PEs). Hence, another experiment was run in which each PE assimilates all the available data (circles in Fig. 5a). The effect on the rmsi is minimal. Its time-mean value over the 3-month period is reduced by less than 1% from 0.906°C to 0.905°C. In contrast, in a third experiment run without imposing the condition that $\mathbf{S}$ be compactly supported [i.e., with $\mathbf{C} \cdot \mathbf{L}\mathbf{L}^T + \mathbf{W}$ in (2b) replaced by $\mathbf{L}\mathbf{L}^T + \mathbf{W}$, as in (1c)], the time-mean rmsi is 1.409°C. The respective effects on the rms residuals (rmsr) of imposing that $\mathbf{S}$ be compactly supported and of distributing the locally supported covariances is very similar (Fig. 5b). The time-mean rmsr is 0.729°C in the case of the compactly supported localized inversions (i.e., application of the Hadamard product and distribution of observations, squares). When compactly supported global inversions are used (application of the Hadamard product to the global representer matrix, circles), the rmsr is 0.727°C. With globally supported global inversions (i.e., no Hadamard product, diamonds), it is 1.309°C.

The results illustrated in Fig. 5 confirm (1) that applying the Hadamard product, thus solving (2b), is a better idea than solving (1c), as investigated in depth by Houtemaker and Mitchell (2001) in the context of a QG model, and (2) that it is safe to distribute the observations between PEs rather than having each PE assimilate them all. The latter confirmation becomes important when it is truly necessary to distribute the as-

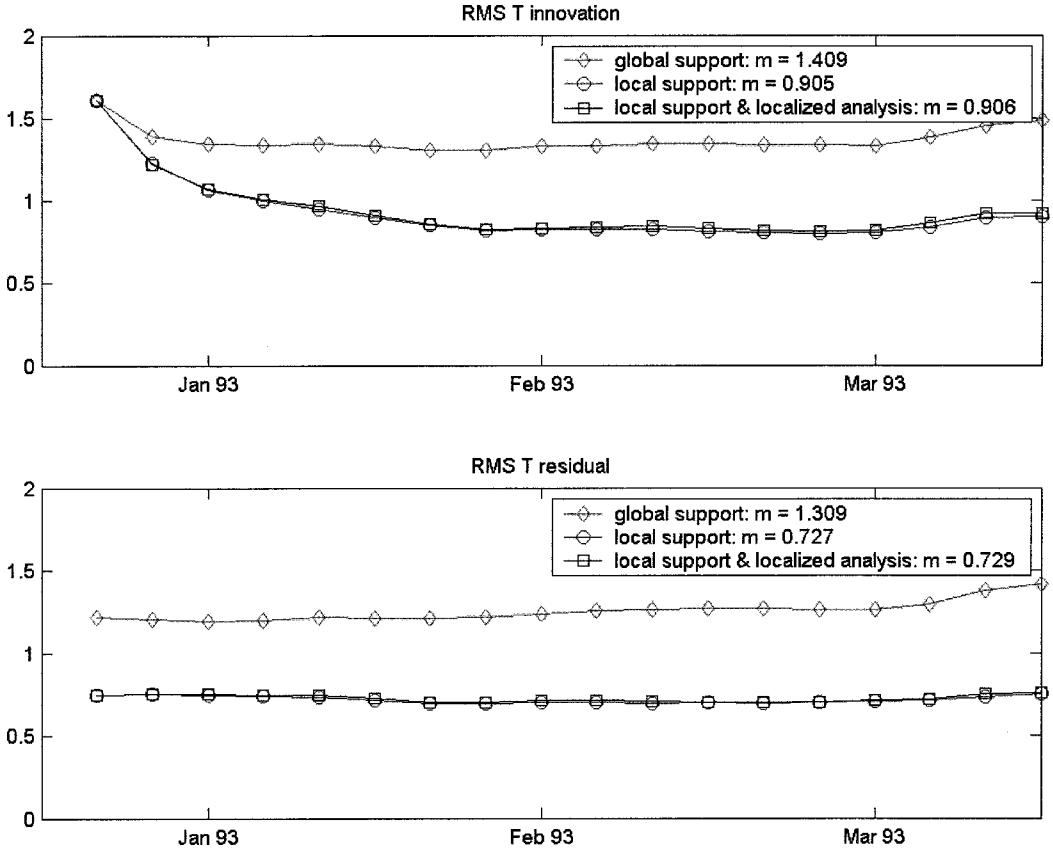

FIG. 5. Evolution of (top) the rms temperature innovation and (bottom) the rms residual during 3-month MvEnKF temperature assimilation runs spanning Jan–Mar 1993. The diamonds, circles, and squares correspond to cases with globally supported global inversions, compactly supported global inversions, and compactly supported localized inversions involving distribution of the observations between PEs, respectively.

similation because there are too many data for each PE to process them all at one time, as when gridded TOPEX/Poseidon altimeter data are assimilated (KR02a).

