ABSTRACT

Title of Document: DATA ASSIMILATION OF THE GLOBAL OCEAN USING THE 4D LOCAL ENSEMBLE TRANSFORM KALMAN FILTER (4D-LETKF) AND THE MODULAR OCEAN MODEL (MOM2)

Stephen G Penny, Doctor of Philosophy, 2011

Directed By:

Dr. Eugenia Kalnay, Distinguished University Professor, Department of Atmospheric and Oceanic Science

Dr. James Carton, Department Chairman, Department of Atmospheric and Oceanic Science

The 4D Local Ensemble Transform Kalman Filter (4D-LETKF), originally designed for atmospheric applications, has been adapted and applied to the Geophysical Fluid Dynamics Laboratory's (GFDL) Modular Ocean Model (MOM2). This new ocean assimilation system provides an estimation of the evolving errors in the global oceanic domain for all state variables. Multiple configurations of LETKF have been designed to manage observation coverage that is sparse relative to the model resolution. An Optimal Interpolation (OI) method, implemented using the Simple Ocean Data Assimilation (SODA) system, has also been applied to MOM2 for use as a benchmark. Retrospective 7-year analyses using the two systems are compared for validation. The oceanic 4D-LETKF assimilation system is demonstrated to be an effective method for data assimilation of the global ocean as determined by comparisons of global and regional 'observation minus forecast' RMS, as well as comparisons with temperature/salinity relationships and independent observations of altimetry and velocity.

DATA ASSIMILATION OF THE GLOBAL OCEAN USING THE 4D LOCAL ENSEMBLE TRANSFORM KALMAN FILTER (4D-LETKF) AND THE MODULAR OCEAN MODEL (MOM2)

By

Stephen G Penny

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Distinguished University Professor Eugenia Kalnay, Co-Chair
Professor and Department Chairman James Carton, Co-Chair
Assistant Professor Kayo Ide
Associate Professor Brian Hunt
Research Assistant Professor Takemasa Miyoshi
Assistant Research Scientist Gennady Chepurin
Associate Professor and Dean's Representative Ning Zeng

Dedication

I dedicate this to the many great teachers I have had in my lifetime. Namely, G. Edgar Parker, James Milbrand, Lynn Stewart Fox, Dave Pruett, Paul Warne, Lynn S. Fichter, Fred Bennett, Jim Carton and Eugenia Kalnay. I have also learned much from professional colleagues and would like to dedicate this in part to them, particularly Jimmy Krozel, Brian Capozzi, Robert Hoffman, and my mentor in the NASA LARSS program, Lawrence L. Green. I'd like to thank the James Madison University Mathematics Department and the NASA LARSS program for inspiring me to pursue a graduate career. And finally, I would like to dedicate this to my parents Greg and Roberta Penny and my brother Jonathan Penny who have always been there for me.

Acknowledgements

I would like to thank Professors Eugenia Kalnay and Kayo Ide for their guidance and assistance with the data assimilation methods, Professor James Carton for guidance in coupling LETKF to a global ocean model, Takemasa Miyoshi for providing the base atmospheric LETKF code with adaptive inflation and for assistance in implementing LETKF on the global ocean model, Gennady Chepurin for assistance in implementing SODA on the global ocean model and guidance on general ocean assimilation, Matthew Hoffman for guidance relating to the assimilation specifically of the Modular Ocean Model MOM2, Professors Brian Hunt and Dianne O'Leary for guidance relating to the theoretical and computational developments of the LETKF algorithm, and Dr. Warren Wiscombe and Professor David Levermore for guidance early on in my graduate career.

Table of Contents

List of Figures

Chapter 1: Introduction

Two data assimilation systems have been implemented, each utilizing the Geophysical Fluid Dynamics Laboratory's (GFDL) Modular Ocean Model (MOM2) on a 360 x 130 x 20 grid. The first is the Simple Ocean Data Assimilation (SODA) system, developed by Carton et al [CCC00]. The second is the 4-Dimensional Local Ensemble Transform Kalman Filter (4D-LETKF) system, developed by Hunt et al [HKS06]. SODA is a state-of-the-art ocean assimilation system, utilizing an Optimal Interpolation (OI) statistical scheme for assimilating historical observations. 4D-LETKF is a next-generation system that utilizes a type of 4D Ensemble Kalman Filter (EnKF). Originally developed for atmospheric applications, here the LETKF system has been adapted for use with the ocean and is now capable of assimilating all historical oceanic temperature, salinity, and ocean current data on record.

A framework mirroring the SODA design was used to build the 4D-LETKF ocean system. This resulted in specifications such as a multi-day analysis cycle, with super-observations collected in 1x1 degree bins. Two methods for performing the analysis updates were designed. First, rather than applying the full adjustment to the initial conditions of the model at each analysis cycle, a scheme used by SODA for incremental analysis updates (IAU) [BT96] was used with LETKF (called LETKF-IAU). IAU was achieved by applying the analysis innovation in small increments at each model integration time step. Second, an iterative dual-pass procedure called Running-in-Place (RIP) was used to re-center the background ensemble and generate a more accurate analysis (called LETKF-RIP). A less computationally intensive

approximation of RIP is called the Quasi-Outer Loop (QOL); both are discussed in [KY08].

This LETKF implementation also includes an adaptive inflation scheme developed by Miyoshi [M11] to adjust inflation to automatically balance the background covariance with the estimated observation error. This is especially useful to prevent a growing ensemble spread in unobserved regions of the ocean that results from the use of a constant inflation parameter. It also helps to account for deficiencies in estimates of the observation errors. Additional considerations for the oceanic application included perturbed wind forcing fields to ensure a non-collapsing ensemble spread, and accounting for landmasses that obstruct the domain of the ocean analysis.

The two assimilation systems are compared, showing the benefit of the next-generation 4D-LETKF system over the current state-of-the-art SODA system. Experiments showing the performance of the two systems during a 7-year period are used to demonstrate this benefit. The key result is an improvement in 4D-LETKF over SODA, particularly in the metric relating the distance of the forecast system states to the observed states, both globally and in specified regions. Further verification against independent observations and theoretical temperature/salinity relationships are given as well.

Data assimilation is the name given in modern geoscience applications to recursive Bayesian estimation, a general approach for estimating an unknown probability density function recursively over time using a combination of measurements and a mathematical model. The true system state is assumed to be an

unobserved Markov process, and the measurements are the observed states of a hidden Markov model.

The procedure for data assimilation is to perform analysis cycles in which past and present observations of the system state are combined with a forecast from the mathematical model to generate a 'best estimate' of the current system state. This 'best estimate' is called the *analysis*. The analysis is then propagated forward in time by the mathematical model, and the result is used as the forecast for the next cycle. Because this forecast is a priori information at the beginning of each analysis cycle, it is also known as the *background*.

The analysis cycle is an application of Bayes theorem. In a theoretical framework, the Fokker-Planck[1] equation would be used to advance the distributions represented by the background and analysis, but to do so in practice is computationally infeasible. If the probability distributions are assumed normal, they can be simplified by representing the distribution parametrically with the mean and covariance. A recursive Bayesian filter that assumes such multivariate normal distributions is known as a Kalman filter. This simplifying assumption is not sufficient for practical applications. It is also not feasible to maintain the true covariance due to the large number of degrees of freedom in the state space of most realistic environmental models. Therefore, these methods utilize various approximations for the covariance.

Data assimilation is most often used for forecasting geophysical systems, particularly weather forecasting and hydrology, and correspondingly the application

[1] The Fokker-Planck equation is discussed in the Appendix

given here is of Numerical Weather Prediction (NWP). However, the applications are far reaching to areas such as tomography (e.g. cloud imaging, medical imaging), trajectory estimation (e.g. for NASA's Apollo program), GPS applications, atmospheric chemistry, air traffic management, mathematical finance, artificial intelligence, and path planning.

1.1 General Problem Statement

Find the trajectory of a dynamical system that best fits a time series of data, given a model for the time evolution of the system and observations of the system state at various times. Specifically, given the ordinary differential equation (ODE), $\frac{d\boldsymbol{x}}{dt} = F(t, \boldsymbol{x})$, where $\boldsymbol{x}$ is an m-dimensional vector representing the system state at time t, and a set of l observations made at times $t=t_1,t_2,t_3,t_4,\ldots$ find the trajectory $\boldsymbol{x}(t)$ that best fits the observations.

There is a further complication that errors exist in both the model and the observations. It is desired to account for as much of the existing error as possible. Typically observation errors come from both instrument error as well as subgrid-scale variability that is not represented in the grid-average values of the model, called the 'error of representativeness' [K03]. The background errors are caused in part by model error, which encompasses errors in the model formulation, the model forcing, model resolution, and numerical rounding errors. These errors are not well quantified, but must be accounted for within any solution approach.

1.2 Reanalysis of the Global Ocean

A retrospective historical data assimilation is typically termed 'reanalysis'. Such reanalysis efforts have been performed for the global ocean, e.g. by Carton et al [CCC00a],[CCC00b], with several others described in [CS08]. The observability condition states that it is possible to uniquely infer the state of a dynamical system from measurements of its inputs and outputs. The primary challenge in performing reanalysis using the historical ocean record is that the observational data sets are extremely sparse. Obviously with sufficient observations the observability condition would be satisfied and the state of the ocean could be determined with great accuracy. But, there is no guarantee that the ocean system satisfies this condition within the timeframe of the historical record, given these limited observations. A second difficulty is that the ocean is highly influenced by surface forcing such as surface wind stress, radiative heat flux, precipitation and evaporation. In stand-alone ocean models, these forcing terms are incorporated as inputs to the model. Thus near the surface, the ocean state will constantly be forced back toward prescribed values, regardless of changes made via data assimilation with a non-coupled ocean model.

The current variety of data assimilation methods used to estimate the physical state of the global ocean (particularly temperature, salinity, currents, sea level) is evident in the nine reanalysis efforts spanning multiple decades as described in [CS08]. Among the nine there are three alternative state estimation approaches. The first approach is the 'no-model' analyses, where temperature or salinity observations update a background provided by monthly climatological estimates. The second approach is a sequential data assimilation analysis, which iterates forward in time from a previous analysis using a numerical simulation of the evolving ocean state

produced by an ocean general circulation model. The third approach is 4D-Var which, in ECCO, uses the initial conditions and surface forcing as control variables to be modified in order to be consistent with the observations as well as a numerical representation of the equations of motion through iterative solution of a single large optimization problem.

Of the cases enumerated by [CS08], Ishii and Levitus began with a first guess of the climatological monthly upper-ocean temperature based on climatologies produced by the NOAA National Oceanographic Data Center. The innovations were mapped onto the analysis levels. Ishii used an alternative 3D-Var approach to do an objective mapping with a smaller decorrelation scale in midlatitudes (300 km) that elongated in the zonal direction by a factor of 3 at equatorial latitudes (to 900 km). Levitus began similarly to Ishii, but used the technique of Cressman and Barnes with a homogeneous scale of 555 km to map the temperature innovation onto a uniform grid.

The sequential approaches can be further divided into those using Optimal Interpolation and the Kalman Filter, and the variational approaches as those using 3D-Var or 4D-Var. Among the nine approaches, INGV and SODA [http://www.atmos.umd.edu/~ocean/] used versions of Optimal Interpolation, while CERFACS, GODAS [http://www.cpc.ncep.noaa.gov/products/GODAS/], and GFDL all used 3D-Var. [CS08]

The ECCO series of assimilation approaches include: ECCO1, ECCO-GODAE [http://www.usgodae.org/], GECCO, ECCO-JPL, ECCO-SIO, ECCO2 [http://ecco2.org/], and OCCA. The ECCO Near Real-Time Ocean Analysis of

Fukumori and Lee and JPL/Caltech is based on the MIT general circulation model. German ECCO (GECCO) is based at the University of Hamburg's Institut fuer Meereskunde. GECCO applied 4D-Var to the decadal ocean estimation problem to cover the full 50-year NCEP/NCAR re-analysis period. This approach provided some benefits including satisfying some conservation laws and the construction of the ocean model adjoint. ECCO2 is a high-resolution global-ocean and sea-ice data synthesis that intends to produce an analysis of all available global-scale ocean and sea-ice data at resolutions that resolve ocean eddies and narrow current systems. [http://www.ecco-group.org/about.htm] JAMSTEC has a 4D-Var data assimilation system developed by the MEXT K7 project using a 1x1º version of the MOM3 ocean model dynamics. [http://www.jamstec.go.jp/e/medid/dias/kadai/clm/clm_or.html]

Operational forecasting systems include the Forecasting Ocean Assimilation Model (FOAM) from the Met Office. The FOAM data assimilation method is based on a version of the analysis/correction scheme devised by [LB91]. This system produces 5-day forecasts and assimilates temperature profile data, altimetry data and surface temperature data on a global 1º grid with 20 vertical levels. High-resolution model configurations are nested inside the global configuration: the Atlantic and Arctic Oceans and the Indian Ocean use 35 km grids; and the North Atlantic, the Mediterranean Sea and the Arabian Sea use 12 km grids. [http://research.metoffice.gov.uk/research/ncof/foam/system.html]

The ECMWF reanalysis system is based on the HOPE-OI scheme. The background field is given by forcing the Hamburg Ocean Primitive Equations

(HOPE) ocean model with daily fluxes of momentum, heat, and fresh water. The observations are assimilated using an Optimal Interpolation (OI) scheme.

Ensemble methods have been applied to the Parallel Ocean Program (POP) ocean model by researchers at NCAR [http://www.image.ucar.edu/DAReS/Presentations/]. In their assimilation, they encountered problems with an under-represented ensemble spread. They initially used identical forcing for each ensemble member, but have since adapted the forcing to include influence from an atmospheric ensemble. [http://www.image.ucar.edu/pub/DART/2010/2010_CESM_Breck_TJH.pdf]

Keppenne et al have developed an Ensemble Kalman Filter and applied it to the Poseidon model using operational data sources of temperature and SSH. [KR08] They use the Bloom [BT96] method of IAU on a uniform 576x538x27 grid. They assumed 0.5ºC error for temperature profiles and 1.4 cm error for SSH. An observation window of 6-days is used for an analysis cycle of 4-days. A thorough review of ensemble analysis methods prior to 2003 is given Evensen in [E03].

Chapter 2: Developing the LETKF System for the Global Ocean

2.1 Motivation

Main Points:

- Naive application of the LETKF algorithm to the ocean model is inadequate to generate a reliable and accurate reanalysis. The ensemble spread collapses for fixed inflation values that are too small, or diverges for fixed inflation values that are too large. Spinup is a problem, especially due to the sparseness of available observations.
- The sparse observation network presents difficulties not typically encountered in synoptic scale atmospheric data assimilation. (As identified by [LKM08], analysis sensitivity to observations increases with lower observation coverage)
- Unlike atmospheric models, ocean models are highly influenced by surface forcing. Thus, adequate representation of forcing and forcing errors is necessary. Land/sea boundaries cause additional complications.
- Depending on the desired application, different configurations of the LETKF assimilation system may be desirable (IAU for smooth reanalysis and model bias correction, RIP for near-term forecasting, QOL for forecasting with computational constraints, Hybrid method for balancing the benefits of all methods and exploring greater solution space)

The main purpose here is to determine a practical framework for implementing an oceanic data assimilation scheme that is easy to implement, is computationally efficient, and that scales well to high dimensional systems (on the order of 10^7) and a wide range of observations (e.g. from 10^2 to 10^5). The sequential

Ensemble Kalman filter (EnKF) is the general approach used. The Local Ensemble Transform Kalman Filter (LETKF) is the specific variant proposed [HKS07], which borrows ideas from the Local Ensemble Kalman Filter (LEKF) of [OH04] and the Ensemble Transform Kalman Filter (ETKF) of [BEM01].

Various spaces are defined: the model space, with dimension *m*; the observation space, with dimension *l*; the ensemble space with dimension *k*. Generally $m >> l >> k$, but this need not be the case. A mapping *H* is defined from the model space to the observations space so that operations may be performed in the observation space. Here *H* is an interpolation operator, but it may also perform physical transformations to convert model parameters to observed quantities. Thus, it may be linear or non-linear, depending on the application.

Using a maximum likelihood approach, a cost function is generated for the background model state (represented by $\boldsymbol{x}^{\mathrm{b}}$) and the current observations (represented by $\mathbf{y}^{\mathrm{o}}$). Namely,

$$J(x) = \left[x - \bar{x}^b\right]^T \left(P^b\right)^{-1} \left[x - \bar{x}^b\right] + \left[y^o - H(x)\right]^T R^{-1} \left[y^o - H(x)\right]. \qquad (1)$$

$\boldsymbol{P}^b$ and $\boldsymbol{R}$ are the covariance matrices for the background and observations, respectively. Observations are assumed to have random errors with zero mean, therefore if $y^o = H\left(x(t)\right) + \varepsilon$, where ε is a Gaussian random variable, then $\boldsymbol{R}$ is the covariance matrix associated with ε, defined $R = E(\varepsilon\varepsilon^T)$. And we assume that the outer product of the perturbations of the ensemble of forecasts about its mean approximates the error covariance matrix of the state estimate. That is, $P^b = (k-1)^{-1} X^b \left(X^b\right)^T$, where the columns of $\boldsymbol{X}^{\mathrm{b}}$ are the perturbations from the

background mean $\bar{x}^b$. The covariance matrices effectively weight the background and observations based on the estimated uncertainty in each.

The desired output of this data assimilation procedure is an analysis ensemble, $\left\{x^{a(i)} : i = 1,2,...,k\right\}$ with the associated mean, $\bar{x}^a = k^{-1}\sum_{i=1}^{k} x^{a(i)}$, and analysis covariance, $P^a = (k-1)^{-1} X^a \left(X^a\right)^T$, where again the columns of $\boldsymbol{X}^a$ are the perturbations from the mean $\bar{x}^a$.

2.2 Local Ensemble Transform Kalman Filter

The method used to determine the analysis ensemble differentiates a number of the data assimilation methods. LETKF is one of a series of methods classified as ensemble square root filters that use deterministic algorithms to generate an analysis ensemble with the required sample mean and covariance. The key advancement offered by LETKF is the combination of localization and transformation applied to the data assimilation process. (Refer to the summary in Chapter 1)

Because there are typically many more state variables than there are ensemble members (by multiple orders of magnitude), the ensemble covariance in this form is rank deficient and has large terms for pairs of points that are spatially distant. Since the values of physical fields at distant locations should not be significantly correlated, the covariance matrix is weighted with decreasing significance as distance from each grid point increases, which gives rise to localized EnKF algorithms. Some of these methods modify the covariance matrix used in the computations, others (e.g., LEKF), perform the analysis locally in space. As a result, the analysis ensemble is no longer

made of only a linear combination of the background ensemble, but instead comprises many different linear combinations of small portions of the background ensemble.

To see the need for transformation, note that by definition $P^b = (k-1)^{-1} X^b (X^b)^T$ has rank at most *k*-1 and is therefore not invertible. However because it is a symmetric matrix, it is 1-to-1 on its column space *S* [HKS06]. *S* is also the column space of $\boldsymbol{X}^b$, the space spanned by the background ensemble perturbations from the mean. However, $\boldsymbol{X}^b$ has been defined so that the sum of the columns is zero (recall it is a set of Gaussian perturbations with mean 0). One can regard $\mathbf{X}^b$ as a linear transformation from a *k*-dimensional space S` onto S. Thus if $\boldsymbol{w}$ is a vector in S`, then $\boldsymbol{X}^b\boldsymbol{w}$ is in S and $x = \bar{x}^b + X^b w$ is the corresponding vector in model space. Therefore, the analysis can be performed in the space S` and transformed back to the space S to determine the appropriate analysis ensemble.

In ensemble methods, the dimension of the solution is limited to the size *k* of the ensemble. Localization improves the situation. "By allowing the local analysis to choose different linear combinations of the ensemble members in different regions, the global analysis is not confined to the *k*-dimensional ensemble space and instead explores a much higher dimensional space." [HKS07] It further reduces correlations between distant locations that are most likely spurious.

EnKF methods tend to underestimate uncertainty in the model, partially due to unaccounted-for model error. As this phenomenon occurs recursively, over time the analysis under-weights the observations. Thus, an ad hoc method of inflation is often used in practice on the background covariance. Current implementations of LETKF

use a multiplicative inflation approach, though [Ka09] has also used additive inflation.

For operational-scale atmospheric applications, the 4D-LETKF algorithm has been implemented with the National Center for Environmental Prediction's (NCEP) Global Forecast System (GFS) model by [SKG07]. This implementation was validated against NCEP's implementation of 3D-Var, called Spectral Statistical Interpolation (SSI). Observations were assimilated every 6 hours on a 192 x 94 x 28 grid (about 500,000 grid points). Local patches of as high as 7 grid points were used. Experiments were run on a Beowulf cluster of 40 3.6 GHz Intel Xeon processors. The assimilation of a typical observational data set (about 330,000 observations) took about 9-10 minutes, 14-16 minutes, and 24-27 minutes for a 40, 60, and 80-member ensemble, respectively. A later implementation was designed and applied to the Earth Simulator AGCM by [MY07] and with the Japan Meteorological Agency (JMA) global model by [MS07] and [MSK10]. This implementation contains numerous advancements, including a distance-based Gaussian localization function [MYE07] and an adaptive inflation procedure [M11], and thus was selected for this study.

2.3 Ocean

The Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model (MOM) is a 3D primitive equation model that was based on work done since the late 1960's, and was first released in 1990 as MOM 1. Improvements were made on the design of this system and it was released in 1995 as MOM2. MOM2 has four prognostic (predicted) variables calculated at all levels: temperature, salinity, and velocity in the zonal and meridional directions. All other state variables are determined from these

prognostic variables, thus these are the primary variables of interest when working with the time evolution of the model. MOM version 2.4 was used for all simulation experiments, utilizing the model parameters as defined in [Carton et al 2000a].

MOM2 uses finite differencing applied to the primitive equations governing ocean circulation: the Navier-Stokes equations using the Boussineq (replacing mean ocean density profile $\rho_0(z)$ by the vertically averaged value $\rho_0 = 1.035$ g/cm^3), hydrostatic (implying the vertical pressure gradients are due to density variations alone, i.e. $\Delta p = \rho\ g\ \Delta z$) and rigid lid approximations (to filter out external gravity waves, though there is an option to solve with a free-surface boundary, or coupled with an atmospheric model). The thin shell approximation is also made. The Coriolis and viscous terms involving vertical velocity in the horizontal momentum equations are ignored. [P96]

The continuous equations used by MOM2 in spherical coordinates are:

$$\boldsymbol{u_t + L(u) - \frac{uv\,tan\phi}{a} - fv = -\frac{1}{\rho_o a \cdot cos\,\phi} p_\lambda + (\kappa_m u_z)_z + F^u} \qquad (2)$$

$$\boldsymbol{v_t + L(v) + \frac{u^2 tan\phi}{a} + fu = -\frac{1}{\rho_o a} p_\phi + (\kappa_m v_z)_z + F^v} \qquad (3)$$

$$T_t + L(T) = (\kappa_h \cdot T_z)_z + \nabla \cdot (A_h \nabla T) \qquad (4)$$

$$S_t + L(S) = (\kappa_h \cdot S_z)_z + \nabla \cdot (A_h \nabla S) \qquad (5)$$

$$w_z = -\frac{1}{a \cdot cos\,\phi}\left(u_\lambda + (cos\phi \cdot v)_\phi\right) \qquad (6)$$

$$p_z = -\rho g \qquad (7)$$

$$\rho = \rho(T, S, p) \qquad (8)$$

where T is potential temperature and S is salinity, u is zonal velocity, v is meridional velocity, w is vertical velocity, p is pressure, ρ is the potential density, g is the mean gravity (980.6 cm/s^2), a is the mean radius of the Earth (6370 x 10^5 cm), κ_m is vertical eddy viscosity (cm^2/s), κ_h is diffusivity coefficient (cm^2/s), A_m is horizontal eddy viscosity (cm^2/s), A_h is horizontal diffusivity coefficient (cm^2/s), F is horizontal friction, and L is advection.

Boundary conditions are supplied at the ocean surface for heat, salt and momentum. At the sides the boundary conditions are set as no-slip no-flux for heat or salt. At the ocean bottom there is a no-flux condition for heat and salt, while boundary conditions on the horizontal velocity may either be free-slip or linear bottom drag. If the model is started from scratch, then initial conditions are required to specify a density structure via potential temperature and salinity with zero velocity (i.e. ocean at rest). Potential temperature T is the temperature of a parcel of water at the sea surface after it has been raised adiabatically from some depth in the ocean. [S05]. Potential temperature is used for two reasons: (1) local stability is dependent on potential density gradients, and (2) both salinity and potential temperature are materially conserved tracers in an adiabatic ocean.

Time integration in MOM is done with the leapfrog scheme but uses an occasional Forward Euler step to eliminate the computational mode. [http://www.ocean-modeling.org/docs.php?page=GFDL-MOM] A restart procedure allows the model to be stopped at each time step, output all prognostic oceanic state parameters, and restart using this data file combined with a Forward Euler finite differencing scheme. While the model uses a 3-timestep scheme, only 2 timesteps are

stored in the restart file. To perform a traditional update of the initial conditions, the two timesteps are differenced, the new model state is updated in the most recent timestep, and the difference is added back on to determine the older timestep. This means that the tendencies are not analyzed. An additional I/O functionality was added to allow a steady forcing term added onto the prognostic equations in order to implement the Incremental Analysis Update (IAU) of [BT96].

The key fields of interest are the prognostic variables. All diagnostic variables are derived from these. Prognostic fields available in the restart option include the following fields for two consecutive model timesteps:

- Temperature (T) in degrees Celsius
- Salinity (S) as a deviation from 0.035 g/cm^3
- Zonal velocity (u) in cm/s
- Meridional velocity (v) in cm/s

MOM2 utilizes a 'memory window' to reduce the use of volatile memory during runtime. This approach places much less restriction on the actual grid size computable with the model. The data for each grid point is read into memory along rows of latitude, by default 3 rows at a time (a larger window may be selected if desired). These rows are then used to calculate central differencing formulas, and from that point, a single row can be read into memory at each new time step, replacing one of the oldest of the previous rows in memory while the other two are shifted in the time index.

Much of the ocean data necessary for computation is read in from databases containing historical data. These include ocean bottom topography, surface wind

information, sea surface temperature, boundary sponge layers, and a variety of other fields.

The MOM2 implementation used in this study uses a 130x362x20 grid with periodic boundary conditions, using an approximate 1º horizontal resolution with higher horizontal resolution near the equator and 15m vertical resolution near the surface. The longitude grid overlaps at the endpoints. Levitus climatology [LB94] of temperature and salinity is used to form a sponge layer on the northern and southern boundaries at about 60ºN and 60ºS. Bottom topography is included. Monthly average surface winds from NCEP reanalysis [K96] centered at the middle of each month were linearly interpolated to the model time.

The main observation datasets prior to the introduction of the Argo floats in 1999 are profile measurements of temperature from mechanical bathythermographs (MBTs), expendable bathythermographs (XBTs), conductivity-temperature-depth devices (CTDs), measurements from thermistors, reversing thermometers and salinity from CTD and station measurements obtained from the World Ocean Database 2005 [http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html]. The XBTs infer depth and drop rate, which make them subject to random systematic errors that increase with depth. [CCC00a]. Additional data includes moored thermistor chains, stations, and ship intake temperatures. A historical record of subsurface temperatures from these sources is provided by [CCC00a] and is reproduced here in Figure 1. Sea surface temperature data were not used in order to reduce the computational costs of the many experiments that were run, while still providing an adequate comparison of the various assimilation approaches.

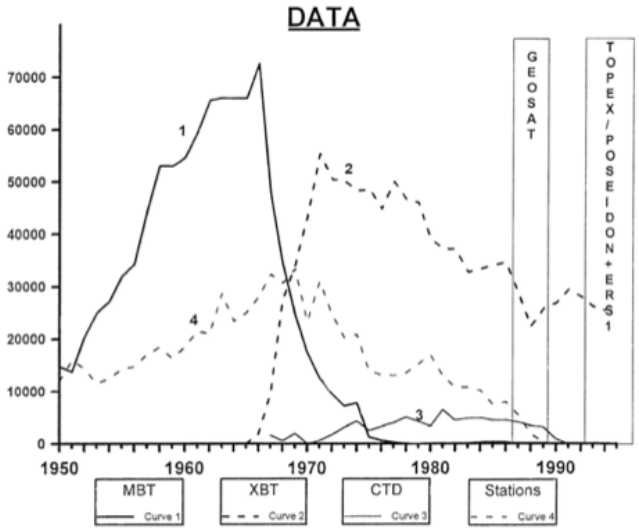

Figure 1. Distribution of combined subsurface temperature observations at 75 m. [CCC00a]

The Argo float network (**Figure 2**) is a globally distributed array of floats that profile temperature and salinity in the upper 2,000 m of the ocean at approximate 10-day intervals. Between 2000 and 2007 the network was brought up to about 3,000 operational free-drifting floats. [http://www.argo.ucsd.edu/] "The float program and its data management system began with regional arrays in 1999, scaled up to global deployments by 2004, and achieved its target of 3000 active instruments in 2007." [RJR09]

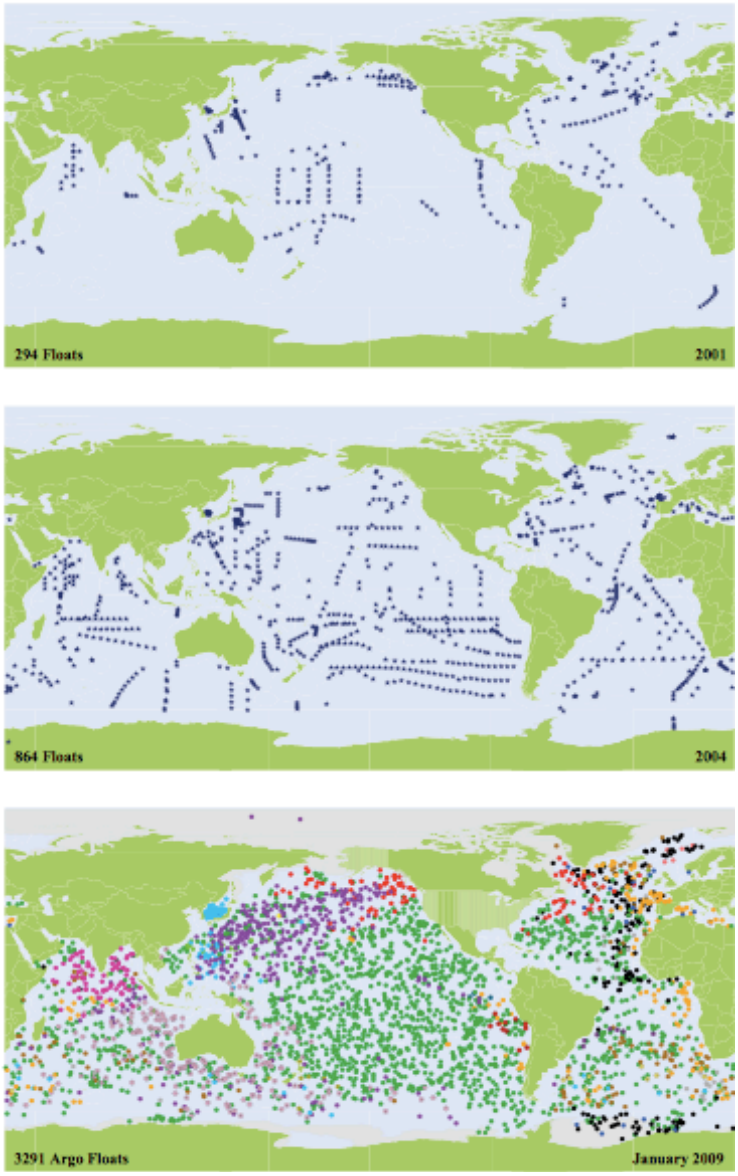

Figure 2. Historical Argo coverage in the years 2001, 2004 and 2009. The colors in the Jan. 2009 plot indicate the various source countries of the floats. [RJR09]

Satellite altimeter sea level is available continuously since 1991 from a number of satellites (T/P, ERS1/2 and Jason). Monthly combined, gridded estimates of sea level obtained from the French Aviso were kept out of the assimilation procedure so that they could be used for validation of the assimilation experiments. The introduction of these satellite data sources is depicted in Figure 1. An example of instantaneous altimetry sea surface height (SSH) data is shown in Figure 3.

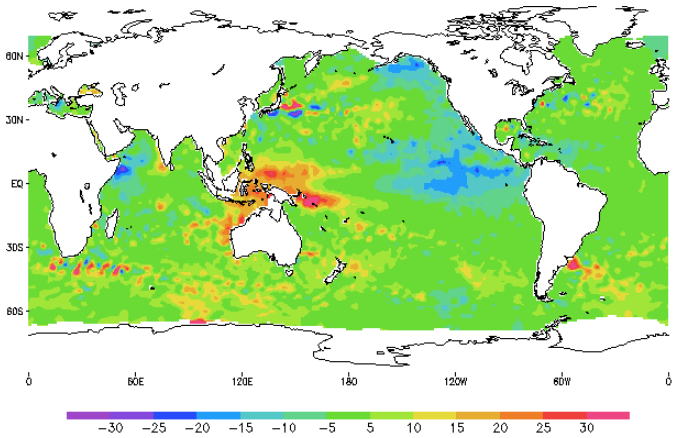

Figure 3. Average Topex altimetry SSH anomaly for January 2000.

2.4 Implementation of the Data Assimilation System

Implementation of the oceanic LETKF data assimilation system was performed in three steps: design, verification and validation, and historical reanalysis. This chapter will focus on the steps required to create a working data assimilation system with LETKF – primarily the design, verification and validation stages. Historical reanalysis is the primary goal of this research, and that discussion will be presented in Chapter 3.

The LETKF data assimilation system is composed of multiple components that interact with each other throughout the analysis cycle. These components include the LETKF System, which prepares and runs the core LETKF algorithm, the ocean model, the observation network, an initialization routine, and an overall control script.

The LETKF System is the encompassing code that prepares all parameters and input data in the format required by the LETKF Core algorithm. The LETKF system was designed and coded by Miyoshi [http://code.google.com/p/miyoshi/]. The model-specific components of the interaction between the model and assimilation algorithm are located here. Many of the adaptations required to use LETKF with an ocean model were performed at this level. Some of the changes included converting

to the MOM2 model grid, changing the vertical coordinates from pressure to fixed depth, designing new filters for observation data (e.g. removing points in lakes and unmodeled ocean areas, removing points outside of the range 60ºS to 60ºN, transferring quality control to a preprocessing step), customizing the localization scheme, outputting additional diagnostic metrics, bypassing computations over land grid points, adapting interpolation procedures.

The LETKF Core contains the primary algorithm as described in [HKS06] and coded by Miyoshi, applied to a model grid point after localization has been performed. Modifications to the LETKF Core algorithm have been researched and are discussed in Chapter 4. However, for this reanalysis effort the LETKF Core has been used without modifications.

The observing network is provided by historical data as a compilation of ship track, XBTs, CTD's, Argo Floats, and satellite data compiled by [CCC00b]. As was performed for SODA, the data is averaged into 1x1 degree bins and vertically interpolated to the MOM2 model levels. Prior to use by LETKF, an estimated observation error is assigned to each individual observation. The data are prepared for input into the LETKF System's required input format. The observation error profiles are discussed in section 2.5.7.

The MOM2 model is used for forward iteration of the system state and is applied one or more times in each analysis cycle, depending on the approach used. Typically, the model is applied to the output ensemble generated by LETKF, and each member is run forward in parallel to generate a forecast that can be compared with the observations during the next 4D-LETKF analysis cycle.

To generate an initial ensemble for the data assimilation system, the model is run from climatological temperature and salinity initial conditions (zero velocity) starting January 1, 1970 using NCEP derived surface boundary conditions and a climatological sponge layer at 60ºN and 60ºS. This run also served as a baseline 'Free-Run' for comparison with assimilated results.

An initialization routine was used to create customized initial ensemble sets. The initial ensemble was created as a linear combination of a base Free-Run corresponding with the experiment start date and various Free-Run run data from past historical dates. Specifically, starting one year prior to the experiment start year, a selection of days was taken at the base day (e.g. Jan. 1), and 25 days before and after the base day. Further ensemble members were added by stepping backward one year at a time until all members were generated for the required ensemble size. The proportion of each historical day used in the linear combination was parameterized as

$$(1 - \alpha_s)\boldsymbol{x}_0 + \alpha_s \boldsymbol{x}_i, \; i = 1, \ldots, k \qquad (9)$$

where $\boldsymbol{x}_0$ is the model state on the base date, $\boldsymbol{x}_i$ is the perturbation state of ensemble member i, and α_s is the proportion of the perturbation state data to use. Though various values were tried, the experiments listed here used $\alpha_s = 0.5$. The initialization routine also prepared the surface wind forcing data for the ensemble members.

The control script was used to initiate and run the data assimilation system. It specified the basic parameters of the analysis experiment, managed all file manipulation and transfer between components, executed all parallel processing, and allowed for halting and restarting of the system during extended experiment runs. The basic process is described in **Figure 4**.

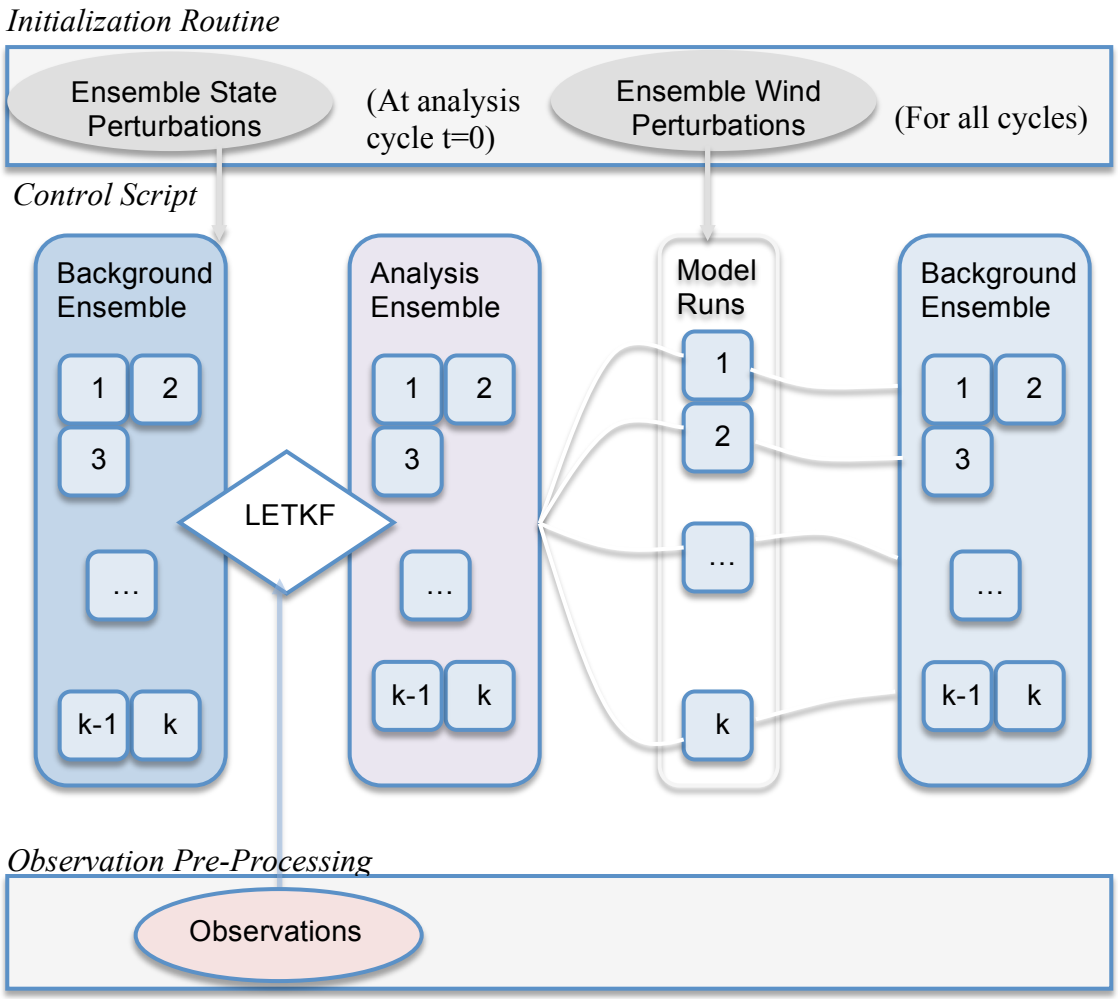

Figure 4. Design diagram of the Data Assimilation System, depicting one analysis cycle of the basic LETKF procedure.

