

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 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 dis-

tance 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.

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.