To gain further insight into how the parallel decomposition impacts the analysis, the temperature analysis increments from a single analysis using all the TAO mooring data on 1 January 1997 are compared in Figs. 6 and 7. For this test, 25 members and 100 PEs were used.

Figure 6a shows the SST increment in the tropical Pacific in the case of a globally supported global solution of (1c). Likewise, Figs. 6b and 6c correspond to the locally supported global solution of (2b) and to the locally supported solution of a distributed version of (2b) [see (9d) in KR01], respectively. Equatorial sections through the temperature increment corresponding to each of Figs. 6a–c are shown in Figs. 7a–c.

The most obvious effect of the Hadamard product of $\mathbf{C}$ and $\mathbf{L}\mathbf{L}^{\mathrm{T}}$ on the assimilation increment is that the latter is tapered away from the equator where no measurements are available (Fig. 6b). The large corrections away from the equator would not be expected on physical grounds and are the result of the spurious long-range background-error covariances discussed in section 3d. The effect of the Hadamard product on the vertical structure of the temperature increment is not as dramatic (Fig. 7b) since the data come from several depths between the surface and 500 m. Yet, even here where there is fairly good data coverage, there are noticeable differences. The differences indicate that spurious long-range correlations also occur along the vertical.

Comparing Figs. 6c and 6b or Figs. 7c and 7b shows that the analysis increments obtained when the analysis calculations are distributed are virtually identical to those obtained with (2), even though the global inversion (2b) is bypassed. Indeed, the rms difference between the SST increments of Figs. 6b and 6c is 1.0×10^{-3} C. Between the equatorial temperature increments of Figs. 7b and 7c it is 6.0×10^{-4} C. Thus, the tremendous computational savings associated with substituting the local $\mathbf{S}$ for the global $\mathbf{S}$ occur with a negligible impact on the quality of the analysis.

b. Scaling

At present, the two main limitations of the parallel MvEnKF are 1) that is scales poorly beyond 100 PEs on the present machine (CRAY T3E) and at the current,

