Technical Report

Implementation and testing of an Ensemble Kalman Filter assimilation system for the Max Planck Institute Ocean General Circulation Model

Olwijn Leeuwenburgh

Royal Netherlands Meteorological Institute (KNMI)
P.O. Box 201
3730 AE, De Bilt, The Netherlands
olwijn.leeuwenburgh@knmi.nl

August 25, 2004

Acknowledgements

The generic EnKF package was developed by Geir Evensen (NERSC). Other important contributions to the design and development of the assimilation system were made by Laurent Bertino (NERSC), Gerrit Burgers and Geert Jan van Oldenborgh (KNMI), and Alberto Troccoli and John Greenaway (ECMWF). Invaluable advice on the MPI Ocean Model and pre- and postprocessing was given by Helmuth Haak and Noel Keenlyside (MPIfM). Discussions with Camiel Severijns (KNMI) are highly appreciated. The ENACT project was funded by the European Union (EVK2-2001-00077). The altimeter products were produced by the CLS Space Oceanography Division as part of the Environment and Climate EU ENACT project and with support from CNES.

Contents

1. Introduction

This report describes the design, implementation and testing of a sequential ocean data assimilation system at KNMI. This work was undertaken as part of the EU-funded project ENACT (Enhanced Ocean Data Assimilation and Climate Prediction), which was established in line with the recommmendation of the European climate research community to undertake a concerted program on enhanced ocean data assimilation and climate prediction. The specific objectives of the project were: 1. to provide a system to assemble and distribute high quality ocean observational data and accompanying atmospheric data, 2. to advance the techniques for assimilating ocean data from multiple sources, and implement state-of-art data assimilation schemes in state-of-art ocean general circulation models, 3. to apply the assimilation schemes and produce multi-model global ocean analyses, 4. to assess the impact of the enhanced analysis schemes by using coupled ocean-atmosphere models together with the ocean analyses to produce seasonal to inter-annual forecasts, and 5. to investigate ocean behaviour, and to quantify the uncertainty in the ocean analyses. These objectives were realised by implementing different data assimilation schemes with several numerical ocean models in order to determine the specific strengths and weaknesses of the schemes.

The task for KNMI has been to implement an Ensemble Kalman Filter (EnKF) assimilation scheme for combining observations with the Max Plack Institut für Meteorologie Ocean Model (MPI-OM), the successor to HOPE-E. KNMI has extensive experience with the HOPE-E model through involvement in the developement of its adjoint and subsequent implementation of the adjoint in a 4D-Var data assimilation scheme [van Oldenborgh et al., 1999; Bonekamp et al., 2001]. MPI-OM (also commonly referred to as HOPE-C) had not been used before at KNMI, but has been used extensively at the MPIfM in climate runs to study long-timescale climate variability and as part of coupled climate and earth-system models. It has not been used before in a data-assimilation system.

KNMI has also been involved closely in developing the ideas and principles of the Ensemble Kalman Filter [e.g. Burgers et al., 1998]. Versions of the EnKF had already been implemented and tested with a linear ENSO model, first in an experimental identical twin setup, and more recently in a realistic assimilation experiment aimed at improving ENSO forecast skill [Leeuwenburgh and Burgers, unpublished manuscript]. Most of the algorithms used in preparing this report can be found in a largely model-independent EnKF package that has recently been released [Evensen, 2003; Evensen, 2004].

During the ENACT project altimetric sea level data from the TOPEX/POSEIDON and ERS missions, as well as temperature and salinity profiles from the WOCE and post-WOCE periods were to be assimilated. These data were pre-processed by CLS and the UK MetOffice repsectively. The validation tests described in this report use simulated altimetric sea level data only.

In Section 2 the main characteristics of the MPI-OM model are reviewed, and modifications that have been implemented are outlined. In Section 3 methods of representing forecast error are discussed. The processing method for altimetry and insitu data is presented in Section 4. Section 5 gives an overview of the EnKF algorithm that has been

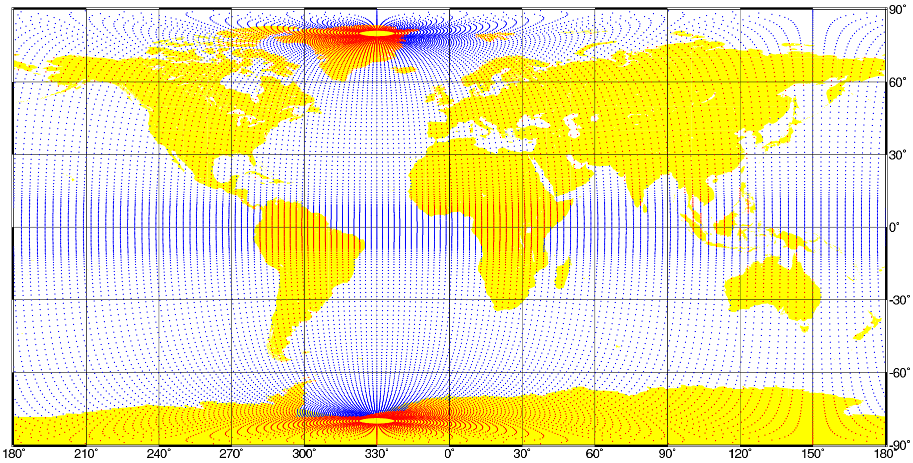


Figure 1: The MPI-OM global grid.

implemented in the current system. In Section 6 the pre- and postprocessing of forcing fields and analysis output are outlined. Section 7 discusses some results from a twin experiment, and Section 8 concludes with a summary. Appendices A and B provide information on Fortran routines and shell scripts that have been developed to run the assimilation on the IBM Cluster 1600 supercomputer system of the ECMWF.

2. The MPI Ocean Model version 1

2.1. Overview

This section will describe the relevant characteristics of the global ocean/sea ice model MPI-OM-1, which has been developed at the Max Plack Institute für Meteorologie . The model equations and parameterizations, as well as results from a 450-year climatologically forced integeration are described in [Marsland et al., 2003], but some relevant properties and characteristics are briefly reviewed here.

The standard global orthogonal curvilinear grid is at a spatial resolution approximating spectral truncation T42, with poles positioned over Greenland and inland of the Weddell Sea to give high resolution in the main sinking areas associated with the THC (Figure 1). The symmetry in their offset means that the equator lies along one of the model's parallels. Additional increase in resolution is achieved by meridional refinement of the grid within 10 degrees of the equator. Horizontal discretization of the primitive equations is on the Arakawa C-grid, while the z-coordinate is discretized on 23 vertical levels.

The main changes in the physics with respect to the previous HOPE versions of the model are in new parameterizations of subgridscale processes, such as a bottom boundary layer slope convection scheme [Beckmann and Döscher, 1997], isoneutral/dianeutral diffusion of tracers [Redi, 1982; Gent et al., 1995; Griffies, 1998], eddy-induced tracer transport [Gent et al., 1995], and optional time dependent penetrative plume convection

[Paluskiewicz and Romea, 1997]. Surface fluxes of heat and momentum are calculated through bulk formulae using prescribed fields of surface wind stress, 10m wind speed, 2m air and dewpoint temperatures, precipitation, cloud cover, and incoming solar radiation. The fluxes are further dependent on the presence of sea ice and snow, which are modeled following [Hibler, 1979]. Additional fresh water forcing is in the form of river runoff and glacier calving.

The model has been run successfully over 450 years with the OMIP climatology forcing and with relaxation to surface salinity only [Marsland et al., 2003]. While in fair agreement with other z-coordinate models, the model produced a slightly low mass flux through Drake Passage, relatively small poleward heat transports, and somewhat weak horizontal subtropical cells in the North Atlantic and Pacific. Sea ice production in the Greenland Sea and Denmark Strait was too strong, while there was too little March sea ice around Antarctica. Deep convection was probably too strong and widespread in both the northern arctic seas and the Weddell Sea despite the increase in spatial resolution in these regions.

2.2. Forcing and relaxation

In this report the OMIP forcing has been replaced by ERA40 daily forcing fields, which were interpolated from the Gaussian T106 grid to the model grid. In addition, a daily ERA40-derived climatology was used during spinup of the model. The ERA40 precipitation in the tropics is known to be too low and was corrected as described by Troccoli and Kållberg [2004]. In both the control and the assimilation runs relaxation of temperature and salinity to monthly Levitus climatology with a 3-year timescale was applied from level 4 (65m) downward, in addition to surface relaxation to monthly surface salinity with the same time-scale, and relaxation to daily observed SST derived from the Reynolds and Smith [1994] data set with a 5-day time-scale (rate equivalent of 200 $Wm^{-2}K^{-1}$). Monthly climatological values were used to represent river runoff of the world's largest rivers, while no glacier calving has been incorporated so far.

2.3. Additional model changes

Several changes have been introduced into the standard MPI-OM-1 model code for data assimilation purposes. These changes include the relaxation to daily surface temperatures, the writing of restart files in a format corresponding to that expected by the assimilation system, specification of exact start and end dates of model runs, and the perturbation of forcing fields and model parameters during a run. The option to relax to surface temperature can be switched on or off on compilation with the appropriate CPP flag. The relaxation coefficient (`CRELTEM`) can be specified in the `OCECTL` file. The length of the run in terms of model days can now be determined by the main program based on the input parameters `LD_START` and `LD_END`. If `LD_START` is less than 0, the start date is calculated from the end date of the previous run which is read from the restart file. The generation of perturbations will be discussed in detail in the next section.