2.5 *Adaptations for the Ocean Environment*

A number of modifications were made to the baseline LETKF implementation to test the benefit of their use in assimilating the global ocean with sparse observation coverage. For the analysis cycle, both the widely used meteorological approach of adjusting the initial conditions according to the analysis and an incremental approach (IAU) [BT96] akin to the nudging method were applied. Various update cycle lengths were used: 5-day, 10-day and 30-day analysis cycles. A variety of ensemble sizes were used: 10-members, 20-members, 40-members. In some cases, a rolling window

of observations that extended beyond the analysis cycle was used to provide loose bounds on the analysis solution. The running-in-place (RIP) method [KY08] was utilized in some cases to accelerate convergence of the filter.

Critical to preventing filter divergence in all cases was the implementation of Adaptive Inflation [M11]. Due to the sparse and constantly shifting nature of the oceanic observation network, this method also required a modification that applied relaxation of the inflation parameter in the instances of accelerating growth in the background covariance.

A nearly unmodified version of LETKF was run to demonstrate the difficulties with using the basic atmospheric LETKF approach for the oceanic domain. **Figure 5** illustrates the result. A 10-day, 40-member ensemble with 5% inflation is shown in comparison to the Root Mean Square Deviations (RMSD) for the free run. While improving the results at first, as the ensemble spread collapses the trajectory actually becomes much worse than the free run in tracking the observations. As the observation error is prescribed in advance, a collapsing ensemble spread effectively negates any contribution to the analysis that the observations might have provided.

Thus the following questions arise: How can one maintain the proper balance between the observation error and the background error so that each contributes an appropriate amount to the analysis? Is it possible to outperform the free-run? Is it possible to perform well enough so that the RMSD between the forecast and observations are steadily improving until within the range of the combined errors of the observations and background? Given that the ocean is a heavily forced system, is

it still possible to correct the state in a way that will not be overwhelmed by the effect of the forcing terms? Many of these questions will be answered in the sections that follow.

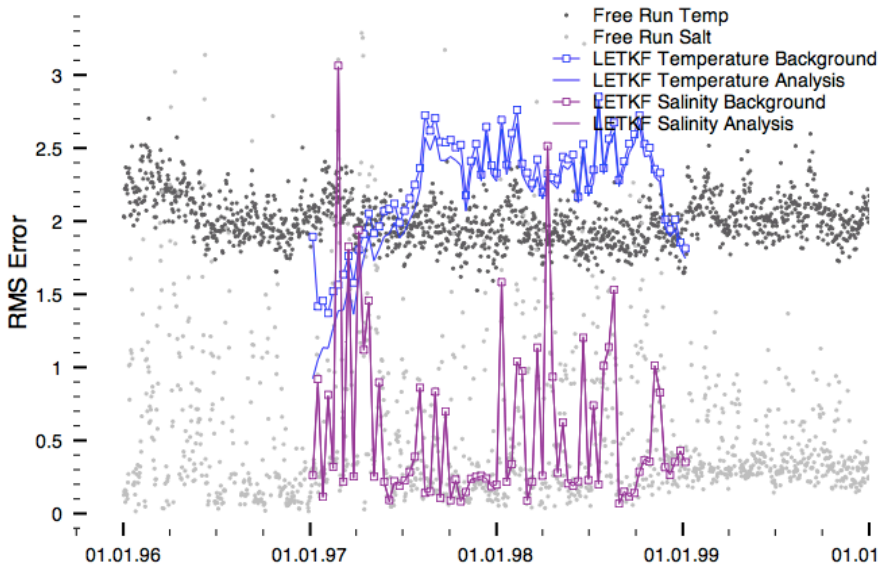

Figure 5. RMS errors for a baseline run of the basic LETKF algorithm blindly applied to the MOM2 model, with only minor adjustments to allow LETKF to run with the ocean environment without model crashes.

2.5.1 Incremental Analysis Update

Applying changes to all prognostic variables in the initial conditions causes temporal discontinuities in the overall historical reanalysis. To avoid this phenomenon, analysis updates were added to the prognostic equations in small increments at each model integration time step. Following the approach of Bloom [BT96], a forecast run is performed and an analysis centered at the middle of he analysis cycle window is computed. Next the analysis increment is divided by {length of analysis cycle in days}*24*60*60 (for days, hours, minutes, seconds) is multiplied by the model time step *dtts (*which equals either 3600 or 7200 seconds) and finally added at each *dtts*-second time step of the MOM2 model.

As discussed in relation to its use by SODA, "This procedure, in effect a form of digital filter, has the advantage of maintaining a nearly geostrophic relationship between the pressure and velocity fields with a minimum excitation of spurious gravity waves. The procedure also reduces bias in the forecast model by 50% relative to the forecast bias when the incremental analysis update procedure is not used." [CG08] Similar effects of bias correction were found in the application of LETKF-IAU.

MOM2 computes transport from one model time step to the next through the combination of diffusion and advection terms for each grid cell *T*,

$$t_{\tau+1} = t_{\tau-1} + 2dt\left[D(T_x) + D(T_y) + D(T_z) - A(T_x) - A(T_y) - A(T_z)\right] \quad (10)$$

where *D*() represent diffusion in the cell, and *A*() represents advection in the cell. The scaled analysis increment *ainc* was added as an additional term to this model integration step,

$$t_{\tau+1} = t_{\tau-1} + 2dt\left[D(Tx) + D(Ty) + D(Tz) - A(Tx) - A(Ty) - A(Tz) + ainc\right], \quad (11)$$

where in order to convert the analysis increment to the timescale of seconds, (where dt = 3600 s)

$$ainc = (\boldsymbol{x}^a - \boldsymbol{x}^b)/(d * 24 * 60^2) \quad (12)$$

Utilizing IAU was also an effective way of accounting for model bias, particularly in the longer analysis cycle (e.g. up to 30-days). A significant hemispherical bias was found to be present in the model forecasts, possibly due to bias in the surface forcing, either having a strong warm bias in the northern hemisphere and cold bias in the southern hemisphere, or vice versa.

2.5.2 Running in Place

As an alternative to the IAU approach, Running-in-Place (RIP) [KY08] was also applied to LETKF, which will be designated LETKF-RIP. This approach more closely mirrors that of the meteorological applications, in that it directly adjusts the initial conditions of the ensemble members. However, the method allows one or more iterations of the analysis cycle over the same time period. It proceeds by applying the weights from the LETKF analysis computed for the end of the analysis cycle to the background at the beginning of the cycle. This is valid if the analysis cycle is short enough to have approximately linear growth of the errors. Because the new background ensemble was influenced by 'future' observations, the background ensemble has thus effectively been re-centered at a more accurate mean state estimate. Small Gaussian noise (calculated as a Gaussian random variable between 0 and 1, scaled by the background mean state multiplied by 10^{-5}) is then added to the new background ensemble members. This has dual benefits of decoupling the background ensemble from the observations that were just used, as well as creating potential for the ensemble to find solutions outside of the linear space spanned by the original ensemble. While this procedure could be iterated many times per cycle, only one additional iteration of the RIP procedure was used for this study.

2.5.3 Managing a Sparse Observation Network

Observations are available on a daily basis, but prior to the introduction of the global Argo observing network the data distribution is very sparse. A representative example of the observation distribution over five days in 1990 is shown in **Figure 6**.

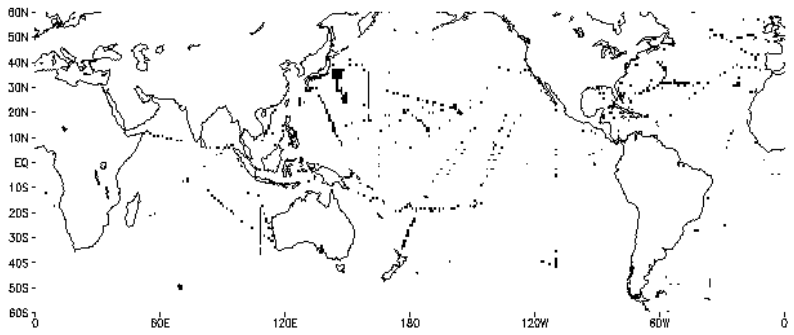

Figure 6. Horizontally binned observed temperature data in a 5-day bin centered at Jan. 1, 1990.

A number of approaches were used to manage this situation temporally. Because SODA used an extended observation window that included observations within +/- 45 days, a similar approach was attempted with LETKF with a 5-day analysis cycle and +/- 25-day window of observations (Figure 7). This resulted in the Extended Observation Window version of LETKF (LETKF-EOW), which is discussed in the appendix. The approach has two drawbacks. First, it reuses the observations numerous times (between 1-11x, depending on quality control measures relating to the distance from the background ensemble mean state to the observed value), which can result in over-fitting the observations and causing a collapse in the ensemble spread. This is not a problem for SODA because its background covariance is constant in time rather than dependent on the analysis. Second, it uses 'future' observations (as does SODA) that occurred after the analysis cycle, which is acceptable for reanalysis but limits the conclusions that can be drawn for using LETKF as a forecasting tool.

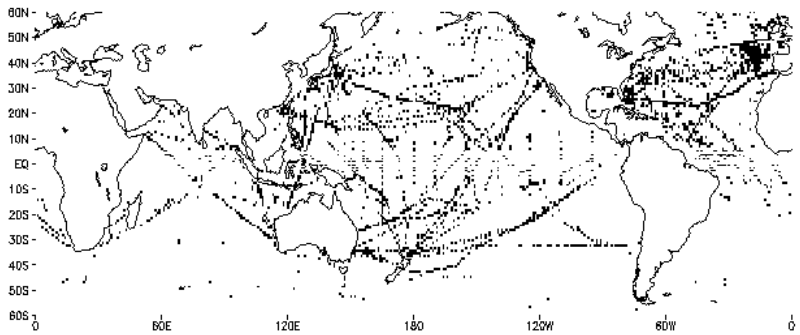

Figure 7. Historical temperature observational coverage, including 5-day analysis cycle and 25-day extended radius before and after the analysis cycle centered on March 14, 2001.

An alternative approach (LETKF-IAU) applied IAU over a longer 30-day analysis cycle window, thus including more observations within the analysis without the potential pitfalls of reusing the same observations across multiple analysis cycles. Finally, the RIP approach was used on a 5-day analysis cycle window. In this approach (LETKF-RIP), each observation is reused once, but random errors are applied to the background ensemble to reduce the chance of over fitting the observations.

Spatially, the localization approach used by LETKF allows the grid points near observations (within a range of the localization radius) to be adjusted toward the observations, with somewhat less impact as the points that are adjacent to the observations (as defined by a Gaussian weighting function). Thus the specific localization approach used plays a large part in how the system utilizes the available observations.

2.5.4 Surface Forcing Fields

Ocean models are highly influenced by forcing that is introduced through the surface and boundaries. This forcing comes in the form of wind stress, radiative heating, and fresh water flux via precipitation and evaporation at the surface. On the boundaries, it

comes through fresh water river deposits from coastal areas and freezing/melting ice in polar regions.

The wind fields demonstrated a significant influence over the evolution of the MOM2 ocean model. This was most evident when observing the perturbed ensemble runs with identical wind fields, in which all ensemble members collapsed toward the nature run within 1 year of simulation time. Thus, an approach was required which would maintain variation in the ensemble members to ensure a non-collapsing ensemble spread.

A collection of wind data was taken from the identical month in randomly selected years, and then iterated forward in time to the end of the experiment period to ensure continuity of the forcing perturbation in time. A percentage of these wind fields was combined with the NCEP reanalysis wind field to create an ensemble of wind fields. This perturbation was added to half of the ensemble members and subtracted from the other half to result in an ensemble mean that was equal to the base nature wind field.

The wind forcing was parameterized as a linear combination similarly to the initial ensemble members,

$$(1 - \alpha_w)\boldsymbol{w}_0 + \alpha_w \boldsymbol{w}_i,\ i = 1, \ldots, k \tag{13}$$

where $\boldsymbol{w}_0$ is wind data from the base year, $\boldsymbol{w}_i$ is the wind forcing of ensemble member i, and α_w is the proportion of the historical perturbation to use. A variety of values were used for α_w, from 0 (no wind perturbations) to 1 (completely different historical winds used for each ensemble member). Large values of α_w caused extreme ensemble spreads near the surface in areas such as the Gulf Stream and Kuroshio Current.

Because the model is sensitive to fluctuations of the surface forcing, and the ensemble was intended to contain slight perturbations of the mean wind field (which was equivalent to the NCEP reanalysis), smaller perturbations were used (α_w=0.1). Additional discussion of this selection is in Section 2.6.

2.5.5 Land/Sea Differentiation

A land/sea map was coded into the LETKF system to ignore non-modeled grid points, including landmasses, lakes, and areas of the ocean not included in the model. There is also the difficulty of applying localization to areas that are not connected water masses. For example, points in the Gulf of Mexico and the Pacific Ocean are usually within the range of the localization radius of grid points off the coast of Mexico. Therefore unless further steps are taken, the background error determined at a grid point in the Gulf of Mexico is affected by the values of grid points and observations in the Pacific, and vice versa.

A basic table-lookup was applied to the Pacific, Gulf of Mexico and Caribbean to identify points that were in separate basins to prevent observations being used from non-local regions. An automated approach that has been developed to apply localization in a general domain with constraints is discussed in Chapter 4. However, this approach has not yet been implemented in the assimilation system.

2.5.6 Extrapolation of Coastal Ocean Data

The three-dimensional interpolation scheme of LETKF caused large errors near the coasts because the land grid points were treated as zero-valued. The simplest solution to this problem would be to drop any observations that occurred at these boundary points. However due to the scarcity of observations, it was preferable to keep as many quality observations as possible. One solution for keeping these observations would be to design an ad hoc interpolation scheme that uses a customized interpolation for each possible configuration of land and ocean points near the coasts. Instead, a general technique was designed to manage these land points, and is described below.

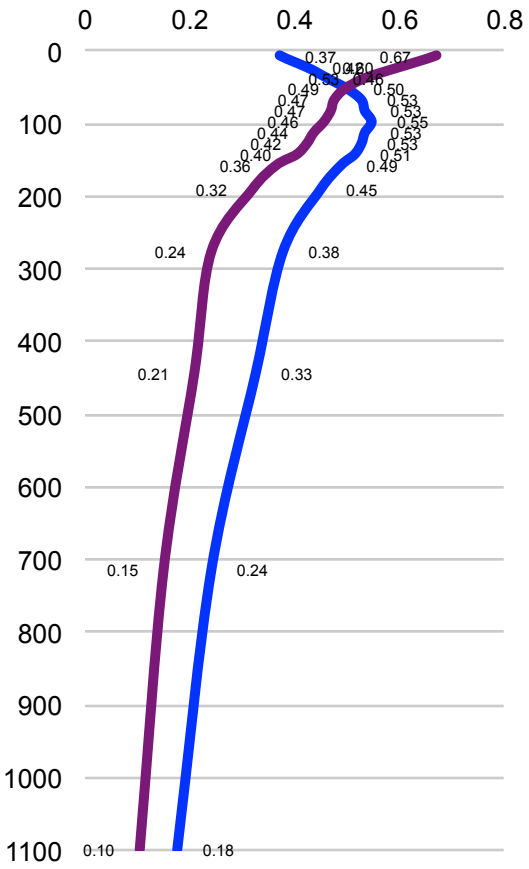

Figure 8. Profiles of scaled standard error used by LETKF for temperature and salinity observations. Values are plotted at model levels up to 1100 m.

The points of the ocean grid and land grid were separated. The ocean grid was then 'shaken' by n gridpoints (for example n=1), in 3 orthogonal directions. In all places where this grid overlapped with the land grid, the values were averaged together, thus creating an extrapolated value on the coastal boundary points. These extrapolated values were then used for interpolation of the model grid to the observation space with the H operator. The pseudo-code is given in the Appendix.

2.5.7 Pre-Processing of Observations

The MOM2 model was used with an approximate grid resolution of 1-degree. However given the sub-grid scale ocean dynamics, patterns in the observational values may not be resolved by the model. In this case, it is preferred to average out these fluctuations so that the model can resolve this generalization of the data. Consequently, the observations were binned and averaged on a 1x1 degree grid.

As input, SODA requires observations on a 1x1 degree grid. Because LETKF interpolates the background state to the observation locations, the binned super-observation approach is not necessary for LETKF to generate an analysis. However reducing the set of observations in a densely covered region to a small set of representative averages, accounting for adjustments in the error characteristics, reduces the computational burden on LETKF.

Observation errors were assigned based on profiles calculated from high-resolution SODA reanalysis, given in **Figure 8**. While instrument error typically increases with depth, mixed layer dynamics often cause a greater degree of representation error that dominates the error characteristics of the observations. The best standard error on modern instrumentation is 0.001º C for temperature and 0.02 psu for salinity [S05]. However, the error profiles calculated from a high-resolution SODA run for temperature super-observations range from 0.25º to 0.85º and for salinity from 0.1 to 1.1 psu, indicating that the representativeness error dominates the error profile.

Because the SODA analysis used observations grouped in 5-day bins, but LETKF used observations binned daily, the prescribed observation error was scaled down (multiplied by 0.625) to account for some reduction in temporal

representativeness error (**Figure 8**). Such a possibility was mentioned in Step 4 of the LETKF algorithm in [HKS06]. This can be thought of as a form of inflation, by increasing the weight of the observations prior to the analysis.

2.5.8 Utilization of Adaptive Inflation for Sparse Observation Networks

Inflation is an empirical approach for countering underestimation of background covariance, a common problem in ensemble data assimilation systems due to the limited sample size of the ensemble, model error, and nonlinearity. If using a perfect model, with a large ensemble size, the forward integration of the computation model should provide enough growing dynamic instability to create the necessary increasing spread in the background ensemble covariance. However, the ocean has lower instability on the larger spatial scales that are captured on the approximately 1x1 degree grid resolution and the 5-10 day analysis cycle within the data assimilation. This is exacerbated by the fact that surface forcing plays a dominant role in the evolution of the ocean state. Furthermore, the highly unstable ocean eddies that dominate the ocean dynamics are not resolved at this grid resolution. [S05]

Due to the sparse distribution of observations in the historical record, the use of a constant inflation parameter would lead to steady growth in the background variance in any region without observation coverage. In the deep ocean, where no observations are present, this would cause unphysically large error and eventual model blow-up (**Figure 9**, Figure 10). In areas of the upper ocean this would cause growing errors in much of the southern hemisphere and some specific areas of the equatorial regions and northern hemisphere, especially prior to the introduction of the Argo network.

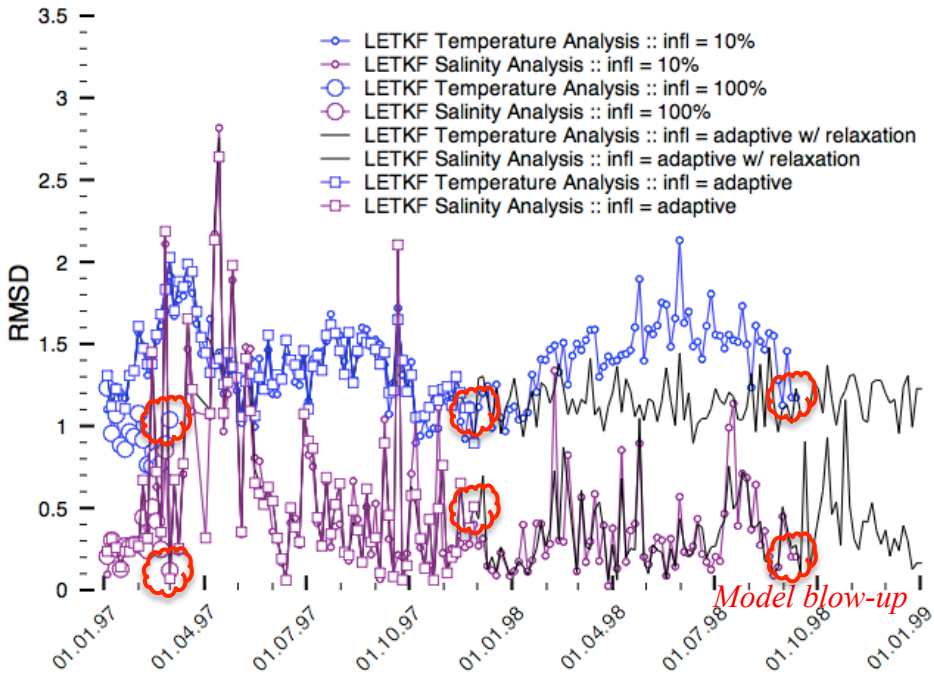

Figure 9. Various levels of inflation for basic LETKF with a 10-member ensemble using 10%, 100% and adaptive inflation initiating at 0%, calculated over the entire domain. For constant inflation parameters, the model blows up after 2 months with 100% inflation. Using 10% inflation merely delays this effect to 21 months. With adaptive inflation the model blows up at 11 months, but with relaxation it is extended indefinitely.

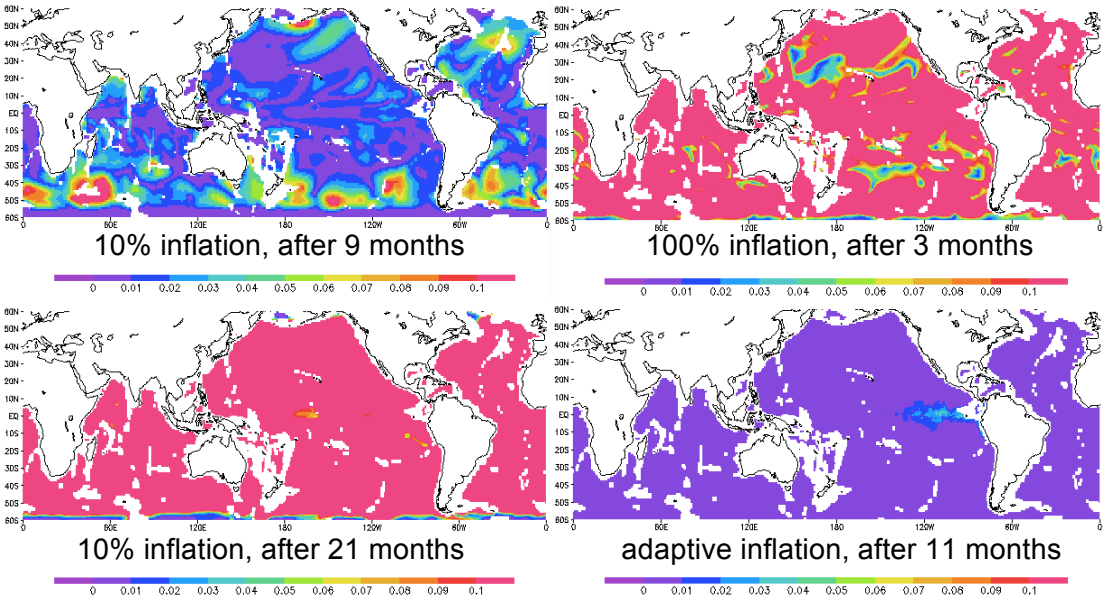

Figure 10. Growth of background spread at 3000 m depth. Shown over 9 months from 1/1/97 to 9/14/97 for 10% inflation and over 3 months (just before model blow-up) from 1/1/97 to 3/6/97 for 100% inflation. Observations are only present to depths of 1000 m, thus the spread will grow unchecked until it is outside of model parameters and causes a model crash. At that rate, the 10% case is expected to cause a model blow-up after about 18 months.

An adaptive inflation procedure, developed by Miyoshi [M11], was chosen to ensure inflation was automatically tuned and applied only to the regions in which observational information was present (Figure 11). The method was designed assuming observation coverage constant in time and had a tendency to generate increasing background covariance in areas where observational coverage was limited in duration. This required that I modify Miyoshi's technique to relax inflation values where the observation coverage dropped from one cycle to the next.

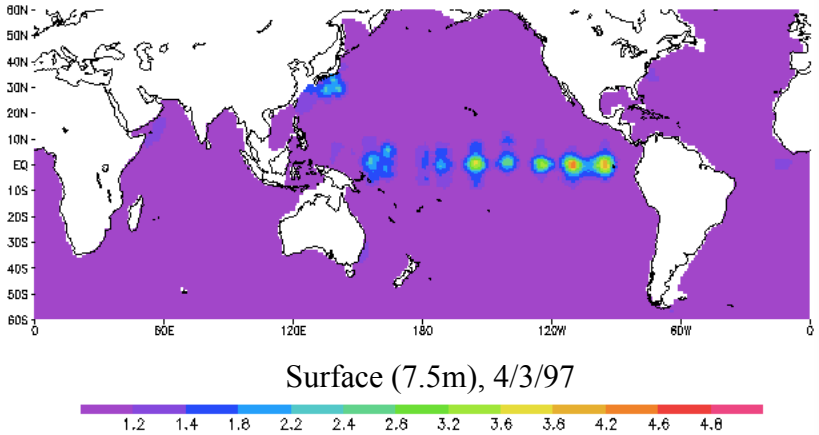

Figure 11. Adaptive inflation after 4 months for 10-member LETKF. (0-400% inflation)

For each grid point, the ratio of analysis ensemble spread to background ensemble spread was computed. A value between 0 and 1 indicates that the spread was reduced by the analysis, which was typical for the first 10 analysis cycles. After about 10 analysis cycles, the analysis usually increased the ensemble spread over the background. When combined with a growing inflation, this tended to cause LETKF to diverge, particularly in areas devoid of observations (as shown in **Figure 12**). This phenomenon could likely be reduced by the use of satellite SST data, particularly because this growth is most prevalent near the surface levels. Though, it does occur at lower depths as well and will likely still require relaxation.

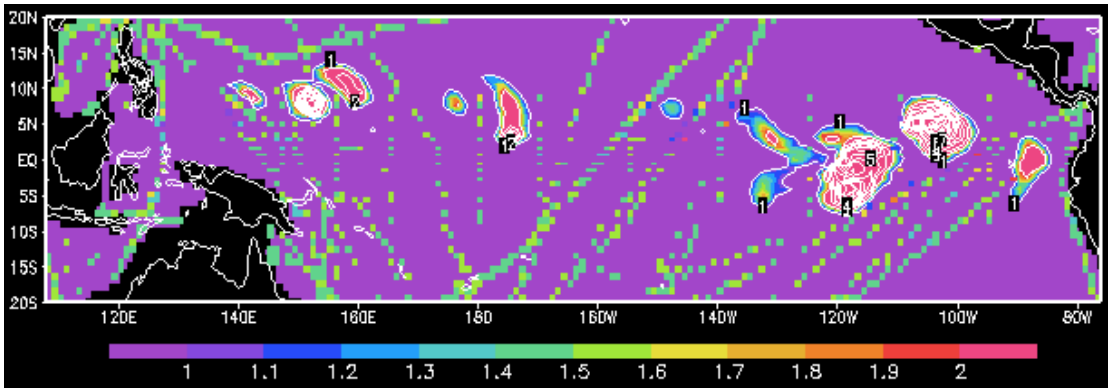

Figure 12. Background ensemble spread of Temperature (ºC) at the surface (shaded contours) and observation locations (gridded) at analysis cycle t=21 (Apr. 13, 1997) in the Equatorial Pacific. As observations appear, they cause positive inflation. If those observations disappear in a later analysis cycle, the inflation continues to compound and generates larger and larger ensemble spreads. The areas with background spread greater than 2 had inflation rescaled to prevent filter divergence.

To counter this problem, the inflation parameter was divided by the ratio of the analysis ensemble spread to background ensemble spread at every grid point for which the latter ratio was greater than 1. This effectively reduced the impact of the inflation at all locations where it was the most likely to cause growing errors, and prevented filter divergence. However, results showed the benefit of the adaptive inflation was greatly reduced, as reflected by an increase in overall RMS error. Thus a limit was introduced that only applied inflation relaxation to areas where the background spread was greater than a specified value. Using a limit of 2ºC (applied at all depths) for the temperature spread was effective in maintaining the benefits of adaptive inflation while still preventing filter divergence.

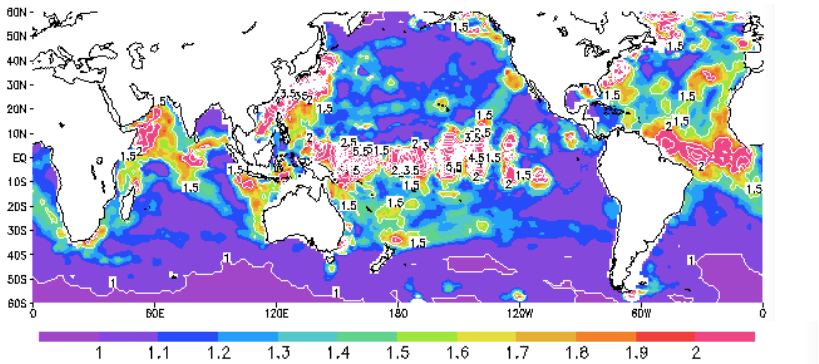

Figure 13. Adaptive inflation parameter. Shaded regions between 1 and 2 (0-100% inflation) and contoured regions from 1 to 10 (0-900% inflation) at 97.5 meters depth, for December 14th 1997.

Results of applying adaptive inflation are shown in **Figure 13** and **Figure 14** for depths of 97.5 m and 443.79 m, respectively, after about 1 year of model time. Values between 1 and 2, representing an inflation of 0 to 100 percent, are shaded. Values from 1 to 10, representing an inflation of 0 to 900 percent, are contoured. Note that the peak inflation occurs in the more dynamically unstable regions, such as the equatorial regions, the Gulf Stream, and the Kuroshio Current. The density of observations available in a particular area also influences the adaptive inflation. As a result, the inflation can be seen to follow some of the ship tracks across the Pacific.

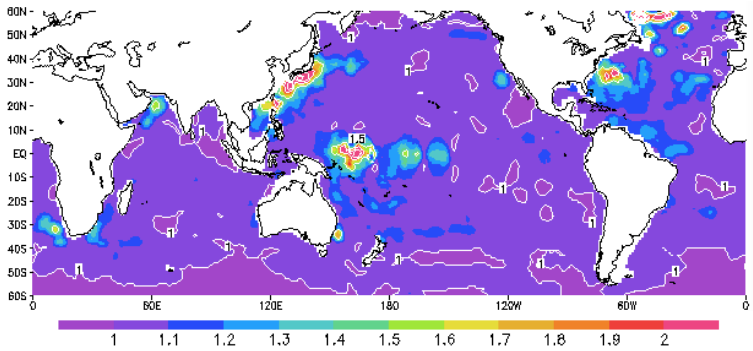

Figure 14. Adaptive inflation parameter. Shaded regions between 1 and 2 (0-100% inflation) and contoured regions from 1 to 10 (0-900% inflation) at 443.79 meters depth, for December 14th 1997.

2.5.9 Localization as a function of Latitude

Before the introduction of the Argo Float observing system, the collection of temperature and salinity observations were isolated to specific regions, primarily dictated by ship tracks. Thus, heavily traveled routes often had many observations, while other regions of the world's oceans had little to no observations for most time periods throughout the history of oceanographic record.

Previous implementations of LETKF used a constant radius localization scheme [MY07] or entered manually according to latitude and depth [HKS06]. For the ocean, a variable localization radius was calculated in LETKF as a function of latitude, with the largest sigma-radius at the equator (about 301 km sigma-radius, maximum extent 1100 km) linearly decreasing to the smallest radius at 60.01º N/S (about 82 km sigma-radius, maximum extent 300 km).

2.6 Verification and Validation

An important part of the development of any large-scale computation system is verification and validation. Some of the experiments used to do this testing are described here. Reading and writing a restart file generated by the model validated the file I/O interface. The initial file was compared against the resulting file, and it was verified that there was no difference in the data. The control script was run without activating LETKF and using identical wind forcing to verify that it matched the control run. When applying IAU, the individual model time steps were output to verify that the analysis was indeed shifting toward the observation at selected points (**Figure 15**).

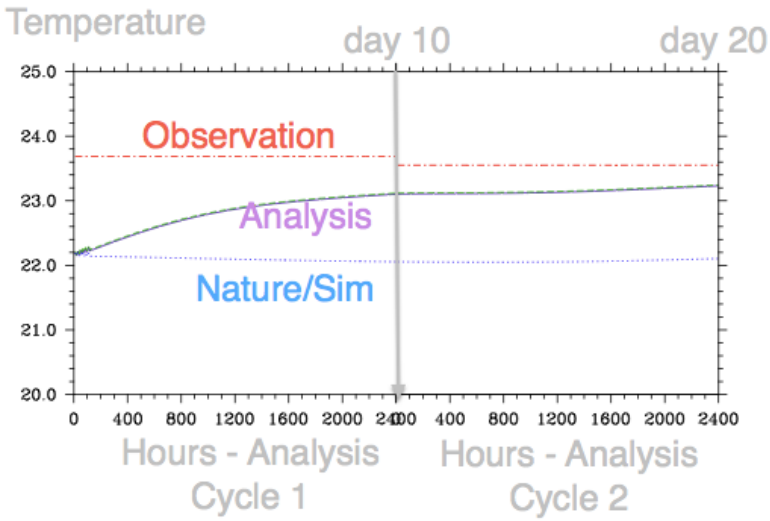

Figure 15. Two consecutive analysis cycles at a selected observation point during the execution of the MOM2 model. Using IAU, the analysis is shown to shift gradually toward the observed value during analysis cycle 1. In analysis cycle 2, a slightly different observed value was recorded. The background for analysis cycle 2 starts where the analysis left off from the previous cycle and continues to shift toward the observed value.

The ensemble forecasts were run with a uniform initial ensemble and various values of the wind forcing ensemble weighting parameter α_w to identify the rate and degree in which initial perturbations grow in the assimilation system. The largest possible values were used for α_s and α_w to examine the behavior at such a level. This uses completely different historical state data for each ensemble member and completely wrong wind data (randomly selected from the historical record) for each ensemble member. At these levels, it is clear the wind forcing dominates, as the error structure from the initial ensemble is almost completely reformed by the second analysis cycle (Figure 16).

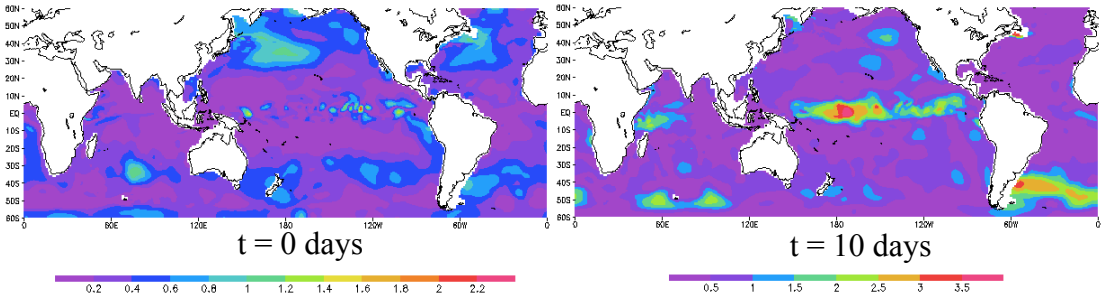

Figure 16. Background spread of Temperature (ºC) at the surface for an initial ensemble (t=0) with a 40-member ensemble generated for Jan. 1, 1990 from historical data sets, with $\alpha_s = 1$, and at the second analysis cycle (t=10 days) with $\alpha_w = 1$.

It was expected that the ensemble would converge given uniform forcing applied to all ensemble members. An ensemble forecast was run to verify this behavior. Using no observations, and the value $\alpha_s = 0.1$ for the initial ensemble perturbations, the ensemble members converged to nearly identical states within 18 months simulation time (Figure 17).

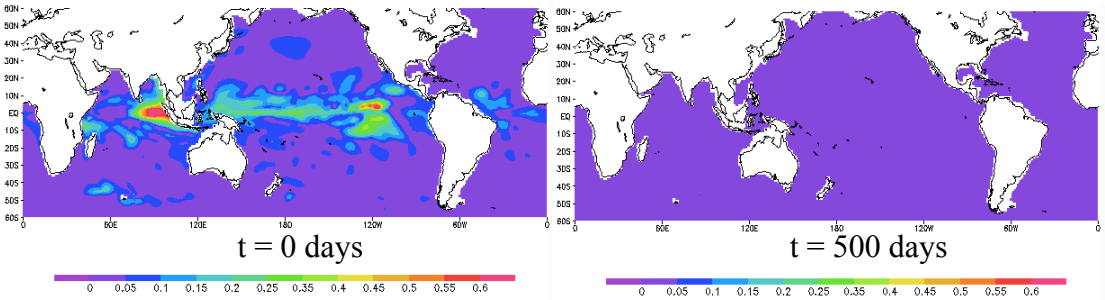

Figure 17. Background ensemble spread of Temperature (ºC) at the surface for a 4-member ensemble generated for Jan. 1, 1997 from historical data sets for t=0 and t=500 days (about 16 months), using $\alpha_s = 0.1$ and $\alpha_w = 0$.

Thus it is apparent that the wind forcing is critical to the model, and an appropriate variety of wind forcing fields are necessary to maintain an ensemble with a spread representative of the true errors of the system. Using $\alpha_s = 0$ and $\alpha_w = 0.1$ are shown in Figure 18. Because the initial spread is 0, this implies the spread depicted here is due entirely to variation in the wind forcing field. Note that the spread generated by the wind forcing in Figure 18 is similar to that generated by the method

for selecting the initial ensemble shown in Figure 17. This value of α_w was chosen for the experiments presented in Chapter 3.