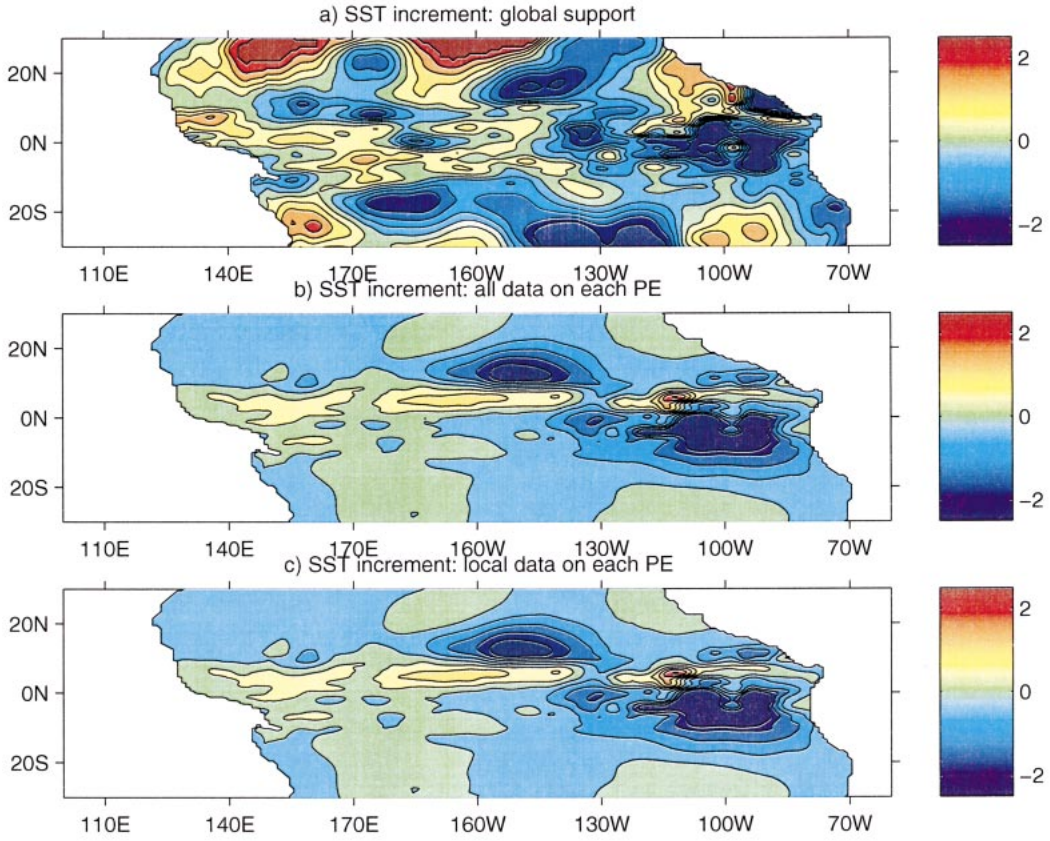

FIG. 6. Tropical portion of the analysis increments for SST (°C). (a) Global inversion without compactly supported covariances. (b) Global inversion with compact support. (c) Distributed inversion with compact support.

relatively coarse, model resolution, and 2) that the maximum ensemble size attainable is dictated by the memory of the individual PEs on a massively parallel processor (MPP) with distributed memory. Since the expected lifetime of a modern supercomputer is about two years, it is unlikely that these limitations will impose the same restrictions by the time the MvEnKF is used with the global OGCM to initialize the NSIPP production forecasts. Therefore, the software engineering approach used to implement the MvEnKF has focused on portability, modularity, and object-oriented design, rather than on optimally using the resources of the current platform.

Figure 8a shows how t_m, the time spent per ensemble member in a 5-day analysis cycle involving the assimilation of TAO temperature data, scales with N_{PE} (diamonds). The dashed curve labeled EnKF perfect extrapolates the value of t_m for 16 PEs in the range from 16 to 256 PEs, assuming linear scaling. According to Amdhal's law, such scaling can never be achieved. Instead, the time per ensemble member used on p PEs is given by $t_m = t_s(f + (1 - f)/p)$, where t_s is the time used per ensemble member on a serial machine and f is the fraction of the operations that must be performed sequentially.

The observed scaling is not easily compared with theory. First, because t_s is unknown. Second, because f depends on N_{PE}. Still, t_m decreases by a mere 16% when N_{PE} doubles from 128 to 256. Rather, t_m decreases by 45% between 16 and 32 PEs. This is indicative of saturation. The horizontal resolution of the Pacific basin version of Poseidon used in these experiments is too coarse for the distribution of its state vector over more than 100 PEs to be optimal. In contrast, the global version of Poseidon to which the MvEnKF will be applied next has enough state variables to warrant its distribution over more than 100 PEs. For reference, the observed and perfect scaling curves are also shown for the UOI. In this case, the saturation becomes apparent with 64 PEs at the current model resolution. For each value of N_{PE}, the UOI timing number is higher than the corresponding MvEnKF number because the latter corresponds to the total time divided by the ensemble size ($m_{\max}$ below).

In Fig. 8b, the largest ensemble size allowed by the individual PE memory on the CRAY T3E-600, $m_{\max}$, is shown as a function of N_{PE}. For each value of N_{PE}, the timing number in Fig. 8a corresponds to running the MvEnKF with $m_{\max}$ ensemble members, so that memory is saturated. Between 16 and 128 PEs, $m_{\max}$ increases