3. Forecast error representation

3.1. Introduction

Forecast uncertainties have three sources: initial conditions, boundary conditions, and model error. Uncertainty in the initial conditions is reflected by the spread of the ensemble at the start of the forecast run. Model error can be associated with everything ranging from limits in numerical precision and the use of finite differences to incorrect constants in parameterizations and the neglect of physical processes. The effect of this variety of error sources can be simulated, for example, by using a range of model parameter values in the different ensemble members, by perturbation of the model fields at forecast time, or by perturbation of the model equations themselves by adding small random errors to the tendencies in the model fields. At this moment only the perturbation of eddy diffusion and viscosity coefficients has been implemented , but it has not been used in the tests described in this report. Uncertainty in the boundary conditions is represented by random perturbations of the ERA40 surface forcing fields. A measure of uncertainty in the prescribed ERA40 fields is obtained by comparison with the NCEP/NCAR reanalysis [Kalnay et al., 1998], under the assumption that it is equally plausible that differences between the two products are due to errors in either one. So far, perturbation of surface wind stress, air temperature, dew point temperature, and solar radiation has been implemented. Two methods by which this can be done are described below.

3.2. Diffusion/convolution

Three-year time series of daily difference fields were used to diagnose proxies for the variance, and spatial and temporal correlation scales of errors at each point in the model grid. The first method for computation of random 2D fields follows the diffusion method [Derber and Rosati, 1989; Weaver and Courtier, 2000]. A spatially correlated 2D random field can be calculated directly by convolution of a field of white noise f_0 with a 2D Gaussian correlation function with specified scales σ_x, σ_y ,

$$f_T(x,y) = \frac{1}{\sqrt{\pi\sigma_x\sigma_y}} \int\limits_A e^{-\frac{(x-x')^2}{\sigma_x^2} - \frac{(y-y')^2}{\sigma_y^2}} \quad f_0(x',y')dA \quad . \tag{1}$$

It can be shown that this is identical to the solution at $t = T$ of the 2D diffusion equation

$$\frac{\partial f}{\partial t} = \kappa_x \frac{\partial^2 f}{\partial x^2} + \kappa_y \frac{\partial^2 f}{\partial y^2} \quad , \tag{2}$$

where $\sigma_x^2 = 4\kappa_x T$ and the initial conditions are $f_{t=0} = f_0$. It turns out that this second method is much faster than the direct convolution. Both methods have been implemented by G. J. van Oldenborgh, KNMI, for the purpose of calculating background covariance matrices for a 4D-var system, but have been adapted here for use with the EnKF. This procedure has the additional advantage over the FFT method suggested by [Evensen,

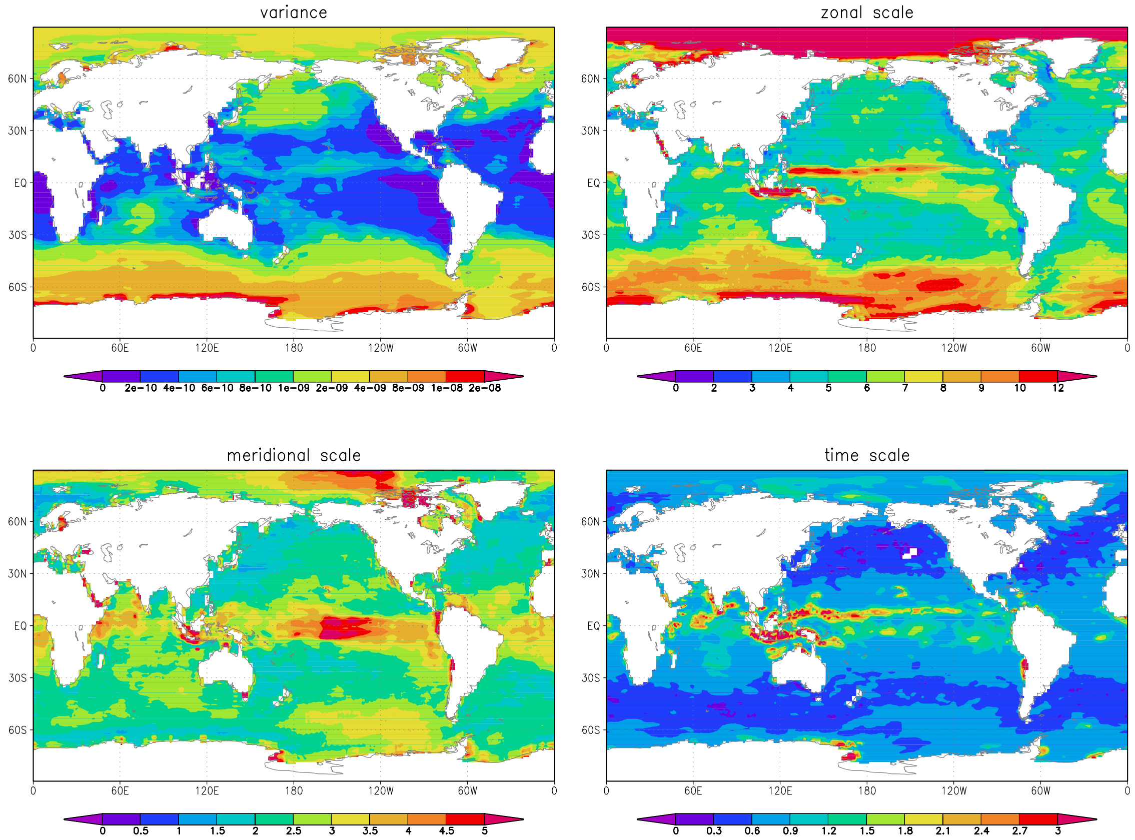


Figure 2: Variance and spatial and temporal correlation scales of zonal windstress perturbations, as determined from ERA40-NCEP daily differences. Spatial scales are in units of spherical degrees, and temporal scales are in days.

2003] that it is possible to specify the spatial error scales independently at each location in the model grid. Figure 2, for example, shows the variance, and zonal, meridional, and temporal scales of zonal wind stress errors based on a comparison of daily difference fields over the period 1992-1994.

It was discovered that extremely large differences exist between air temperatures in the ERA40 versus the NCEP/NCAR renalysis in the polar regions, the Weddell Sea and the Arctic Ocean in particular. This may be associated with the prescribed presence or absense of land or sea ice in these regions in the reanalyses, but no definite explanation has been found so far. Therefore it was decided to limit the perturbation standard deviations for air temperature and dew point temperature to a maximum value of 4°C.

The random field generation is done by the routine `convol.F90` which is called at the start of every model day and called by the main program `MPIOM.F90`. It takes as input the zonal and meridional spatial scales of the errors. The resulting random spatial field should then be rescaled with the proper error variance times a rescaling factor which should be determined on forehand seperately for each forcing variable, since it depends

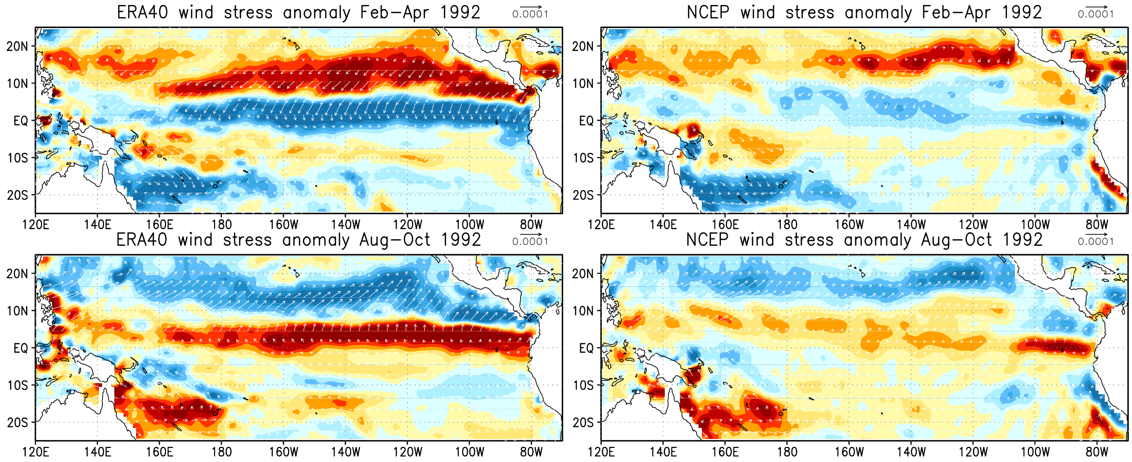

ERA40 wind stress anomaly Feb–Apr 1992
NCEP wind stress anomaly Feb–Apr 1992
ERA40 wind stress anomaly Aug–Oct 1992
NCEP wind stress anomaly Aug–Oct 1992
0.0001
20N
10N
EQ
10S
20S
120E
140E
160E
180
160W
140W
120W
100W
80W

Figure 3: ERA40 and NCEP/NCAR reanalysis surface wind stress anomalies and their convergence during high and low phases of the seasonal cycle.

on the specific spatial error scales of each variable. The random number generator is initialised using the current calendar date, the time, and the system clock, in order to avoid repeating the random seed at successive steps in the assimilation cycle.

