An Ensemble Recentering Kalman Filter with an Application to Argo Temperature Data Assimilation into the NASA GEOS-5 Coupled Model

Christian L. Keppenne[1, 2]

[1]*Global Modeling and Assimilation Office*
Code 610.1, NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA

[2]*Science Systems and Applications inc.*
10210 Greenbelt Road, Suite 600, Lanham, Maryland 20706, USA

Correspondence address:
Christian Keppenne
email: christian.keppenne@nasa.gov
telephone: 011-1-301-6145874
mail: Code 610.1. NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

October 14, 2013

Abstract

A two-step ensemble recentering Kalman filter (ERKF) analysis scheme is introduced. The algorithm consists of a recentering step followed by an ensemble Kalman filter (EnKF) analysis step. The recentering step is formulated such as to adjust the prior distribution of an ensemble of model states so that the deviations of individual samples from the sample mean are unchanged but the original sample mean is shifted to the prior position of the most likely particle, where the likelihood of each particle is measured in terms of closeness to a chosen subset of the observations. The computational cost of the ERKF is essentially the same as that of a same size EnKF.

The ERKF is applied to the assimilation of Argo temperature profiles into the OGCM component of an ensemble of NASA GEOS-5 coupled models. Unassimilated Argo salt data are used for validation. A surprisingly small number (16) of model trajectories is sufficient to significantly improve model estimates of salinity over estimates from an ensemble run without assimilation. The two-step algorithm also performs better than the EnKF although its performance is degraded in poorly observed regions.

Keywords:
Data assimilation; Kalman filter; ensemble Kalman filter; Particle filter;
Coupled data assimilation

1. Introduction

Since its introduction by *Evensen* [1994, 1996] in the context of a quasigeostrophic ocean model, the ensemble Kalman filter (EnKF) has gained wide acceptance among the atmosphere and ocean modeling communities as a viable data assimilation technique. In an EnKF, the prohibitive cost associated with evolving the model-background error-covariance matrix according to the traditional Kalman filter [*Kalman*, 1960] formulation is avoided. Rather, the background-error covariance propagation is replaced with the concurrent integration of an ensemble of model trajectories. The forecast-error covariance statistics needed to compute the Kalman gain are then estimated from the statistical distribution of the ensemble of model states. The EnKF uses the ensemble mean to estimate the true state of the dynamical system. The posterior ensemble variance serves to estimate the analysis error variance.

The sample mean is the most likely forecast if the ensemble is normally distributed. A normal distribution is often assumed in climate system data analysis, either for convenience or for lack of a better assumption. However, it is known from nonlinear dynamical system theory and confirmed from observations that clustering around several likely forecast solutions can occur. For example, the Kuroshio is in a multiple equilibrium state in which meandering and straight paths are equally plausible but the intermediate mean path is unstable. Because of this, the central forecast (defined as the ensemble member closest to the mean in terms of RMS difference) is sometimes used to estimate the true state.

Another problem associated with using the sample mean to estimate the true state is that while the individual ensemble members are dynamically balanced states, their mean generally is not. In a multivariate EnKF in which certain prognostic model variables are updated as observations of other variables are assimilated, the resultant analysis increments can lead to one or more ensemble members becoming dynamically unstable when the analyzed ensemble is advanced further in time.

Particle filters are a class of estimation methods that do not estimate the true state with the sample mean. Instead, the true-state estimates obtained with particle filters are statistically plausible, dynamically balanced states. The original particle filter algorithm, sample importance resampling [SIR: *Gordon et al.*, 1993], approximates the true state with a weighted sum of particles (ensemble members) where the weights are proportional to the respective probability of each particle. In nonlinear systems, SIR can estimate the true system state more accurately than the EnKF when the number of particles is large enough. Nevertheless, for the small ensemble sizes typically used in GCM ensemble prediction, *i.e.* a few tens to a few hundred ensemble members, the EnKF is generally more accurate [*Weerts and El Serafy*, 2006]. Besides, the weighted-particle SIR estimate is just as likely to be unstable as the EnKF ensemble mean. If, however, the most likely particle is used to estimate the state, the estimate will generally be stable and balanced. Note however that an even larger number of particles will be required for such as scheme to be accurate.

The purpose of this note is to introduce a two-step ensemble recentering Kalman filter (ERKF) analysis procedure that combines advantages from the EnKF and SIR schemes. It is shown that the ERKF can produce a more accurate state estimate than a same-size EnKF with dramatically less particles than a typical SIR filter would require. Our demonstration uses the NASA Global

Modeling and Assimilation Office (GMAO) global earth observing system (GEOS) integrated ocean data assimilation system (iODAS) applied to the assimilation of Argo temperature profiles into the OGCM component of the GEOS-5 coupled system. The ERKF is introduced in Section 2. The next two Sections contain overviews of the CGCM and of the data assimilation system. The validation experiments are discussed in Section 5. Our conclusions follow in Section 6.