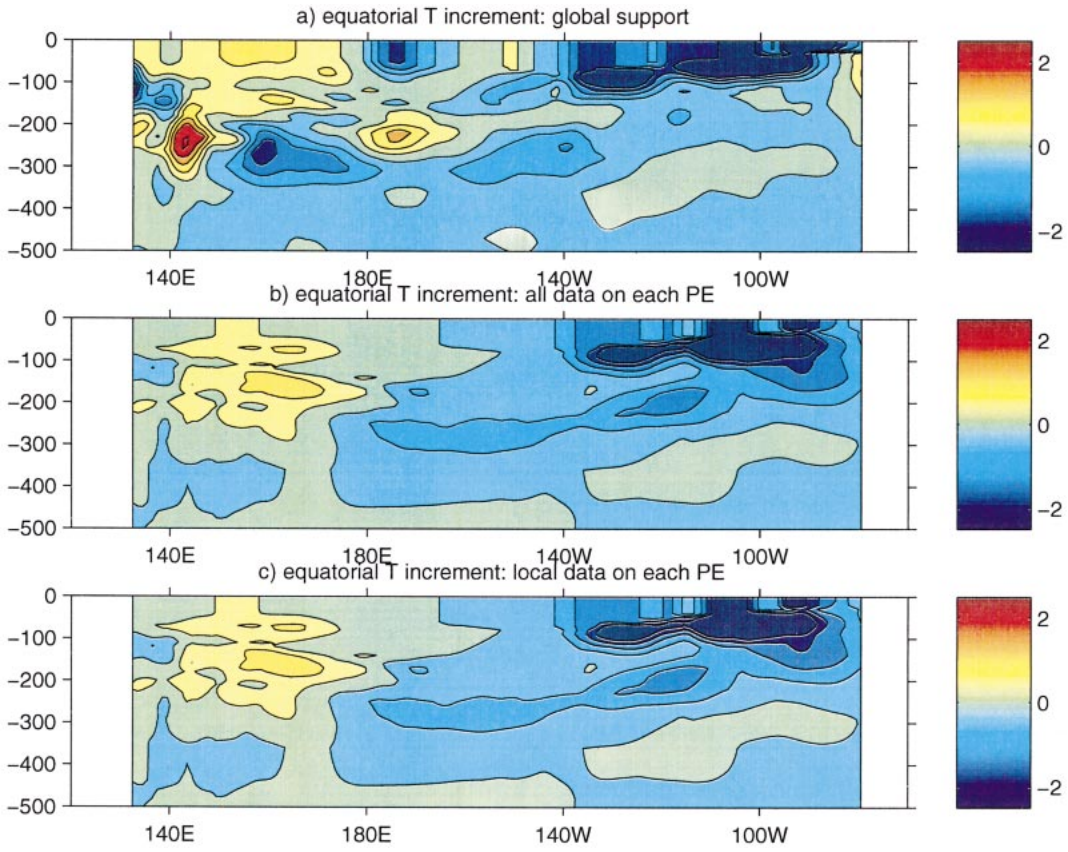

FIG. 7. Same as in Fig. 6 except for the equatorial temperature increments.

approximately linearly from 6 to 36. On 256 PEs, $m_{\max}$ is 46.

To increase $m_{\max}$ for given N_{PE}, one could simultaneously run several small ensembles on smaller PE partitions rather than a single ensemble on a large partition. However, this would require a communication mechanism not currently supported by the GEMS library. Alternatively, running the MvEnKF on a platform with globally addressable memory would also allow larger ensemble sizes. The 40-member ensembles used in section 4c and in KR02a,b achieve a good compromise between accuracy and keeping the cost of the data assimilation within acceptable limits.

c. Application

As mentioned in section 4a and to verify that the various MvEnKF components are working properly, an experiment was run in which TAO temperature data are assimilated every 5 days into Poseidon for January 1993 to March 1993 using a 40-member ensemble distributed on 256 PEs. The initial ensemble configuration in this experiment is such that the state vector of each ensemble member corresponds to the final state of a different multiyear Poseidon run. The different runs vary in the forcing anomalies used in each run to simulate the forcing errors (see section 3f).

For reference, a run without assimilation and one in which the data are assimilated using the UOI are initialized with the initial ocean state corresponding to the MvEnKF central forecast (ensemble member closest to the mean in terms of rms distance in the phase space spanned by the model state variables) at the beginning of the experiment.

After the 3-month assimilation period, the central forecast from the MvEnKF run is used to initialize a 12-month hindcast run of Poseidon forced with climatological winds, SSTs, and heat fluxes, and without temperature assimilation. The hindcast is a test of the improved ''memory'' of the subsurface ocean state following the assimilation of the TAO data. Two similar hindcasts are initialized with the states of the UOI and control runs at the end of the assimilation period.

Figure 9 shows the evolution of the spatial-mean temperature anomaly at the TAO-mooring locations, in the observations as well as in the MvEnKF, UOI, and control runs. The anomalies shown are with respect to the mean seasonal cycle calculated at each mooring and at each measurement depth for the 1990s. For the MvEnKF run, the anomalies are those of the central forecast.