3.3. Empirical Orthogonal Functions

A second method of calculating 2D random fields that has been implemented is to use random combinations of Empirical Orthogonal Functions (EOFs) of the difference fields, as was done by [Robert and Alves, 2004]. Since higher EOFs typically contain increasingly smaller scales, using a limited number of EOFs provides a natural way of imposing a limit on the spatial scales of the perturbation fields. A combined EOF decomposition was performed of a 2-year record of all the above difference fields simultaneously, after normalizing each variable with its standard deviation and after removal of all variability with periods shorter than 20 days by application of a Loess smoother. The combined EOF decomposition results in error patterns that are temporally coherent between the different forcing variables. This will decrease the possibility that random perturbations in e.g. wind stress and air temperature act in such a way that their effects on surface temperature cancel, which would be less effective in increasing the ensemble spread. The first 4 EOFs, representing more than 40% of the total variance, correspond to annual and semi-annual modes describing a seasonal difference in the strength and position of the ITCZ bewteen ERA40 and NCEP (see Figure 3). This signal can also be recognized in Figure 2 in for example the zonal scales. This signal consitutes a predictable bias in the forcing, which can most likely be ascribed to the NCEP/NCAR reanalysis which has a much weaker ITCZ. Therefore we repeated the EOF decomposition after first removing the annual and semi-annual cycles. The first EOF of the remaining record now only represents about 7% of the total variance.

In their single 10-day ensemble run with an OGCM Robert and Alves (2004) used a temporally constant perturbation of wind stress. Here we have implemented an easy method (Evensen, 2003) to enforce time correlation between random daily perturbations with time scales corresponding to those found in the ERA-NCEP time series. Daily perturbations Δp_k are calculated as the weighted sum of the previous perturbation and a Gaussian distributed random number r

$$\Delta p_k = \alpha \Delta p_{k-1} + \beta r$$

with $\alpha = e^{-1/\gamma}$ and $\beta = \sqrt{1 - \alpha^2}$. The resulting time series p_k can be characterized by an exponentially decreasing correlation function with e-folding scale γ. It is possible to manually modify the temporal scales if desired. One may for example wish to include perturbation time scales reminiscent of westerly wind bursts, which have been associated with the onset of ENSO, or rather increase time scales to enable the wind perturbations to trigger oceanic Rossby and Kelvin waves. The most consistent method to do this however is to recalculate the EOFs after filtering the time series over the appropriate time scales.

Care should be taken that perturbations do not cause the forcing variable to take on unphysical values. For example, the incoming solar radiation and precipitation can not be negative, while the fractional cloud cover should remain between 0 and 1. A simple way to prevent this problem from occurring that was implemented for solar radiation is to reset all negative values after perturbation to 0. In order to assure that the mean is not changed all values greater than twice the mean should all so be cut off. This leaves a distribution of perturbations that is no longer Guassian, since the tails have been cut off. An alternative method to create perturbations for non-Gaussian variables is described by Bertino et al. [2002], and involves the use of a transformation function that maps the variable's distribution to a (nearly) Gaussian shape. Perturbation takes place in this transformed space after which the inverse function is used to map the variable back to physical space. This approach could be used to generate additional perturbations for cloud cover and precipitation.

4. Observations

4.1. Altimetric data selection and preparation

The alimetry products used in this report are the along-track ERS-1, ERS-2, and TOPEX/POSEIDON SLA data provided by the CLS Space Oceanography Division . This section will describe the process of selection and pre-processing of these altimetric data for use in the assimilation.

Given some specified time window around each assimilation time, the appropriate repeat cycles from the ERS and T/P missions from which data are to be selected are determined . Each track is then subsampled within certain geographical limits with an adaptive step size, based on the local size of the model grid cells. The rationale behind this is that only variability on spatial scales greater than a certain number n times the local grid size is actually resolved by the model. A Loess filter [Cleveland and Devlin, 1988; Schlax and Chelton, 1992] with cutoff wavelength $n\Delta x$ is applied along the track to determine the unresolved variability in a root-mean-square sense over all available cycles. This 'representation error' is subsequently added to the instrument noise (2.5 cm for T/P and 5 cm for ERS) to obtain a total error estimate at the sample point, after which the procedure is repeated at the next sample point which is located a distance $n\Delta x/2$ further along the track. (An alternative method to select altimetric data and obtain the corresponding error estimates was suggested by [Appeldoorn and van Oldenborgh, 2003].)

During the assimilation of both the altimetry and insitu data errors between points seperated horizontally in space are generally considered uncorrelated. It has been made possible to include along-track correlated error in the altimetry, since it is known that geographically correlated orbit error with a dominant wavelength of one orbital revolution still remains in the data.

A subsequent step is required in preparation for assimilation that associates each data point with a corrsponding grid cell in the model. While this is very easy for regular lat-lon model grids, more care is required for the irregular grid of MPI-OM. In absense of a mapping function relating lat-lon coordinates to model indices, the model grid point nearest to the observation is first found with the help of a search table, which contains strings of indices of model grid cells contained within regular lat-lon boxes. This way the nearest grid point can be determined by calculating the distances to only a limited number of points. In order to establish which grid cell actually contains the observation (the *pivot point*), the vectors from this nearest grid point to the four surrounding grid points, as well as to the observation are determined. From the directions of these five vectors, one can determine the pivot point. The model equivalent of the observation will eventually be determined as a distance-weighted sum of the model values in the four grid cells surrounding the observation. It is required that all of these four grid cells are 'wet points' with depths greater than 2000m in order to remove observations from shallow water that may contain large errors due to the poor quality of tidal correction models in these regions.

4.2. Insitu data

The insitu dataset has been prepared at the UK Met Office and contains reprocessd temperature and salinity observations from the World Ocean Database 2001 [Levitus, 2001], and additional CTD data from the WOCE hydrographic program, Greg Johnson's CTD data set, BMRC/CSIRO XBT data from 1997 onward, and GTSPP data from Jan 2000. The processing of these data for use in ENACT has been documented by Ingleby and Huddleston [2003].

The data selection largely follows the procedure set out in the previous section. Again, data are selected from a time window around the analysis time. As before, there is a minimum depth requirement for the four surrounding model grid cells. Currently only data above 800 m depth are assimilated since no mechanisms are in place yet to create substantial model ensemble spread at larger depths. The accuracy of salinity measurements is currently set to 0.05 PSU, while for temperature the depth-dependent functional form for the background errors of the Levitus subsurface analysis is used. This report does not address the testing of insitu data assimilation.

5. The Ensemble Kalman Filter

5.1. Introduction

The objective of sequential methods to assimilate data is to get a best estimate of the state of the climate system based on all available information up to and including the analysis time. Thus no data beyond the analysis time are used and past analyses are not updated when the data at the analysis time become available. The application of sequential methods lies therefore mostly in now- and forecasting where the analysis serves as the initial condition for a forecast run. The link between sequential filters and variational methods, which use the adjoint of the model to propagate information backwards in time, is made by the so-called 'smoothers' [Rauch et al., 1965; van Leeuwen, 2001].

The Ensemble Kalman Filter was first formally proposed by Evensen [1994]. It uses an ensemble of different model realisations to represent the uncertainty in the model forecast. This Monte Carlo-type of approach is the fundamental difference with the Kalman Filter, which is the optimal estimator for linear models, and the Extended Kalman Filter (EKF), which uses linarized dynamics at the analysis time to propagate the covariances forward in time, usually in combination with some form of reduced-space representation. The ensemble representation of model error covariances has two main advantages over other approaches. First, no linearization of the dynamics is required, so that non-linear effects on the mean are properly incorporated, and secondly, the use of a finite ensemble size N provides a convenient way to limit the computational cost. It can be shown that the error in empirical model covariance estimates is proportional to $1/\sqrt{N}$, and that the ensemble representation of model error will asymptotically approach the true error distribution if the ensemble size goes to infinity. It has been found empirically that a minimum ensemble size of about 100 members is required to properly represent cross-covariances between state variables in intermediate-complexity ocean and atmosphere models, while larger ensemble sizes have so far typicaly proven too computationally expensive for larger models.

5.2. Implementation

The standard formulation of the EnKF algorithm is given in [Evensen, 1994] and [Burgers et al., 1998]. The practical implementation used here largely follows the modified version described in [Evensen, 2003], with minor modifications suggested by the comment of Kepert [2004]. An alternative implementation of the so-called *local analysis* scheme follows the procedure outlined in [Houtekamer and Mitchell, 1998], and [Kepenne and Rienecker, 2001]. Two recent implementations [Evensen, 2004] of ensemble 'square-root' algorithms are introduced as well. A short recap of the analysis equation is given below, and modifications with respect to the standard algorithm are indicated.

Following the notation of [Evensen, 2003], the standard analysis equation can be written

$$\mathbf{A}^a_{i+1} = \mathbf{A}_{i+1} + \mathbf{P}_{i+1}\,\mathbf{H}^T\,(\mathbf{H}\,\mathbf{P}_{i+1}\,\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{D}_{i+1} - \mathbf{H}\,\mathbf{A}_{i+1})\;, \tag{3}$$

If an ensemble is used, the model error covariances are represented by the spread of the model ensemble $\mathbf{A} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_N)$,

$$\mathbf{P} = \frac{\mathbf{A}'\mathbf{A}'^T}{N-1}\;, \tag{4}$$

where $\boldsymbol{\psi}_i$ contains the state vector of ensemble member i, and primes indicate anomalies with respect to the ensemble mean. Evensen [2003] proposed a modification of the standard EnKF algorithm in which the observation error covariance matrix $\mathbf{R}$ is computed from the observation perturbations. Dropping the time indices, the above equation can then be rewritten as

$$\mathbf{A}^a = \mathbf{A} + \mathbf{A}'\mathbf{A}'^T\mathbf{H}^T(\mathbf{H}\mathbf{A}'\mathbf{A}'^T\mathbf{H}^T + \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T)^{-1}(\mathbf{D} - \mathbf{H}\,\mathbf{A})\;. \tag{5}$$

The observation error covariance matrix has thus been represented by the covariances between the observation perturbations $\boldsymbol{\Upsilon}$. The advantage of this is that the inverse of $(\mathbf{H}\mathbf{A}'\mathbf{A}'^T\mathbf{H}^T + \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T)$ can be computed in numerically very efficient way, using the SVD of $(\mathbf{H}\mathbf{A}' + \boldsymbol{\Upsilon})$, provided that $\mathbf{H}\mathbf{A}'\boldsymbol{\Upsilon}^T = 0$. However, it was shown by Kepert [2004] that this 'recycled' use of the observation perturbations in both $\mathbf{D}$ and $\mathbf{R}$ will formally lead to collapse of the ensemble in the common situation where the size of the ensemble N is less than $m/2+1$, where m is the number of observations. While the truncation of the SVD of $(\mathbf{H}\mathbf{A}' + \boldsymbol{\Upsilon})$ may prevent this collapse, it was shown that even with proper tuning the results remain inferior to the standard EnKF algorithm. In order to saveguard against possible collapse, a second set of observation perturbations can be used to represent $\mathbf{R}$. This does not solve the fact that the condition $\mathbf{H}\mathbf{A}'\boldsymbol{\Upsilon}^T = 0$, although statistically true for an infinite ensemble, will not be met exactly, and it was found in a simple 1D test case that small-scale noise will still remain in the analysis [Jeff Kepert, personal communication]. Alternative solutions to preventing collapse include improved perturbation sampling, and alternative algorithms based on the ensemble square-root filters which avoid observation perturbations altogether.