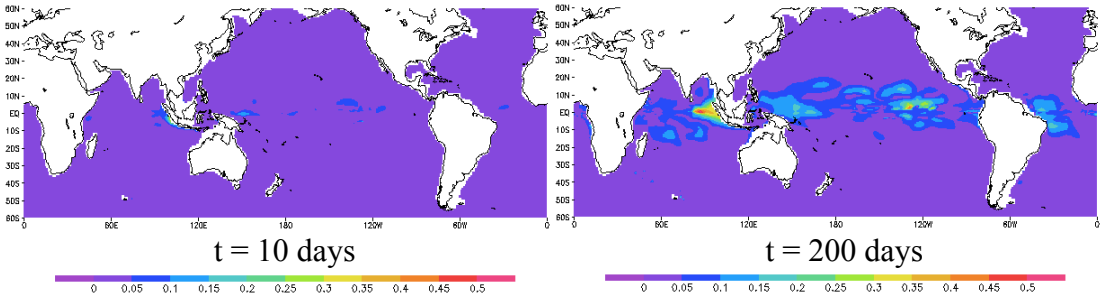

Figure 18. Background ensemble spread of Temperature (ºC) at 100 meters for a 4-member initial ensemble generated for Jan. 1, 1990 from historical data sets for t=10 and t=200 days (about 7 months), using $\alpha_s = 0$ and $\alpha_w = 0.1$.

Single observation experiments in several isolated areas were used to test the impact of the observations on the assimilation system, as shown in **Figure 19**. The horizontal localization radius can be seen to reduce with distance from the equator. The second panel in **Figure 19** shows the impact of one observation profile throughout the top 500 m along the equator.

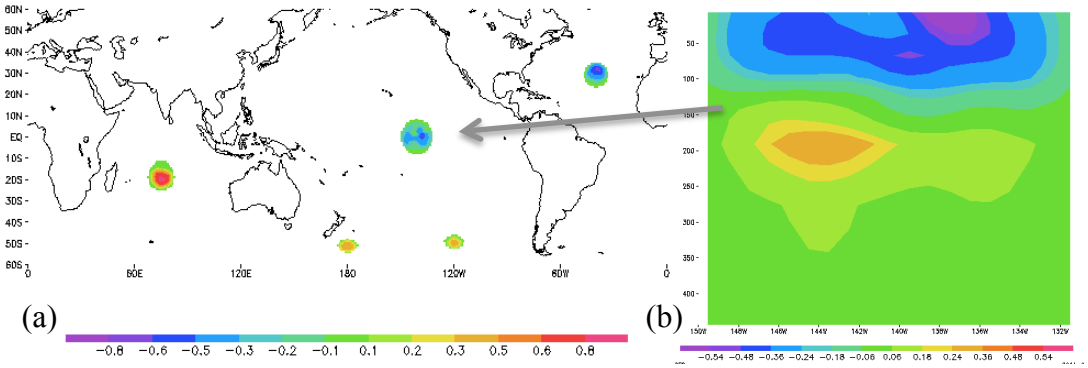

Figure 19. (a) Analysis increments of temperature (ºC), shown at the surface for 5 synthetic observations at grid coordinates: (181,109), (220, 65), (321, 99), (241, 12), and (76, 42). The varying radii reflect the latitude-dependent localization radius. (b) Top 500 m vertical cross section of the analysis increment at the equator for point (220, 65) in the Pacific from 150W to 130W.

The Bayesian approach to data assimilation is most effective when the uncertainty in both the model state estimate and observations is well quantified [HKS06]. However, there is unknown error in both the observation network and the

model. Error in the background state estimate usually comes either from uncertainty in the initial conditions generated by the previous analysis or model error. For example, a prominent hemispheric model bias was noted on the initial conditions of the data assimilation (after a 27-year spinup) when compared with the temperature observation data (Figure 20). This hemispheric bias is acknowledged but is restricted primarily to the surface fields (though other forms of model bias likely exist in the model) and should be corrected in part by the data assimilation.

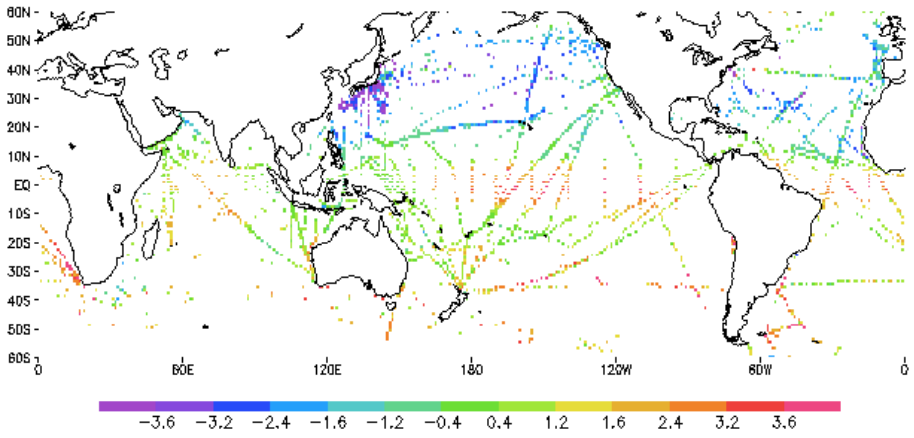

Figure 20. Observation minus background Temperature (ºC) at the surface for initial ensemble mean, centered at free-run values for January 3, 1997. There is a clear warm model bias in the northern hemisphere and cold model bias at the equator and in the southern hemisphere.

The phenomenon shown in Figure 20 is present over time, yet oscillates in the northern and southern hemispheres. This is shown in Figure 21. The northern hemisphere starts too hot, while the southern hemisphere is too cold. At the middle of the calendar year, the bias switches – the northern hemisphere becomes too cold and the southern hemisphere too hot. This pattern continues to oscillate, to varying degrees, year after year. Each year the winters are too warm and the summers are too cold.

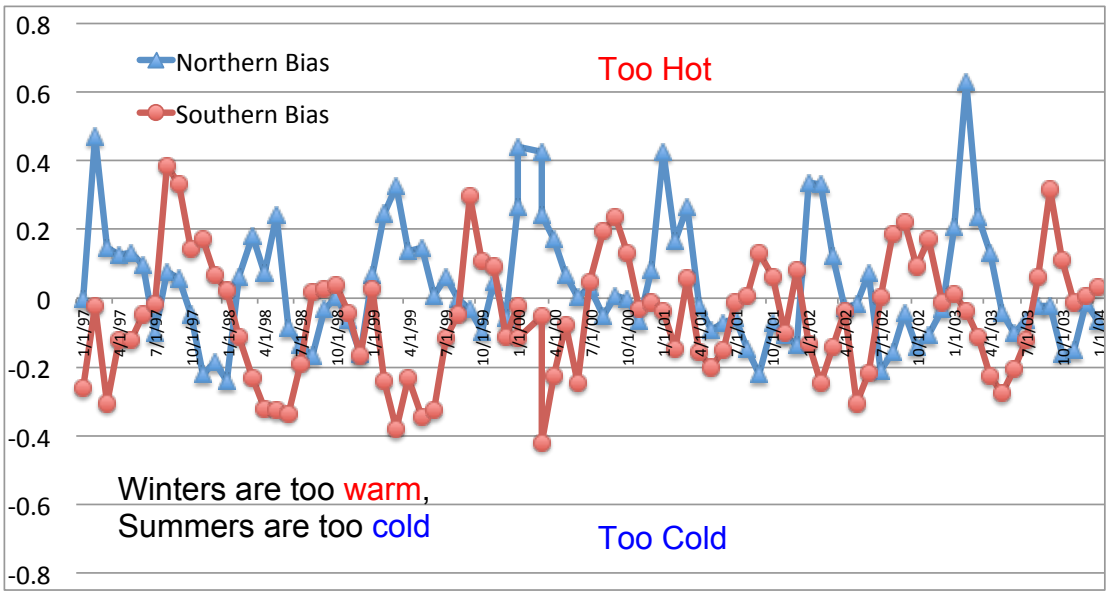

Figure 21. Observation minus background Temperature (ºC) at the surface from 1997 to 2004 for the Northern (10ºN to 60ºN) and Southern (60ºS to 10ºS) hemispheres.

The equatorial model bias is shown in Figure 22. While it also has an oscillating pattern, it seems to be less regular. Overall, there is a pattern of dips that hit their bottoms around August-October of each year. The exception is during the 1997-98 ENSO.

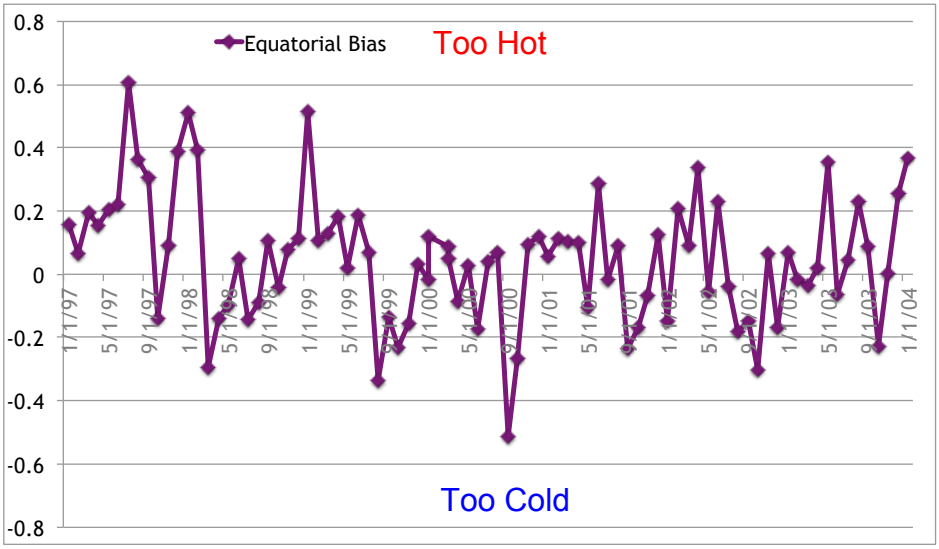

Figure 22. Observation minus background Temperature (ºC) at the surface from 1997 to 2004 for the equatorial region (10ºS to 10ºN). There is clearly an oscillation, but it is not as predictable as the mid-latitudes.

2.7 Conclusions for Chapter 2

When developing and implementing LETKF, it was important to balance many factors to obtain an effective assimilation system. For example, the optimal length of the analysis cycle window is related the degree of nonlinearity in the model. The growth in background error covariance when propagating the nonlinear model should approximately match the magnitude of the observation errors. If it does not, then inflation is needed to grow the background covariance to a suitable magnitude. The size of the localization window is dependent on the number of ensemble members available and also on the degree of nonlinearity in the system.

In general, the stronger the nonlinearity in the system, the more ensemble members are needed, the smaller the localization that should be used, the shorter analysis cycle windows that should be used, and the less need for artificial inflation.

Forcing adds an additional complication. Due to the strong impact that surface forcing has on the ocean model, if uniform forcing is used the entire ensemble tends toward a single outcome. Thus perturbations were added to the wind fields, taken as a small proportion of historical winds added onto the baseline reanalysis winds. However, it would be preferable to use the atmospheric reanalysis ensemble members directly to populate the forcing for the oceanic reanalysis ensemble. And ultimately, assimilating a coupled model is preferable to allow the nonlinear dynamics to evolve naturally in an integrated earth system.

From the results of numerous experiments running various configurations of LETKF, the inflation values resulting from the adaptive inflation seemed to give an indication of the quality of the analysis. Whether the result or the cause, high inflation tended to be correlated with over-fitting of the data. This was most prevalent in the

LETKF-EOW method, which required additional safeguards to prevent unphysical analysis increments caused by large inflation. The LETKF-RIP method resulted in much lower inflation values and required no additional tuning of the analysis increment, indicating that the results may be closer to the truth.

Thus, it may be possible to use the adaptive inflation as a metric for identifying when the ensemble methodology is insufficient for the system being assimilated. For example, it may indicate that the forcing is not varying sufficiently, or that the analysis cycle is too short, or that more ensemble members are needed, or that multiple models should be used to generate the ensemble. Of course, there are many possibilities.

Surface wind data will be taken from past atmospheric reanalysis ensembles, rather than perturbed reanalysis winds. Beyond that work, coupled ocean/atmosphere models are the obvious next step.

Chapter 3: Comparing SODA and LETKF using a Global Ocean Model

3.1 Abstract

Many advancements have been made in data assimilation in the meteorological community, but they have been slow to make their way to oceanic applications. The oceanic applications differ from meteorological applications in several important ways. The ocean observation network is extremely sparse relative to the atmospheric network and thus presents a severely underdetermined problem. Approaches that leverage the information available in the data most effectively are predicted to be most successful. Circulation in the ocean is strongly controlled by the surface wind field and thus instabilities play a less important role on the basin-scale.

The leading methods of data assimilation applied to the atmosphere are 4D-Var and variants of the Ensemble Kalman Filter (EnKF) [E94]. The EnKF benefits from the fact that it generates evolving estimates of the background error covariance. In addition, the EnKF does not use an adjoint of the linearized ocean model and is thus considerably easier and cheaper to implement. There is an additional problem that the oceanic implementation of 4D-Var is likely unstable for large time windows as one shifts to high eddy resolution because there are many local minima. For these reasons, implementation of LETKF has been explored rather than 4D-Var for the global ocean problem. Additionally, implementations of LETKF have been shown to outperform 4D-Var [YK11].

Main Points

- LETKF-RIP forecasts outperform SODA in O-F RMSD for temperature and salinity. This is shown on a global, regional, and vertical (level-by-level) scale.
- LETKF-IAU forecasts outperform or are on par with SODA in comparisons with independent observations, including equatorial zonal current velocity and altimetry. LETKF-IAU analyses outperform SODA in O-A RMSD.
- LETKF is capable of generating an analysis using a small fraction of the number of super observations used by SODA per cycle, and without using future data.
- Run-time is obviously more costly with LETKF, given that it is an ensemble method. The additional run-time is proportional to the number of ensemble members used.
- LETKF maintains an evolving estimate of the forecast error covariance, and information about the analysis error.

3.2 Introduction

The purpose of this section is to detail quantitative and qualitative comparisons made between the LETKF approach and the SODA benchmark. Two implementations of LETKF, LETKF-IAU and LETKF-RIP are presented, along with a Free-Run of the model for comparison. A third implementation, LETKF-EOW is discussed in the appendix.

3.3 Methodology

GFDL's MOM2 global ocean model was implemented with both the LETKF and SODA data assimilation systems to examine performance over multi-year historical periods spanning January 1997 to January 2004. Only temperature and salinity profiles were assimilated in these cases. SST, velocity, and altimetry observation data were not assimilated.

The model was started at rest from climatological conditions and run from 1970 to 2005 to create adequate initial conditions for various runs. Monthly surface wind stress and sea surface temperature (SST) data were used from the NCEP reanalysis. Climatology was used for sea surface salinity.

3.4 Simple Ocean Data Assimilation (SODA)

The Simple Ocean Data Assimilation (SODA) system was developed by Carton et al [CCC00a] and originally implemented on the MOM2 ocean model. SODA uses a variant of the Optimal Interpolation (OI) assimilation method, minimizing the mean square difference between the model and observations. This leads to the interpolation equation for the analysis

$$x_k^a = \tilde{x}_k^f + K_k[y_k^o - H\tilde{x}_k^f], \tag{14}$$

assuming that there exists a bias g_k^f:

$$x_k^t = \left(x_k^f - g_k^f\right) - e_k^f \tag{15}$$

in which the bias-corrected model state is defined as:

$$\tilde{x}_k^f = (x_k^f - g_k^f). \tag{16}$$

Using the observation operator H_k, the observation error is defined

$$e_k^o = y_k^o - H_k x_k^t, \tag{17}$$

and includes measurement error, error of representativeness, and error due to unresolved physics. With bias being updated by the following equations, using μ equal to 0.5:

$$g_k = \mu g_{k-1} \tag{18}$$

$$\tilde{g}_k = g_k - L_k[y_k^o - H_k \tilde{x}_k^f]. \tag{19}$$

The following equations are solved locally in a series of patches, with alpha equal to 0.7:

$$K_k = (1-\alpha) P_k^f H_k^T \left[H_k P_k^f H_k^T + R_k \right]^{-1} \tag{20}$$

$$L_k = \alpha P_k^b H_k^T \left[H_k P_k^b H_k^t + H_k P_k^f H_k^T + R_k \right]^{-1} \tag{21}$$

where,

$$P_k^f = \left\langle e_k^f \left(e_k^f \right)^T \right\rangle \tag{22}$$

$$P_k^o = \left\langle e_k^o \left(e_k^o \right)^T \right\rangle \tag{23}$$

$$P_k^b = \left\langle \left(g_k^f - \left\langle g_k^f \right\rangle \right) \left(g_k^f - \left\langle g_k^f \right\rangle \right)^T \right\rangle. \tag{24}$$

SODA uses estimated forecast errors that are varied with latitude and depth, but the errors are not updated in time based on the model state evolution as with an EnKF.

The modern SODA implementation uses the Parallel Ocean Program (POP) global ocean model [http://www.atmos.umd.edu/~ocean/PDF/ccsm-2003.pdf], which shares its historical evolution with precursors of the MOM model [http://climate.lanl.gov/Models/POP/]. Though SODA was once implemented with MOM2 [CCC00a], requiring extensive modification of the model code, the system

has gone through a number of modifications and improvements since that time and has been developed into an external modular package. It was therefore necessary to create a new implementation of the modern SODA with MOM2.

Observation data was vertically interpolated to the MOM2 model levels. Correlation data was also interpolated from 40 POP vertical levels to 20 MOM2 vertical levels. To operate on the MOM2 forecast data, the model grid must be interpolated to the uniform 1x1 degree grid used by SODA. Surface height must be calculated by MOM2 diagnostic routines to input to SODA. Climatology from Levitus was used to estimate the temperature and salinity relationship, and was also interpolated to the MOM2 vertical levels. MOM2 salinity was converted from a perturbation value to an absolute salinity value in psu.

3.5 Experiment parameters

The following parameters for the LETKF-RIP analyses were used: a 40-member ensemble, perturbation of the initial ensemble α_s is 0.5 (defined in Section 2.4), perturbation of the surface wind field α_w is 0.1 (defined in Section 2.5.4), the σ_b value [M11] indicating time smoothing for adaptive inflation is 0.001, inflation relaxation (Section 2.5.8) is limited to areas where the analysis spread for temperature is greater than 2º C. A 5-day analysis cycle was used. For LETKF-IAU, identical parameters were used with the exception of a 20-member ensemble and a 30-day analysis cycle.

The LETKF analyses were not sensitive to variations in the parameter α_s. The spread of the initial ensemble was quickly reduced after a few analysis cycles. There was some impact over time to the surface levels when using a large wind perturbation

α_w. This created increasing ensemble spread in the Gulf Stream and Kuroshio Current after months of simulation time.

Various configurations of LETKF were used (some additional results are reported in the Appendix). The ensemble sizes ranged from 40, 20 or 10 members. The analysis cycles ranged from 30, 10 or 5 days. Adaptive inflation and inflation relaxation were used with all runs. In some cases, shown in an appendix, an extended window was used (LETKF-EOW) to incorporate information from observations outside the current analysis cycle. This approach mirrored that of SODA's treatment of observations.

The observations were binned as super-obs in 1x1-degree horizontal grid, vertically interpolated to MOM2 vertical levels. For SODA the observations were placed in 5-day bins, while for LETKF the observations were placed in 1-day bins to better utilize the 4D-capability of LETKF. SODA used a 45-day radius window of observations combined and centered at the analysis time. LETKF-RIP and LETKF-IAU used observations only within the analysis cycle window.

For these experiments, LETKF uses a localization range of 1100 km at the equator that decreases linearly to 300 km at +/- 60º latitude. The localization range is divided by approximately 3.65, giving an effective sigma radius of 300 km and 82 km, respectively, with respect to the Gaussian localization weighting function. In the vertical, a radius of 300 m was used, giving an effective sigma radius of 82 m.

SODA is not an ensemble method. While LETKF used an input ensemble with perturbations applied to the input members and the surface forcing, SODA used an initial background from the model spin-up equivalent to the LETKF ensemble

mean. SODA used the baseline NCEP reanalysis for surface forcing. LETKF used perturbed forcing fields with the mean equivalent to the NCEP reanalysis used by SODA. As a result, LETKF is accounting for some degree of error in the model's forcing fields, which can have a significant effect on the model's forecasts.

3.6 Analysis Cycle

The analysis cycle differed slightly among the various configurations of LETKF and SODA. The differences are highlighted below in **Figure 23**, **Figure 24**, and **Figure 25**.

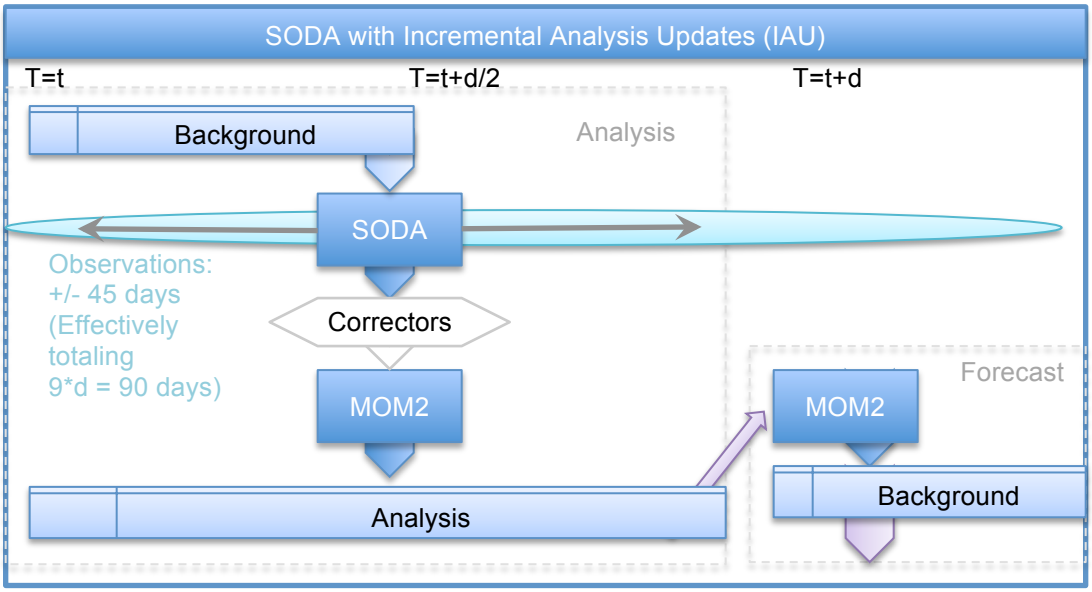

Figure 23. Schematic diagram of the SODA analysis cycle. For SODA, the analysis cycle length is d=10 days. The forecast is run for 5 days, analysis increments are used to generate 'correctors' to the model integration via a forcing term. Finally, a 10-day model run is performed incorporating these correction terms, providing initial conditions for the next 5-day forecast.

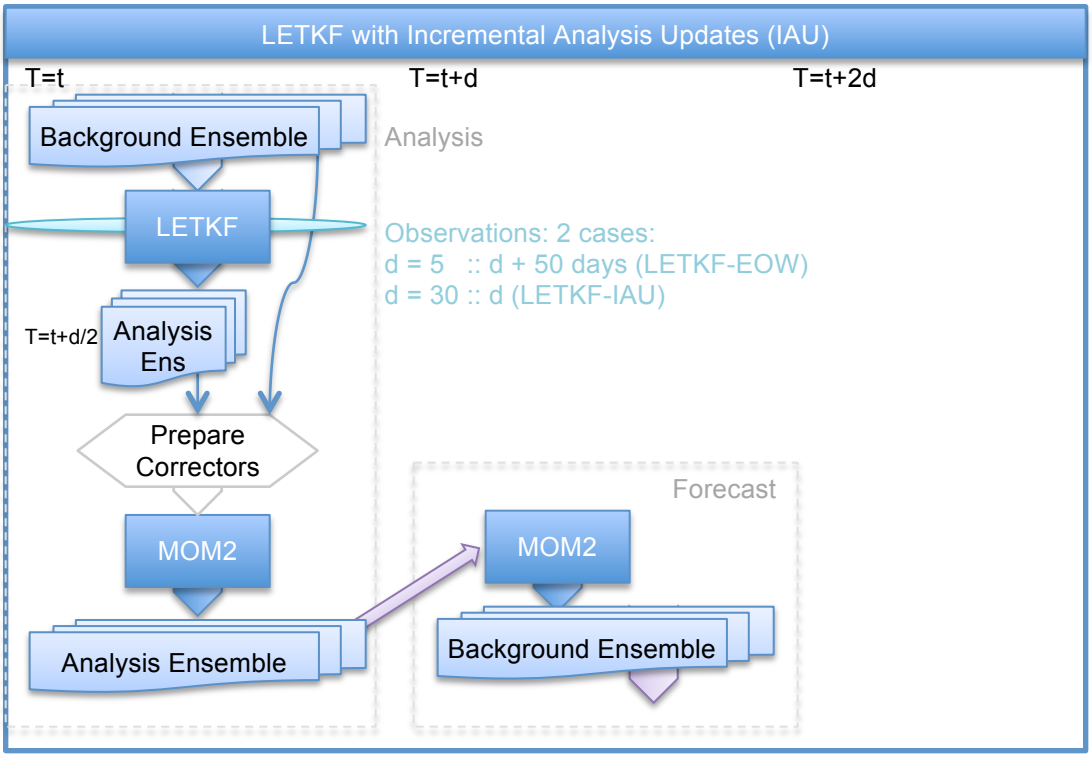

Figure 24. Schematic diagram of the LETKF analysis cycle using IAU. Guided by the approach of SODA, correctors are calculated by differencing the analysis centered in the analysis cycle and the corresponding background. These values are added incrementally to the model integration via a forcing term.

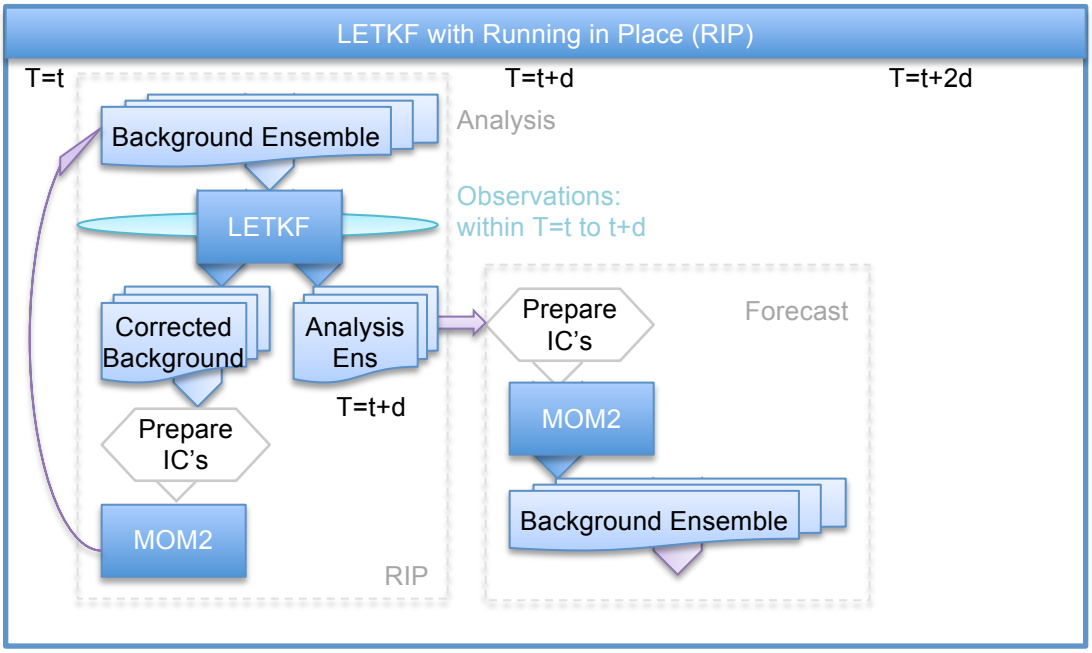

Figure 25. Schematic diagram of the LETKF analysis cycle using RIP. LETKF produces an analysis at the end of the cycle and a corrected background at the beginning of the cycle. The corrected background is used to repeat LETKF.

3.7 Experiments

While numerous combinations of ensemble sizes, analysis cycle lengths, and configurations of LETKF were used, the results from only three will be presented here. The following experiments will be compared in the subsequent sections:

Table 1. Titles for experimental results throughout the remainder of the report.

Title	Description
Free Run	The MOM2 model run from 1970 to 2004 with NCEP surface forcing. It is compared to temperature and salinity profiles for reference.
SODA	An assimilation of temperature and salinity profiles for 7 years from Jan. 1997 to Jan. 2004 using SODA with NCEP surface forcing in a 10-day analysis cycle.
LETKF-IAU	Assimilation of T and S profiles for 7 years from Jan. 1997 to Jan. 2004 using LETKF with IAU, a 20-member ensemble, with perturbed NCEP winds, and a 30-day analysis cycle.
LETKF-RIP	Assimilation of T and S profiles for 7 years from Jan. 1997 to Jan. 2004 using LETKF with RIP, a 40-member ensemble, with perturbed NCEP winds, and a 5-day analysis cycle.
LETKF-EOW	Assimilation of T and S profiles for periods 1997-1998 and 2001-2003 using LETKF with IAU, 40-member ensembles (except where otherwise noted), with perturbed NCEP winds, a 5-day analysis cycle, and an Extended Observation Window of +/- 25 days beyond the analysis cycle. (Results shown in the Appendix)

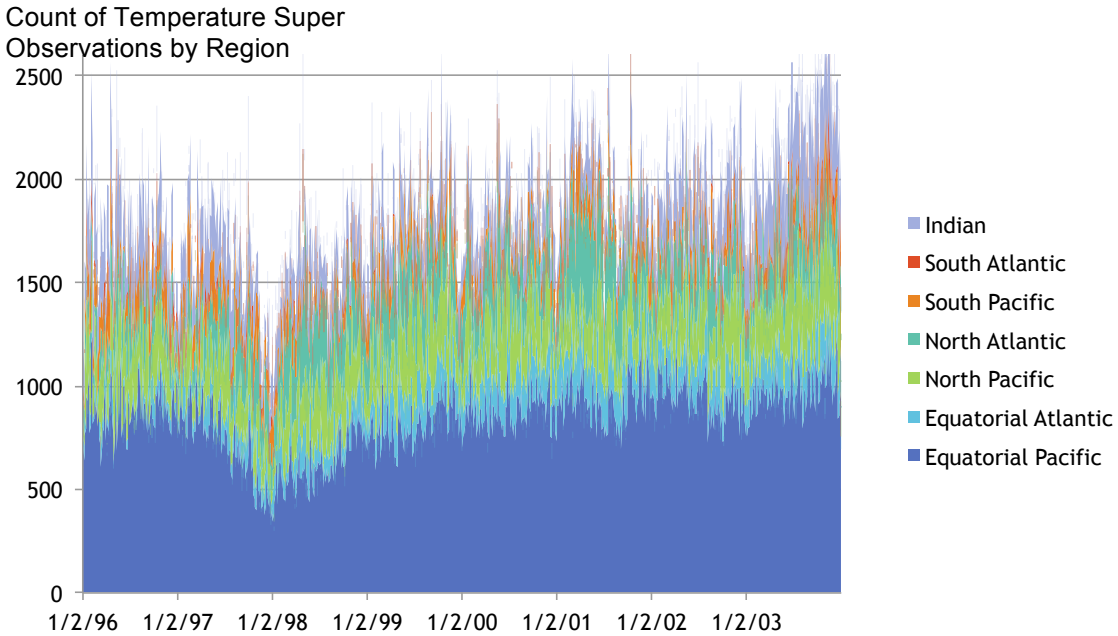

Figure 26. Count of super observations of temperature by region, for all depths. (Regions are shaded as a proportion of the total).

Results are presented focusing on the 7-year experiment runs showing long-term performance of the systems. These results include the period from 1997-98 during a significant El Niño, for which temperature and salinity observation data are relatively sparse. In addition, they include the introduction of the Argo float observing system, from 2001-2003. Overall, the experiment period contains a wide range of observation network conditions. This can be seen for temperature in Figure 26 and for salinity Figure 27. From the period spanning 1996 to 2004, there is an increase in Temperature observations, concentrated primarily in the Equatorial and North Pacific. However, during this period there is a dramatic increase in the number of Salinity observations, particularly concentrated in the Indian and North Atlantic Oceans.

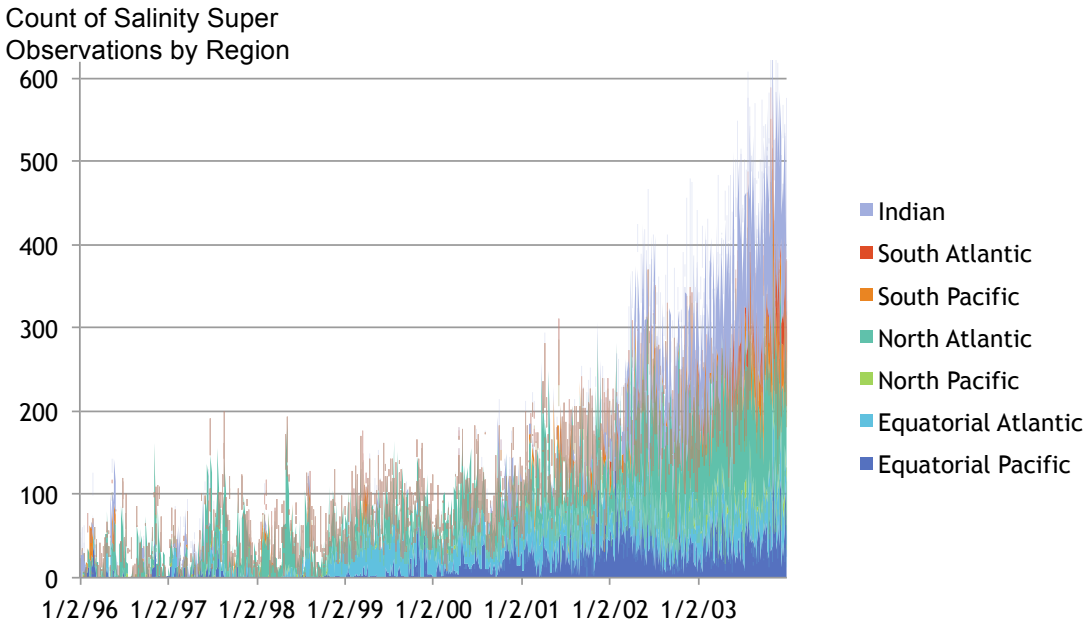

Figure 27. Count of super observations of salinity by region, for all depths.

3.7.1 Observed minus Forecast Results

A key measure of the performance of the data assimilation systems is the Root Mean Square Deviation (RMSD, sometimes called Root Mean Square Difference or Root Mean Square Distance) between the Forecast (F) and the Observations (O). The RMSD is a common measure of the aggregate difference, or residuals, between a model and observed values that the model attempts to predict. This is usually described as the (O-F). While comparing the difference between the forecast and observed values, it gives an estimate of the actual error between the forecast and the truth. [See Appendix for additional discussion] **Figure 28** shows the (O-F) for a 'Free Run' of the model starting from a 20-year spin-up using detrended NCEP surface forcing. Results are shown aggregated over all levels in which super observation profiles were available, typically ranging from 0 to 1,000 m depth. This figure gives an upper-limit on the acceptable RMSD of any of the data assimilation results. The Free-Run temperature RMSDs are maintained at a fairly consistent level due to the effect of surface forcing and a relatively large dispersed sample of observations (compared to the number of salinity observations). There are large fluctuations in the salinity observations due to a very small sample size and uneven dispersion, which gradually improves as the number of salinity observations increases over 5x during this period.

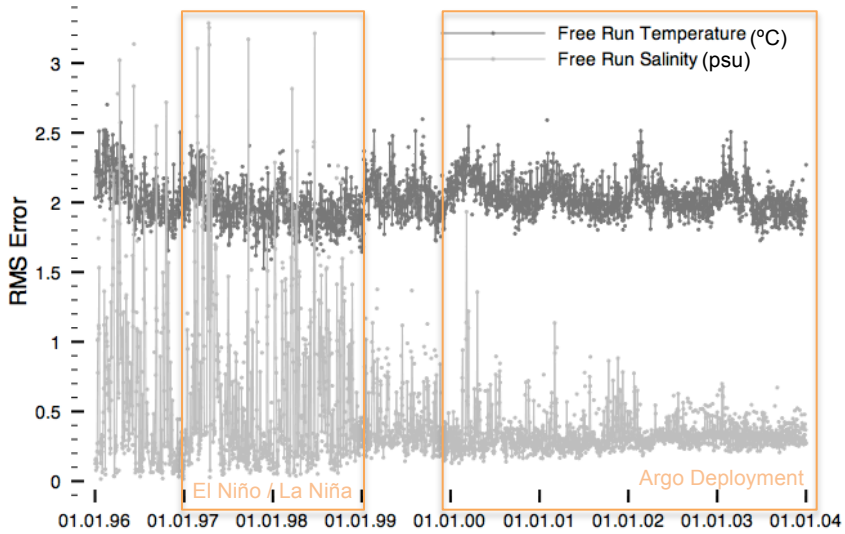

Figure 28. RMSD of temperature and salinity in a Free Run of the MOM2 model from 1996 through 2003, calculated for super observations at all levels.

In comparison, **Figure 29** shows a steady improvement in time using LETKF-IAU, a conservative configuration of LETKF (incorporating IAU, 20-member ensemble, 30-day analysis cycles, adaptive inflation) starting in January 1997. The assimilation produces a slow spin-up, but gradually reduces the (O-F) RMSD as the cycles proceed. This is particularly true as the number of salinity observations goes through the 3rd or 4th doubling. By 2002 the reanalysis (O-F) RMSDs are consistently below the Free-Run values.

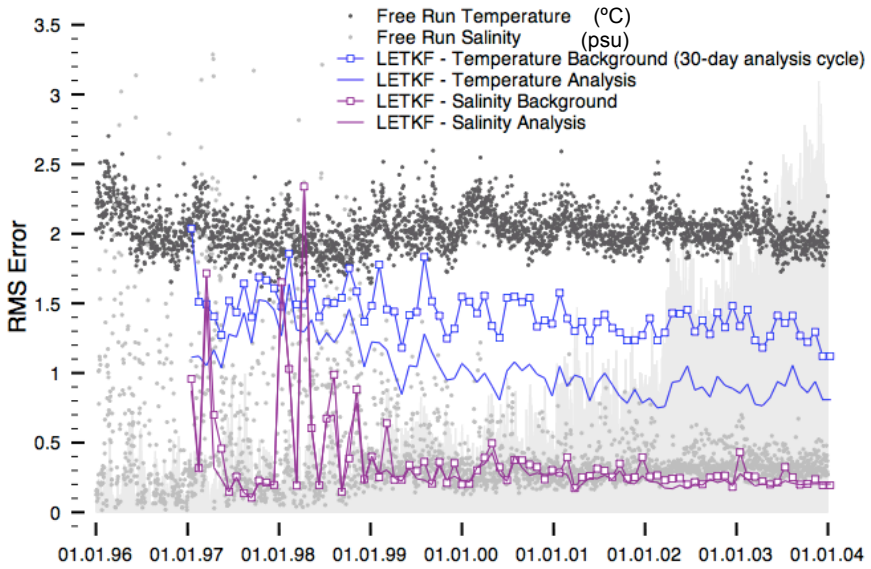

Figure 29. Comparing RMSDs in Temperature and Salinity background (O-F) and analysis (O-A) for 30-day analysis cycle, 20-member LETKF with IAU, versus a Free Run of the MOM2 model. The salinity observation counts are scaled (divided by 200) and shown as the shaded background for reference.