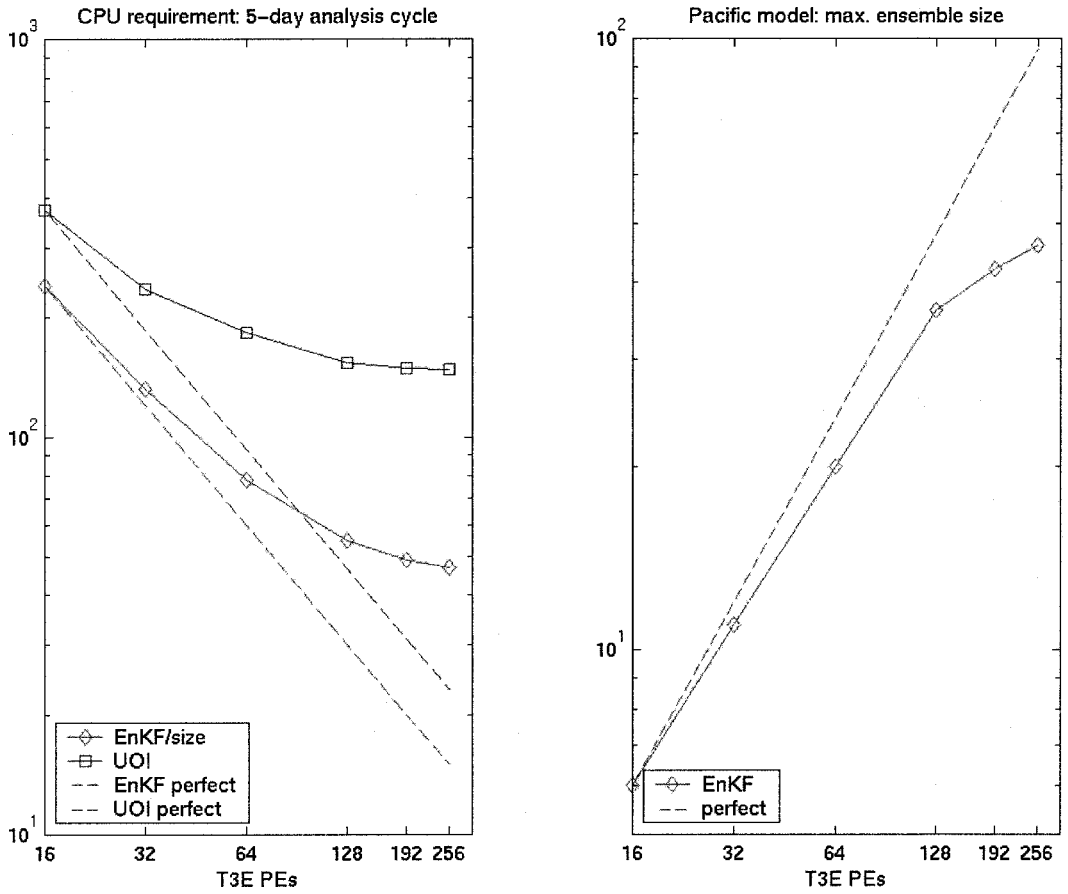

FIG. 8. (left) Time per ensemble member required to complete one 5-day analysis cycle when TAO temperature data are assimilated (t_m in text). The curves labeled perfect correspond to an unattainable linear scaling. (right) Largest ensemble size possible as a function of N_{PE} (m_{max} in text) on the CRAY T3E-600.

Initially, the UOI and MvEnKF runs have the same positive bias as the control run, since the same initial ocean state is used in the three runs. The temperature bias in the equatorial region—loosely defined here as the difference between the spatial-mean temperature anomaly in the model and the mean anomaly in the observations—is a common feature of most OGCMs. It is the result of the thermocline layer becoming too diffuse in response to the numerical diffusion applied to model for long timescale dissipative processes. Therefore, the initial effect of the temperature assimilation is to tighten the thermocline along the equator. Once this adjustment has taken place, the assumption of unbiased forecast errors is approximately satisfied and further improvements caused by the assimilation are mainly the result of correcting the unbiased error component. After a few assimilation cycles, the mean anomalies from the MvEnKF and UOI runs are close to the corresponding observed anomalies during the period with temperature assimilation.