The scheme of [Houtekamer and Mitchell, 1998] remains closer to the standard EnKF and also avoids collapse. The main innovation here has been the use of compactly supported covariance functions to reduce the effects of spurious long-range correlations due to the finite size of the model ensemble. It was noted by Gaspari and Cohn [1999] that the so-called Schur product of two covariance functions is also a covariance function, and they suggested taking the Schur product of an empirical covariance function with a compactly supported (space-limited) function to ease the computational burden in data-analysis algorithms. Their function Eq. (4.10) has been implemented here. It was argued by Houtekamer and Mitchell [1998] that it is reasonable to take the Schur product after application of the measurement functional $\mathbf{H}$ such that explicit calculation of $\mathbf{P}$ itself can be avoided. A second approximation to the standard EnKF that can be used now is the so-called local-analysis. This results from the fact that observations further than a certain distance will now be uncorrelated with the analysis grid point. The analysis can thus take place independently for each grid column (already true for the 'global analysis'), but using only the nearest observations. A similar effect can be achieved by a modification to the [Evensen, 2003] scheme by downweighting innovations that are further away from the analysis grid point, and this has been implemented here. Since the standard EnKF algorithm uses the true $\mathbf{R}$, the matrix $\mathbf{H}\ \mathbf{P}\ \mathbf{H}^T + \mathbf{R}$ needs to inverted. A singular value decomposition is implemented here (Houtekamer and Mitchell [1998] use a Cholesky decomposition while Keppenne and Rienecker [2001] use the LU decomposition).

In a recent paper Evensen [2004] implemented a so-called determinitic square root algorithm [Bierman, 1977; Heemink et al., 2001; Tippett et al., 2003], and found that this algorithm peformed equally well with low ($m < N$) and full ($m = N$) rank representations of the observation error covariance matrix for a one-dimensional linear advection model. All ensemble square root algorithms use the standard Kalman Filter analysis equation to update the ensemble mean. The ensemble anomalies to the analyzed mean are obtained as matrix square roots of the analysis error covariance matrix

$$\mathbf{A}^{a'}\mathbf{A}^{a'^T} = \mathbf{A}'(\mathbf{I} - \mathbf{A}'^T\mathbf{H}^T(\mathbf{H}\mathbf{A}'\mathbf{A}'^T\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{A}')\mathbf{A}'^T$$

No perturbation of the observations is required. It can be shown that if $\mathbf{A}^{a'}$ is a valid root, then so is $\mathbf{A}^{a'}\mathbf{U}$, where $\mathbf{U}$ is any $m \times m$ orthogonal matrix for which $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$. Various choices for transformation of the matrix square roots have led to alternative algorithms [Bishop et al., 2001; Anderson, 2001; Whitaker and Hamill, 2002]. The implementation by Evensen [2004] uses no transformation, but applies a fast decomposition of the innovation error covariance matrix.

5.3. Practical issues

Some of the practical issues associated with the implementation of the generic NERSC EnKF package for a specific numerical model will now be discussed. The horizontal and vertical dimensions of the model grid should first be specified in `mod_dimensions.F90`. The prognostic variables of the models should then be identified and listed in `mod_states.F90`. The routine `m_consistency_check.F90` checks that the model forecast fields do not exceed certain physically unrealistic limits, and is dependent on the model used. Depending on whether a specific type of observation has already been accounted for, `m_modstate_point.F90`, which finds the model equivalent to each observation, may need to be changed. The choice between local and global analysis is indicated in the input file `assimilation.in`. A few more points to consider when starting from scratch are mentioned in the *Implementation Guide for the Generic EnKF Package* which can be found on the NERSC EnKF website.

6. Pre- and post-processing

Several external packages were used in the preparation of this report. The standard input and output format for the MPI-OM-1 model is the EXTRA format, developed at the Meteorologisches Institut der Universität Hamburg. This format can be read in a Fortran program by

```
READ(10) IDATE,ICODE,ILEVEL,NSIZE
READ(10) (FIELD(ISIZE),ISIZE=NSIZE) .
```

The Procedural INterface for Grib formatted Object (PINGO) package was developed at the Deutsches Klimarechenzentrum (Waszkewitz et al., 1996) and provides many useful tools for manipulation of EXTRA formatted files. The main postprocessing tools used for this report are contained in the Standard Ocean Model Postprocessor (STOMPP) package (V. Gayler, 2001). This package integrates the Spherical Coordinate Remapping and Interpolation Package (SCRIP) (P. Jones, 1997) with the Grid Analysis and Display System (GRADS) (B. Doty, University of Maryland). The SCRIP package contains software used to generate interpolation weights for the remapping of fields between irregular grids in sperical geometry. In this report the area-integrated conservative remapping scheme is used throughout to interpolate 2D and 3D model output fields to the ENACT common grid, with longitudes defined from 0°to 359°E and latitudes between 89°S and 89°N. 3D fields were additionally interpolated in the vertical to the 33 standard levels of the World Ocean Database, using a cubic-spline interpolation routine from the Numerical Recipes package (Press et al., 1992). The ENACT common output format is NetCDF (Rew et al., 1997).

7. Testing and validation

7.1. Model behaviour

Figure 4 compares the mean of the model sea level over the 7 year control run period 1993-1999 with the Mean Dynamic Topography (MDT) prepared by M.-H. Rio and F. Hernandez (CLS). A value of 30cm has been subtracted from the zero gobal-mean MDT. Model gradients are generally a little smoother across the western boundary current regions and their extensions (e.g. the northern hemisphere subtropical-subpolar gradients). While otherwise the broad-scale model mean compares very well to the MDT, the model mean has been in used so far to obtain sea level anomalies for comparison with the altimetric anomalies.

Figure 5 shows the standard deviation of model sea level variability determined over the 7 year control run period 1993-1999. This figure can be compared with variability maps calculated from altimetric anomalies [e.g., Ducet and Le Traon, 1999]. The most notable shortcoming of the model is the severe lack of variability in all the western boundary current regions, such as the Gulf Stream and the Kuroshio, as well as in the Aghulas Retroflection Zone and along the path of the ACC. The tropical and equatorial current

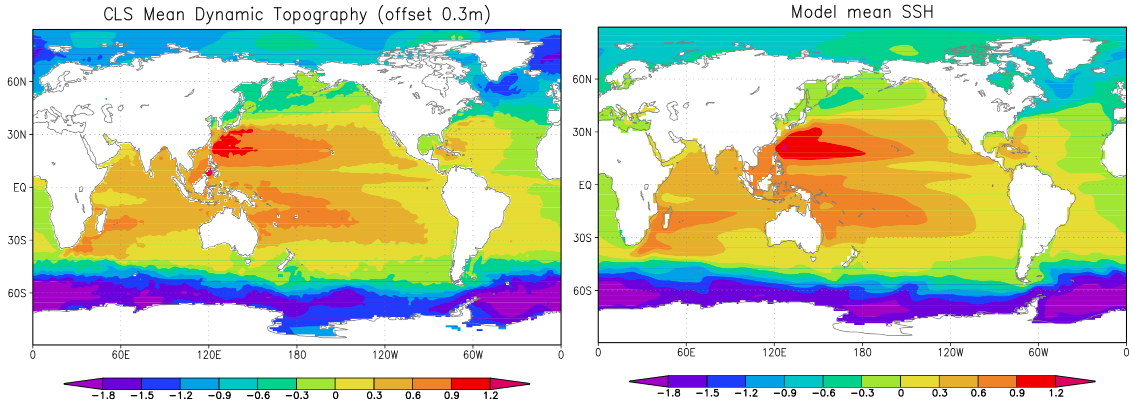


Figure 4: CLS Mean Dynamic Topography (offset 0.3m) calculated over the period 1993-1999 from altimetry, and the mean SSH from MPI-OM over the same period.