Figure 29 shows the 15-day 'free-forecast' O-F in comparison with the LETKF-IAU analysis O-A. However, the adjustment made to the model state in the analysis process is not that large. Due to the gradual approach of LETKF-IAU, the analysis increment is added in small parts to each model integration step during a second 'forced-forecast' over the same time period. The forced-forecast resulting from use of IAU is shown in Figure 30. SODA uses the same incremental analysis update approach, with its 5-day free-forecast shown as the 'background'.

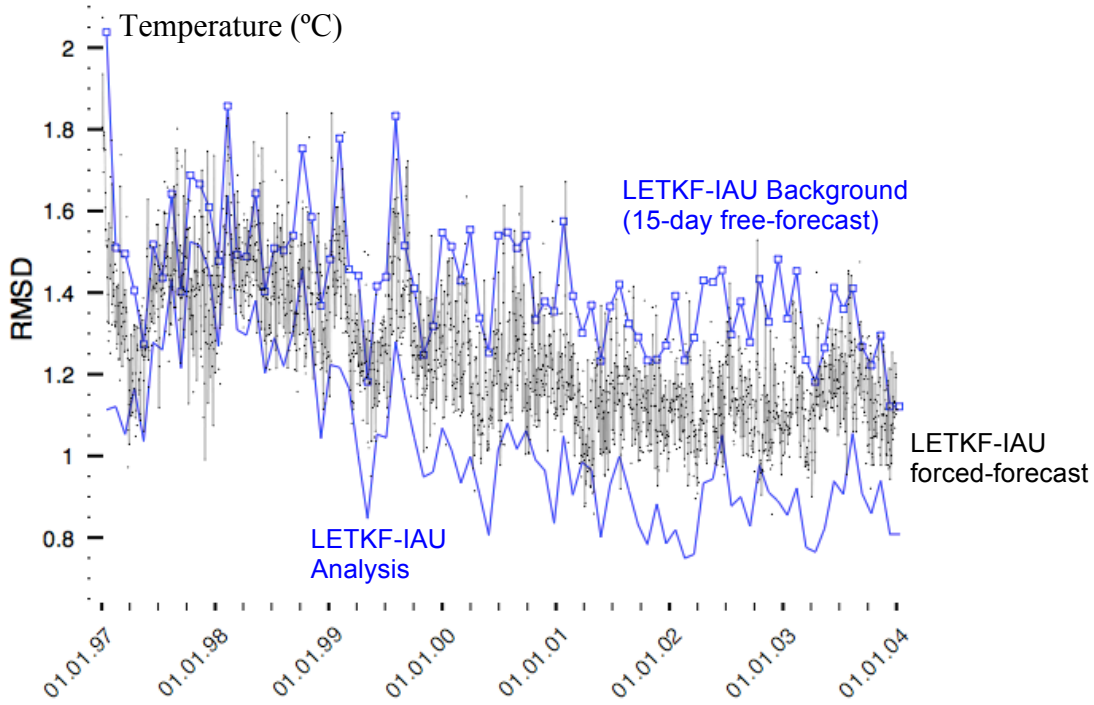

Figure 30. Comparing RMSDs in the 15-day Temperature free-forecast (O-F), the corresponding LETKF-IAU analysis (O-A) at the center of the 30-day analysis cycle, and the resultant forced-forecast (plotted daily) after adding on the small analysis increments to each model integration step.

LETKF-IAU was run with 10, 20 and 40 members to test the sensitivity of the analysis results to the ensemble size. The adaptive inflation values from the end of the 20-member case were used as the initial conditions for the inflation for the 10-members case to test the spinup of the inflation parameter. Typically, the RMSD for each ensemble size is expected to follow the relation: RMSD(LETKF-IAU40) < RMSD(LETKF-IAU20) < RMSD(LETKF-IAU10). This relationship is achieved after about 3-4 years in the experiments, as shown in Figure 31. Thus two conclusions that can be drawn from this experiment are that (1) the O-F and O-A RMSD for LETKF-IAU using various ensemble sizes are relatively similar, and (2) with the given parameters, the adaptive inflation procedure takes about 3-4 years (~40 analysis cycles) to settle on its inflation values.

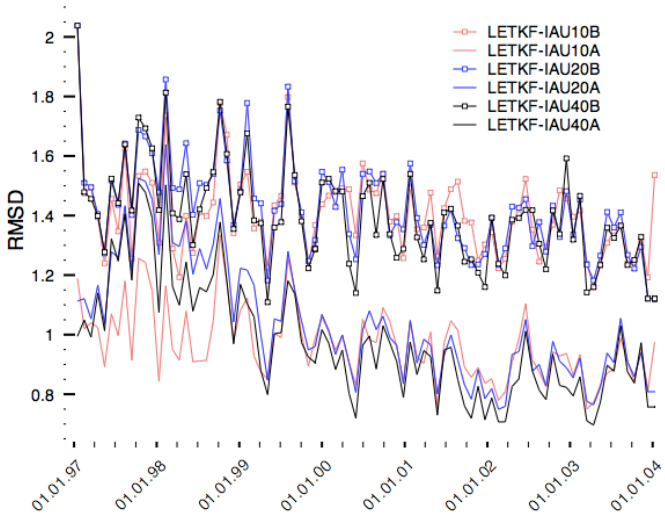

Figure 31. Comparing RMSDs for Temperature (O-F) and (O-A) with LETKF-IAU using 10, 20 and 40 ensemble members. To accelerate spin-up, the 10-member case used the inflation from the end of the 20-member case as its initial inflation values. It takes about 3-4 years for the 10-member case to take its expected place as the 'lowest' performer'.

The Free-Run is compared to SODA, LETKF-RIP and LETKF-IAU in Figure 32.

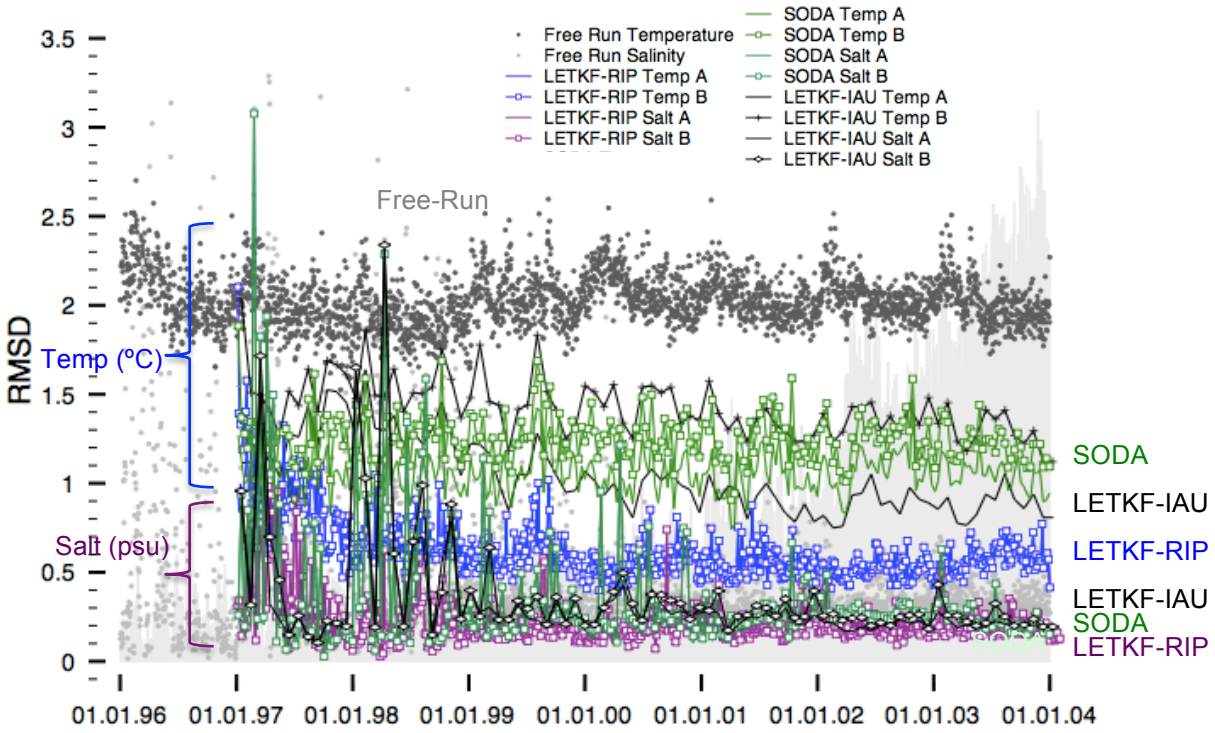

Figure 32. The LETKF-IAU from **Figure 29** tracks the SODA well after a longer spinup. This is done *without* reusing observations. The LETKF-RIP (reusing observations once) quickly spins up and outperforms relative to temperature and salinity errors. The count of super observations is shown in the background by the filled areas for temperature (light gray) and salinity (dark gray) and measured by the second y-axis. The same color scheme will be used for all remaining RMSD plots.

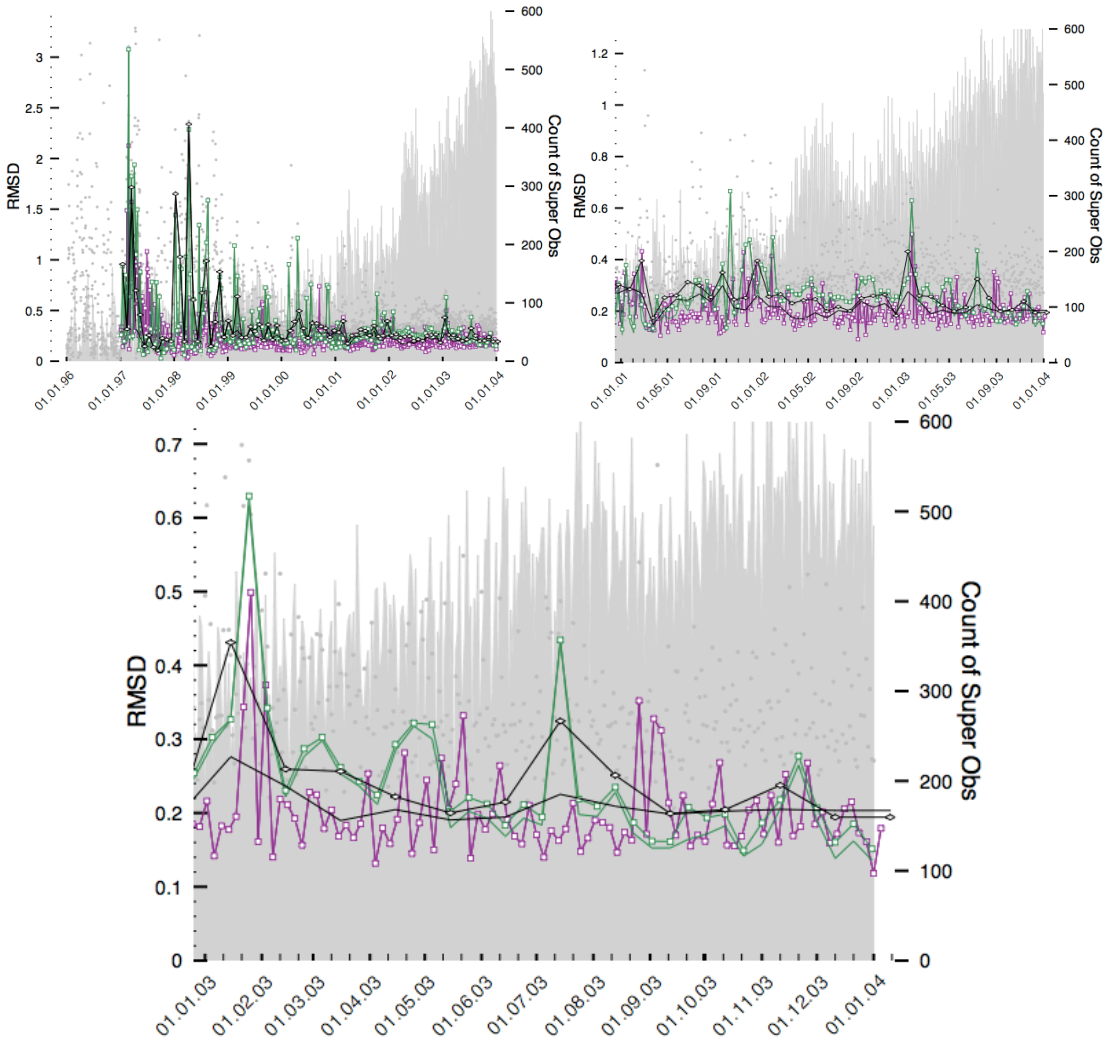

Figure 33. Comparing RMSDs for Salinity (O-F) and (O-A) with Free-Run, LETKF-RIP, LETKF-IAU, and SODA. Periods shown are 7-years from 1997-2004, 3-years from 2001-2004, and 1 year Jan. 2003 – Jan. 2004.

Due to the complexity of the data, the trends are indicated with a 12-month moving average in Figure 34 and Figure 35.

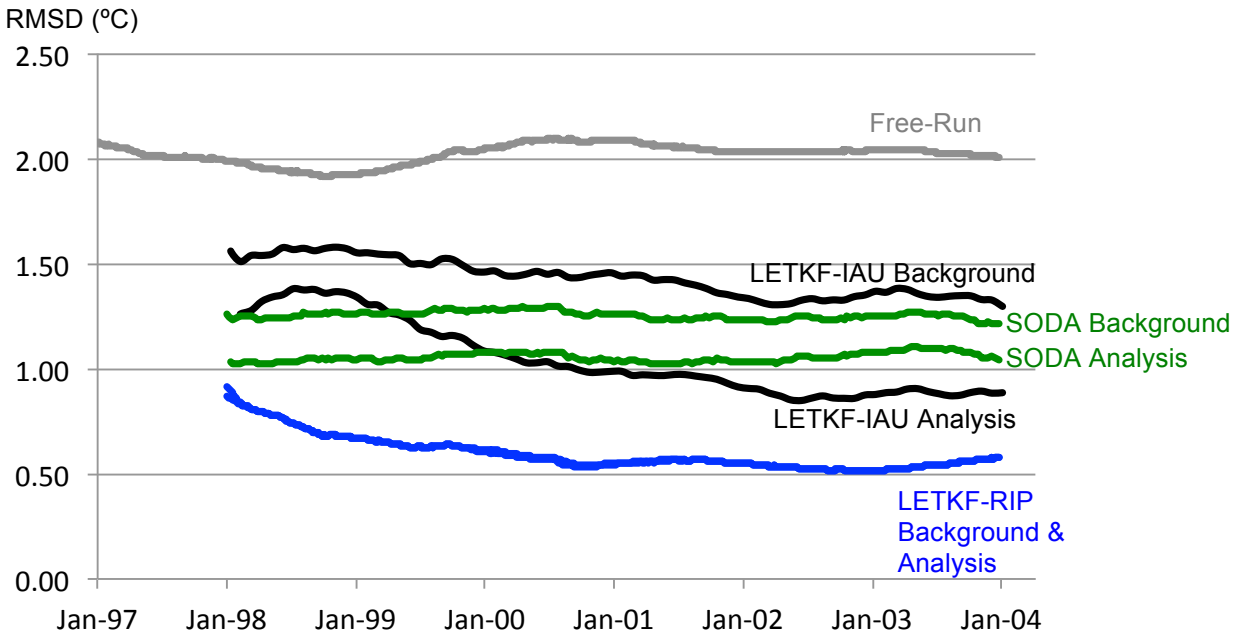

Figure 34. 12-month moving average of Free-Run LETKF-IAU, LETKF-RIP and SODA Temperature (ºC) RMSD for O-F (Background) and O-A (Analysis)

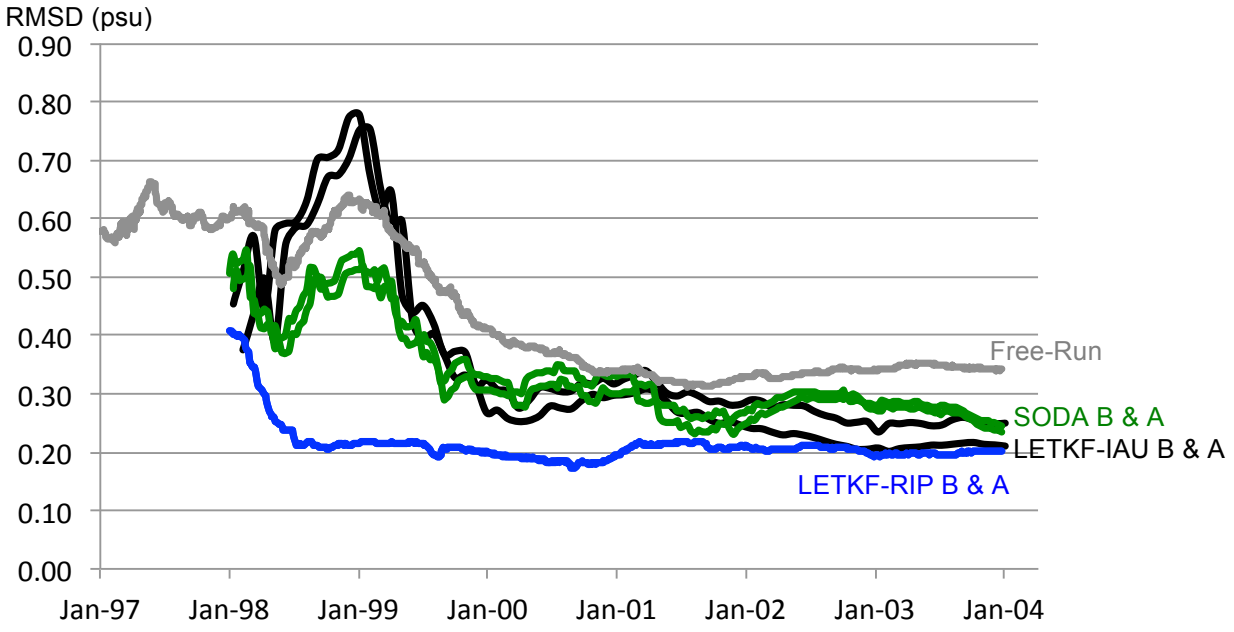

Figure 35. 12-month moving average of Free-Run LETKF-IAU, LETKF-RIP and SODA Salinity (psu) RMSD for O-F (Background) and O-A (Analysis)

The results for the LETKF-RIP method are broken down regionally and by level the following figures: Figure 36 (at the surface, 100 m and 500 m depths), Figure 37 (disjoint regions together comprising the entire global ocean), and Figure 38

(dynamically active sub-regions of the North Pacific and the North Atlantic, respectively the Kuroshio current and the Gulf Stream).

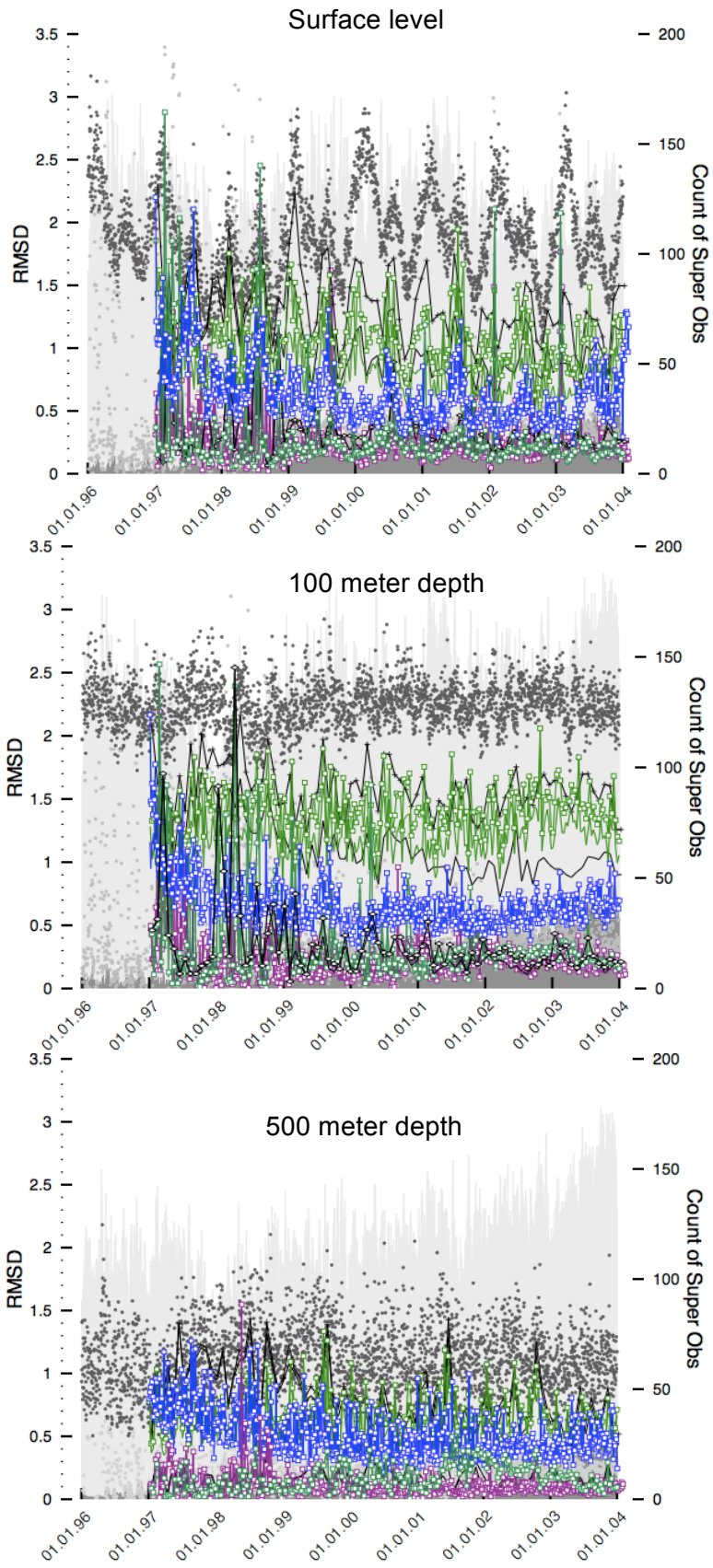

Figure 36. LETKF-RIP RMSD at the surface, 100 meters, and 500 meters for temperature (ºC) and salinity (psu). LETKF-RIP performs best at the 100 meter depth as it is less influenced by the model bias at the surface and has more dynamic variability than the deeper levels. Though the prescribed observation error used for the LETKF analysis was determined independently from a high-resolution SODA analysis, the free-run reflects a similar error profile. (The magnitude of the average RMSD increases toward 100 meters where the representativeness error is the highest, and then decreases rapidly down to 500 meters where the representativeness error is much lower.) This could potentially give guidance to determining an appropriate error covariance to use for the observations in different regions and depths due to representativeness error.

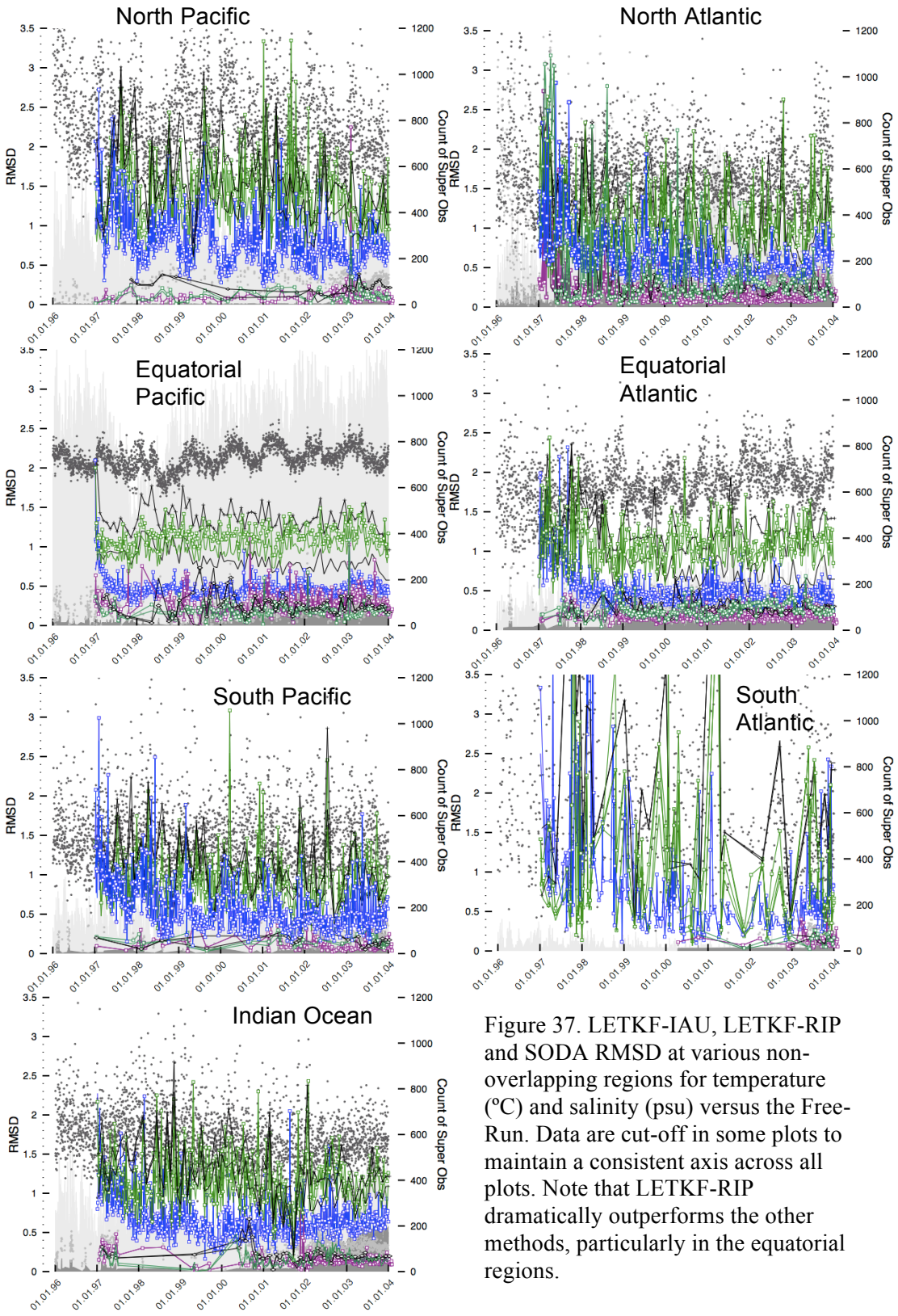

Figure 37. LETKF-IAU, LETKF-RIP and SODA RMSD at various non-overlapping regions for temperature (ºC) and salinity (psu) versus the Free-Run. Data are cut-off in some plots to maintain a consistent axis across all plots. Note that LETKF-RIP dramatically outperforms the other methods, particularly in the equatorial regions.

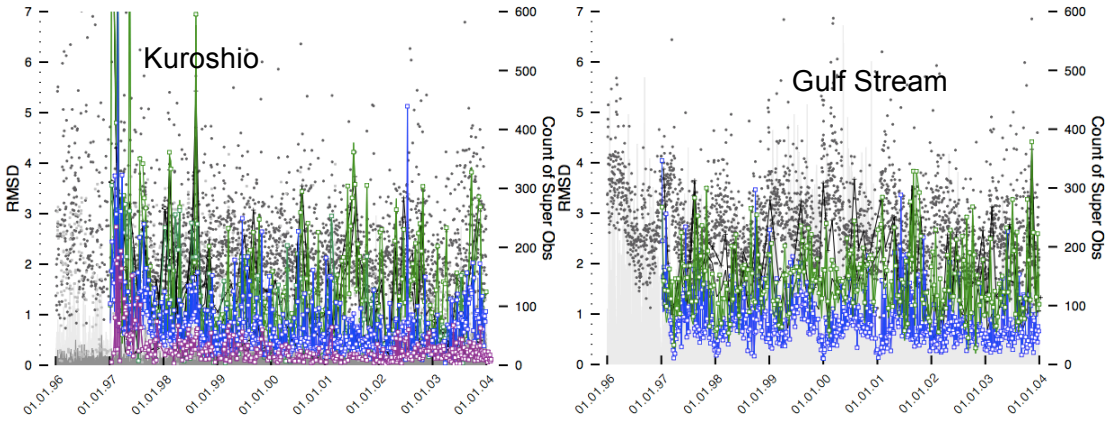

Figure 38. LETKF-RIP and SODA RMSD for the sub-regions Kuroshio (sub-region of the North Pacific) and Gulf Stream (sub-region of the North Atlantic). Note the axes are different than the previous figure. No salinity data is available in the Kuroshio region during this period.

In Figure 36, there is a clear seasonal bias in the Free-Run data at the surface, as identified previously, and is likely related to bias in the model or the model surface forcing. The RMSD for all methods are reflective of this phenomenon, though there is still a clear improvement shown by LETKF-RIP over the other approaches. The best performance is seen at the 100 m depth.

A regional breakdown is given in Figure 37. Again the hemispheric model bias is present and seen in the northern and southern areas of both the Pacific and the Atlantic oceans. These model biases are reflected in the RMSD of the various methods, though there is still a clear improvement in the LETKF-RIP over the remaining methods. While there is also some bias evident for the Free-Run in the equatorial regions, the clearest example of LETKF-RIP outperforming the other methods is seen in these areas.

Thus we see that LETKF-IAU, even without reusing observations, is competitive with the SODA method. Because SODA has undergone significant tuning throughout its development, it is presumed that the same conclusion would be true when comparing to any straightforward OI or 3D-Var reanalysis approach.

Furthermore, with careful reuse of the observation data as in LETKF-RIP, spin-up can be accelerated. In the following sections, results of these and additional experiments will be explored.

3.7.2 Altimetry and Thermocline Heat Content

Sea surface height is closely related to the depth of the thermocline, particularly in the tropics [CGX96]. The relationship typically breaks down in the mid-latitudes. Because the altimetry observations of sea level have not been used in the assimilation, they are used here for verification of the results from each of the assimilation methods. The cross-correlation between altimetry and the integrated heat content of the top 300 meters is shown in **Figure 39** for the period 1997-2002. In all methods, high correlation (> 0.8) is shown throughout most of the equatorial Pacific. Further, there is relatively high correlation (> 0.5) in the Equatorial Atlantic, Equatorial Indian, North Atlantic and some areas of the North and South Pacific.

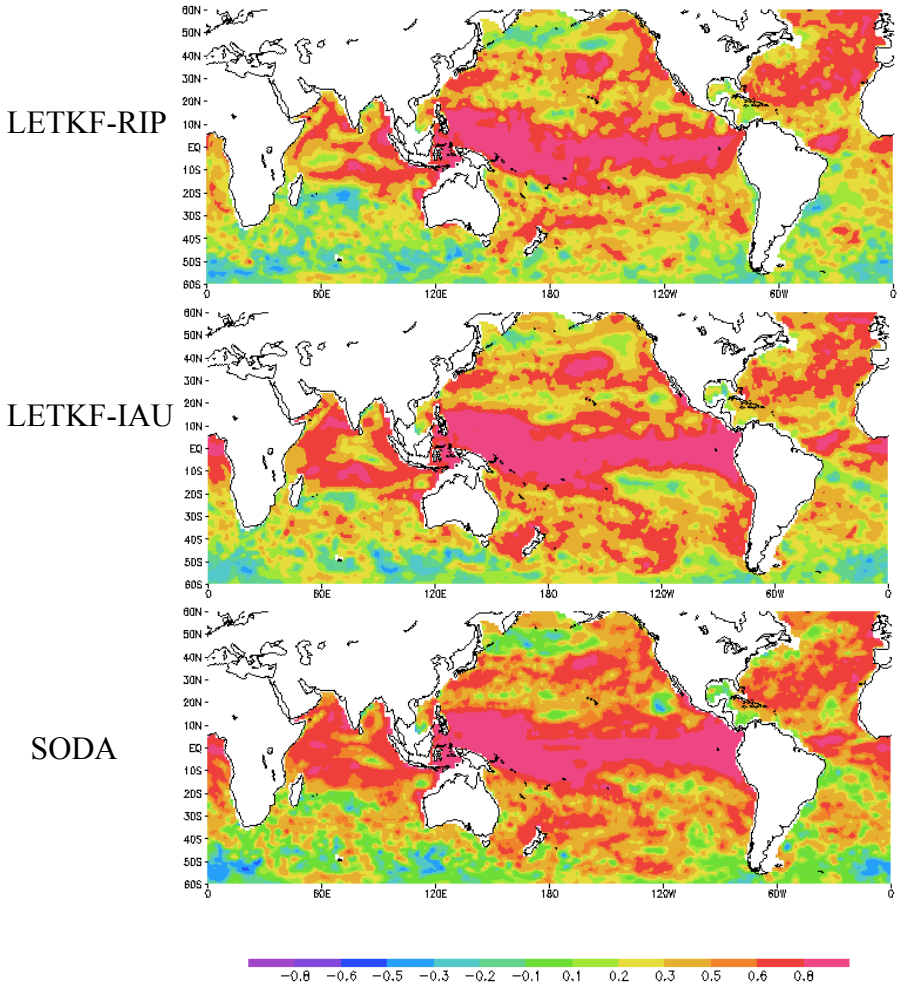

Figure 39. Monthly average top 300m analyzed heat content correlated with altimetry sea level during 1997-2002. Altimetry is calculated as cm perturbations and heat content as vertically integrated temperature perturbations from the time mean.

The first and second Fourier terms of the altimetry and vertically integrated heat content (300m) are shown in **Figure 41**. As would be expected based on the correlation patterns, there is a strong matching pattern in the equatorial areas.

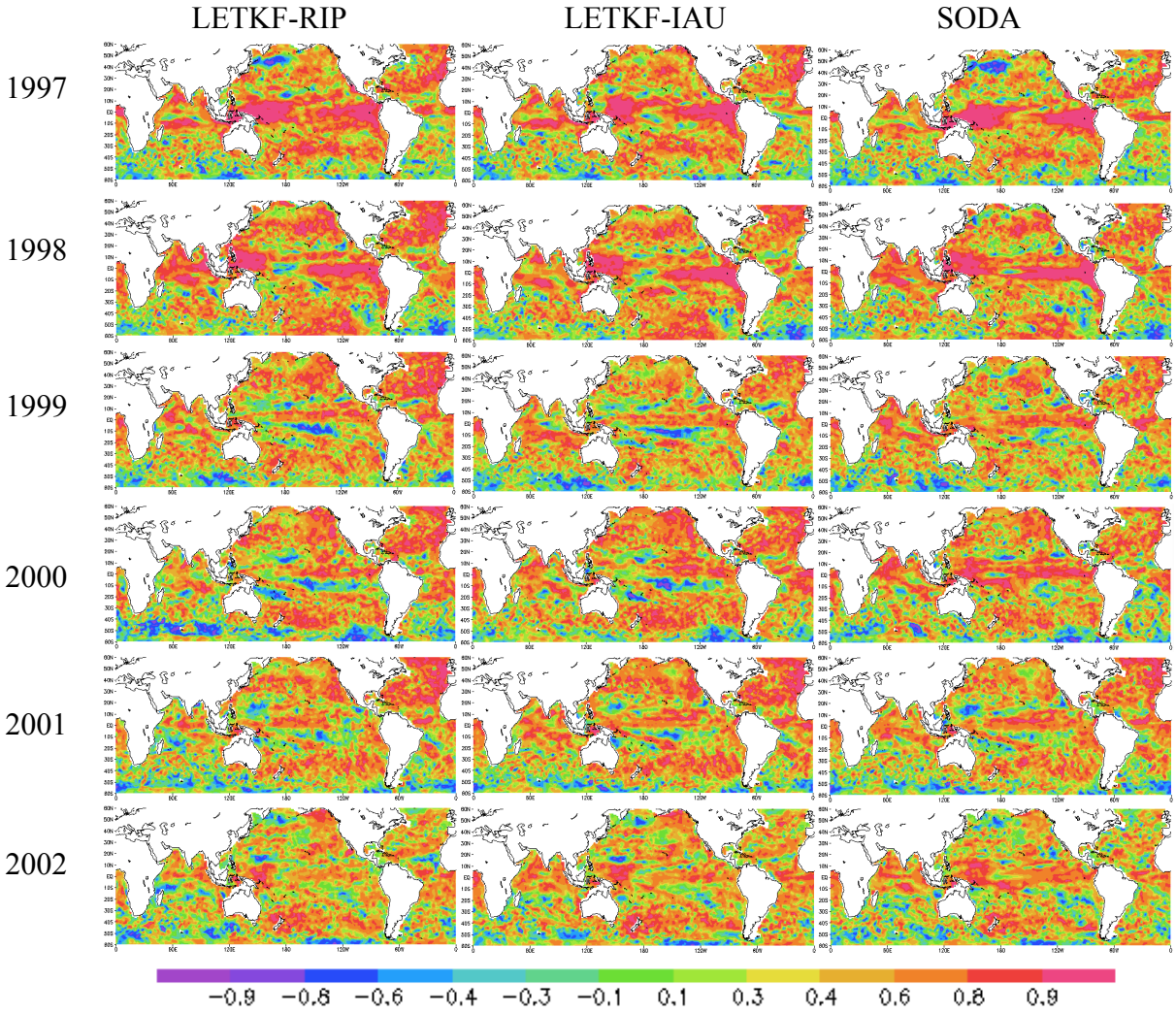

Figure 40. Monthly average top 300m analyzed heat content correlated with altimetry sea level shown for every year from 1997 to 2002. Altimetry is calculated as cm perturbations and heat content as vertically integrated temperature perturbations from the time mean.

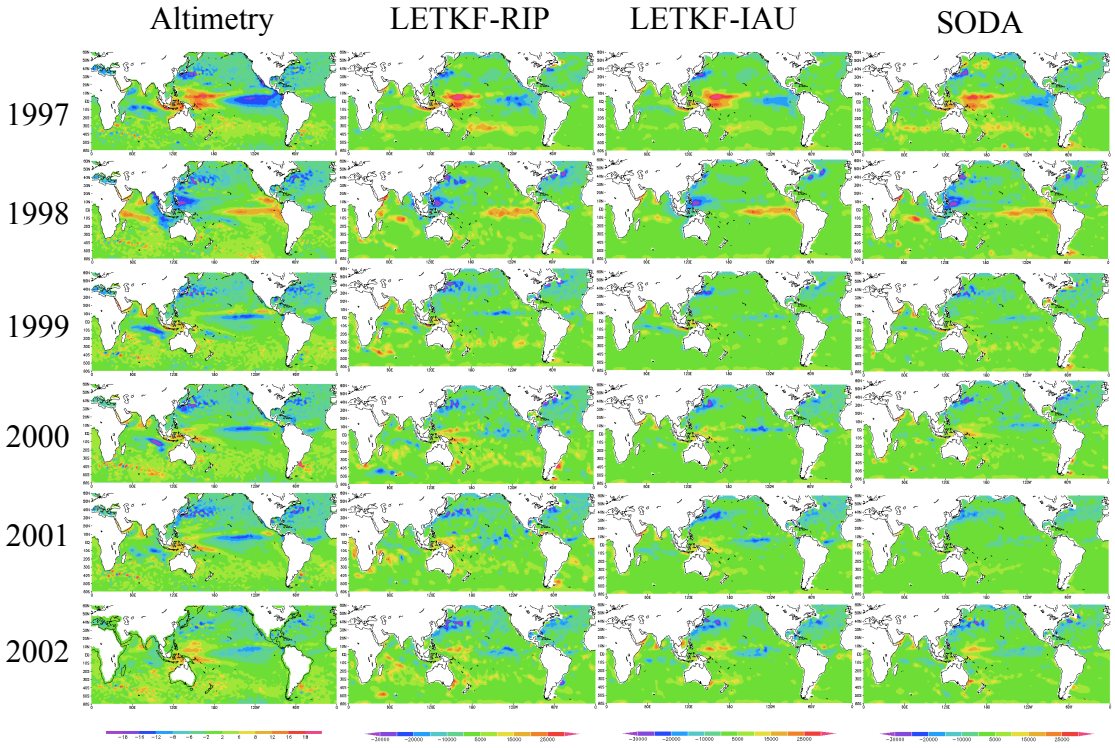

Figure 41. First (cos) Fourier terms for monthly averaged altimetry and 300 m heat content anomaly over the 6 years during 1997-2003. Altimetry is shown as cm perturbations and heat content as vertically integrated temperature perturbations from the time mean.