When the temperature assimilation ceases, the UOI and MvEnKF runs start drifting back towards the warm conditions of the control run. However, even after a year with climatological forcing and no data assimilation, the positive bias of the MvEnKF run (diamonds) is only 29% that of the control run (squares). At that point, the level of bias seen in the UOI run is 72% that seen in the control. This result suggests that, even though the assimilation of the TAO temperature data using either method has a positive impact on the forecast-model bias for temperature, the better performance of the MvEnKF can partly be attributed to the underlying multivariate correction in which not only T, but also S, u, and v, are updated. Results from runs conducted under different experimental conditions and in which T only is updated (i.e., the MvEnKF is replaced by a univariate temperature EnKF) have confirmed this assumption. The two other likely reasons for the improvement over the UOI are that the MvEnKF allows for the anisotropy of the background-error covariances and that it evolves the error-covariance estimates together with the model flow.

Figure 10 shows how the ratio of the rms ensemble spread for temperature (hereafter, Enrms) to the rmsi evolves during the 3-month period with temperature assimilation. The Enrms is calculated before each analysis, after interpolation of the model temperature field to the

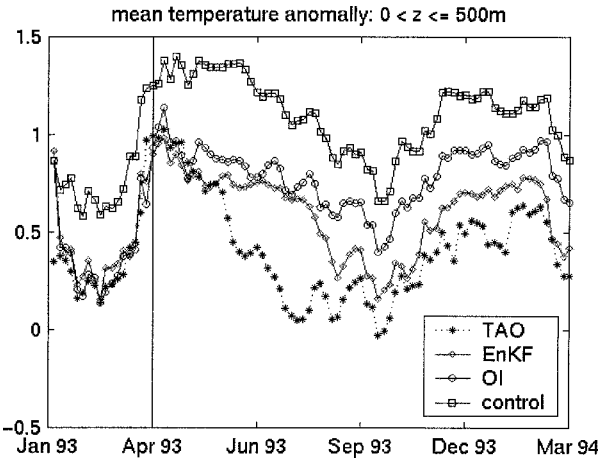

FIG. 9. Evolution of spatial-mean temperature anomaly at the TAO mooring locations during the 3-month period with TAO temperature data assimilation and during the 1-yr hindcast period without temperature assimilation. The mean anomalies shown correspond to the TAO observations (dotted line), the MvEnKF run (diamonds), the UOI run (circles) and the free-model control run (squares).

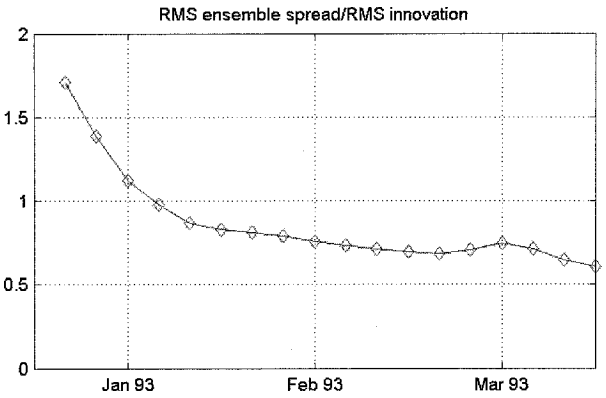

FIG. 10. Evolution of the ratio of the rms ensemble spread for temperature (taken at the location of each measurement) to the rms temperature innovation during the 3-month period with temperature assimilation. The diamonds indicate the time of each analysis.

location of each TAO measurement. It estimates the rms forecast error for temperature. Its ratio to the rmsi is thus an indicator of the how well the ensemble spread accurately predicts the rms error.

The initial value of the rms ratio is 1.73. The ensemble initialization method used in this experiment is roughly equivalent to randomly picking m initial conditions from a time series of model states. Thus, the ensemble mean is initially close to the model climatology and the fact that the Enrms is initially larger than the rmsi reveals that the initial errors for temperature have less amplitude than the climatological variability.