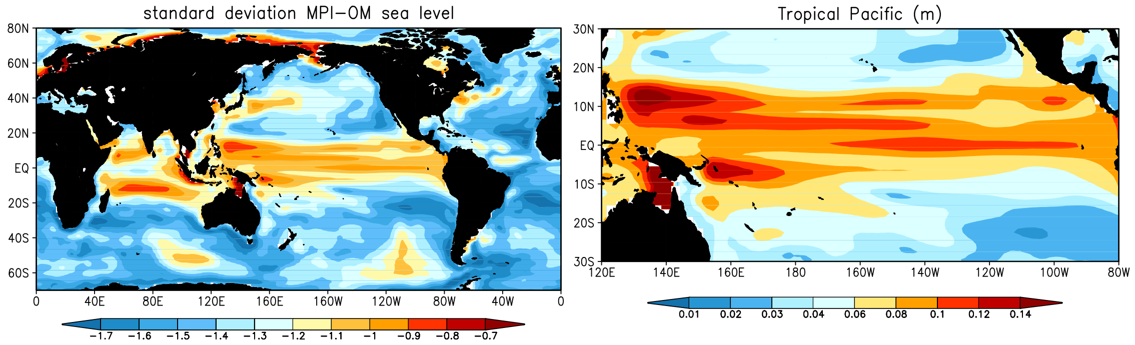


Figure 5: Standard deviation of global model sea level variability (log10), and of Tropical Pacific sea level variability (m).

systems are represented much better due to the increased resolution there, and SLA variance reaches levels comparable to that seen in the observations. Other notable features in the observations, such as the variability associated with the Azores Current and the Loop Current in the Gulf of Mexico are not clearly seen in the model. As is common in low resolution models, the Gulf Stream separation point is too far north. The amplitude of the annual cycle in sea level (not shown) is fairly well reproduced, except in a zonal band south and east of South Africa where model amplitudes are rather low.

7.2. Twin experiment

The assimilation system is tested here using a twin experiment setup. The true ocean state is defined by a forward run of the ocean model using unperturbed NCEP/NCAR reanalysis forcing fields. A background (or 'first-guess') estimate of the ocean state is obtained by running the ocean model over the same period forced by unperturbed ERA40 forcing fields. A plot (not shown) of the true and background states at the start of the

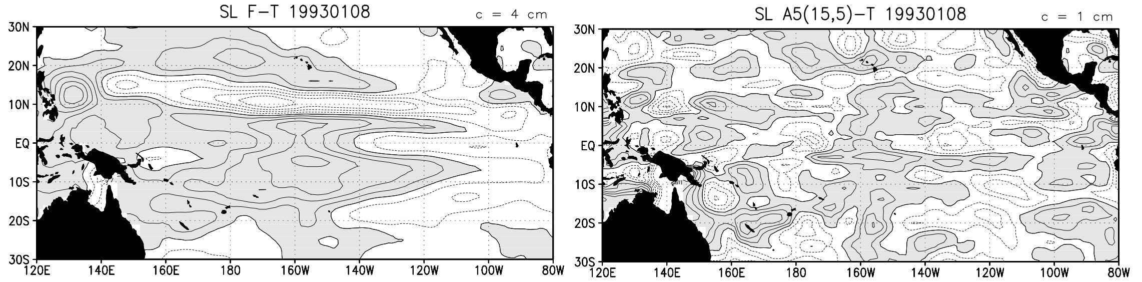
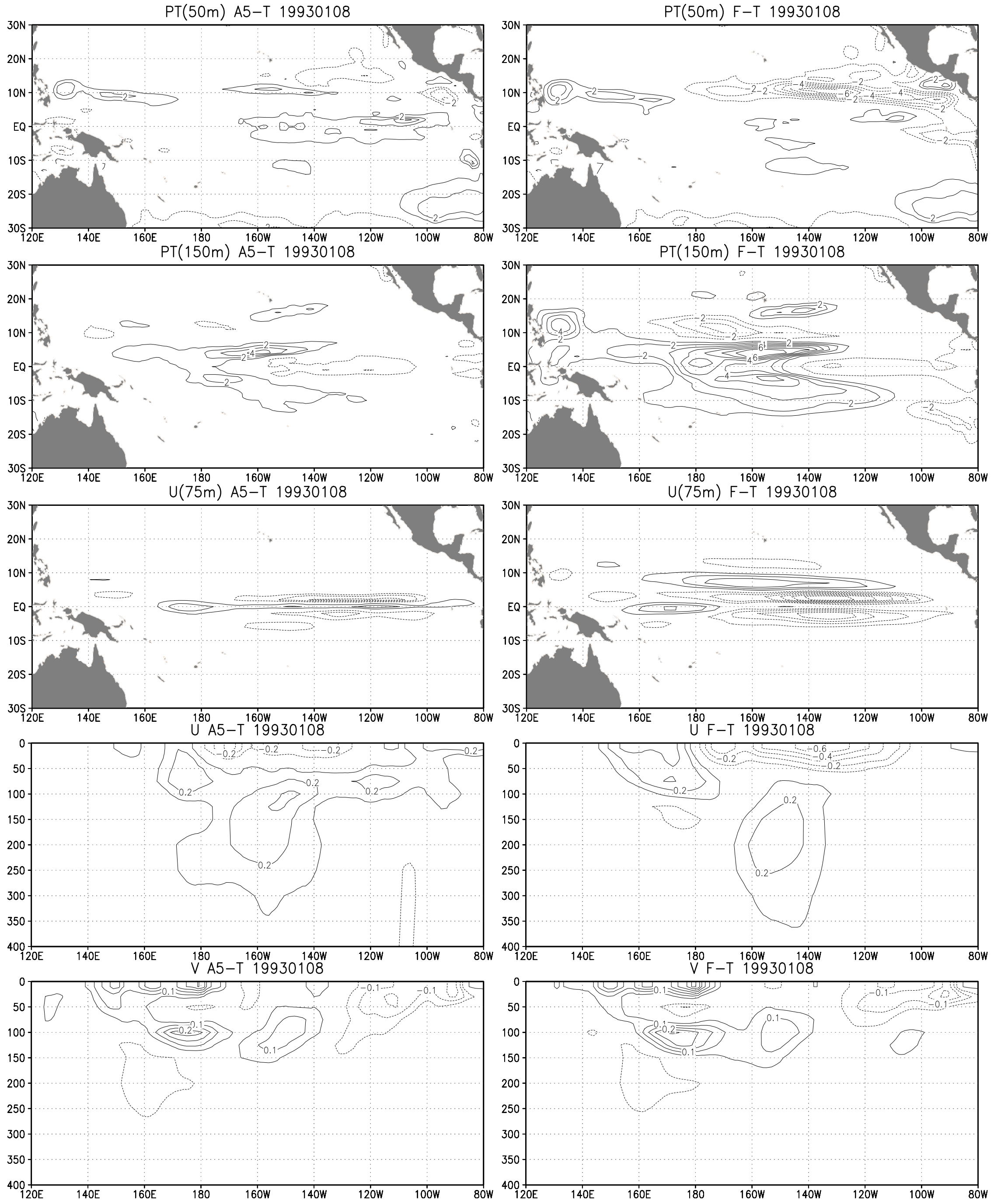


Figure 6: Forecast and analyzed sea level versus the truth after a single assimilation. Positive values are shaded, negative contours are dashed.

assimilation run shows that the 2 solutions have departed significantly as a result of using different forcing fields. A 112-member ensemble was subsequently run, starting again from the same initial state as the background run, but now the best-guess (ERA40) forcing fields were perturbed using the EOF method. A single assimilation step was performed with both the traditional stochastic analysis algorithm (A2) and a version of the deterministic square-root algorithm (A5). Local analysis was used, selecting all data within a distance of 15°and 5°in zonal and meridional directions from the analysis point respectively. All calculations were performed on the IBM high-performance cluster at the ECMWF in Reading using 40 processors for the ensemble runs. In Figure 6 the sea level forecast, as well as the resulting analysis from a single assimilation step with A5, are compared with the truth (the figure for A2 is almost identical to that for A5). Since the uncertainty in the forecast is very large (corresponding to a large ensemble spread) at the start of the assimilation run, almost all weight is put on the observations. Despite observation errors of 2.5 to 5 cm, differences between the analysis and the truth are smaller than 2 cm almost everywhere.

The potential value of sea level assimilation for correction of subsurface variables is illustrated by the plots in Figure 7 which again compare the forecast and analysis with the truth. All these analyses were obtained using the full 112-member ensemble. The forecast and analysis for potential temperature are investigated at 50m and 150m depth. The forecast contains errors of up to 7 degrees at both levels, but located in different regions of the domain. The near surface forecast temperatures at the base of the mixed layer are too low in a zonal band lying in the eastern half of the Pacific centered at about 10°N, while forecast temperatures at 150m depth at the base of the thermocline are too high in the central Pacific in a zonal band centered at 5°N, and in a large region south of the equator extending down to 10°S. Nearly depth-uniform positive forecast errors are found just north-east of Mindanao in the western Pacific. All these errors in the temperature field correspond to forecast errors of the corresponding sign in the sea level (compare Figure 6). The assimilation can be seen to have corrected most of the temperature errors at both levels. The errors in the west Pacific are completely corrected at 150m, but only partly at 50m depth. This example illustrates the potential for sea level assimilation to correct for baroclinic error structures.


PT(50m) A5−T 19930108

PT(50m) F−T 19930108

PT(150m) A5−T 19930108

PT(150m) F−T 19930108

U(75m) A5−T 19930108

U(75m) F−T 19930108

U A5−T 19930108

U F−T 19930108

V A5−T 19930108

V F−T 19930108

Figure 7: Comparison of forecast and analyzed subsurface states with the truth after a single assimilation.