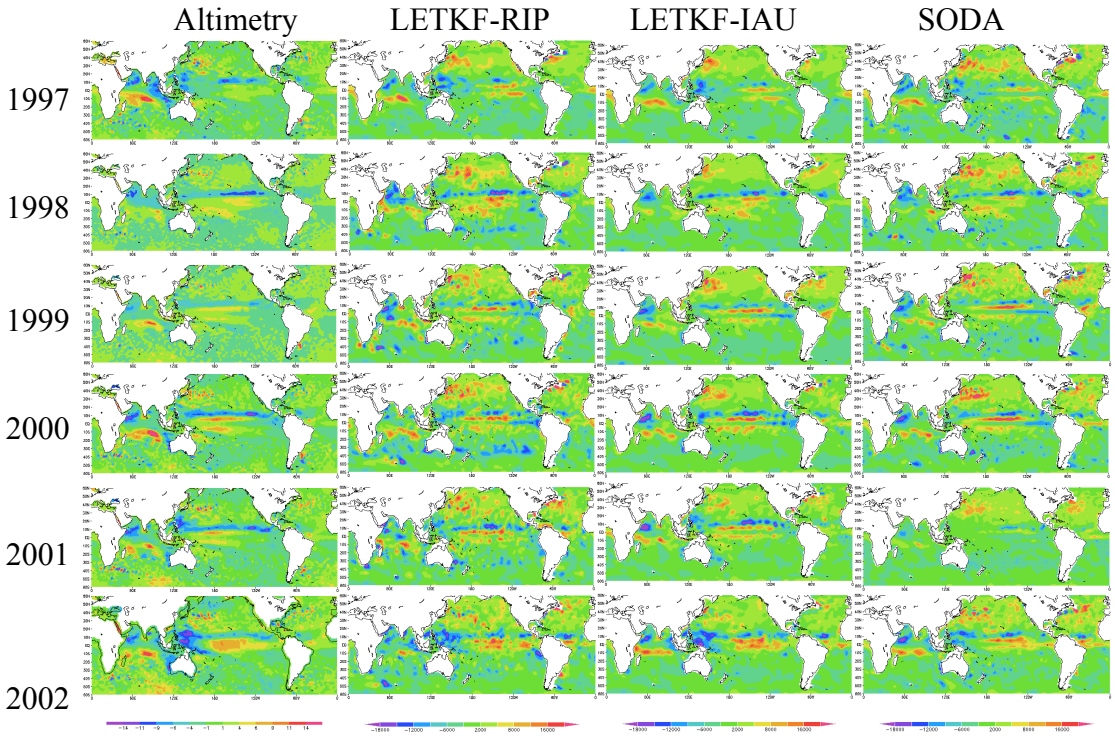

Figure 42. Second (sin) Fourier terms for monthly averaged altimetry and 300 m heat content anomaly over the 6 years during 1997-2003. Altimetry is shown as cm perturbations and heat content as vertically integrated temperature perturbations from the time mean.

3.7.3 Temperature-Salinity Relationship

Because temperature and salinity determine the density at a fixed pressure, an understanding of ocean circulation can come from examining the flow of water masses via the balance of temperature and salinity at various depths. Temperature and salinity obtained at the surface is usually retained by water masses as they are subducted into the ocean interior. The water masses are named and described in **Figure 43** based on their temperature-salinity relationship.

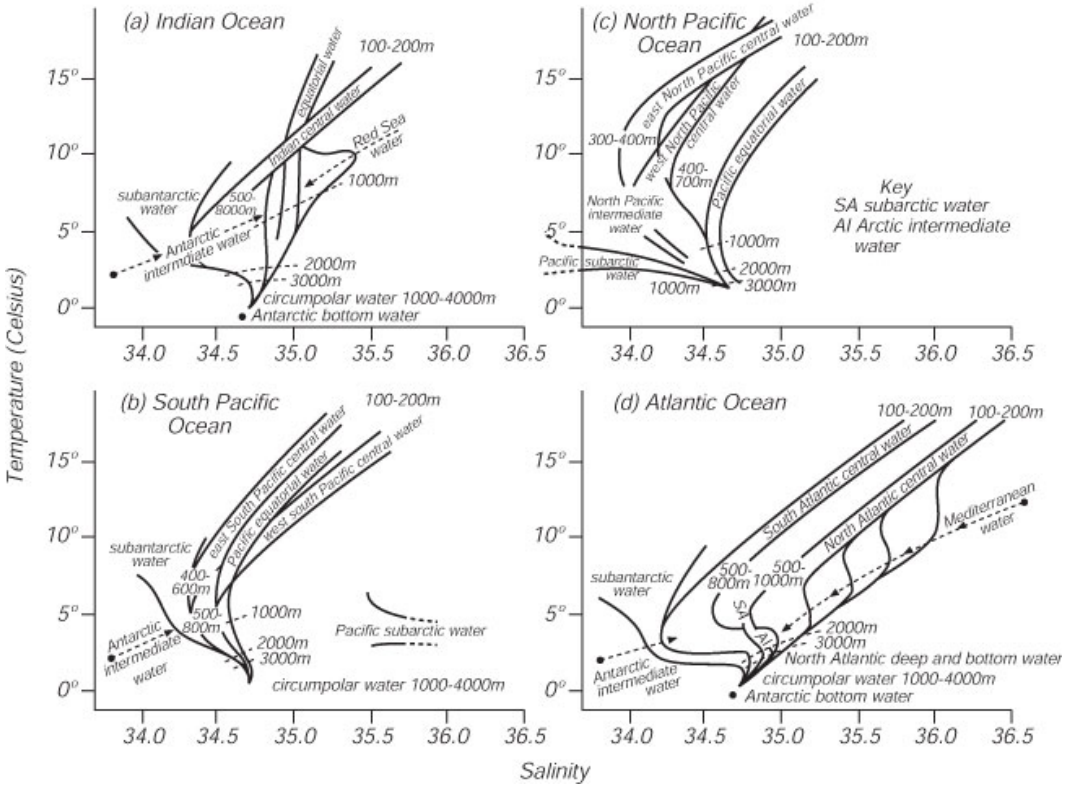

Figure 43. Temperature-Salinity relationships of water masses in various ocean basins (Tolmazin 85).

An investigation of selected water masses is shown for LETKF and SODA in **Figure 44**. Water masses identified theoretically in **Figure 43** are shown to be present in the analyses of both LETKF and SODA.

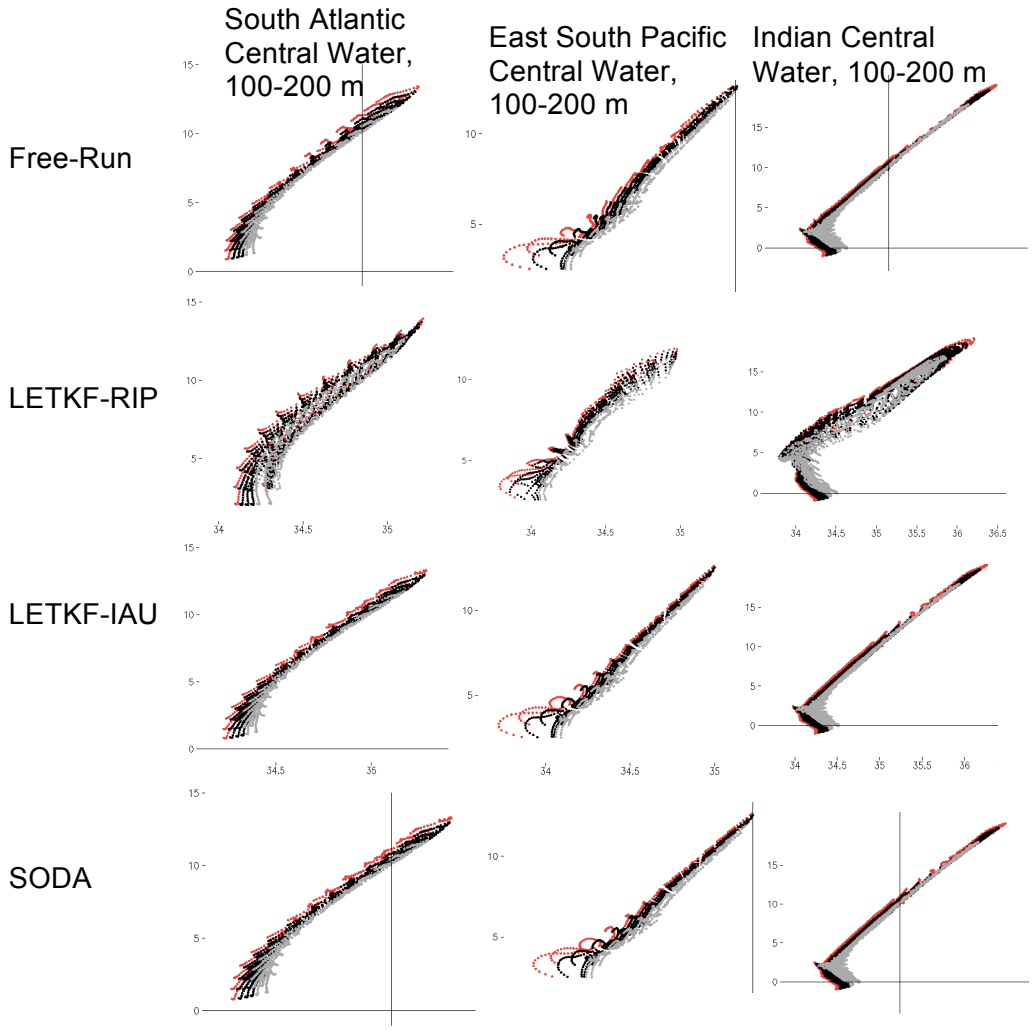

Figure 44. Temperature-Salinity relationships for overlapping analysis cycles LETKF-RIP (1/12/2001) and SODA (1/10/2001). Points are colored by vertical levels.

3.7.4 Temperature, Salinity and Velocity at the Equator

Differences in temperature and salinity anomalies from the SODA time mean are shown in **Figure 45** and **Figure 46**, respectively. Over time, LETKF-RIP establishes warmer areas in the east-central Pacific and east Atlantic compared with the other methods. There is a large difference between LETKF-RIP and SODA salinity analyses. On average, LETKF-RIP shows a much fresher Indian Ocean and saltier Pacific Ocean at the equator. Few salinity observations were present in the Pacific over the first half of the experiment period, thus most of this impact was due to covariance between temperature and salinity as identified by the ensemble. There was

a large increase in salinity coverage throughout the Indian Ocean by 2003, which coincides with an increase in salinity, thus indicating a potentially under-predicted salinity during the earlier period. Inflation is currently applied uniformly across all variables. It is expected variable-dependent inflation will remedy this situation. LETKF-IAU fell between LETKF-RIP and SODA.

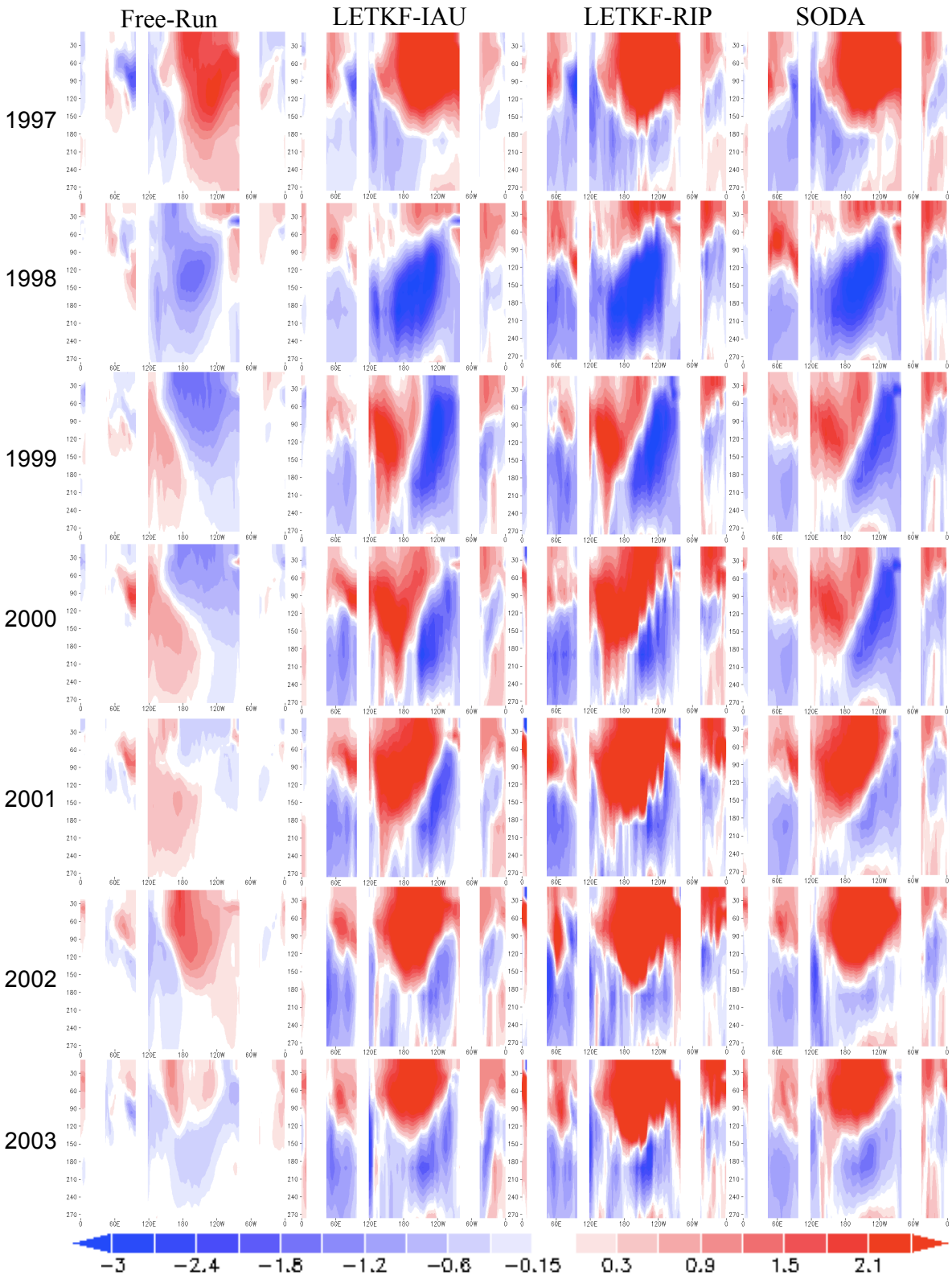

Figure 45. Average yearly temperature anomaly at the equator in the top 300 m during 1997-2003 for Free-Run, LETKF-IAU, LETKF-RIP, and SODA, versus Free-Run time mean.

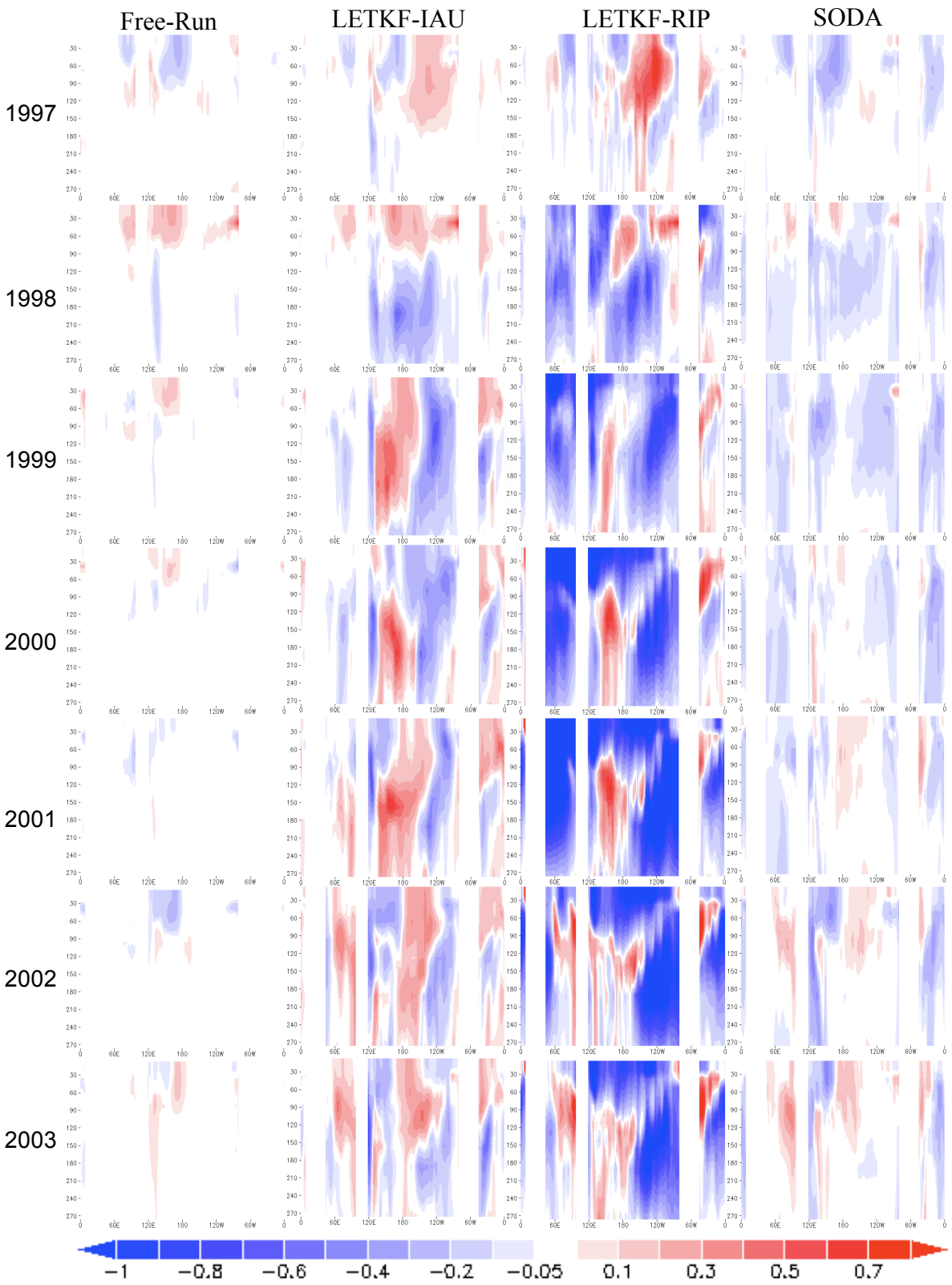

Figure 46. Average yearly salinity anomaly at the equator in the top 300 m during 1997-2003 for Free-Run, LETKF-IAU, LETKF-RIP, and SODA, versus Free-Run time mean.

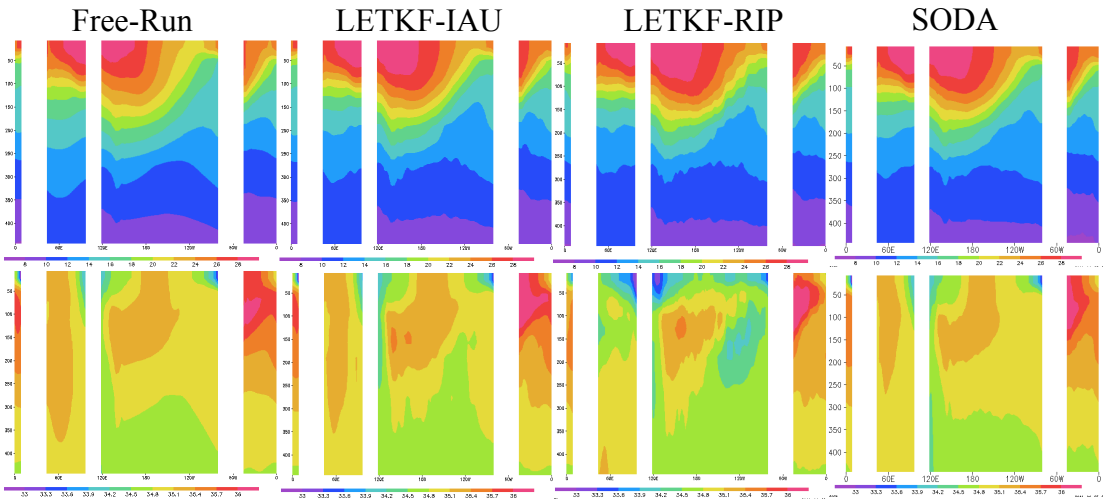

Figure 47. Mean temperature and salinity at the equator in the top 500 m over 1997-2003 for the Free-Run, LETKF-IAU, LETKF-RIP, and SODA.

The following are independent velocity observations compared with analysis results from LETKF and SODA. The velocity observations were not used in the analyses. Acoustic Doppler Current Profiler (ADCP) observations are averaged every 5-days. LETKF-RIP, LETKF-IAU, SODA and a Free-Run control are shown for comparison. The spatial and temporal patterns of LETKF-IAU appear to most closely match that of the ADCP observations. The wind forcing for the LETKF-RIP was changed in mid-2000, which may account for the stronger currents in the upper 90 m. It is clear in **Figure 48** that while the new wind forcing may improve the magnitude, it has negative effects on the zonal currents in relation to the observed values. LETKF-IAU and SODA share two features that may account for better matching of the observed currents. First, is the use of IAU, which applies a steady forcing rather than a sudden change to the system state at each analysis time – a possible disruption to the currents. The second is the use of a much larger window of observations, LETKF-IAU using a 30-day window and SODA using a 90-day window. Because LETKF-RIP is capable of making frequent adjustments, 4x per SODA analysis cycle and 12x per LETKF-IAU analysis cycle, it has higher frequency oscillations.

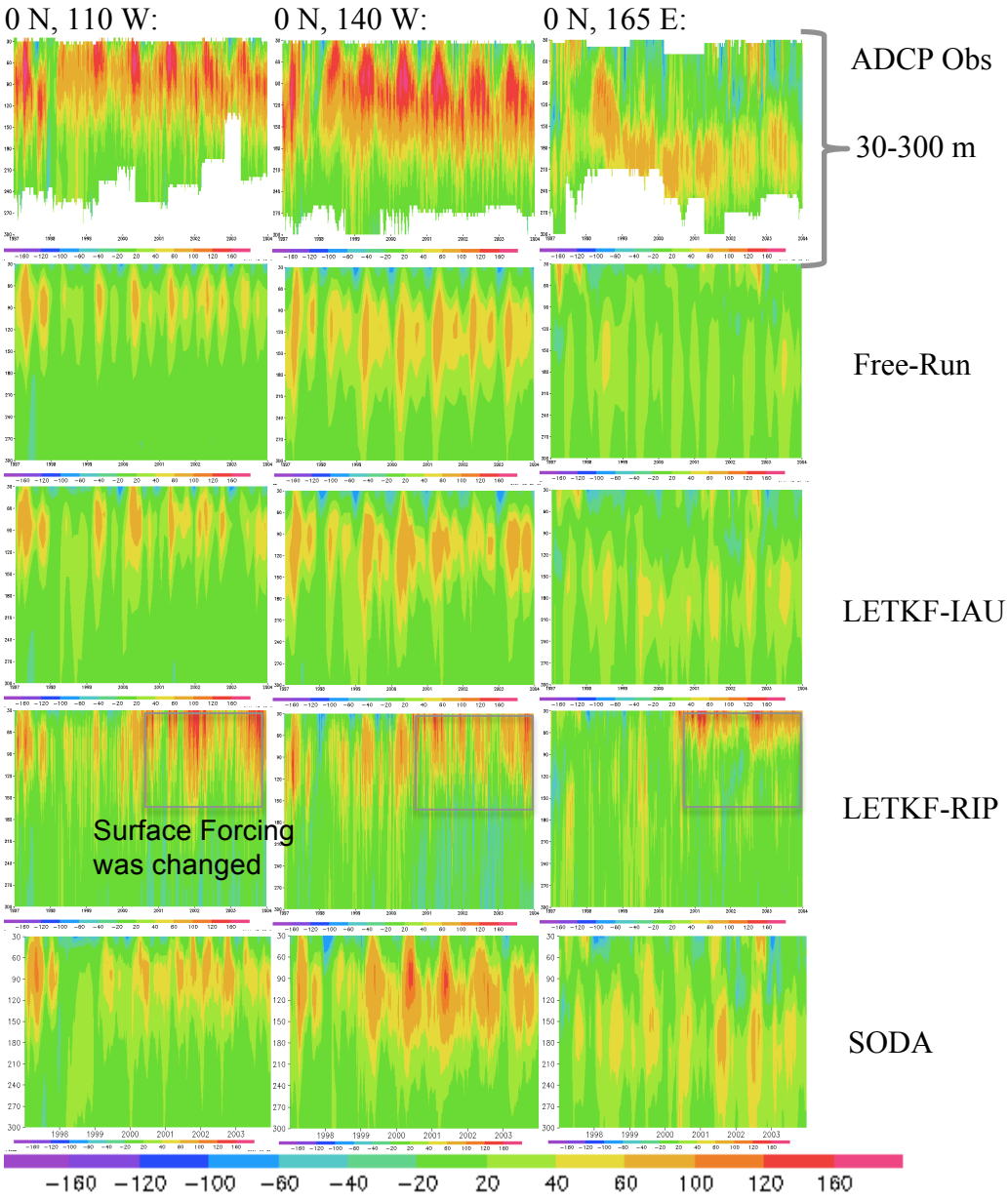

Figure 48. Zonal velocity (cm/s) in the top 300 m from Jan. 1 1997 to Jan. 1 2004.

3.7.5 Station Data

Time series results are shown for the top 500 meters at two stations: the Bermuda Station (30.55,-63.5) and the Aloha station (24.75,-158.0), for the observed values, LETKF-RIP, LETKF-IAU, and SODA (**Figure 49**).

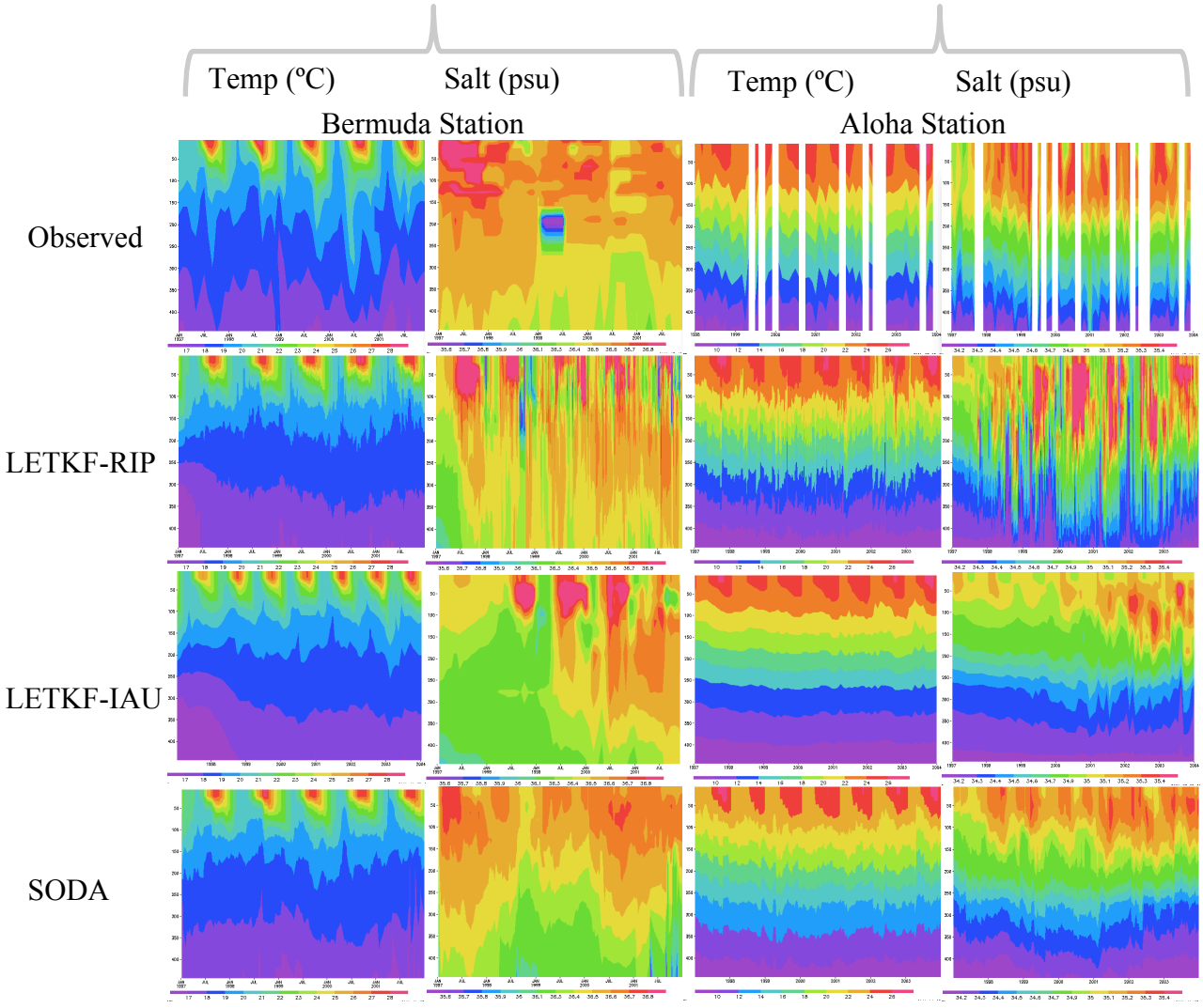

Figure 49. Temperature and Salinity at Station S (Bermuda; (30.55,-63.5)) from Jan. 1997 to Dec. 2001, and Aloha Station (24.75,-158.0) from Jan. 1997 to Dec. 2003 for the top 500 m.

The model is forced with climatological sea surface salinity. Thus the observations are providing the only influence of the true salinity state over time. The primary sources of observations are of temperature, and the methods adjust salinity relative to covariances between temperature and salinity state variables. Station S was

a recurring observation of both temperature and salinity within the experiment period and thus the results in **Figure 49** indicate the influence of this observation on the analysis. The state of model adjustment to the inclusion of the salinity data can give an estimation of the model spin-up. Due to the 30-day analysis cycle used, LETKF-IAU experiences a greater spin-up time (on the order of years). That can be compared to LETKF-RIP, which spins up over a few months.

3.7.6 Error estimation and Inflation

Small perturbations in the analysis ensemble grow nonlinearly during the forecast (using model integration). The areas with larger spread (standard deviation) represent greater uncertainty in the forecast.

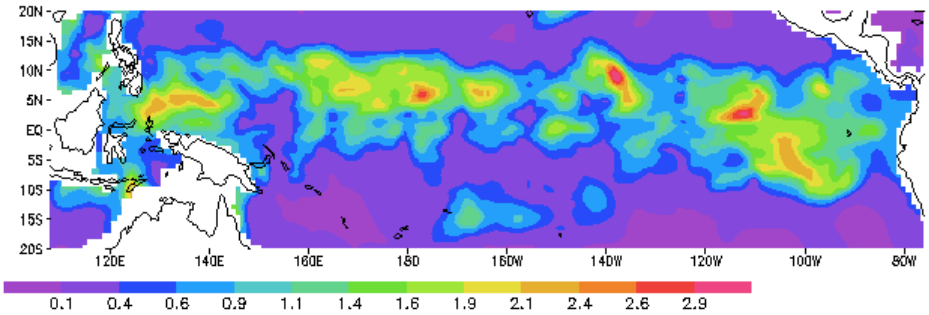

Figure 50. Background temperature ensemble spread at the surface in the Equatorial Pacific for Dec. 2 1997, LETKF-RIP.

As discussed in Section 2.6, the variation in the wind forcing fields has an impact on the ensemble spread at the 100 meter depth. Compare the ensemble spread at 100 meters using LETKF-IAU (Figure 51) with that of the perturbed wind forcing in Figure 18. There is similar growth in spread, particularly in the eastern Equatorial Pacific and Equatorial Atlantic.

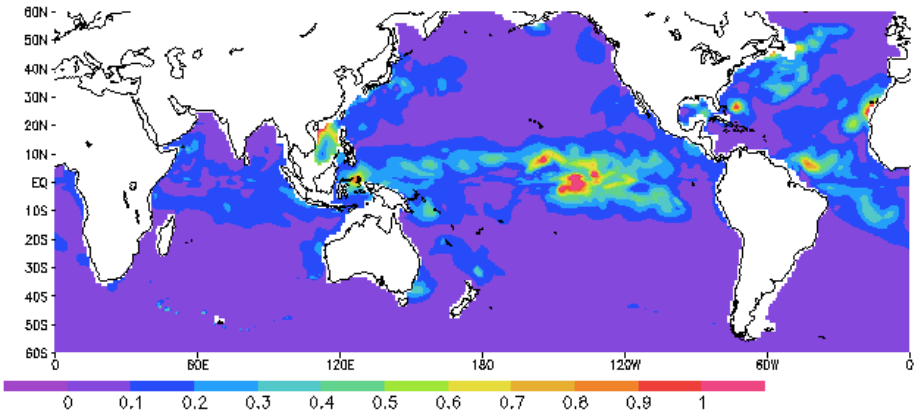

Figure 51. Background temperature ensemble spread at 100-meter depth for Sep. 3 1999, LETKF-IAU.

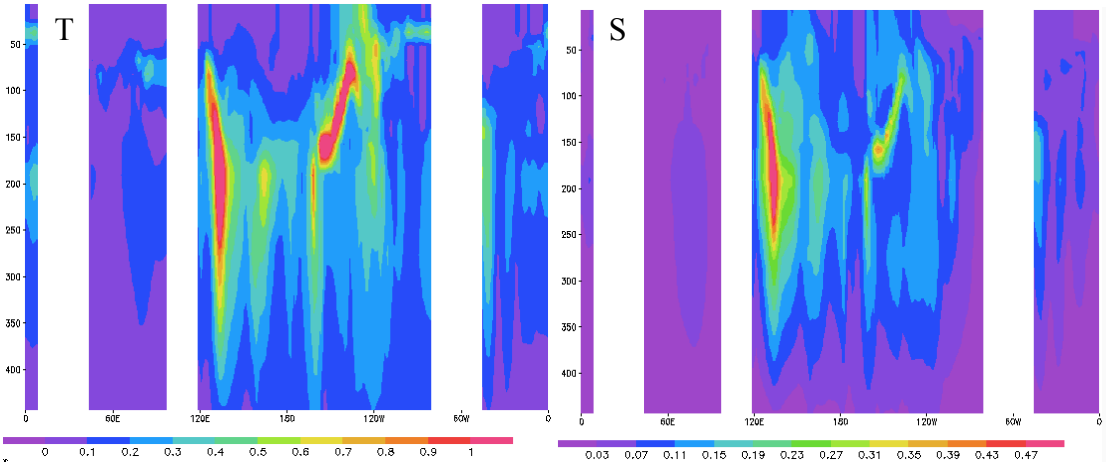

Figure 52. Background temperature and salinity ensemble spread at the equator from 0-500 meters for Sep. 3 1999, LETKF-IAU.

Larger inflation values indicate that the growth of the ensemble spread due to the model is not large enough in these locations, and artificial inflation of the ensemble spread is necessary to find an appropriate balance between forecast and observation errors.

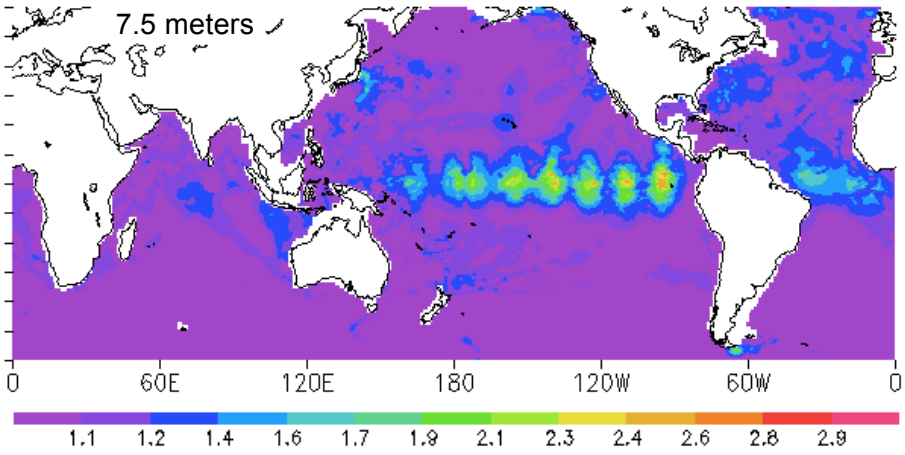

Figure 53. Inflation values generated by adaptive inflation at the surface on April 1, 2001, for the LETKF-RIP assimilation commencing Jan 1997 (values from 1 to 3 are equivalent to 0-200% inflation). Adaptive inflation typically decreases with depth.

Because the wind forcing variability has a dominant effect on the ensemble spread, an experiment was run to identify the dynamic variability that resulted from the coupled model/assimilation process. **Figure 54** shows this experiment using LETKF-EOW with α_w=0. The ensemble spread tends to be largest in areas devoid of observations. The uncertainty in these regions is grown by inflation.