The ratio of the Enrms to the rmsi drops to a value of 0.98 between the first and fourth analyses. Hence, the ensemble spread is in good agreement with the actual errors at the time of the fourth assimilation. After the fifth analysis the rms ratio keeps declining, albeit more slowly than between the first and fifth analyses. Moreover, the rate at which the rms ratio declines between two successive analyses is roughly constant after the fifth analysis. A plausible explanation is that the process-noise model does not account for enough variability since the system noise is represented solely by modeling the errors in the surface wind stress and heat flux forcing (section 3f). In order for the rms ratio to stay close to unity until the end of the experiment, it will be necessary to include a term to simulate the model errors in the system-noise representation. As mentioned in section 3f, such a system-noise representation is in development. Without it, the data assimilation would eventually fail if the experiment were continued for several years. The failure would result from the MvEnKF gradually attributing less weight to the observations even though the forecast does not become gradually more accurate after the rmsi has reached a steady state. In this example, the rmsi stays close to 0.8°C after about eight analyses (Fig. 5a).

5. Summary

This article describes the MvEnKF design and its parallel implementation for the Poseidon OGCM. A domain decomposition whereby the memory of each PE contains the portion of every ensemble member's state vector that corresponds to the PE's position on a 2D horizontal lattice is used. The assimilation is parallelized through a localization of the forecast-error covariance matrix. When data becomes available to assimilate, each PE collects from neighboring PEs the innovations and measurement-functional elements according to the localization strategy. The covariance functions are given compact support by means of a Hadamard product of the background-error covariance matrix with an idealized locally supported correlation function. In EnKF implementations involving low-resolution models, one has the freedom to work with ensemble sizes on the order of hundreds or thousands. Rather, with the state-vector size of approximately 2 million variables considered here, memory, communications between PEs, and operation count limit the ensemble size. In most instances, 40 ensemble members distributed over 256 CRAY T3E PEs are used.

Besides the details of the observing system implementation, the impact of the background-covariance localization on the analysis increments is discussed, as well as performance issues. To confirm that the data assimilation system is working properly, the discussion also includes results from an initial test run in which the MvEnKF is used to assimilate TAO temperature data into Poseidon.

Some issues that must be addressed to improve the MvEnKF are the deficiency of the system-noise model, which only accounts for forcing errors, the problem of ensemble initialization, which can be addressed using a

perturbation-breeding approach, and the memory limitations inherent with running the MvEnKF on a MPP with distributed memory. On a machine with globally addressable memory, the memory-imposed constraints would be less severe. Fortunately, the modular, object-oriented approach used to develop the MvEnKF allows an easy port of the implementation from the CRAY T3E to almost every distributed-memory or shared-memory parallel computing architecture.

Acknowledgments. We thank Max Suarez for his guidance on how to use GEMS and Anna Borovikov for her help with the datasets. We are also thankful to two anonymous reviewers for raising several important issues as well as for making numerous useful suggestions. This contribution from NSIPP was supported by NASA's Physical Oceanography Program under RTOP 622-48-04. The computing resources were provided by the NASA Center for Computational Sciences. The TAO data were provided by the NOAA/Pacific Marine Environment Laboratory's TAO Project Office.

REFERENCES

Atlas, R., R. Hoffman, S. Bloom, J. Jusem, and J. Ardizzone, 1996: A multiyear global surface wind velocity dataset using SSM/I wind observations. *Bull. Amer. Meteor. Soc.,* **77,** 869–882.

Bennett, A., 1992: *Inverse Methods in Physical Oceanography.* Cambridge University Press, 346 pp.

Blanchet, I., C. Frankignoul, and M. Cane, 1997: A comparison of adaptive Kalman filters for a tropical Pacific Ocean model. *Mon. Wea. Rev.,* **125,** 40–58.

Bloom, S., L. Takacs, A. DaSilva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.,* **124,** 1256–1271.

Borovikov, A. M., and M. Rienecker, 2002: Multivariate forecast error covariance for an ocean model estimated by Monte-Carlo simulation. Preprints, *Symp. on Observations, Data Assimilation, and Probabilistic Prediction,* Orlando, FL, Amer. Meteor. Soc., 82–85.

——, ——, and P. Schopf, 2001: Surface heat balance in the equatorial Pacific Ocean: Climatology and the warming event of 1994–95. *J. Climate,* **14,** 2624–2641.

Burgers, G., P. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.,* **126,** 1719–1724.

Cane, M., A. Kaplan, R. Miller, B. Tang, E. Hackert, and A. Busalacchi, 1996: Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.,* **101C,** 22 599–22 617.

Chen, D., S. Zebiak, A. Busalacchi, and M. Cane, 1995: An improved procedure for El Niño forecasting. *Science,* **269,** 1699–1702.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Dee, D., and A. Da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.,* **124,** 269–295.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99C,** 10 143–10 162.