The forecast error in the zonal geostrophic current velocity at 75m depth has a banded pattern, reflecting the zonal character of the equatorial current sytem. The southern and northern westward flowing branches of the South Equatorial Current are too strong, while the eastward flowing North Equatorial Counter Current is too strong as well. These errors are corrected to a large extend in the analysis, although the assimilation has lifted the Equatorial Under Current a bit too high, resulting in a positive band of analysis errors along the equator. Since geostrophic balance is not maintained on the equator, the Equatorial Under Current can be adjusted through sea level only by correction of the basin-wide equatorial surface slope, which is difficult with a local analysis scheme. The vertical distribution of velocity corrections on the equator is displayed in Figure 7 as well. Near-surface forecast errors in the zonal current velocity are mostly corrected in the analysis, but a large region of subsurface errors between 75m and 250m depth is merely displaced a few degrees eastward. These results suggest that additional assimilation of temperature profiles may be needed to better correct the subsurface flow field below the mixed layer. More details on the twin expriment can be found in [Leeuwenburgh, in preparation].

8. Future improvements

Some improvements can now be envisioned to the current assimilation system. The use of covariance localisation methods and optimal data selection criteria are discussed by Leeuwenburgh [in preparation]. An additional method that has been proposed in past assimilation studies [e.g. Alves et al., 2001] to reduce the impact of shock-effects after the analysis step is to add the analysis increment gradually over the succeeding integration. A more optimal initial ensemble can possibly be obtained by sampling the dominant eigenvectors from a very large ensemble with optimized conditioning [Evensen, 2004]. This should become possible when a very long control run has been completed. The currently imposed temperature and salinity relaxation can probably be relaxed, and replaced with assimilation of subsurface temperature and salinity profiles and sea surface temperature fields. Ensemble spread can be increased by perturbation of the currently unchanged forcing fields such as precipitation, cloud cover and wind speed using the method described by [Bertino et al., 2003], or alternatively, by additional perturbation of forecast fields, model tendencies, or model parameters such as diffusion and viscosity constants. The magnitude and impact of contributions on different timescales to the total forcing error should be investigated to obtain an optimal error correlation scale. The assimilation of additional observables such as ice cover and thickness can be considered for applications where higher-latitude dynamics become of interest. The validation of the assimilation results should be improved by implementing, in addition to root-mean-square error estimates, some of the verification scores proposed for ensemble forecasting, such as probabilistic skill scores, reliability diagrams, the relative operating characteristic, and rank histograms.

References

1. Alves, J. O. S., K. Haines, and D. L. T. Anderson, Sea level assimilation experiments in the Tropical Pacific, *J. Phys. Oceanogr.*, 31, 305-323.

2. Anderson, J., An ensemble Adjustment Filter for data assimilation, *Mon. Wea. Rev.*, 129, 2884-2903, 2001.

3. Appeldoorn, G., and G. J. van Oldenborgh, 4DVar assimilation of subsurface and altimetry observations in the HOPE OGCM adjusting surface fluxes, unpublished manuscript, 2003.

4. Beckmann, A., R. Döscher, A method for improved representation of dense water spreading over topography in geopotential-coordinate models, *J. Phys Oceanogr.*, 27, 581-591, 1997.

5. Bertino, L., G. Evensen, and H. Wackernagel, Sequential data assimilation techniques in oceanography, *International Statistical Review*, 71, 223-242, 2003.

6. Bierman, G. J., *Factorization methods for discrete discrete sequential estimation*, Academic Press, 241 pp, 1977

7. Bishop, C., B. Etherton, and S. Mujamdar, S., Adaptive sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical aspects, *Mon. Wea. Rev.*, 129, 420-436, 2001.

8. Bonekamp, H., G. J. van Oldenborgh, and G. Burgers, Variational assimilation of TAO and XBT data in the HOPE OGCM, adjusting the surface fluxes in the tropical ocean, *Geophys. Res.*, C106, 16693-16709, 2001.

9. Burgers, G., P. J. van Leeuwen, and G. Evensen, Analysis scheme in the Ensemble Kalman Filter, *Mon. Wea. Rev.*, 126, 1719-1724, 1998.

10. Cleveland, W. S., and S. J. Devlin, Locally weighted regression; An approach to regression analysis by local fitting, *J. Am. Stat. Assoc.*, 83, 596-610, 1988

11. Derber, J., and A. Rosati, A global oceanic data assimilation system, *J. Phys Oceanogr.*, 19, 1333-1347, 1989.

12. Ducet, N., P. Y. Le Traon, and G. Reverdin, Global high-resolution mapping of the ocean circulation from TOPEX/POSEIDON and ERS-1/2, *J. Geophys. Res.*, 105, 19477-19498, 2000.

13. Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res*, C99, 10143-10162, 1994.

14. Evensen, G., The ensemble Kalman Filter: theoretical formulation and practical implementation, *Ocean Dyn.*, 53, 343-367, DOI 10.1007/210236-003-0036-9, 2003.

15. Evensen, G., Sampling strategies and square root analysis schemes for the EnKF, *submitted to Ocean Dyn.*, 2004.

16. Gayler, V. Documentation of the Standard Ocean Model PostProcessor version 2.0beta, Max Planck Institut fuer Meteorologie, Hamburg, September 2001.

17. Gaspari, G., and S. E. Cohn, Construction of correlation functions in two and three dimensions, *Q. J. R. Meteorol. Soc.*, 125, 723-757, 1999.

18. Gent, P. R., J. Willebrand, T. McDougall, and J. C. McWilliams, Parameterizing eddy induced tracer transports in ocean circulatin models, *J. Phys Oceanogr.*, 25, 463-474.

19. Griffies, S. M., The Gent-McWilliams skew flux, *J. Phys Oceanogr.*, 28, 831-841, 1998.

20. Heemink, A. W., M. Verlaan, and A. J. Segers, Variance reduced Ensemble Kalman filtering, *Mon. Wea. Rev.*, 129, 1718-1728, 2001.

21. Hibler, W. D., A dynamic thermodynamic sea ice model, *J. Phys Oceanogr.*, 9, 815-846, 1979.

22. Houtekamer, P. , and H. Mitchell, A sequential Ensemble Kalman Filter for atmospheric data assimilation, *Mon. Wea. Rev.*, 129, 123-137, 2001.

23. Ingleby, B., and M. Huddleston, Ocean profile processing and quality control, Dataset version 1.0 Document version 1.0, UK Met Office, 2003.

24. Jones, P. W., A user's guide for SCRIP: A Sperical Coordinate Remapping and Interpolation Package, version 1.4, Los Alamos National Laboratory, University of California, 1997.

25. Kalnay, E. et al., The NCEP/NCAR 40 -year reanalysis project. *Bull. Amer. Meteorol. Soc.*, 77, 437-471, 1996.

26. Kepert, J. D., Comment on "The Ensemble Kalman Filter: theoretical formulation and practical implementation", *submitted to Ocean Dyn.*, 2004.

27. Keppenne, C. L., and M. M. Rienecker, Initial testing of a massively parallel Ensemble Kalman Filter with the Poseidon isopycnal ocean general circulation model, *Mon. Wea. Rev.*, 130, 2951-2965, 2002.

28. Leeuwenburgh, O., Assimilation of along-track altimetry into an ocean model ensemble, in preparation.

29. Leeuwenburgh, O., and G. Burgers, Data assimilation with the Ensemble Kalman Filter in forced and coupled versions of a linear ENSO model, *unpublished manuscript*, 2004.

30. Marsland, S., H. Haak, J. H. Jungclaus, M. Latif, and F. Röske, The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates, *Ocean Modelling*,5, 91-127, 2003.

31. Paluskiewicz, T., and R. D. Romea, A one-dimensional model for the parameterization of deep convection in the ocean, *Dyn. Atmos. Oceans*, 26, 95-130, 1997.

32. Redi, M. H., Oceanic isopycnal mixing by coordinate rotation, *J. Phys Oceanogr.*, 12, 1154-1158, 1982.

33. Rew, R. G. Davis, S. Emmerson, and H. Davies, NetCDF user's guide for FORTRAN, Unidata Program Center, University Corporation for Atmospheric Research, Boulder Colorado, 1997.

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

35. Robert, C., and O. Alves, Tropical Pacific ocean model error covariances from Monte-Carlo simulations, manuscript, 2003.

36. Schlax, M. G., and D. B. Chelton, Frequency domain diagnostics for linear smoothers, *J. Amer. Stat. Assoc.*, 87, 1070-1081, 1992.

37. Troccoli, A., and P. Kållberg, Precipitation correction in the ERA-40 reanalysis, ERA-40 Project Report Series No. 13, ECMWF, 2004.

38. van Oldenborgh, G. J., G. Burgers, S. Venzke, C. Eckert, and R. Giering, Tracking down the ENSO delayed oscillator with an adjoint OGCM, *Mon. Wea. Rev.*, 127, 1477-1496, 1999.

39. Waszkiewitz, J., P. Lenzen, and N. Gillet, The PINGO package(Procedural INterface for Grib formatted Objects) Version 1.1, Technical Report. No. 11, Deutsches Klimarechenzentrum, Hamburg, December 1996.

40. Weaver, A., and P. Courtier, Correlation modelling on the sphere using a generalized diffusion equation, *Q. J. R. Meteorol. Soc.*, 127, 1815-1846, 2001.

41. Whitaker, J., and T. Hamill, Ensemble data assimilation without perturbed observations, *Mon. Wea. Rev.*, 130, 1913-1924, 2002.