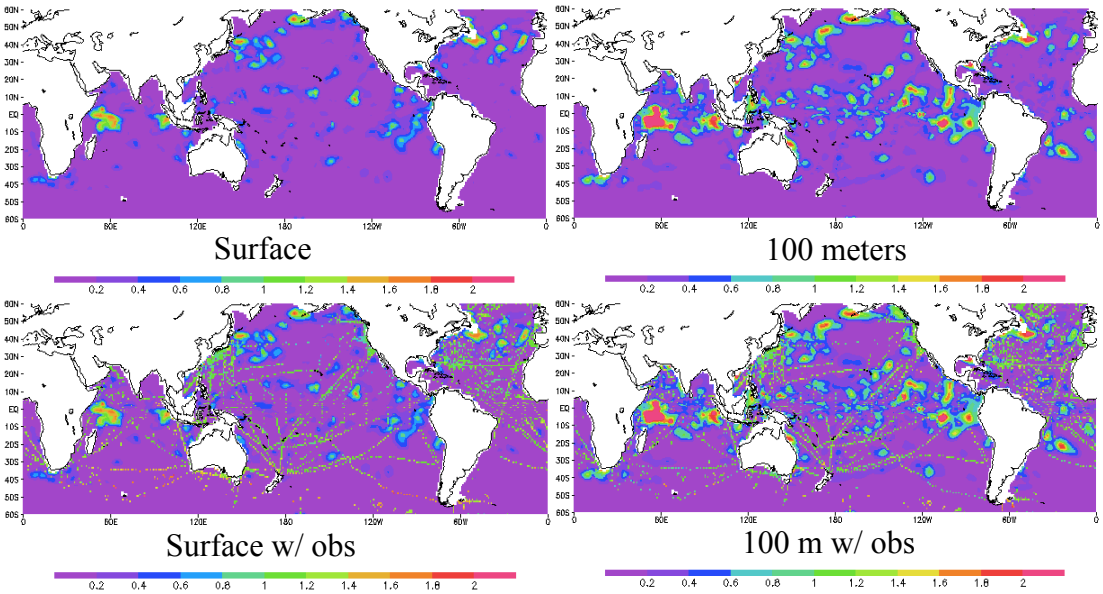

Figure 54. Ensemble Temperature (ºC) spread at the surface and 100 meters on Feb. 2, 1998, for the LETKF-EOW assimilation commencing Jan 1997. This case uses α_w=0, thus eliminating any impact on the spread from wind forcing.

3.8 Computational Performance

LETKF scales approximately linearly in the number of observations used. Thus, the use of the extended observation window for LETKF-EOW caused an excessive slowdown in runtime, as each analysis cycle used about 10x more observations than contained within the analysis cycle alone.

Another consideration is the analysis cycle length and the repetitive execution of LETKF in a given timeframe. For example, using a 5-day analysis cycle, one would perform 6-7 analyses in a month. While with a 30-day analysis cycle, only one analysis would be performed. Though the number of observations used may be identical, there is overhead to perform multiple cycles and thus the longer analysis cycle typically performed faster.

With LETKF-RIP there is continual repetition of the analysis cycles. In the experiments presented here, only one cycle of RIP was used. However, that has the impact of effectively doubling the runtime of LETKF.

In regards to any of the LETKF approaches, statistical accuracy can be sacrificed for faster runtimes by reducing the ensemble size. Cutting the ensemble size in half approximately cut runtime in half as well. It has been shown for LETKF-IAU that ensemble sizes as small as 10-members have performed comparably to 40-members for assimilation using the MOM2 system. Longer analysis cycle windows would typically require larger ensemble sizes to account for the greater opportunity for errors to grow with the model dynamics. The forced nature of the LETKF-IAU approach, as well as the use of surface forcing in general with ocean models, may mitigate the effect of such error growth.

Results are shown in **Figure 55** and **Figure 56** for the main experiments reported in this work. This is not intended to be a formal benchmarking of these assimilation approaches. The LETKF and SODA systems were run on different machines, with different architectures, and different compilers. The timings reported here are the actual times required to run the main experiments presented in this chapter and are given simply as a means of conveying the relative differences in computational effort for each approach.

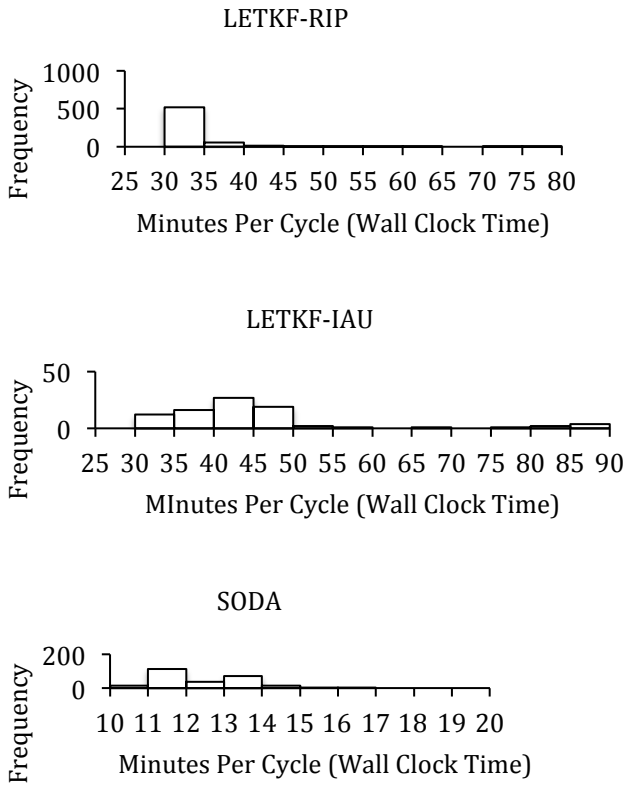

Figure 55. Wall Clock Time for analysis cycles of LETKF-RIP, LETKF-IAU and SODA. LETKF-RIP was parallelized on 40 processors, LETKF-IAU on 20 processors. SODA was run on a single processor.

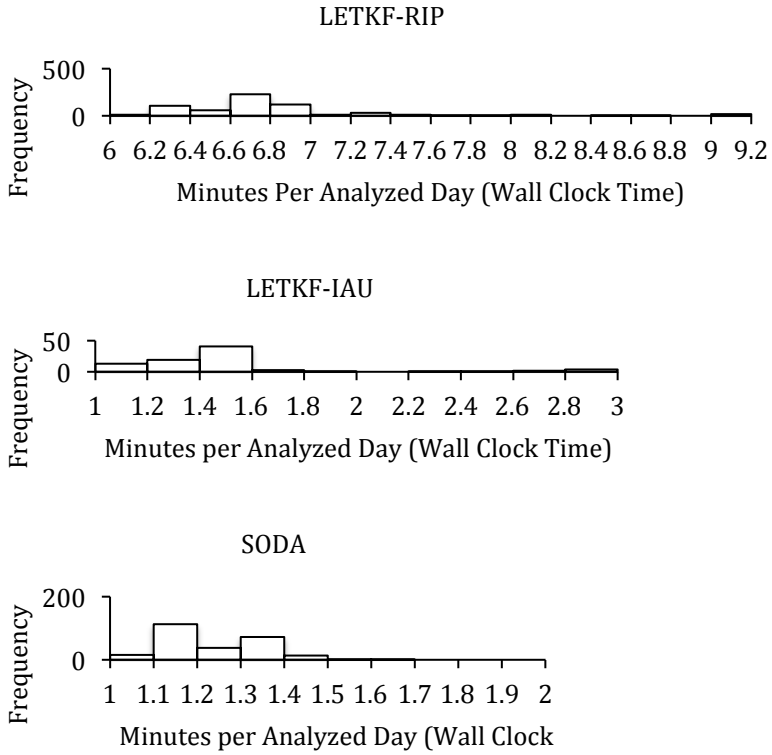

Figure 56. Due to the use of various analysis cycle, an average estimate of Wall Clock Time per analyzed day is reported here for a more accurate comparison. LETKF was parallelized on 20 (-IAU) or 40 (-RIP) processors. SODA was run on a single quad-core processor.

3.9 Conclusions for Chapter 3

LETKF-RIP outperforms SODA in the metric measuring RMS error between observations and forecast (O-F), and observations and analysis (O-A). LETKF-IAU outperforms SODA in terms of O-F when using an extended observation window, and performs on par with SODA after sufficient spin-up when limited to using observations only once.

LETKF provided much more adjustment to the salinity field than SODA. Based on comparisons with individual station observations as well as overall RMSDs, it appears LETKF, while noisier, is more accurately representing the salinity field.

Comparison to independent observations of equatorial velocity indicate that LETKF-IAU seems to best capture the temporal pattern at 3 specific locations in the equatorial Pacific. In general though, the model appears to underestimate the magnitude of the currents, which may be due to partially to the lack of model resolution and partially to the particular surface forcing used. LETKF-RIP, LETKF-IAU and SODA are also approximately equal in their upper level heat content correlation with satellite altimetry.

Due to the application of a steady forcing term to the model prognostic equations at all analyzed grid points, the IAU approach limited growth of nonlinear dynamical errors that are critical to establishing a representative ensemble. For this reason, IAU required much larger inflation to be effective. However, IAU countered effects of surface forcing and model bias so as to allow for longer analysis cycles. Therefore, there may be some circumstances where using IAU may be advantageous and preferable to RIP. However, RIP has the benefits that it allows the solution space to explore outside of the linear combination of ensemble members, and it only adjusts initial conditions and therefore allows the model more freedom in developing more realistic nonlinear growth of errors.

The choice of analysis cycle window length is related to the resolution of the forecasting model, the dynamics of the model, and the frequency and coverage of available observation data. Further study is required to identify the optimal analysis cycle length for capturing various dynamical features of the ocean. A combination of multiple analysis cycle lengths may be appropriate.

Ultimately, the perturbed surface wind fields forcing each ensemble member should be replaced with an atmospheric analysis ensemble. Results using various sources of surface forcing should be compared (e.g. NCEP and ECMWF). Potentially, a system that utilizes surface forcing from various sources could be advantageous.

The model will be upgraded from the basic lower fidelity MOM2 model to a more modern high fidelity model that resolves eddy dynamics. Greater variations in the model makeup may prove beneficial, either through variations in model parameters or via the use of completely different modeling schemes for subsets of the ensemble. The potential for a hybrid combination of LETKF and a 3D-Var variant such as SODA will be explored. This will allow benefits of both methods to be realized simultaneously.

Also, the observation set should be expanded with the inclusion of SST, altimetry, and velocity data, which were not used for the present analysis. Satellite data is numerous and will naturally increase run-time of the LETKF assimilation system. However, methods have been developed which analyze the impact of observations, which would allow clever selection of which observation to assimilate [KMK11]. Additional concepts in Chapter 4 will allow the selection of the nearest lowest-error observations. Many additional ideas for future work are listed in Chapter 5.

Chapter 4: New Algorithms and Designed Advances for the LETKF

4.1 Introduction

Some enhancements to the oceanic data assimilation system were developed but not yet implemented in the production system. These enhancements will be implemented in future versions of the oceanic 4D-LETKF data assimilation system. These include customization of the localization procedure utilizing computational geometry methods, reformulation of the LETKF core algorithm to provide faster performance and computational feasibility to using a non-diagonal observation covariance matrix **R**, as well as further treatment of observation error within the framework of a data assimilation system.

4.2 External Customized Localization and Preprocessing

As stated in [HKS06], the choice of which observations to use for the analysis at each grid point during localization is up to the user of the method. Therefore, a number of customized localization schemes were designed an implemented in a stand-alone test package. The main components of this package were the Boost library and the Computational Geometry Algorithms Library (CGAL), both written in C++. These were combined with I/O routines in a new package written in Fortran 90 to interface with the data files produced by the LETKF assimilation system and the MOM2 ocean model.

In this study, a horizontal localization radius was varied by latitude, and should be varied by depth as well. A generalized approach to localization is needed. For example, a technique was used to identify points in all regions that were occluded

by land. A general A* heuristic graph search algorithm was applied to a search graph generated over the ocean grid points of the MOM2 model grid using the model's ocean grid points as vertices and connecting adjacent points in the grid as edges. For each ocean grid point, a shortest path is computed to every observation within a localization radius using the A* graph search. The heuristic for the A* search is the great circle distance to the target observation. The shortest path will necessarily avoid land obstacles, both horizontally and vertically. The exact great circle distance path is also computed between the grid point and each observation. Using a tolerance parameter, *tol*, the ratio of the shortest path to the great circle distance can be used to identify paths that make too much deviation (e.g. around continents versus around islands). Paths that make no deviations around non-ocean grid points will have a ratio close to 1 (the ratio will not be exactly 1 due to the discrete nature of the search grid). This shortest path can be weighted based on the density of the water mass within which it is searching. Therefore, the localization range can follow areas of constant density.

A Delaunay triangulation of the observations, including the land points as obstacles, allows a fast lookup of the nearest observations to any given grid point. It may also be used to generate a Voronoi diagram, which can be used to quickly extrapolate point data to closest surrounding regions.

The spatial range search version of the k-nearest neighbor search is used with the Gaussian weighting function to scale observation errors in both space and time. By using the resultant observation errors as a distance metric, the *k*-nearest neighbors in the observation error space are found (independently for each grid point). Thus the

observations used in the LETKF analysis can be limited only to the k most accurate set of observations (i.e. the closest in the observation error space) for each grid point. For data rich areas, this will result in a small physical (and temporal radius for LETKF-EOW). For data sparse areas, this will result in a larger radius, limited by a maximum physical spatial localization value.

To create flexibility in the design and testing of localization methods, as well as greater efficiency and improved run-times of the LETKF system, the custom localization procedures were implemented externally to the LETKF system. This required the design of a procedure with the LETKF system that allowed for the input of a general customized localization for each grid point. For each grid point, the observation indices are input to identify which observations to use for the local assimilation. This approach requires very little computation during the run-time of LETKF.

4.3 Reformulation of LETKF Algorithm

Due to the scarcity of observations in the ocean assimilation application, the general assumption in [HKS06] that $k << l << m$ is not necessarily satisfied. The LETKF core algorithm is separated into 9 steps in [HKS06]. Steps 4, 5, 6 and 7 are performed locally at each grid point, thus approximately $O(m)$ times per analysis cycle. The main thrust of this reformulation is to remove the computation of (1) the eigenvalue decomposition, and (2) any computations of matrix inverses. Both are known to be costly operations, $O(k^3)$ floating point operations, and the latter is associated with numerical errors. [H02][http://www.cs.umd.edu/~oleary/c660/660mxfacthand.pdf] In situations where the number of observations in a localized region is large, the method

of Nearest Neighbors can be used to reduce the set of observations to the $n_l \leq k$ most relevant observations (after scaling observations spatially and temporally by Gaussian weighting factor, considering the statistics on the innovations).

4.3.1 Reformulation of LETKF Core Algorithm Step 5

Using the Sherman-Morrison-Woodbury (SMW) identity, applied to step 5 of the LETKF algorithm, we can eliminate the need to calculate the inverse of the matrix $\boldsymbol{R}$ as well as eliminate the overall inverse calculation of step 5:

(For reference, step 4 is to compute the k x l matrix $\boldsymbol{C} = (\mathbf{Y}^{b})^{T}\,\mathbf{R}^{-1}$)

Step 5 of LETKF is [HKS06]:

$$\tilde{P}^{a} = \left(\frac{(k-1)}{\rho} I + (Y^{b})^{T} R^{-1} Y^{b} \right)^{-1} \tag{25}$$

The SMW identity is given by,

$$\left(A + UCV\right)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}. \tag{26}$$

So by SMW,

$$\tilde{P}^{a} = \frac{\rho}{(k-1)} I - \frac{\rho}{(k-1)} (Y^{b})^{T} \left(R + \frac{1}{(k-1)} Y^{b} (Y^{b})^{T} \right)^{-1} Y^{b} \frac{\rho}{(k-1)}. \tag{27}$$

If we let,

$$B = \left(R + \frac{1}{(k-1)} Y^{b} (Y^{b})^{T} \right)^{-1} Y^{b}, \tag{28}$$

so that,

$$\tilde{P}^{a} = \frac{\rho}{(k-1)} I - \frac{\rho}{(k-1)} (Y^{b})^{T} B \frac{\rho}{(k-1)}, \tag{29}$$

then solving the linear system

$$\left(R + \frac{\rho}{(k-1)} Y^b (Y^b)^T\right) B = Y^b \tag{30}$$

for B will allow us to compute

$$\tilde{P}^a = \frac{\rho}{(k-1)} I - \frac{\rho^2}{(k-1)^2} (Y^b)^T B. \tag{31}$$

Since $\boldsymbol{Y}^b$ is k x l, computing $\boldsymbol{B}$ amounts to solving l linear systems of the form $\boldsymbol{A}\mathbf{x}=\mathbf{b}$, typically at a cost of $l^3/3$. Therefore depending on the relationship between k and l, e.g. if $3ck > l$, where c is the constant multiplier for the inverse computation or eigenvalue decomposition, both of cost $O(k^3)$, then this method may also be preferable. For example, if using an ensemble of size k=40, then this approach would be advantageous as long as the number of observations in a local region around the grid point was less than 120c. For the ocean, most areas satisfy this criterion, or can be made to satisfy it using super observations, selecting an appropriately small localization radius, or reducing the observation set to a subset of the most influential. Alternatively, the method of computation can be chosen at runtime independently for each grid point based on the relative size of k and l.

Making this modification to a basic implementation of the LETKF algorithm with the 3D-Lorenz model [K03] (with $m = k = l = 3$) in MATLAB gave just under 10x speedup over the inverse computation in the Step 5 process, and a 13% improvement over the eigenvalue decomposition inverse computation (which is used in the current LETKF implementation). If Step 7 is reformulated as below, then computing $\boldsymbol{R}^{-1}$ is not necessary anywhere in the algorithm and solving the linear system in Step 4 can be eliminated, replaced elsewhere in the algorithm by matrix-vector multiplications. It should be noted that $\boldsymbol{R}$ is simplified in the current LETKF

implementation as a diagonal matrix, implying no correlation between observations. However, correlations between observations do exist in real-world observing networks. This reformulation makes the inversion of $\boldsymbol{R}$ unnecessary and allows the use of a non-diagonal (though likely sparse) covariance matrix $\boldsymbol{R}$ computationally feasible.

4.3.2 Reformulation of LETKF Core Algorithm Step 6

At present, an eigenvalue decomposition, $O(k^3)$, is used to compute Steps 5 and 6. Though expensive for Step 5 compared with the alternate method proposed above, it is also utilized for the matrix symmetric square root computation in Step 6.

$$W^a = \left[(k-1)\tilde{P}^a\right]^{1/2} \tag{32}$$

Several alternative efficient methods exist for determining a matrix square root. The first is a modified Cholesky decomposition with pivoting (pivoting is required due to the matrix being semi-positive definite rather than positive definite) which would provide a unique triangular matrix as the weight matrix $\boldsymbol{W}^a$.

Tests with a basic implementation of the LETKF algorithm using a Lorenz96 model (with m =40, k =20, l =10) in Matlab suggested a 12% speed improvement when combining the Step 5 reformulation with a Cholesky decomposition in Step 6, as compared to using the eigenvalue decomposition to compute both Steps 5 and 6. Of course, the matrix square root computed by Cholesky is triangular, while the symmetric square root computed using the eigenvalue decomposition is symmetric and minimizes the mean square distance between $\boldsymbol{W}^a$ and the identity matrix. The latter may be preferable for smoothness between weight matrices of adjacent local grid points.

Another is the family of iterative methods derived from Newton's method for the matrix square root. These can be used to find the principal square root of $\boldsymbol{W}^{a}$ (a numerically stable version proposed by Higham [H86]), by producing the solution of the iterative matrix equation,

$$F(X) \equiv X^2 - (k-1)\tilde{P}^a = 0 \quad . \tag{33}$$

Further, iterative method can be seeded by previous iterations or results from neighboring grid points, potentially speeding convergence. Newton's method is $O(n^3 \log 2)$ per step, with local quadratic convergence. However, it has the drawbacks that there is instability and lack of global convergence. For those reasons, a variety of reformulations of Newton's iterative method have been derived that improve stability. The Denman-Beavers (DB) square root iteration is one such incarnation [DB76]. It is given by,

$$\begin{aligned} Y_0 &= A \\ Z_0 &= I \\ Y_{k+1} &= \frac{1}{2}\left(Y_k + Z_k^{-1}\right) \\ Z_{k+1} &= \frac{1}{2}\left(Z_k + Y_k^{-1}\right) \end{aligned} \tag{34}$$

For this iteration, $\mathbf{Y}_k$ converges quadratically to $\mathbf{A}^{1/2}$ and $\mathbf{Z}_k$ converges quadratically to $\mathbf{A}^{-1/2}$. Implementing using LU factorization, for symmetric positive definite matrices the operation counts are the same as using Cholesky factorizations. [H97] A stable variant that uses Cholesky with pivoting is given in [H97] as 'Algorithm 2'. Using one of these iterative methods and assuming the diagonal simplification of covariance matrix **R**, stages 5 and 6 can be combined using a single iterative solution method for computing $\boldsymbol{W}^{a}$.

Because $\tilde{P}^a = \left(\frac{(k-1)}{\rho} I + (Y^b)^T R^{-1} Y^b \right)^{-1}$, and $W^a = \left[(k-1)\tilde{P}^a \right]^{1/2}$, the DB iteration leads directly to

$$W^a = (k-1)^{1/2} \left(\frac{(k-1)}{\rho} I + (Y^b)^T R^{-1} Y^b \right)^{-1/2}. \tag{35}$$

The benefits and drawbacks of many approaches to computing the matrix square root are discussed in [H97]. If any of these are deemed viable as alternatives to the symmetric square root computation, the total of these modifications may provide significant speedup to the overall algorithm.

4.3.3 Reformulation of LETKF Core Algorithm Step 7

Stage 7 performs the computation,

$$\overline{w}^a = \tilde{P}^a \left(Y^b \right)^T R^{-1} \left(y^o - \overline{y}^b \right), \tag{36}$$

which is straightforward if $\boldsymbol{R}$ is diagonal. However, if it is not than this is better computed as the linear system,

$$Rz = \left(y^o - \overline{y}^b \right), \tag{37}$$

solving for z with sparse l x l matrix $\boldsymbol{R}$, and then computing from right to left to ensure only matrix-vector multiplications,

$$\overline{w}^a = \tilde{P}^a \left(Y^b \right)^T z. \tag{38}$$

In this formulation, the inverse computation is avoided and replaced by a simple linear system of the form **Ax**=**b**.

4.4 Observation Error

Within the LETKF system, there is a mechanism for quality control of input observations. When using raw observations, such a check is necessary to avoid adjusting the analysis toward an observation that is far from the mean background. LETKF uses a multiple of the observation error (e.g. 5.0-10.0 in the oceanic implementation) as a measure of the allowable distance from the mean background to incorporate this observation data. However, as the ocean data has been quality controlled against climatology and averaged into super observations, the quality of the observations is considered fairly good. As a result, the more likely explanation of the background mean being far from the observation is either a poor forecast or systematic model error. However, if the observation is kept, the combined large distance from the mean with the small prescribed observation error will cause an excessive correction to the background and potential filter divergence. For this reason, it is preferable to inflate the observation error and retain the observation. The analysis can then be pushed more gradually toward the observed values without an unnecessarily large analysis increment.

Such an approach is used explicitly in the current ocean implementation for LETKF-IAU and LETKF-RIP, and is achieved in effect for LETKF-EOW by the extended window of observations. For the latter case, because low weight is applied to temporally distant observations via increased observation error, these observations are used when the more accurate observations are too far from the background mean to be retained by the LETKF quality control mechanism.

Raw observations were binned into 1x1 degree super observations for the ocean analysis. In many locations the bins contain only a few observations. However

in other areas the observation coverage is very dense. The variance of the observed values within these bins gives a measure of the representativeness error within each bin. Therefore this variance information should be used to modify the prescribed error for each super observation.

Chapter 5: Conclusions and Future Research

Three versions of the Local Ensemble Transform Kalman Filter (LETKF-IAU, LETKF-RIP and LETKF-EOW) have been implemented and tested against a Free-Run baseline and an Optimal Interpolation system (SODA) benchmark. All assimilation methods improved upon the Free-Run baseline. Further, LETKF implementation now provide a means of generating error estimates for the analyzed or forecasted ocean state.

As noted by [http://www.image.ucar.edu/pub/DART/2011/JLA_sec_seattle.pdf], ensemble filters are optimal and exact when the following conditions are satisfied: the model is linear, the observation operator is linear, the observation error is Gaussian, the ensemble size is sufficiently large, and the filter is a variation of EAKF [A01]. For weather and geophysical applications, these conditions are most certainly not satisfied, and thus any application of an ensemble filter to such an area is the result of approximations to these conditions. This can lead to problems such as underestimating ensemble variance and overestimating correlations within the system. These problems have been mitigated by the use of inflation and localization in most EnKF systems (LETKF uses an adaptive inflation approach). However, these are ad hoc techniques designed to compensate for applying EnKF methods to cases in which the method's assumptions are violated. The effect of these approximations should be studied, as well as developing and implementing methods that rely on more general initial assumptions, as in [S95], [HS96] who utilize stochastic differential equation approaches.

LETKF-RIP performed best (defined by RMSD) in areas with the highest observation coverage, particularly if those areas had higher dynamic instability. LETKF-RIP performed better than SODA, with lower RMSD, in all regions with an adequate level of observation coverage. In regions and vertical levels with very limited observation coverage, LETKF-RIP was on par with SODA. However, it is noted that areas with higher observational coverage are typically the more dynamically active areas. The lower coverage areas, such as the deep ocean, have much lower variability. This may explain why the performance of LETKF was better in the areas with higher coverage.

Due to the sparse nature of the ocean observing system, reuse of the observations in the reanalysis was necessary to produce reasonable results. Both the SODA and LETKF-RIP systems reuse portions of the observation data to produce a smoothed reanalysis. Using its extended observation window SODA reuses each observation with varying weights approximately 9 times. Utilizing a similar extended window, LETKF-EOW reuses each observation temporally about 5 times. Using the RIP method with a single iteration, LETKF-RIP uses each observation 2 times. The LETKF-IAU and the basic LETKF use each observation only once.

For a reanalysis effort, it may be preferable to utilize a combination of the approaches presented. During the earlier historical period of very low observational coverage (which typically also had larger observational errors), an extended window of observations can be used to provide sufficient guidance to the dynamical model. Upon the induction of the Argo float system, a conversion to the RIP method would

be advantageous, and ultimately as the observation coverage increases further a single pass of LETKF may be sufficient.

Extending the analysis cycle to larger windows made more noticeable the effects of systematic model/forcing bias. In this case 'noticeable' means the magnitude of the bias errors were of a similar magnitude to the prescribed observations errors. Using the IAU, these effects were mitigated.

It is anticipated that the post-2007 complete observation coverage provided by the Argo float network (consisting of over 3000 floats) will greatly improve the LETKF analysis. Results showed that sparse observation data coverage had the most detrimental impact on analysis quality. However, as observation coverage increases, the computational runtime of the LETKF system also increases. Preliminary results show the Temperature RMSD for LETKF-RIP oscillating around 0.6 ºC between 2004-2007 (Figure 57).

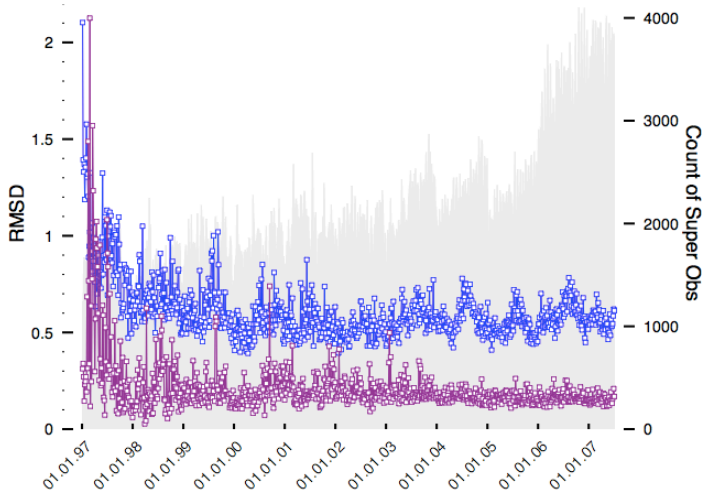

Figure 57. Preliminary results extending LETKF-RIP to the high-coverage Argo era, achieved by the end of 2006. Temperature RMSD have leveled off and salinity continues to improve as the global count of temperature super observations doubles from around 2000 in 2003 to around 4000 in 2007.

Results from this work enlightened many future areas of research in oceanic data assimilation. The remainder of this section describes some of those research topics.

Implementation of the algorithmic enhancements discussed in Chapter 4 will improve the speed of the LETKF analysis performed at each model grid point, thus multiplying the performance benefits by the number of degrees of freedom within the model. Implementation of the computational geometry pre-processing discussed in Chapter 4 will allow for greater flexibility in the localization scheme used by LETKF. Localization plays a critical part in the assimilation process and a well-designed localization method may provide significant improvements to the overall analysis product. The localization distance is closely tied to the length of the analysis cycle, the flow of the dynamics, the nonlinearity of the system, and the size of the ensemble. Practically speaking, a well-chosen localization scheme can reduce computations by limiting the number of observations used for the analysis at each grid point to only the most influential.

The LETKF system will be extended to more advanced ocean models with higher resolution model grids (the most natural extension being an upgrade from the MOM2 to the MOM4 model). As the resolution of the models increase, the smaller scale instabilities within the ocean will be resolved. The method for incorporating observations into the analysis must be revisited to determine whether to continue using binned super-observations or revert to raw observations. If binned, the appropriate scale for binning these observations must be determined. Along with the

upgraded model, all available datasets will be assimilated, such as SST, SSH, and SSS.

The adaptive grids of finite element methods present great opportunities for ensemble data assimilation. Finite element meshes can be designed to fit the error estimation and observation coverage of the system. Further, multiple meshes can be used for different subsets of the ensemble.

There is a tradeoff between increasing resolution and increasing the number of ensemble members. Keeping computational resources fixed, increasing model grid resolution will necessarily require a reduction in ensemble size. Various methods balancing this issue have been proposed, such as multi-resolution assimilation [YKH08], Quasi-Outer Loop (QOL) assimilation [KY08], and hybrid methods [WHW07].

Due to the limitations in computational resources, the ensemble size must remain fairly small. While the ensemble method is effective, the accounted-for uncertainty is limited to the space of linear combinations of ensemble members on a local scale. The 3D-Var and OI schemes do not suffer from this limitation. Thus there is potential for a hybrid system that utilizes the prediction of background error covariance from the ensemble method while using 3D-Var to explore the space outside of the ensemble space. A hybrid system may be developed that combines the benefits of the ensemble assimilation scheme with a variant of 3D-Var or SODA's OI approach.

One approach to handling the sparseness of the data prior to the Argo era would be to utilize climatology as an 'observation' at all points in which the

observations are absent, using historical variability as an error estimate for these 'observations'. Used in conjunction with adaptive inflation, this could ensure that the analyses remain within the realm of physically realistic solutions. This is not unlike hybrid methods that utilize similar information by combining ensemble covariance with a constant covariance matrix $\boldsymbol{B}$. Alternatively, a climatology perturbation could be subtracted from the ensemble perturbations and added as an additional ensemble member at each analysis cycle. Thus, as the ensemble mean diverges from the climatology, the background covariance would automatically increase. This could potentially reduce the reliance on inflation.

Further, the observation network is constantly improving due to increased coverage and upgraded technology. While much attention has been given to evolving forecast error covariance in ensemble forecasting, far less attention has been given to the evolving observation error covariance, though there has been some work [LKM], [KLM], [DBC05]. Effort should be made to analyze and estimate the changing nature of the observation data, specifically the spatial and temporal variations of the error profiles. With a highly trusted model, the observation outliers can be estimated from the ensemble spread in the EnKF. However, the observation error itself is an input to LETKF for each individual observation used in the assimilation. Thus, it may be preferable to do an independent statistical analysis of the observations, the sample variance in selected areas, regional and depth-dependent bias trends, and spatial and temporal correlations among observations. Alternatively, the online method proposed by [KLM] may be implemented and validated with such an independent statistical analysis.

Forecasting capability is limited due to the large influence of surface forcing and forcing error on the ocean models. In future work, the perturbed wind forcing ensemble will be replaced by an ensemble atmospheric assimilation product. Due to the strong dependence on surface forcing, and the presence of growing errors on multiple time scales [HKC09], an obvious extension is to a coupled atmospheric/oceanic assimilation system. This has implications for potentially extending predictions of the ENSO cycle beyond current temporal limitations.

A focused effort should be conducted to identify the optimal analysis cycle length, ensemble size, and model resolution for a desired decorrelation scale and a given set of computational resources. There is support for running multiple loops of LETKF for a number of reasons. (1) Using a nested loop that does a larger low-resolution ensemble and interpolates the analysis weights to a smaller high-resolution ensemble. (2) Using the Quasi-Outer Loop (QOL) to do a less-computationally intensive version of RIP. (3) Combining all of the above: Using a longer-range, low-resolution analysis cycle as an outer loop to capture the slow-moving dynamics and estimate model bias, while using a short-range, high-resolution analyses to do short-term predictions of detailed dynamics, while incorporating estimated model bias and long-term analysis weights.

Model bias was found to be an influencing factor, particularly near the surface. An examination of model bias and the bias correction facilitated by the IAU procedure will be examined. An adaptation of the methods of Danforth [DKM07], and Li [LKM09], will be applied to account for model bias.

As the error structure of the observation data set becomes more complicated, from a constant observation error used in previous work, to the vertical profile used in this work, to an adaptive grid-based estimates of observation error, a new metric for measuring analysis quality is required. Rather than simply observe the RMS distance between the model and observations, a metric should be developed that accepts the estimated observation error, and potentially the estimated forecast error, to identify whether the analysis and forecast are within the specified range of statistical parameters in both the observation space and the model space. Ideally, an estimate of the continuous error over the field would be preferred, possibly weighted in areas with the greatest dynamic instability so as to indicate better initial conditions for forecasts. Once such metrics are developed, they can then be incorporated back into data assimilation systems to further correct analyses.

The varying levels of temporal dynamics must be analyzed to allow for potential improvements in predicting large-scale oscillations such as ENSO and utilizing these predictions to improve long-term forecasting of small-scale oscillations such as eddy dynamics. This should be commenced by investigating techniques to address both slow and fast degrees of freedom. [http://www.cimms.caltech.edu/workshops_dir/w-ipam/presentations/posters/Hartmann/hartmann_poster.pdf]

A suite of automated diagnostic/analysis tools should be developed to accelerate the verification and validation of assimilation results. As with Desroziers' diagnostics [DBC05] that were developed into online analysis improvements [KLM], new automated diagnostics will likely lead to improvements in analysis methods.

Parameter variations for LETKF should be performed and documented in an organized matrix. In particular, the relationship between the length of the analysis cycle and the range of the localization is dependent on the ocean model dynamics and the modes that one attempts to capture. But, many other parameters should be examined, including the initial ensemble spread composition, the selection of the surface forcing, the choice of adaptive inflation forgetting factor, the amount of relaxation applied to the adaptive inflation, the amount of error added to each ensemble member for LETKF-RIP, independent vertical and horizontal localization for each variable, and the observation error.

In conclusion, the implementation of 4D-LETKF for the oceanic domain is a versatile and effective system for assimilating sparse observations with a coarse resolution model. While the number of ocean observations are increasing, particularly due to the introduction of the Argo floats and satellite measurements of surface parameters such as SST, SSH, and SSS [http://aquarius.nasa.gov/], increasing the model resolution will again render these observations relatively sparse relative to the model dynamics. For that reason, regardless of the increases in observation coverage, robust methods that handle sparse observations will always be needed.

Appendices

Fokker-Planck Equation

The term *Fokker-Planck Equation* refers to the work of two physicists on Brownian motion, namely A.D. Fokker [F14] and Max Planck [P17]. Kolmogorov obtained the same equation in his fundamental paper on markov processes [Ko31]. He referred to it as the second fundamental differential equation and it has since become known as the Kolmogorov forward equation, while his first fundamental equation is now called the Kolmogorov backward equation. Kolmogorov was not familiar with the papers of Fokker and Planck in 1931 but from 1934 he referred to the Fokker-Planck equation, though his backward equation had not previously appeared. [D89] [http://jeff560.tripod.com/f.html]

The general form of the Fokker-Planck equation can be derived in terms of "the evolution of a non-stationary probability distribution from a defined initial condition, or in terms of the evolution of the conditional probabilities for a stationary random process." [http://www.pma.caltech.edu/~mcc/Ph127/b/Lecture17.pdf]

For calculating the time-dependent probability distribution of an N-dimensional vector $\boldsymbol{x}$, the stochastic differential equation (SDE) is:

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t)d\boldsymbol{W}_t, \tag{39}$$

where $\boldsymbol{X}_t$ is an N-dimensional random variable and $\boldsymbol{W}_t$ is an M-dimensional Wiener process. The probability density $f(\boldsymbol{x},t)$ for the random variable $\boldsymbol{X}_t$ satisfies the Fokker–Planck equation with the drift coefficient $\boldsymbol{\mu}$ (representing the drift of the function

mean), and the diffusion coefficient $\boldsymbol{\sigma}$ (representing the diffusion of the function's standard deviation).

This equation can also be expressed as the sum of a Lebesgue and an Itō integral in the integral equation,

$$\boldsymbol{X}_{t+\delta} = \boldsymbol{X}_t + \int_t^{t+\delta} \boldsymbol{\mu}(\boldsymbol{X}_s, s)ds + \int_t^{t+\delta} \boldsymbol{\sigma}(\boldsymbol{X}_s, s)d\boldsymbol{W}_s. \quad (40)$$

This equation characterizes the behavior of the continuous time stochastic process X_t. This may be interpreted: in a time interval [t, t+δ], $\boldsymbol{X}_t$ changes its value by a normally distributed amount with expectation $\boldsymbol{\mu}(\boldsymbol{X}_t, t)\,\delta$ and standard deviation $\boldsymbol{\sigma}(\boldsymbol{X}_t, t)\,\delta$.

Because the increments of a Wiener process $\boldsymbol{W}_t$ are independent and normally distributed, the change in $\boldsymbol{X}_t$ is independent of the past behavior of the process. This stochastic process $\boldsymbol{X}_t$ is thus a Markov process. There also exist more general SDEs where the coefficients $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ depend on both current and previous values of the process $\boldsymbol{X}_t$ and possibly on values of other processes as well. In that case the solution process is called an Itō process.

Note: Depending on the application, the general equation is also sometime written in this way,

$$\frac{\partial f}{\partial t} = -\sum_{i=0}^{N} \frac{\partial}{\partial x_i}\left[D_i^1(x_1, \ldots, x_N)f\right] + \sum_{i=0}^{N}\sum_{j=0}^{N} \frac{\partial^2}{\partial x_i \partial x_j}\left[D_{ij}^2(x_1, \ldots, x_N)f\right] \quad (41)$$

where D^1 is the Itō drift vector and D^2 the diffusion tensor resulting from the presence of stochastic force,

$$D_i^1(\boldsymbol{x}, t) = \mu_i(\boldsymbol{x}, t), \quad (42)$$

$$D_{ij}^2(\boldsymbol{x}, t) = \frac{1}{2}\sum_k \sigma_{ik}(\boldsymbol{x}, t)\sigma_{jk}(\boldsymbol{x}, t). \quad (43)$$

Pseudo-code for coastal extrapolation

```
set mask to 0
set vcnt to 0
*Create a buffer that is larger than the map grid size
set buf3d(1:nlon+buffer,1:nlat+buffer,1:nlev+buffer) to 0

*Flutter grid up, down left, right and diagonal within range of buffer. (Use v3d to
store the boundary data during computation, then add it to v3d0 to get new grid)

*Put the data in the middle of the buffer
buf3d(buffer:nlon+(buffer-1),buffer:nlat+(buffer-1),buffer:nlev+(buffer-1)) = v3d0

*Sum all the values on this gridpoint that have water in them
do i=0,buffer
  do j=0,buffer
    do k=0,buffer
      where(v3d is ocean .and. buf3d(1+i:nlon+i,1+j:nlat+j,1+k:nlev+k) is positive)
          v3d= v3d+ buf3d(1+i:nlon+i,1+j:nlat+j,1+k:nlev+k)
          vcnt = vcnt + 1

 *If it intersects land (kmt<lev), then it's a boundary point (on the land side).
 *Average the flutter values to get an approximate extrapolation value. (being careful
not to divide by zero...)
 where(vcnt is positive)
    v3d= v3d/ vcnt

*Add back on the pre-existing values
 v3d = v3d0 + v3d
```

Analysis of RMSD as a performance metric

Root Mean Square Deviation (RMSD) is a common metric for examining the difference between a model and observations of the system that model is trying to predict. It is an aggregate measure of many local deviations, or residuals, between the observations and the model forecast. Because of its derivation, outliers have a larger impact on the RMSD.