——, and P. van Leeuwen, 1996: Assimilation of GEOSAT altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.,* **124,** 85–96.

Fukumori, I., and P. Malanotte-Rizzoli, 1995: An approximate Kalman filter for ocean data assimilation—An example with an idealized Gulf-Stream model. *J. Geophys Res.,* **100C,** 6777–6793.

Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.,* **125,** 723–757.

Gelb, A., Ed.,1974: *Applied Optimal Estimation.* MIT Press, 374 pp.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Hamill, T., and C. Snyder, 2000: A hybrid ensemble Kalman filter-3D variational analysis scheme. *Mon. Wea. Rev.,* **128,** 2905–2919.

Horn, R., and C. Johnson, 1991: *Topics in Matrix Analysis.* Cambridge University Press, 615 pp.

Houtekamer, P., and H. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.,* **129,** 123–137.

Ji, M., and A. Leetmaa, 1997: Impact of data assimilation on ocean initialization and El Niño prediction. *Mon. Wea. Rev.,* **125,** 742–753.

——, ——, and V. Kousky, 1996: Coupled model forecasts of ENSO during the 1980s and 1990s at the National Meteorological Center. *J. Climate,* **9,** 3105–3120.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.,* **D82,** 35–45.

Keppenne, C., 2000: Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Wea. Rev.,* **128,** 1971–1981.

——, and M. Rienecker, 2001: Design and implementation of a parallel multivariate ensemble Kalman filter for the Poseidon ocean general circulation model. NASA Tech. Rep. Series on Global Modeling and Data Assimilation, NASA Tech. Memo-2001-104606, Vol. 21, 35 pp. [Available online at http://nsipp.gsfc.nasa.gov/pubs/tech_rpts/KR_Techmemo2001.pdf.]

——, and ——, 2002a: Multivariate assimilation of altimetry into an OGCM with diagnostic sea surface height using the ensemble Kalman filter. Preprints, *Symp. on Observations, Data Assimilation and, Probabilistic Prediction,* Orlando, FL, Amer. Meteor. Soc., 158–163.

——, and ——, 2002b: Assimilation of temperature into an isopycnal ocean general circulation model using a parallel Ensemble Kalman filter. *J. Mar. Sys.,* in press.

Konchady, M., A. Sood, and P. Schopf, 1998: Implementation and performance evaluation of a parallel ocean model. *Parallel Comput.,* **24,** 181–203.

Lermusiaux, P., and A. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.,* **127,** 1385–1407.

Lorenc, A., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.,* **109,** 701–721.

McPhaden, M., and Coauthors, 1998: The Tropical Ocean-Global Atmosphere observing system: A decade of progress. *J. Geophys. Res.,* **C103,** 14 169–14 240.

Mitchell, H., and P. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.,* **128,** 416–433.

Niiler, P., and E. Kraus, 1977: One-dimensional models of the upper ocean. *Modeling and Prediction of the Upper Layers of the Ocean,* E. Kraus, Ed., Pergamon, 143–172.

Pacanowski, R., and S. Philander, 1981: Parameterization of vertical mixing in numerical models of the tropical oceans. *J. Phys. Oceanogr.,* **11,** 1443–1451.

Pham, D., J. Verron, and M. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Sys.,* **16,** 323–340.

Reynolds, R., and T. Smith, 1994: Improved global sea-surface temperature analyses using optimum interpolation. *J. Climate,* **7,** 929–948.

Schaffer, D., and M. Suarez, 2000: Design and performance analysis of a massively parallel atmospheric general circulation model. *Sci. Prog.,* **8,** 49–57.

Schopf, P., and A. Loughe, 1995: A reduced-gravity isopycnic ocean model: Hindcasts of El Niño. *Mon. Wea. Rev.*, **123,** 2839–2863.

Seager, R., M. Blumenthal, and Y. Kushnir, 1995: An advective mixed layer model for ocean modeling purposes. Global simulation of surface heat fluxes. *J. Climate*, **8,** 1951–1964.

Verlaan, M., and A. Heemink, 1997: Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrol. Hydraul.*, **11,** 349–368.

Yang, S., K. Lau, and P. Schopf, 1999: Sensitivity of the tropical Pacific Ocean to precipitation induced freshwater flux. *Climate Dyn.*, **15,** 737–750.