Appendix A: List of auxiliary routines

name	use	function
`grib2ext.job`	`grib2ext.job`	convert GRIB to EXTRA
`forcing.F`	`forcing` *days*	interpolate Gaussian grid to MPI-OM grid
`lev2t43.F`	`lev2t43` *days*	interpolate Levitus grid to MPI-OM grid
`getdecorxyt.f`	`getdecorxyt`	compute decorrelation scales
`eof5t43.f`	`eof5t43`	combined EOF decomposition
`convol.f`	`call convol`*(pert,sig,ran,mtd,nx,ny)*	2D random fields by convolution/diffusion methods
`calcdate.f`	`calcdate` *date days*	compute calendar dates
`julianday.f90`	`julianday` *date*	convert to Julian days
`get_restart.f90`	`get_restart` *ensemble restart member*	extract restart from ensemble file
`add_restart.f90`	`add_restart` *restart ensemble member*	add restart to ensemble
`estats.f90`	`estats` *dir N days date*	ensemble statistics
`readsun_res.f`	`readsun_res < readsun_res.in`	extract altimetry data
`prep_altimetry.f90`	`prep_altimetry` *data_file*	prepare EnKF input
`prep_insitu.f90`	`prep_insitu` *type date1 date2*	read insitu data and prepare EnKF input
`make_table.f90`	`make_table`	grid search table
`ncinter.F` (main program)	`interpol.x` (requires `interpol_in`)	interpolation to common grid
`nchopeC.F` (idem)	`nchopeC.x4.x` (requires `inp.files`)	store output in NetCDF

Table 1: Auxiliary routines that were used in this report with the MPI-OM model and the EnKF package.

Appendix B: Description of the control script

The assimilation is controlled by the job script `enkf.job`. The header of the script contains the LoadLeveler instructions:

```
#-------------------------------------------------#
#                 QUEUE OPTIONS                   #
#-------------------------------------------------#
#
#@ shell            = /bin/ksh
#@ job_type         = serial
#@ class            = ns
#@ job_name         = enkf
#@ output           = $(job_name).out
#@ error            = $(job_name).out
#@ notification     = error
#@ resources        = ConsumableCpus(1)
                      ConsumableMemory(900mb)
#@ cpu_limit        = 2:00:00
#@ wall_clock_limit = 2:00:00
#@ data_limit       = unlimited
#@ stack_limit      = unlimited
#@ queue
```

Apart from running the model ensemble, all parts of the assimilation cycle are performed sequentially on a single processor. The analysis step itself could also be done in parallel at some point.

The parameters that control the assimilation cycle are set next, starting with the model run. Begin and end date can be explicitly specified, or, if the length of the model run `NDAYS` is greater than zero, the end date will be determined from the starting date, the length of the model runs, and the number of assimilation steps `NSTEPS`. The ensemble size is set with `NMEMB`. At the moment, the number of nodes required to run this size ensemble must still be set manually further on in the script. One can also specify the forcing type (ERA40 or climatology). If restart files are available, `RESTART` should be set to `true`. Individual parts of the script (such as running the model, assembling the output from the ensemble members, data selection, analysis, and post-processing) can also be run seperately, by setting the appropriate switches to `true` or `false`.

```
#-------------------------------------------------#
#              CONTROL PARAMETERS                 #
#                                                 #
#------------------- Model -----------------------#

MODEL=mpiom.x

BGNDATE=19920901
ENDDATE=19920930
NSTEPS=1
NDAYS=30
NMEMB=64
FTYPE=ERA40
RESTART=.true.
RUNMOD=.true.
WAIT=.false.
NSLP=60
DSLP=120
ASSEMBLE=.false.
```

The type of observation to be assimilated can be indicated next. For sea level anomalies (SLA) the cutoff wavelength of the low-pass filter is currently calculated by `readsun_res.f` as `SAMP` times the local grid size. `WDAYS` is the half-width of the time window around the analysis time from which observations are used. `LAT` and `LON` can be used to indicate the geographical region from which data are to be selected.

```
#--------------- Observations --------------------#

SLA=.false.
SPAN=100.0
SAMP=2

TEM=.false.
SAL=.false.

WDAYS=2.5
LAT1=-80.0
LAT2=+80.0
LON1=0.0
LON2=360.0
```

If the local analysis scheme is used, the parameters `RADIUS` and `MAXOBS` indicated the influence radius around the analysis point and the maximum number of nearest observations respectively.

```
#----------------- Analysis ----------------------#

ANALYZE=.false.

LOCAL=.true.
RADIUS=3000000.0
MAXOBS=1000

#-------------- Post-processing ------------------#

POST=.false.
```

If the model has to be run (`RUNMOD=.true.`), the appropriate forcing files will first need to be copied from ECFS. Once this has been done, they will only need to be updated when a new calendar year is entered. The script uses the file `forcing.id` to check if the forcing files that are present correspond to the current model year. The appropriate Reynolds SST file is copied to `SURTEM`, the field used for relaxation of surface temperature.

```
#--------------------------------------------------#
#        COPY FORCING FILES FROM ECFS              #
#--------------------------------------------------#

FID=notdefined

if [ ! -d ${FORCEDIR} ]; then
  mkdir -p ${FORCEDIR}
fi

cd ${FORCEDIR}

if [ $FTYPE = CLIM ]; then
  if [ $YEAR -eq $BGNYEAR ]; then
    if [ -f forcing.id ]; then
      FID=‘cat forcing.id‘ ; export FID
    fi
    if [ $FID != climatology ]; then
      ecp ec:/nlp/forcing_ecmwf/era40_clim.tar ./
      tar xvf era40_clim.tar
      rm -f era40_clim.tar
      mv GISST SURTEM
      echo climatology > forcing.id
    fi
  fi
elif [ $FTYPE = ERA40 ]; then
  if [ -f forcing.id ]; then
    FID=‘cat forcing.id‘ ; export FID
  fi
  if [ $FID != $YEAR ]; then
    ecp ec:/nlp/forcing_ecmwf/era40_${YEAR}.tar ./
    tar xvf era40_${YEAR}.tar
    rm -f era40_${YEAR}.tar
    echo $YEAR > forcing.id
    ecp ec:/nlp/oisst_v2/OISST_V2.${YEAR} SURTEM
  fi
else
  echo " ** no correct forcing type defined ** "
  exit
fi
```

The ensemble can be run in multiple steps, in case only a certain limited number of nodes can be claimed. For example, a 128 member ensemble can be run by sequentially running two 64 member ensembles. In the script this is achieved as a loop over `NRUNS=2` steps.

```
NRUNS=1
IMEMB1=1
IMEMB2=$NMEMB
if [ $NMEMB -gt 64 ]; then
  NRUNS=2
  IMEMB2=64
fi

IRUN=1
while [ $IRUN -le $NRUNS ]; do
#--------------------------------------------------#
#         RUN ENSEMBLE IN NRUNS STEPS              #
#--------------------------------------------------#
NENS=‘expr \$IMEMB2 - \$IMEMB1 + 1‘
NODES=‘expr \$NENS \/ 8  + 1‘
if [ ‘expr $NODES \* 8 - \$NENS‘ -eq 8 ]; then
  NODES=‘expr \$NENS \/ 8 ‘
fi
echo " First and last member, nodes: ",$IMEMB1,...
       $IMEMB2, $NODES
```

The model ensemble is run in Multiple Programs Multiple Data (MPMD) parallel mode (see e.g. the manuals *RS/6000 SP: Practical MPI Programming*, p.137, and *IBM Parallel Environment for AIX, Operation and Use, Volume 1*, p.30). LoadLeveler and command instructions for the parallel task are produced by the script in the part reproduced below. The command file `ensemble.cmd` contains the actual commands to be executed in parallel (run all `NMEMB` models).

```
hlp="#@"
cat > ${EXPDIR}/ensemble.job << EOF
## AIX: script to run ensemble in MPMD parallel mode
$hlp shell        = /bin/ksh
$hlp account_no   = spnlocda
$hlp error        = ensemble.out
$hlp output       = ensemble.out
$hlp notification = error
$hlp class        = np
$hlp job_type     = parallel
$hlp resources    = ConsumableCpus(1)
                    ConsumableMemory(400Mb)
$hlp node         = ${NODES}
$hlp total_tasks  = ${NENS}
$hlp cpu_limit    = 0:40:00
$hlp wall_clock_limit =0:40:00
$hlp queue

poe -pgmmodel mpmd -cmdfile ensemble.cmd

EOF
```

The parameters for the model run are written to the file `OCECTL`. Most values have been chosen based on experience with the model at the MPIfM in Hamburg. This includes values for the eddy diffusivity and viscosity of $1 \cdot 10^{-2}$. The relaxation parameters `CRELSAL` and `CRELTEM` were decided on by the ENACT project. The parameter `GAMD` sets the temporal correlation scale (in days) of eddy diffusivity and viscosity perturbations and is not used when it's value is less than 0. `DZW` contains the layer thicknesses of the model.

```
#--------------------------------------------------#
#              WRITE OCECTL FILE                   #
#--------------------------------------------------#

cat > OCECTL  << EOF
```

```
 &OCECTL
 DT      = 2160.,
 CAULAPTS= 0.,
 CAULAPUV= 0.004,
 CAH00   = 1000.,
 AUS     = 0.,
 AV0     = 1.E-2,
 DV0     = 1.E-2,
 CWT     = 5.E-4,
 CSTABEPS= 0.030,
 DBACK   = 1.E-5,
 ABACK   = 1.E-4,
 CRELSAL = 1.06E-8,
 CRELTEM = 2.5E-6,
 CDVOCON = 0.10,
 IMEAN   = 1,
 LD_START= ${START},
 LD_END  = ${END},
 EXPTID  = 'MPIOM_CS1',
 IENS    = ${IMEMB},
 GAM     = 7.0,
 GAMD    = -1.0  /

 &OCEDZW
 DZW = 20.,20.,25.,25.,25.,25.,25.,30.,45.,60.,90.,
 120.,150.,180.,210.,250.,300.,400.,500.,600.,700.,
 900.,1400. /

EOF
```