When applied to two elements x and y, the various means satisfy:

$$H(x,y) \leq G(x,y) \leq E(x,y) \leq RMS(x,y), \tag{44}$$

representing the Harmonic, Geometric, Arithmetic, and Room Mean Square, respectively. Using this information, we may convert the forecast and truth to the observation space for comparison.

If we let $F=H\mathbf{x}^f$, $T=H\mathbf{x}^t$, $O=\mathbf{y}^o$, and $\mathbf{e}^t = (T-F)$, $\mathbf{e}^f = (O-F)$, $\mathbf{e}^o = (O-T)$, then $\mathbf{e}^t = \mathbf{e}^f - \mathbf{e}^o$ represents the error between the truth and the forecast, and

$$E(\mathbf{e}^t) = E(\mathbf{e}^f - \mathbf{e}^o) = E(\mathbf{e}^f) - E(\mathbf{e}^o) \leq E(\mathbf{e}^f) + |\beta^o| \leq RMS(\mathbf{e}^f) + |\beta^o|, \quad (45)$$

$$\text{or, } E(\mathbf{e}^t) \leq RMS(\mathbf{e}^f) + |\beta^o| \quad (46)$$

That is, within an error of β^o (the mean observation bias), the RMSD bounds the mean error of the forecast. If there were sufficient observations to satisfy the observability condition, then this would guarantee the accuracy of the forecast. Unfortunately that is not typically the case for a sparsely observed system, and further investigation must be done to verify that the assimilation system is generating accurate forecasts.

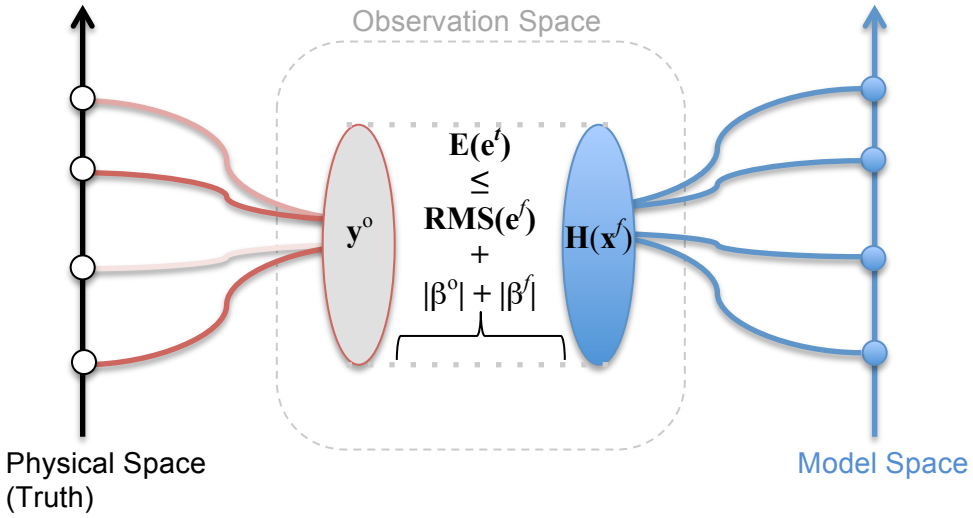

Figure 58. Diagram of the comparison between truth and forecast made using the RMSD measure.

Decorrelation Scale Length

In the spatial dimension, the decorrelation scale length, sometimes called the *e*-folding scale length, is analogous to the localization radius used by LETKF. A latitude-dependent fixed value was used based on previous studies, e.g. the method of ISHII described in [CS08], and adjusted slightly based on experimental results. In some cases, the temporal decorrelation length has been defined as the length of time before the autocorrelation function switches regimes from positive to negative. The temporal decorrelation length is tied closely to the length of the analysis cycle window used by LETKF.

Spatial decorrelation scale lengths tend to vary proportionately to the local Rossby deformation radius for ocean color data in the North Atlantic [http://remotesensing.whoi.edu/~david/decorr/decorr_text.html]. In the study [RRC06], yearly SST decorrelation scale lengths were between 70 and 125 days. [GK96] find that the Geosat altimeter data indicate a spatial decorrelation scale of 85 km and a temporal *e*-folding scale of 34 days in the Southern Ocean, particularly focused on the Antarctic Circumpolar Current (ACC). Work is currently underway to identify spatial decorrelation scales based on statistical analysis of Argo data [http://www.euro-argo.eu/content/download/21530/310958/file/17.%20Lorna%20McLean.pdf].

In the instance that the LETKF analysis window is larger than the temporal decorrelation scale, additional measures must be put into place to ensure proper treatment of the observations. For example, temporal scaling should be applied to the errors for observations within the analysis window in this case. It should be noted that because both the true physical system and the model are both dynamical systems, it is

possible that both have different spatial and temporal decorrelation scales. For a method such as LETKF-IAU, which is heavily forced by analysis increments throughout the forecast stage, the model's temporal decorrelation scale length may be lengthened.

LETKF with the Extended Observation Window (LETKF-EOW)

Due to the sparse distribution of observations throughout the ocean, particularly before the implementation of the Argo network, a rolling window of observations was used extending weeks before and after the analysis cycle. Reusing the observations has the effect of adding weight to the observations in the analysis. In essence, this has an effect similar to inflation, a common technique used to compensate for the effects of model error, the small sample size of the ensemble, and the linear approximations used by the Kalman filter. For this reason, a Gaussian weighting function was applied to the observation errors, increasing the error from the endpoints of the analysis window to +/- 25 days beyond the assimilation window. A 25-day window was chosen rather than SODA's 45-day window to balance the added benefit of the extended window with computational run-time.

The temporal error scaling was applied as: $r_s = r_o e^{\frac{1}{4}\left(\frac{d_o - d_a}{\sigma_t}\right)^2}$, where r_s is the scaled observation error, r_o is the original prescribed observation error, d_o is the date of the observation and d_a is the date of the analysis, and σ_t is the sigma scale of the Gaussian weighting. The inverse of r_s is applied in the LETKF algorithm, effectively weighting the observations far from the analysis time very low in the analysis. This

weighting is applied similarly to the spatial weighting already implemented within LETKF.

With $\sigma_t = 10.0$, this allowed for a range of weighting applied to r_o between 1 and 4.8. Thus observations that were +/- 25 days from the analysis time had approximately 5-times the prescribed error compared with the analysis cycle containing the observation, prior to spatial weighting. After applying spatial weighting, this prescribed error increased yet further. Thus the observations at the tail ends of the window are only a very loose constraint on the analysis. Yet, they provide the benefit of keeping the model within the range of observed quantities in an otherwise very sparsely observed system.

It should be noted that in the theory of linear Kalman filters, observations should not be reused, following a common approach in ocean data assimilation [ICG97]. For this reason, the LETKF-IAU and LETKF-RIP methods are preferred. A number of other ocean assimilations have taken this approach, however, including [CCC00a] [KR08] [OBG08] [HXB08] [SH09] [DT10]. While the observations are indeed reused by LETKF-EOW, they are applied within the given analysis cycle, against a different background time for each analysis, and with different prescribed observation errors.

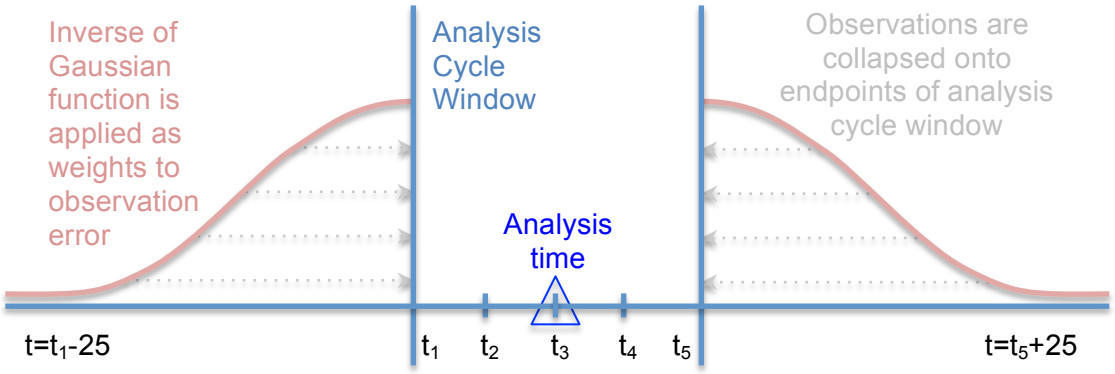

Figure 59. Diagram of Extended Observation Window used by LETKF-EOW and how the weights are applied to the observation errors. Because the errors on these observations are much larger after this weighting process, they have little impact on the analysis and serve only to maintain the trajectory within a range (e.g. keeping temperature within +/- 5 ºC) of the near past and near future observed values.

In some regions there were numerous overlapping observations at various times within the extended window. For regions in which the background was close to the observations at the analysis time, this weighted most the impact from observations that occurred within the analysis cycle. However for regions in which the background was far from the observations, the quality control in LETKF would drop these observations (because they were too far from the mean state value) and instead the observations outside of the analysis cycle window with larger prescribed error would be the only observations to impact the analysis in this region.

The original approach used in LETKF was to remove any observations that were 5 standard deviations of **R**, ($5\sigma_o$), away from the mean background state. If the assimilation settles on the correct trajectory then this is an adequate form of quality control. However, if that is not the case then this 'quality control' actually removes the highest quality observations and retains only observations with large error. Unfortunately, if kept these observations could cause a trajectory shift that is too large for the filter to handle. In the LETKF-IAU and LETKF-RIP methods, this approach

was modified. Instead of removing observations that did not satisfy the $5\sigma_o$ requirement, the value of the observation error σ_o was increased to satisfy the requirement instead, thus allowing the higher quality observations to be retained but only gradually pushing the filter toward the observed trajectory.

In a sense, the extended window acts as a loose boundary condition on the LETKF analysis. The start and end of the analysis cycle window are constrained by the observations extending beyond the analysis cycle window, with increased error applied to those observations.

Prior to the implementation of the extended time window, experiments showed large growth in RMS observation minus background [o-b] from one cycle to the next. This is primarily due to the sparse and changing nature of the observation network. While the analysis corrects the areas at locations for one cycle, in the next cycle the locations of the observations are different and have not yet been corrected, thus registering a larger [O-F] RMSD. In comparison, SODA uses an extended time window and the background RMSD closely follows the analysis RMSD.

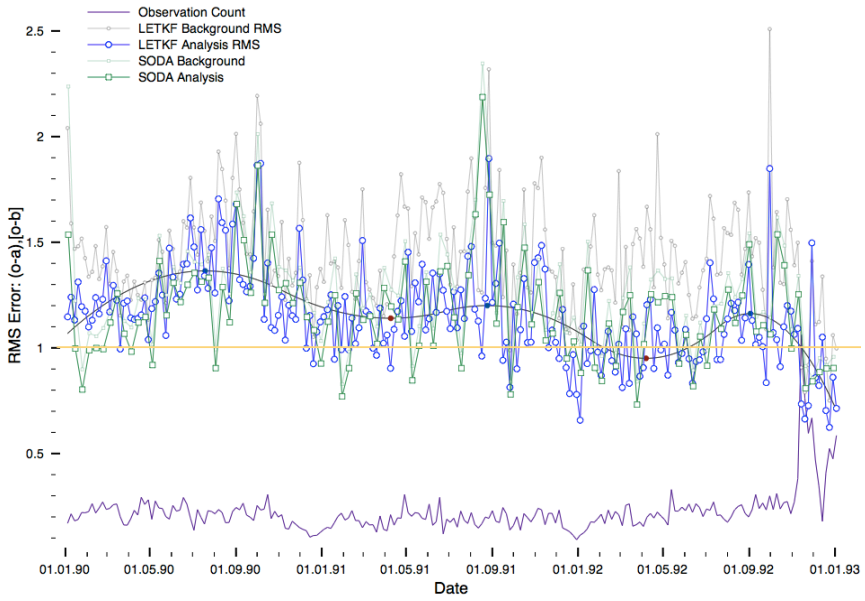

Figure 60. RMSDs in Temperature between observations and the background and analysis fields, with observation error at 1º C. The overall trend of the observations minus LETKF analysis RMSD was computed via a spline method. The observation count is scaled as a percentage of 10,000.

The 1997-1998 El Niño was the strongest on record, and had one of the fastest onsets. Effects of the event were seen as early as August-October 1997. [http://www.pmel.noaa.gov/tao/elnino/faq.html] Thus, the first period of interest that is to be examined spans from January 1, 1997 to January 1, 1999. The salinity observations during this period provide relatively low coverage compared with the temperature observation coverage.

The following figures report results using a configuration of LETKF (LETKF-EOW) that closely mirrors SODA. Both use an extended window of observations extending beyond the analysis cycle. Both use the IAU method of incremental updates within the model's prognostic equations. LETKF uses a 40-member ensemble with perturbations applied to each member's wind forcing. In these cases, SODA assimilates only temperature and relaxed observation errors (causing an improvement in SODA's temperature RMSD, but a higher salinity RMSD). **Figure 61**

and **Figure 62** show the (O-F) and (O-A) RMSDs for Temperature and Salinity. The global RMSD for temperature is lower overall for LETKF versus the SODA baseline. The RMSD for salinity fluctuates during spin up, then eventually levels off. Data points are shown at the analysis time; at day 3 of the 5-day analysis cycle for LETKF and at day 6 of the 10-day analysis cycle for SODA. Only observations corresponding to the day of the analysis were used for reporting the RMS departures.

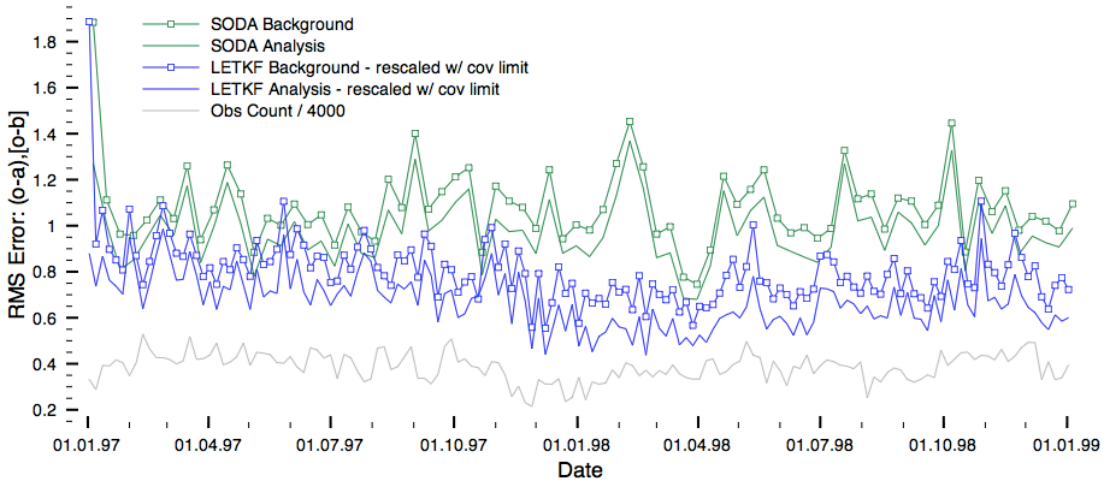

Figure 61. Temperature RMSDs for SODA and LETKF-EOW.

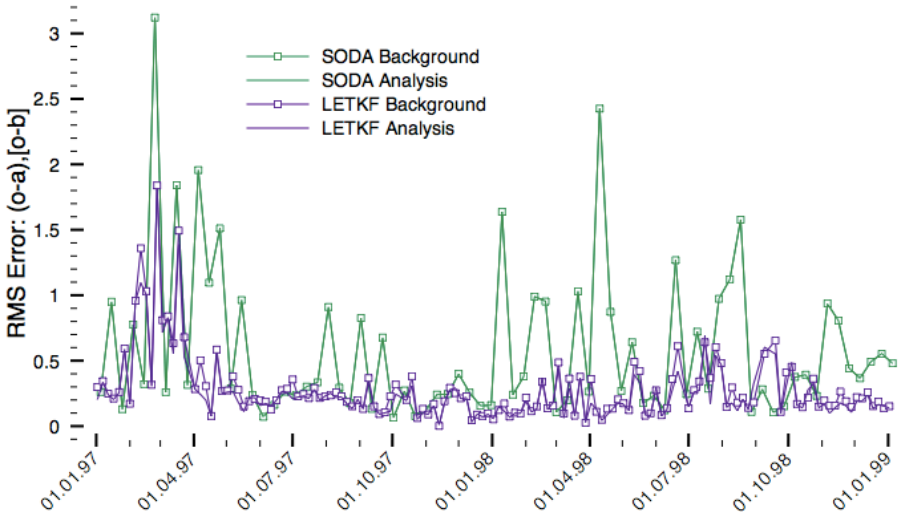

Figure 62. Salinity RMSDs for SODA and LETKF-EOW.

As shown in **Figure 63**, the primary benefit of LETKF-EOW is realized in the data-rich equatorial areas. There is some marginal improvement in the North Pacific. In other areas in which historical observation coverage is sparser, LETKF-EOW performed proportionately with SODA.

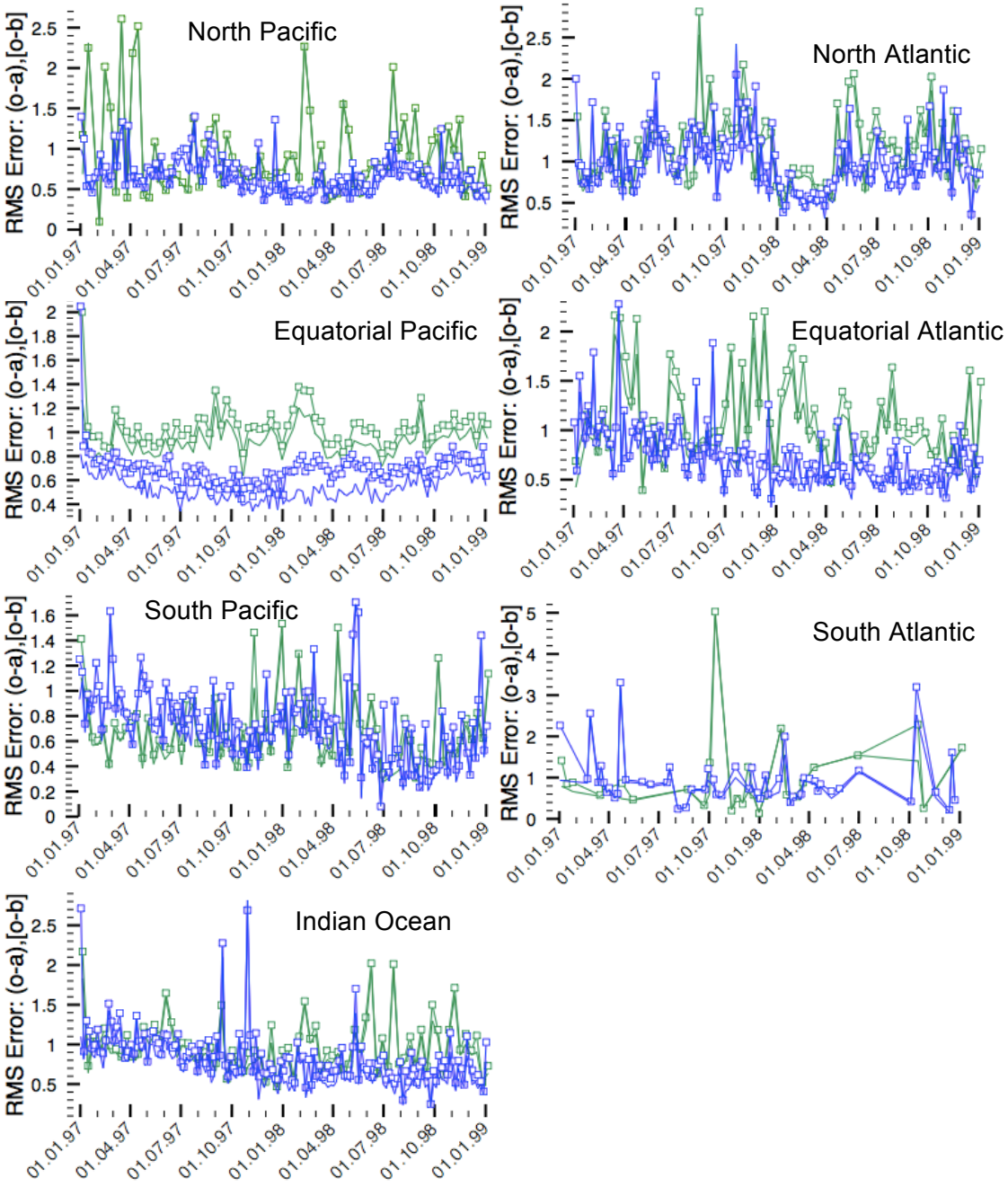

Figure 63. Regional breakdown of RMS temperature differences for SODA and LETKF-EOW. (notation as in Figure 61)

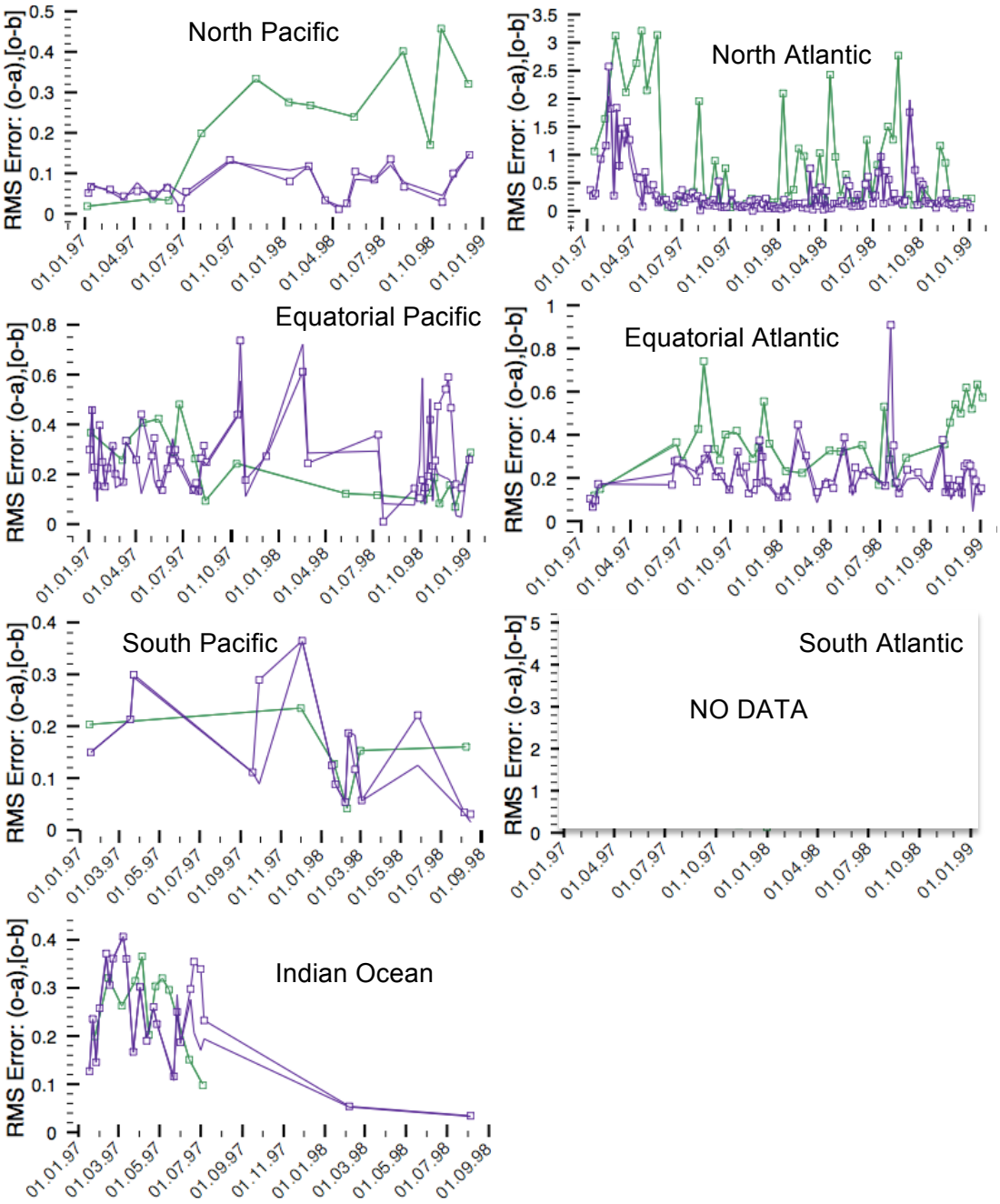

Figure 64. Regional breakdown of RMS salinity differences for SODA and LETKF-EOW

Ensemble methods typically perform better when the ensemble size is large enough to adequately sample the dimensions of instability in the model. However, there is a tradeoff to increasing the ensemble size, as it requires either increasing computational runtime of the overall assimilation procedure, or reducing the

resolution of the model to maintain this runtime. A variety of ensemble sizes have been used to test the assimilation quality, with results shown in **Figure 65** and **Figure 66**. As one would expect, the errors improve when increasing the size of the ensemble. Notably, there is still a competitive advantage with LETKF even when using only 10 ensemble members.

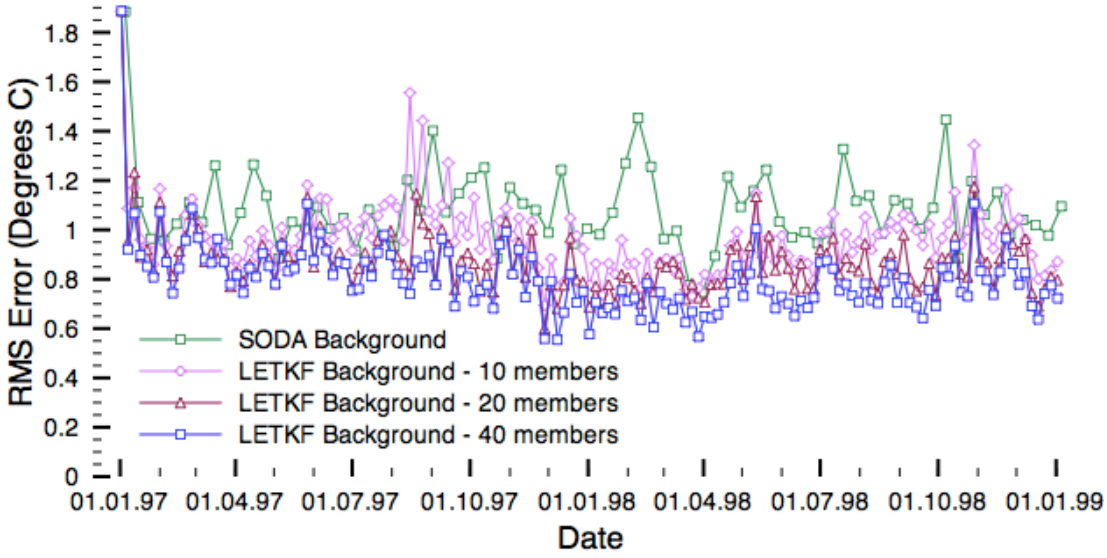

Figure 65. RMSDs in Temperature (o-b) for SODA and LETKF-EOW using 10, 20 and 40 ensemble members.

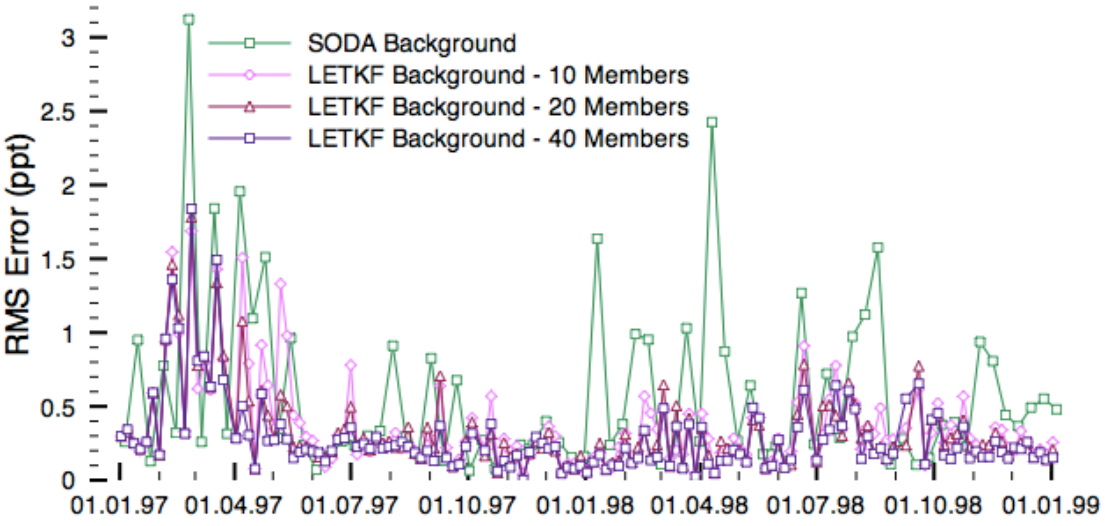

Figure 66. RMS errors in Salinity (o-b) for SODA and LETKF-EOW using 10, 20 and 40 ensemble members.

With the configuration used so far, both LETKF-EOW and SODA were set up for solving a reanalysis problem. In the interest of addressing the forecasting problem, a variety of additional configurations were used for both methods. In **Figure 67**, both methods were run with a 1-sided window of observations, subtracting all future observation data from the analyses. Using half the observations resulted in naturally higher RMSDs, but still results show relatively even performance between the two methods.

Figure 67. RMSDs in Temperature (o-b) for SODA and LETKF-EOW using a 1-sided observation window, to simulate forecasting.

Next, the analyses generated using the two-sided extended observation window were used to initialize 1-month forecasts starting from each analysis time. **Figure 68** shows both the short-term forecasts corresponding with the analysis cycle, and the 1-month forecasts generated by those analyses. As is expected, the longer forecasts have higher RMS error. Using a 10-member ensemble, LETKF and SODA show similar performance.

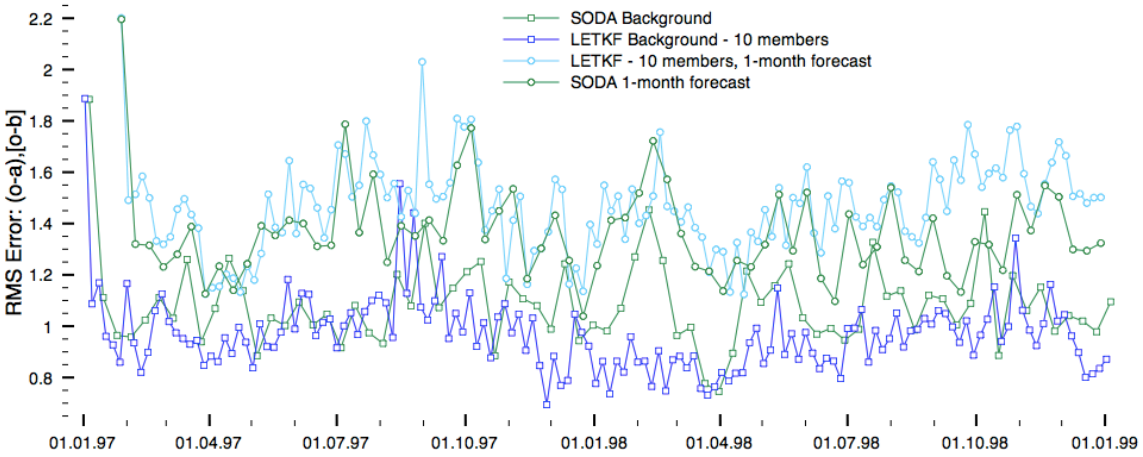

Figure 68. Comparing RMSDs in Temperature (o-b) for SODA and LETKF-EOW using 1-month forecast starting from respective backgrounds generated from 2-sided extended observation window.

Argo deployments began in the year 2000. Thus the period from 2001-2003 is studied to determine the impact of growing observation coverage on the oceanic LETKF assimilation system. Though not yet utilized, the LETKF system has the ability to assimilate the Argo velocity observation data as well as the presently assimilated temperature and salinity. As can be seen in **Figure 69**, there is a dramatic increase in global salinity observations, concentrated primarily in the North Atlantic and Indian Oceans.

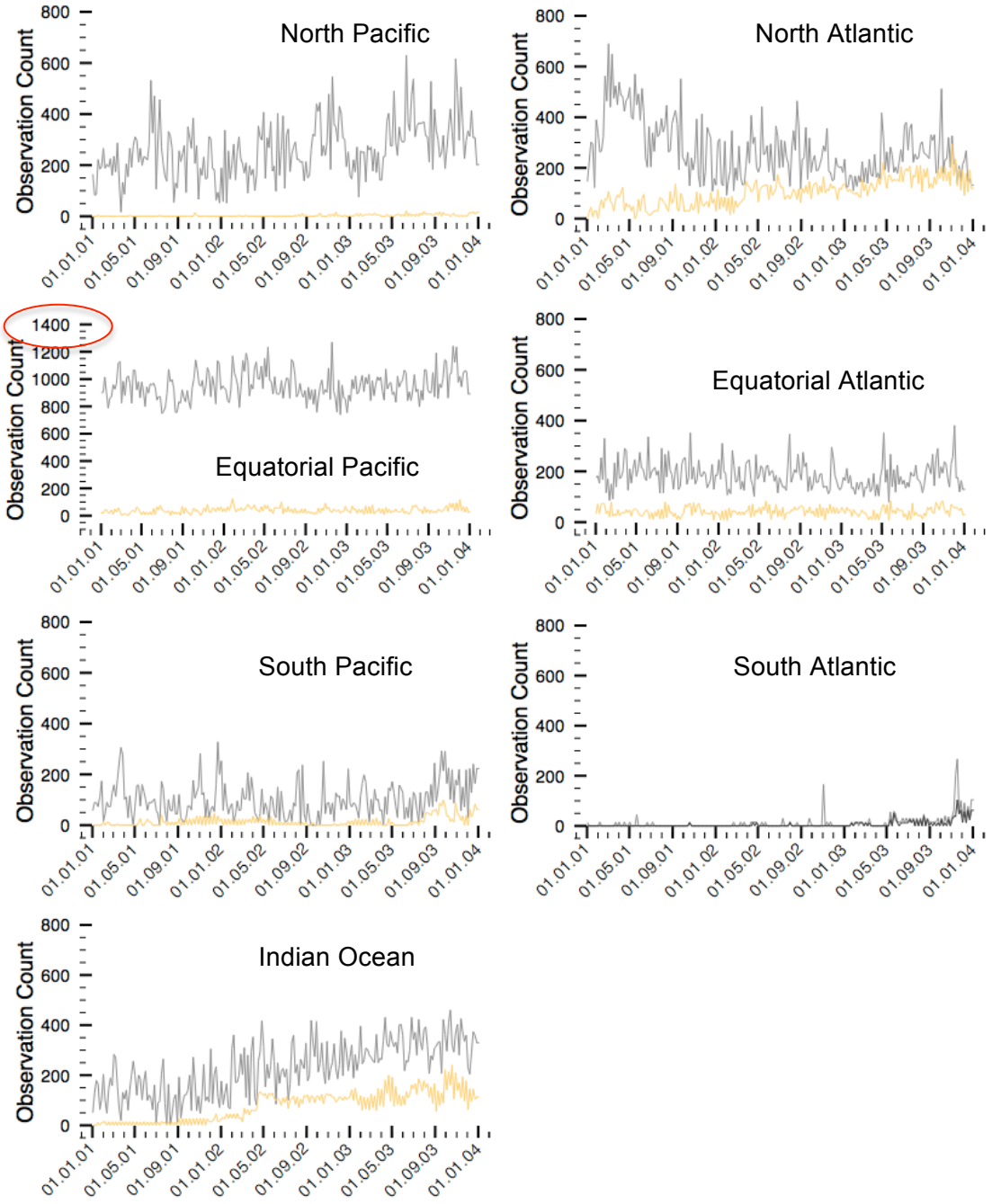

Figure 69. Regional breakdown of Observation Counts used by LETKF analysis.

For this experiment, the SODA experiment was run continuously from 1997 to 2004. The LETKF analysis was started at the beginning of 2001 due to the additional computational time needed to run LETKF-EOW.

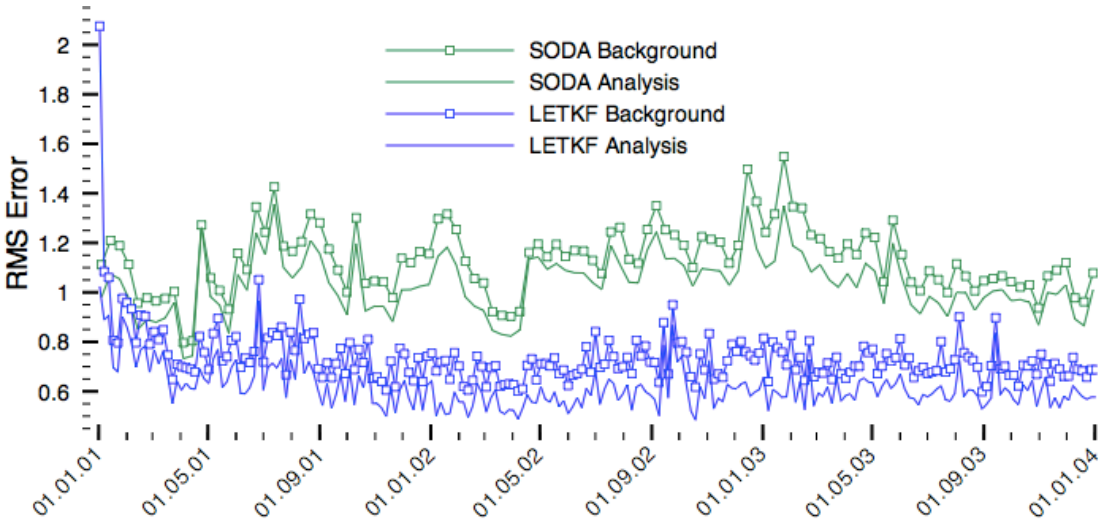

Figure 70. Global RMS temperature differences for SODA and LETKF-EOW from January 1, 2001 to January 1, 2004.