During a production run, the sequential tasks should proceed only after a the model forecast has been completed. In the following part it is checked if the ensemble run has finished by counting the number of restart files produced. If not all members have finished yet, the command 'sleep DSLP' is executed where DLSP is the number of seconds specified in the control part of the script. This process is repeated a maximum number of NSLP times.

```
#--------------------------------------------------#
#         MONITOR PROGRESS ENSEMBLE RUNS           #
#--------------------------------------------------#

MCOUNT=0
TCOUNT=0

while [ $MCOUNT -lt $NENS ]; do
  MCOUNT=0
  IMEMB=$IMEMB1
  while [ $IMEMB -le $IMEMB2 ]; do
    if [ $IMEMB -lt 10 ]; then
      MID=00${IMEMB}
    elif [ $IMEMB -lt 100 ]; then
      MID=0${IMEMB}
    else
      MID=${IMEMB}
    fi
    if [ -s ${EXPDIR}/tmp${MID}/restart${END}.uf ];
    then
      MCOUNT=‘expr $MCOUNT + 1‘
    fi
    IMEMB=‘expr $IMEMB + 1‘
  done
  if [ $MCOUNT -lt $NENS ]; then
    echo "restarts not ready yet. only " $MCOUNT
         " members for now"
    echo "going to sleep for 1 minute"
    TCOUNT=‘expr $TCOUNT + 1‘
    if [ $TCOUNT -gt $NSLP ]; then
      echo "waited " $TCOUNT "steps, stop the job"
      exit
    else
      sleep $DSLP
    fi
  fi
done
```

If the auxiliary files such as depths.uf are not yet present, they are copied from ECFS. If the analysis time falls within the time limits of phases C or G of the ERS-1 mission, or within those of the ERS-2 or TOPEX/POSEIDON missions, the appropriate raw data files are also copied from ECFS. Using the corresponding Julian day, the required data are retrieved from the files and used as input for prep_altimetry which creates the file observations.uf, used by the EnKF. The analysis step is subsequently started if ANALYSIS has been set to true.

```
#--------------------------------------------------#
#                 ALTIMETRY DATA                   #
#--------------------------------------------------#

if [ ! -d $OBSDIR ]; then
  mkdir -p $OBSDIR
fi
cd $OBSDIR

ecp ec:/nlp/observations/depths.uf ./
ecp ec:/nlp/observations/mbathy.uf ./
ecp ec:/nlp/observations/levels.uf ./
ecp ec:/nlp/observations/newpos.uf ./
ecp ec:/nlp/observations/latlon.table ./
ecp ec:/nlp/observations/dlxyp.uf ./

ERS=.false.
TPX=.false.
rm -f *.dat *.out observations.uf fort.99

if [ $END -ge 19921006 ] && [ $END -le 19931223 ];
then
 if [ ! -s altimetry.ers1c ]; then
 ecp ec:/nlp/observations/ ...
 sla_ers1_phasec_005_018_xxc.bin ./altimetry.ers1c
 fi
 ERS=.true.
 ERSFILE=altimetry.ers1c
fi
if [ $END -ge 19950324 ] && [ $END -le 19950514 ];
```

```
then
 if [ ! -s altimetry.ers1g ]; then
 ecp ec:/nlp/observations/ ...
 sla_ers1_phaseg_031_032_xxc.bin ./altimetry.ers1g
 fi
 ERS=.true.
 ERSFILE=altimetry.ers1g
fi
if [ $END -ge 19950516 ] && [ $END -le 20020304 ];
then
 if [ ! -s altimetry.ers2 ]; then
 ecp ec:/nlp/observations/ ...
 sla_ers2_001_071_xxc.bin ./altimetry.ers2
 fi
 ERS=.true.
 ERSFILE=altimetry.ers2
fi
if [ $END -ge 19921004 ] && [ $END -le 20020404 ];
then
 if [ ! -s altimetry.tpx ]; then
 ecp ec:/nlp/observations/sla_tp_002_351_xxc.bin
 ./altimetry.tpx
 fi
 TPX=.true.
fi

JDAY=$(${BINDIR}/julianday $END 1)
JDAY=${JDAY}.0

if [ $ERS = .true. ]; then
  echo $ERSFILE > readsun_res.in
  echo ers.dat >> readsun_res.in
  echo $JDAY >> readsun_res.in
  echo $WDAYS >> readsun_res.in
  echo $LAT1 >> readsun_res.in
  echo $LAT2 >> readsun_res.in
  echo $LON1 >> readsun_res.in
  echo $LON2 >> readsun_res.in
  echo $SPAN >> readsun_res.in
  echo $SAMP >> readsun_res.in
  ${BINDIR}/readsun_res < readsun_res.in > ...
  readsun_res.out

  ${BINDIR}/prep_altimetry ers.dat
  cat fort.99 > prep_altimetry.out
# echo "number of ERS measurements = \c"
# cat prep_altimetry.out | wc -w
fi
if [ $TPX = .true. ]; then
  echo altimetry.tpx > readsun_res.in
  echo tpx.dat >> readsun_res.in
  echo $JDAY >> readsun_res.in
  echo $WDAYS >> readsun_res.in
  echo $LAT1 >> readsun_res.in
  echo $LAT2 >> readsun_res.in
  echo $LON1 >> readsun_res.in
  echo $LON2 >> readsun_res.in
  echo $SPAN >> readsun_res.in
  echo $SAMP >> readsun_res.in
  ${BINDIR}/readsun_res < readsun_res.in >> ...
  readsun_res.out
  ${BINDIR}/prep_altimetry tpx.dat
  cat fort.99 >> prep_altimetry.out
# echo "total number of measurements = \c"
# cat prep_altimetry.out | wc -w
fi

#cat > prep_altimetry.in << EOF
#  &input
#  sla_name='ers.out' /
#EOF

mv observations.uf ${ENKFDIR}/
cp prep_altimetry.out ${ENKFDIR}/
if [ -s prep_altimetry.out ]; then
  mv prep_altimetry.out ssh.${END}
fi

if [ $ANALYZE = .true. ]; then
cd $ENKFDIR
if [ -s observations.uf ]; then
  ${ENKFSRC}/EnKF
  mv ensembleF.uf forecast.${END}
  mv ensembleA.uf analysis.${END}
else
  mv ensembleF.uf forecast.${END}
  cp forecast.${END} analysis.${END}
fi
fi
```

A similar procedure as for the altimetry follows for the selection and pre-processing of insitu data. Since these data are stored in monthly files, data may need to be extracted from two files for assimilation times near the beginning or end of a calendar month. Data files that are no longer needed are removed.

Post-processing proceeds in two phases. First, the EXTRA formatted analysis output is interpolated in the vertical to the standard output levels and stored in NetCDF format, using the program `nchopeC.x4.x`. The resulting file is the input for the SCRIP-driven interpolation to the regular common output grid with `interpol.x`.

```
#-------------------------------------------------#
#               POST_PROCESSING                   #
#                                                 #
#---------- put output in NETCDF file ------------#

ecp ec:/nlp/files/weto.ext4 ${FILEDIR}/
ecp ec:/nlp/files/gila.ext4 ${FILEDIR}/
ecp ec:/nlp/files/giph.ext4 ${FILEDIR}/
ecp ec:/nlp/files/amsuo.ext4 ${FILEDIR}/
ecp ec:/nlp/files/amsue.ext4 ${FILEDIR}/

cd $ENKFDIR
cat > inp.files << EOF
  &input_files
  expt_title='${EXPID}_${END}'
  gila='${FILEDIR}/gila.ext4'
  giph=’${FILEDIR}/giph.ext4’
  weto=’${FILEDIR}/weto.ext4’
  amsuo=’${FILEDIR}/amsuo.ext4’
  amsue=’${FILEDIR}/amsue.ext4’
  toffsetyr=$YEAR
  outputfile=’${ENKFDIR}/analysis${END}.nc’
  NTIM=1
  NVAR=8
  file_list(1)=’mean_tho.ext4’
  file_list(2)=’mean_sao.ext4’
  file_list(3)=’mean_uko.ext4’
  file_list(4)=’mean_vke.ext4’
  file_list(5)=’mean_wo.ext4’
  file_list(6)=’mean_zo.ext4’
  file_list(7)=’mean_taux.ext4’
  file_list(8)=’mean_tauy.ext4’
  var_list(1)=’PT’
  var_list(2)=’S’
  var_list(3)=’U’
  var_list(4)=’V’
  var_list(5)=’W’
  var_list(6)=’SL’
  var_list(7)=’taux’
  var_list(8)=’tauy’ /
EOF

cp ${POSTDIR}/nchopeC.x4.x .
nchopeC.x4.x

if [ -s analysis${END}.nc ]; then
  rm -f mean_*.ext4
fi
```

```

#-------- interpolation to regular grid ----------#

cat > varlist_in << EOF
&variable_list
  varlist(1)=’PT’
  varlist(2)=’S’
  varlist(3)=’U’
  varlist(4)=’V’
  varlist(5)=’W’
  varlist(6)=’SL’
  varlist(7)=’taux’
  varlist(8)=’tauy’ /
EOF

cat > interpol_in << EOF
&interpol_inputs
    datafile_in=’${ENKFDIR}/analysis${END}.nc’
    scripfile=’${SCRIPDIR}/ ...
    rmp_hopeCht43_to_enact_conserv.nc’
    datafile_out=’${ENKFDIR}/analysis_reg${END}.nc’
    maskfile=’mask.nc’
    sectionfile=’section.nc’
    gradientfile=’gradient.nc’
    order=’second’
    maskcalc=.TRUE.
    showgrad=.FALSE.  /
EOF

cp ${STOMPPDIR}/interpol.x .
interpol.x
```