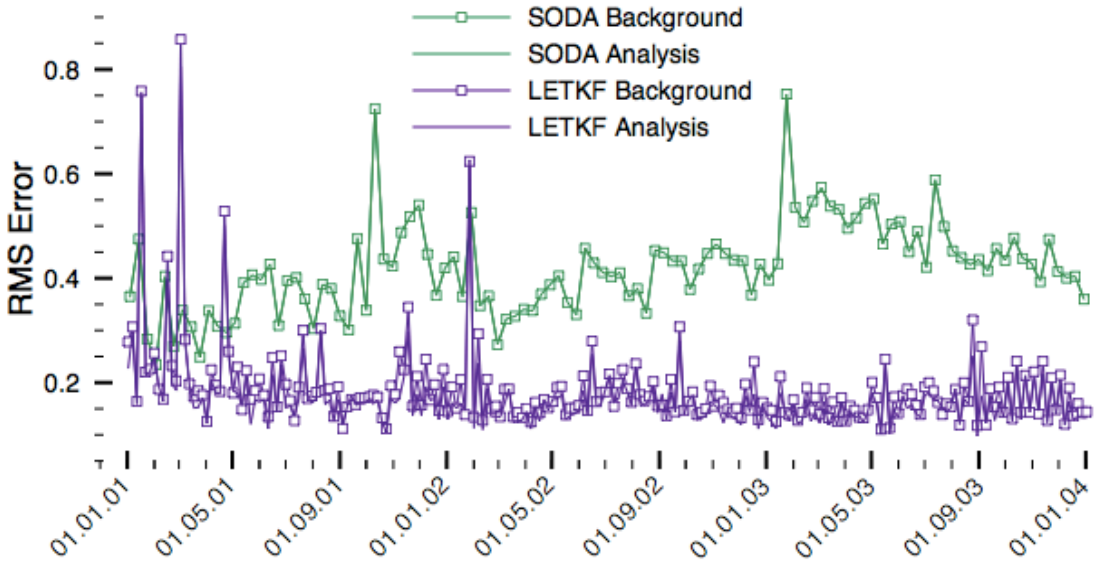

Figure 71. Global RMS salinity differences for SODA and LETKF-EOW from January 1, 2001 to January 1, 2004.

Amid a growing number of observations as shown in **Figure 69**, particularly the dramatic increase in salinity observations in the North Atlantic and Indian Oceans, the RMSDs remain at fairly steady levels. The increase in both temperature and salinity observations in the Indian Ocean has allowed for a large improvement in the RMSDs in this region. Observation coverage in the South Atlantic is negligible; the results in this region are not expected to be of any sufficient quality. Though,

there is some improvement with the small increase in coverage in the last half of 2003.

In the regional results, there are notable improvements in RMSD due to increased observational coverage. For the temperature RMSDs, this is particularly the case in the equatorial Atlantic, southern Atlantic beginning mid-2003, and Indian Ocean beginning in late 2001. This improvement is evident in all regions for salinity RMSD.

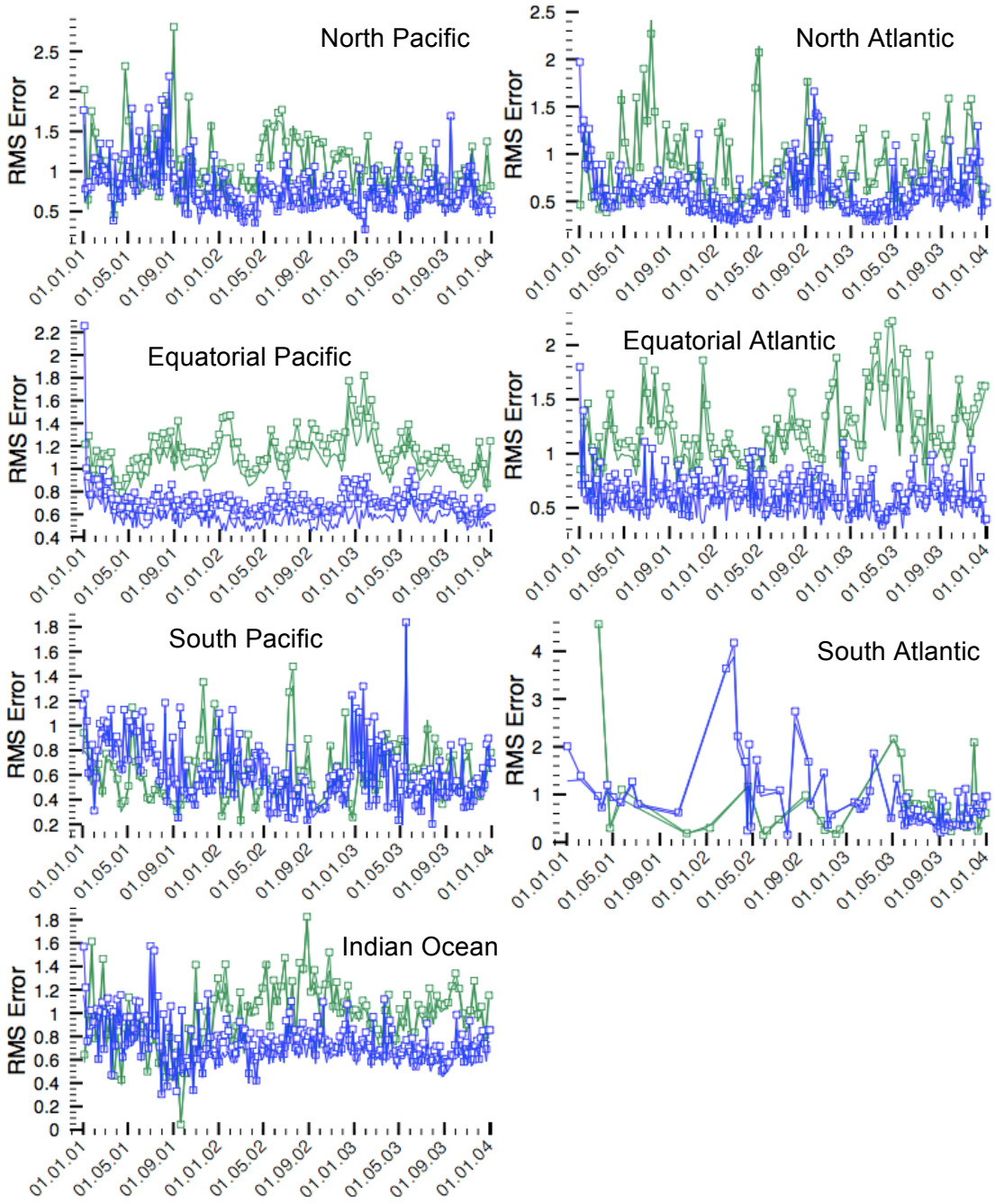

Figure 72. Regional breakdown of temperature RMSD for SODA and LETKF-EOW. (notation as in Figure 61)

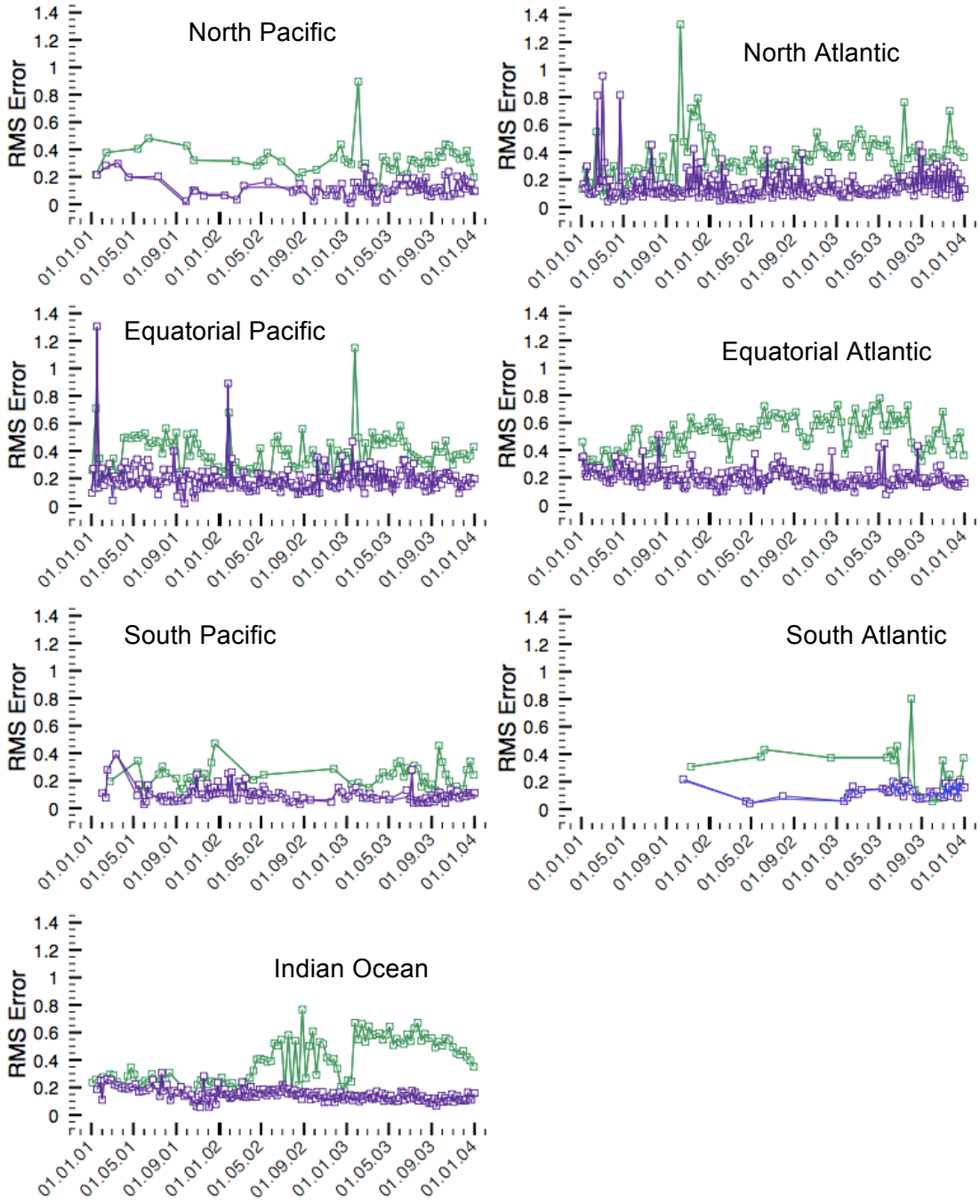

Figure 73. Regional breakdown of salinity RMSD for SODA and LETKF-EOW

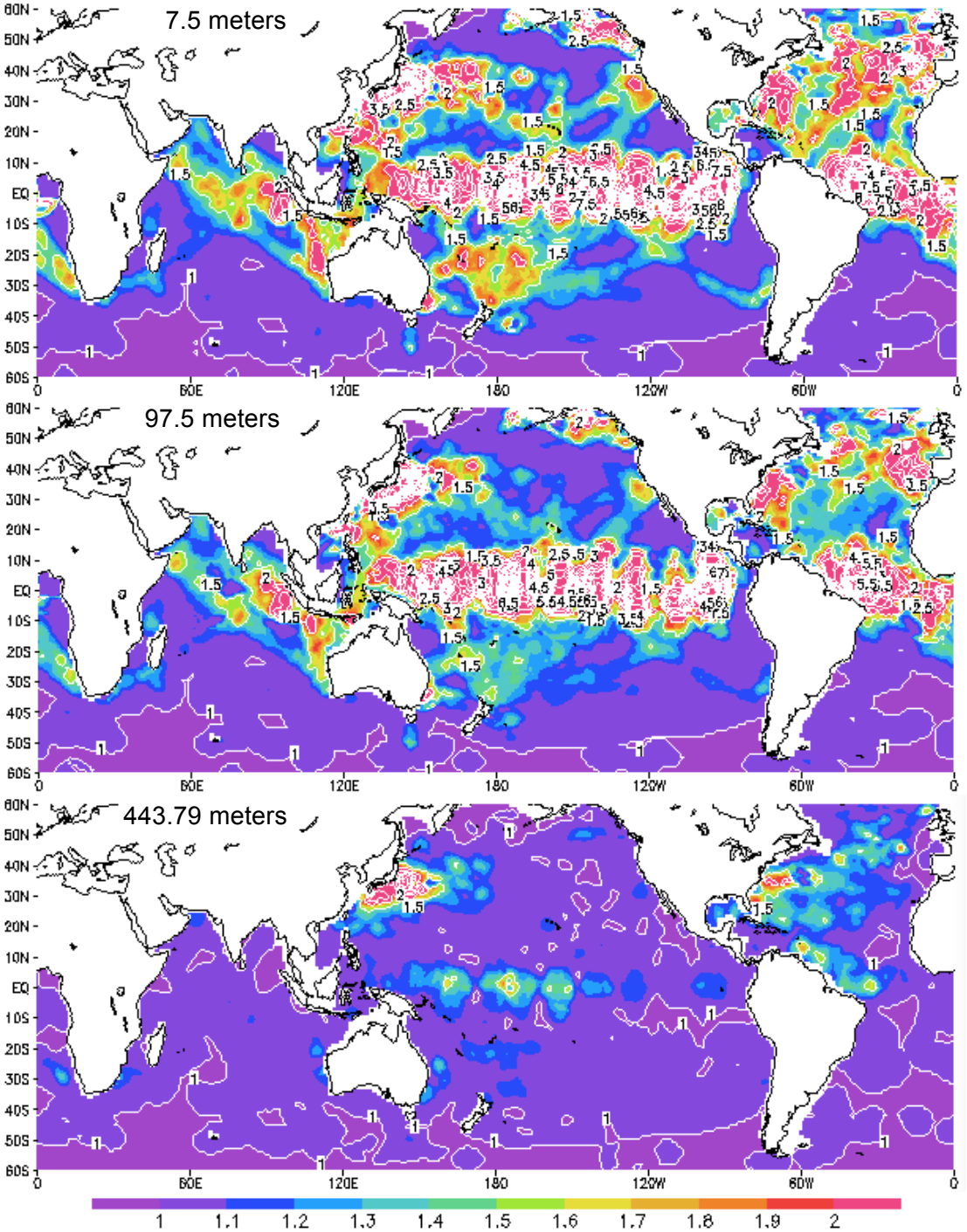

Figure 74. Inflation values generated by adaptive inflation at selected depths on January 3, 2002, approximately 1 year after initial experiment time using LETKF-EOW.

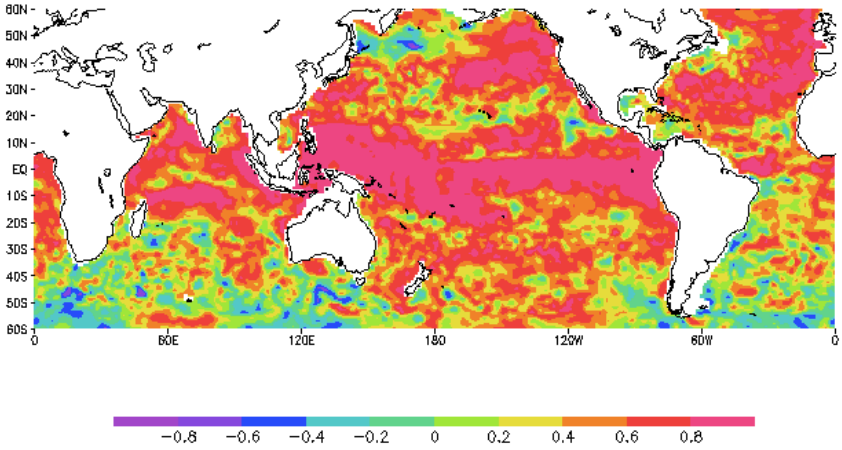

Figure 75. 300 m analyzed heat content correlated with altimetry during 1997-98 ENSO with LETKF-EOW.

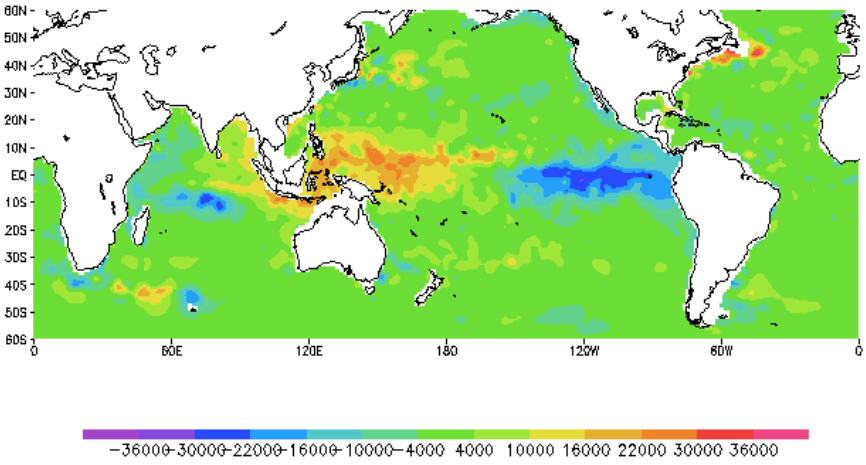

Figure 76. First Fourier term for monthly averaged 300 m vertically integrated heat content anomaly during 1997-98 ENSO with LETKF-EOW analysis.

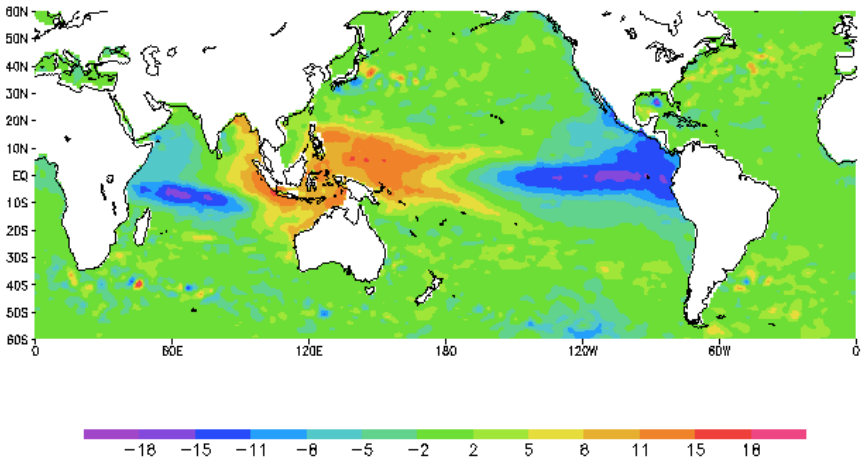

Figure 77. First Fourier term for altimetry during 1997-98 ENSO.

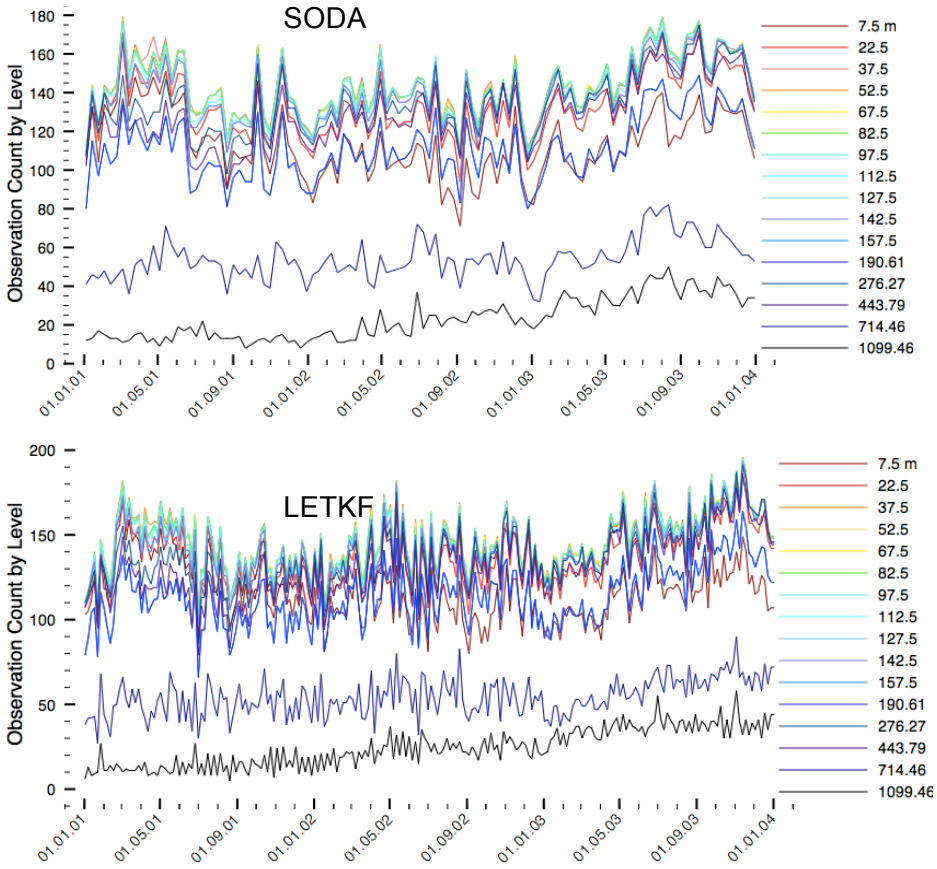

Figure 78. Vertical level breakdown of temperature observation counts at the analysis times of SODA and LETKF.

The results in **Figure 78** indicate that areas with relatively lower observation coverage perform worse than those areas with relatively higher coverage. The largest improvements in RMSDs for LETKF-EOW over SODA occur in the depths where the observation coverage is greatest. The depths in which SODA outperforms LETKF-EOW have less than half the coverage of the depths in which LETKF-EOW outperforms SODA. The areas in which LETKF-EOW performs best typically have greater dynamic instability.

Glossary of Terms

Analysis	The model state resulting from the data assimilation procedure after statistically combining the Background and Observations.
Analysis Cycle	The processes making up a single sequential data assimilation step.
Analysis Cycle Window	The duration of an analysis cycle. For the ocean, usually denominated by *d* days.
Background	The best available initial guess of the system state used by the assimilation procedure. When the term 'Background' is used, it typically refers to the beginning of the analysis cycle.
Ensemble Space	The *k*-dimensional space spanned by the *k* ensemble members.
Ensemble Spread	Standard deviation of the ensemble members relative to the ensemble mean.
Error of Representativeness	Subgrid-scale variability that is not represented in the grid-average values of the model.
Extended Obs Window	A range extending beyond the duration of the analysis cycle to include observations that fall before and after the analysis cycle window.
Forecast	Synonymous with 'Background', but implies the guess was generated from some forecasting method such as a

	computational model. When the term 'Forecast' is used, it typically refers to the Background for the next analysis cycle.
Incremental Analysis Update	A procedure that applies a small fraction of the analysis innovation to the model prognostic equations at each model time step (1 second in MOM2) via a forcing term.
Model Space	The *m*-dimensional space spanned by the product of *n* model grid points and *v* model state variables.
O-A	The difference in the observed value and the analyzed value interpolated to observation space.
O-F	The difference in the observed value and the forecast value interpolated to observation space.
Observability Condition	The condition that determines whether there is sufficient information present in the system inputs and outputs to uniquely determine the system state.
Observation	A single point in physical space that consists an estimated value of one parameter of the true system state and an error associated with that estimate.
Observation Space	The *l*-dimensional space spanned by the *l* observations made at during any given analysis cycle.
Quasi-Outer Loop	A procedure that iterates within the analysis cycle as with Running-in-Place, but which only corrects the

	ensemble mean without changing the structure of the ensemble perturbations.
Running-in-Place	A procedure that iterates within the analysis cycle to explore the space outside of the linear combination of ensemble members.
Spread	Standard deviation of the ensemble, either referring to the analysis ensemble, or to the background/forecast ensemble.
Super Observation	A spatially and/or temporally averaged quantity representing 1 or more observations used for the purpose of smoothing the observed values.

Glossary of Mathematical Quantities

$\boldsymbol{B}$	The constant covariance matrix used by methods such as 3D-Var
k	The dimension of the ensemble space (i.e. the ensemble size)
l	The dimension of the observation space
m	The dimension of the model space
$\boldsymbol{P}^a$	The covariance matrix of the analysis errors.
$\boldsymbol{P}^b$	The covariance matrix of the background errors.
$\boldsymbol{R}$	The covariance matrix of the observation errors.
$\boldsymbol{x}$	A state vector in model space
$\boldsymbol{X}^a$	A matrix of state vectors representing perturbations from the analysis ensemble mean
$\boldsymbol{X}^b$	A matrix of state vectors representing perturbations from the background ensemble mean
$\boldsymbol{y}$	A state vector in observation space
$\boldsymbol{Y}^b$	A matrix of state vectors representing perturbations from the ensemble mean in observation space
α_s	The weighting parameter for the proportion of historical fields to use in construction of the initial ensemble (between 0 and 1)
α_w	The weighting parameter for the proportion of historical fields to use in constructing the surface wind forcing ensemble (between 0 and 1)

Acronyms

3D-Var	Three Dimensional Variational
4D-LETKF	Four Dimensional Local Ensemble Transform Kalman Filter
4D-Var	Four Dimensional Variational
ADCP	Acoustic Doppler Current Profiler
CTD	Conductivity-Temperature-Depth
DB	Denman-Beavers Iteration
EnKF	Ensemble Kalman Filter
EOW	Extended Observation Window
GFDL	Geophysical Fluid Dynamics Laboratory
IAU	Incremental Analysis Update
LETKF	Local Ensemble Transform Kalman Filter
MATLAB	Matrix Laboratory
MBT	Mechanical Bathythermograph
MOM2	Modular Ocean Model, version 2.4b
MOM4	Modular Ocean Model, version 4.1
NASA	National Aeronautics and Space Administration
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
OI	Optimal Interpolation
QOL	Quasi-Outer Loop
RIP	Running in Place

RMSD	Root Mean Square Deviation (Difference/Distance)
SDE	Stochastic Differential Equation
SMW	Sherman-Morrison-Woodbury
SODA	Simple Ocean Data Assimilation
SST	Sea Surface Temperature
XBT	Expendable Bathythermographs

Bibliography

[A01] Anderson, J., 2001: *An ensemble adjustment filter for data assimilation*. Monthly Weather Review, 129.

[BEM01] Bishop, C. H., B. J. Etherton, and S. J. Majumdar. *Adaptive Sampling with Ensemble Transform Kalman Filter. Part I: Theoretical Aspects*. Mon. Wea. Rev., 129, 420-436, 2001.

[BT96] Bloom, S.C., Takacs, L.L., da Silva, A.M., Ledvina, D. *Data Assimilation using Incremental Analysis Updates.* Monthly Weather Review, vol. 124, p. 1256-1271, 1996.

[CS08] Carton, J.A., and A. Santorelli, 2008: *Global upper ocean heat content as viewed in nine analyses*, J. Clim., 21, 6015–6035.

[CGX96] Carton, J.A., B.S. Giese, X. Cao, and L. Miller, 1996: *Impact of TOPEX and thermistor data on retrospective analyses of the tropical Pacific Ocean*, *J. Geophys. Res.* , **101**,14,147-14,159.

[CCC00a] Carton, J.A., Chepurin, G., CAO, X., Giese, B. *A Simple Ocean Data Assimilation Analysis of the Global Upper Ocean 1950-95 Part I: Methodology*. Journal of Physical Oceanography, Volume 30, p. 294-309, 2000.

[CCC00b] Carton, J.A., Chepurin, G., CAO, X., Giese, B. *A Simple Ocean Data Assimilation Analysis of the Global Upper Ocean 1950-95 Part II: Results*. Journal of Physical Oceanography, Volume 30, p. 311-326, 2000.

[CG08] Carton, J.A., Giese B.S., *A Reanalysis of Ocean Climate Using Simple Ocean Data Assimialtion (SODA)*. Monthly Weather Review, Volume 136, p. 2999-3016, August 2008.

[DKM07] Danforth, C. M., E. Kalnay, T. Miyoshi. *Estimating and Correcting Global Weather Model Error.* Monthly Weather Review, 135, No. 2, 281299 (2007).

[DT10] Deng, Z., Tang, Y., Assimilation of Argo Temperature and Salinity Profiles using a bias-aware localized EnKF system for the Pacific Ocean. 2010. [http://web.unbc.ca/~ytang/Version_July052010.pdf]

[DB76] Denman, E. D., Beavers, A. N., *The matrix sign function and computations in systems*. (1976), Applied Mathematics and Computation 2 (1): 63–94

[DBC05] Desroziers, G., L. Berre, B. Chapnik, and P. Poli. *Diagnosis of observation, background and analysis-error statistics in observation space*.

[D89] Dynkin, E. B., *Kolmogorov and the Theory of Markov Processes*. Annals of Probability, 17, (1989), p. 823.

[E94] Evenson, G., *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*. J. Geophys. Res., 99 (C5), 10 143–10 162.

[E03] Evenson, G., *The Ensemble Kalman Filter: Theoretical formulation and practical implementation*. Ocean dynamics, 2003.

[F14] Fokker, A.D. "Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld" Annalen der Physik 43, (1914) 810-820

[HH05] Harlim, J., Hunt, B. *Local Ensemble Transform Kalman Filter: An Efficient Scheme for Assimilating Atmospheric Data*. IPST, UMD 2005.

[H86] Higham, N. J. *Newton's Method for the Matrix Square Root*. Mathematics of Computation, 46 (174), pp. 537-549, April 1986.

[H02] Higham, N.J., *Accuracy and Stability of Numerical Algorithms*. 10.1137/1.9780898718027.ch14, SIAM, 2002.

[H97] Higham, N.J., *Stable Iterations for the Matrix Square Root*. Numerical Algorithms 15 (2). p.227-242, 1997.

[HS96] Hobbs, S. L., Sritharan, S. S. (1996), *Nonlinear Filtering of Stochastic Reacting and Diffusing Systems*, from N. Gretsky, J. Goldstein and J.J. Uhl, editors, Probability and Modern Analysis, Marcel Dekker.

[HKC09] Hoffman, M. J., E. Kalnay, J. A. Carton, and S.C. Yang (2009), *Use of breeding to detect and explain instabilities in the global ocean*, Geophys. Res. Lett., 36, L12608, doi:10.1029/2009GL037729

[HXB08] Huang B., Xue, Y., Behringer, D., 2008. Impacts of Argo salinity in NCEP Global Ocean Data Assimilation System: The tropical Indian Ocean. J.Geo.R.,113, doi:10.1029/2007JC004388

[HKK04] Hunt, B.R., Kalnay, E., Kostelich, E.J., Ott, E., Patil, D.J., Sauer, T., Szunyogh, I., Yorke, J.A., Zimin, A.V. *Four-Dimensional Ensemble Kalman Filtering*. Tellus 56A (2004), 4. April 2, 2004.

[HKS06] Hunt, B.R., Kostelich E.J., Szunyogh, I. *Efficient Data Assimilation for Spatiotemporal Chaos: A Local Ensemble Transform Kalman Filter*. arXiv:physics/0511236 v1 28 Nov 2005. Dated May 24, 2006.

[ICG97] Ide, K., P. Courtier, M. Ghil, and A. Lorenc. *Unified notation for data assimilation: Operational, sequential and variational*. J. Meteor. Soc. Japan, 75, 181-189. 1997.

[K03] Kalnay, E. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, 2003. Chapter 5.

[K96] Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, Bull. Am. Meteorol. Soc., 77, 437 – 471.

[KLM] Kalnay, E., Li, H., Miyoshi, T., Adaptive Estimation of Background and Observation Errors within Local Ensemble Transform Kalman Filter.

[KMY07] Kalnay, E., Li, H., Miyoshi, T., Yang, S., Ballarrera-Poy, J. *4D-Var or Ensemble Kalman Filter*? 2006.

[KY08] Kalnay, E., Yang, S.C., Accelerating the spin-up of Ensemble Kalman Filtering. [http://www.atmos.umd.edu/~ekalnay/AcceleratingSpinupEnkf_QJRMS.pdf]

[Ka09] Kang, J.S., Carbon Cycle Data Assimilation using a Coupled Atmosphere-Vegetation Model and the Local Ensemble Transform Kalman Filter. Ph.D. thesis at the University of Maryland, 2009.

[KR08] Keppenne, C.L., Rienecker, M.M., Jacob, J.P., Kovach, R. *Error Covariance Modeling in the GMAO Ocean Ensemble Kalman Filter*. Monthly Weather Review, American Meteorological Society, vol. 136, p. 2964, August 2008.

[Ko31] Kolmogorov, A.N. "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung," Math. Ann. 104, (1931), 415-458.

[KMK11] Kunii, M., Miyoshi, T., Kalnay, E. *Data Denial Experiments for Tropical Cyclone SINLAKU*. (2011) http://www.weatherchaos.umd.edu/group_log/data/y1011/110307_weatherchaos_kunii.pdf

[LTC05] Lang, S., Tao, W.K., Cifelli, R., Olson, W., Halverson, J., Rutledge, S., Simpson, J., *Improving Simulation of Convective Systems from TRMM LBA: Easterly and Westerly Regimes.* Journal of the Atmospheric Sciences, November 2005.

[LB94] Levitus, S., Boyer, T., *World Ocean Atlas 1994*, vol. 4, Temperature, NOAA Atlas NESDIS, vol. 4, 129 pp., NOAA, Silver Spring, Md, 1994.

[LKM09] Li H., E. Kalnay, T. Miyoshi, C. M. Danforth. *Accounting for Model Errors in Ensemble Data Assimilation*. Monthly Weather Review. 137, No. 10, 3407-3419 (2009) doi:10.1175/2009MWR2766.1

[LKM] Li, H., Kalnay, E., Miyoshi, T., Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter.

[LKM08] Liu, J., Kalnay, E., Miyoshi, T., Cardinale, C., *Analysis Sensitivity Calculation with an Ensemble Kalman Filter*. Quarterly Journal of the Royal Meteorological Society, QJ-08-0083.R1

[LB91] Lorenc, A.C., Bell, R.S. and MacPherson, B. *The Meteorological Office analysis correction data assimilation scheme*. Quart. J. Roy. Meteor. Soc., 117, 59–89, 1991.

[MS07] Miyoshi, T. and Y. Sato, 2007: *Assimilating Satellite Radiances with a Local Ensemble Transform Kalman Filter (LETKF) Applied to the JMA Global Model (GSM)*. *SOLA*, 3, 37-40. doi:10.2151/sola.2007-010

[MYE07] Miyoshi, T., S. Yamane, and T. Enomoto, 2007: Localizing the Error Covariance by Physical Distances within a Local Ensemble Transform Kalman Filter (LETKF). *SOLA*,3, 89-92. doi:10.2151/sola.2007-023

[M11] Miyoshi, T., *The Gaussian Approach to Adaptive Covariance Inflation and its Implementation with the Local Ensemble Transform Kalman Filter*. Mon. Wea. Rev., in press. 2011.

[MSK10] Miyoshi, T., Y. Sato, and T. Kadowaki, 2010: *Ensemble Kalman filter and 4D-Var inter-comparison with the Japanese operational global analysis and prediction system*. *Mon. Wea. Rev.*, 138, 2846-2866. doi:10.1175/2010MWR3209.1

[MY07] Miyoshi, T., Yamane, S. *Local Ensemble Transform Kalman Filtering with an AGCM at a T159/L48 Resolution*. Mon. Wea. Rev., 2007.

[OBG08] Oke P.P., Brassington, G. B., Griffin, D. A. Schiller, A., 2008. The bluelink ocean data assimilation system (BODAS). Ocean Modelling,

21, 46-70.

[OHS04] Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E.J., Corazza, M., Kalnay, E., Patil, D.J., Yorke, J.A. *A Local Ensemble Kalman Filter for Atmospheric Data Assimilation*. Tellus, December 10, 2004.

[P96] Pacanowski, R.C. *MOM2, Version 2.0 Beta, Documentation User's Guide and Reference Manual*. GFDL Ocean Technical Report 3.2, Nov. 7, 1996.

[PK04] Peña, M., and E. Kalnay (2004), Separating fast and slow modes in coupled chaotic systems, Nonlinear Processes Geophys., 11, 319 – 327.

[P17] Planck, M. "Ueber einen Satz der statistichen Dynamik und eine Erweiterung in der Quantumtheorie", Sitzungberichte der Preussischen Akadademie der Wissenschaften (1917) p. 324-341

[RJR09] Roemmich, R., Johnson, G.C., Riser, S., Davis, R., Gilson, J., Owens, W.B., Garzoli, S.L., Schmid, C., Ignaszewski, M. The Argo Program: Observing the Global Ocean with Profiling Floats. NOPP Special Issue, Excellence in Partnering Award Winners. Oceanography, Vol. 22, No. 2. p.34-43, June 2009.

[RRC06] Romanou, A., Rossow, W.B., Chou, S-H. *Decorrelation Scales of High Resolution Turbulent Fluxes at the Ocean-Surface*. [http://earth.esa.int/workshops/venice06/participants/1012/paper_1012_romanou.pdf]

[SH09] Smith G.C. and Haines K, 2009. Evaluation of the S(T) assimilation method with Argo dataset. Q.J.R. Meteorol. Soc. 135:739-756.

[S95] Sritharan, S. S. (1995). *Nonlinear Filtering of Stochastic Navier-Stokes equations*, from T. Funaki and W.A. Woyczynski, editors, Nonlinear Stochastic PDEs: Burgers Turbulence and Hydrodynamic Limit, Springer-Verlag, pp. 247–260.

[S05] Stewart, R.H., Introduction to Physical Oceanography. 1997-2005. [http://oceanworld.tamu.edu/resources/ocng_textbook/contents.html]

[SKG07] Szunyogh, I., Kostelich, E.J., Gyarmati, G., Kalnay, E., Hunt, B.R., Ott, E., Satterfield, E., Yorke, J.A. *A Local Ensemble Transform Kalman Filter Data Assimilation System for the NCEP Global Model*. Tellus, April 18, 2007.

[BAG06] T.P. Boyer, J.I. Antonov, H.E. Garcia, D.R. Johnson, R.A. Locarnini, A.V. Mishonov, M.T. Pitcher, O.K. Baranova, I.V. Smolyar, 2006. World Ocean Database 2005. S. Levitus, Ed., NOAA Atlas NESDIS 60, U.S. Government Printing Office, Washington, D.C., 190 pp., DVDs.

[T87] Twomey, S. *Iterative Nonlinear Methods for Tomographic Problems*. Journal of the Atmospheric Sciences, Vol. 44, No. 23, 1987.

[WHW07] Wang, X., Hamill, T.M., Whitaker, J. S., Bishop, C.H. *A Comparison of Hybrid Transform Kalman Filter-Optimum Interpolation and Ensemble Square Root Filter Analysis Schemes.* Monthly Weather Review, Volume 135. DOI: 10.1175/MWR3307.1, 2007.

[YK11] Yang, S-C and E. Kalnay, 2011: *Handling nonlinearity and non-Gaussianity in Ensemble Kalman Filter*. submitted to a special collection "Intercomparisons of 4D-Variational Assimilation and the Ensemble Kalman Filter", Mon. Wea. Rev.

[YKH08] Yang, S-C, E. Kalnay, B. Hunt, N. Bowler, 2008b: Weight interpolation for efficient data assimilation with the Local Ensemble Transform Kalman Filter, Quart. J. Roy. Meteor. Soc., under revision.