2. The Two-Step Analysis Procedure

Let us consider a nonlinear model, $\boldsymbol{M}$, that evolves an approximation, $\boldsymbol{x}$, of the true system state, $\boldsymbol{x}_t$, subject to a forcing term, $\boldsymbol{f}$. We define an observation operator, $\boldsymbol{H}$, that maps the true system state to an observation vector, $\boldsymbol{y}$. Then, the system that defines the estimated state evolution can be written:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{M}(\boldsymbol{x}, \boldsymbol{f}) + \boldsymbol{q}, \tag{1a}$$

$$\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x}_t) + \boldsymbol{r}, \tag{1b}$$

$$E((\boldsymbol{x} - \boldsymbol{x}_t)(\boldsymbol{x} - \boldsymbol{x}_t)^T) = \boldsymbol{P}, \qquad E(\boldsymbol{q}\boldsymbol{q}^T) = \boldsymbol{Q}, \qquad E(\boldsymbol{r}\boldsymbol{r}^T) = \boldsymbol{R}, \tag{1c}$$

where $\boldsymbol{E}$ is the usual expectation operator.

We want to optimally estimate $\boldsymbol{x}_t$ given information about $\boldsymbol{x}$, $\boldsymbol{y}$ and their error distributions. The Kalman filter [*Kalman*, 1960] tells us that the estimate that minimizes the analysis error variance is

$$\boldsymbol{x}^a = \boldsymbol{x}^f + \boldsymbol{P}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^f\right), \tag{2}$$

where $\boldsymbol{x}^f$ is the prior state estimate (model forecast).

In an EnKF, $\boldsymbol{P}$ is estimated from the statistical distribution of an ensemble of model forecasts, $\boldsymbol{x}_i, i = 1,\ldots,n,$ each evolved according to (1a), so that

$$\boldsymbol{S} = \left\{\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \cdots, \boldsymbol{x}_n - \bar{\boldsymbol{x}}\right\}, \tag{3a}$$

$$\boldsymbol{P} \approx \frac{1}{n-1}\boldsymbol{S}\boldsymbol{S}^T, \tag{3b}$$

where the overbar denotes the sample mean. The analysis for $\boldsymbol{x}_i$ follows from substituting $\boldsymbol{x}_i$ for $\boldsymbol{x}^f$ in (2).

The first (recentering) step of the ERKF works by finding the most likely sample, $\boldsymbol{x}_p$, such that

$$\sigma_p = \sum_k \left(\frac{y_k - \boldsymbol{H}_k\boldsymbol{x}_p}{r_k}\right)^2, \tag{4}$$

is minimized, where the k subscript runs over a chosen subset of the latest batch of observations, and r_k and $\boldsymbol{H}_k$ are the measurement error and observation operator for the kth datum. A pre-analysis increment, $\boldsymbol{\Delta}_p = \boldsymbol{x}_p - \bar{\boldsymbol{x}}$, is then applied to each ensemble member. This places the ensemble mean at the original "location" of $\boldsymbol{x}_p$ without affecting our estimates of $\boldsymbol{P}$ and $\boldsymbol{R}$ since the sample distribution relative to its mean is not changed (Figure 1).

The second step uses (2) and (3) to compute the usual EnKF analysis from the ensemble of $\boldsymbol{x}'_i = \boldsymbol{x}_i + \boldsymbol{\Delta}_p, i = 1,\ldots,n$, thus producing a set of analysis increments,

$$\boldsymbol{\Delta}_i = \boldsymbol{P}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}'_i + \boldsymbol{r}_i\right), \qquad i = 1,\ldots,n, \tag{5}$$

so that the total increment applied to ensemble member i is $\boldsymbol{\Delta}_p + \boldsymbol{\Delta}_i$.

Note that the observation vector used to identify $\boldsymbol{x}_p$ could be different from that used in the Kalman filter update (5).

3. The GEOS-5 Coupled Model

The GEOS-5 CGCM couples the NASA GEOS-5 atmospheric model [*Rienecker et al.*, 2008] and GFDL's Modular Ocean Model version 4 (MOM4) and the Los Alamos CICE sea ice model [*Hunke et al.,* 2013] using the Earth System Modeling Framework [ESMF*: Hill et al*., 2006]. Ocean-atmosphere coupling is done through a physical interface layer of constant depth (2m). Atmospheric fluxes of heat and fresh water are applied at the top of this layer and oceanic turbulent fluxes are parameterized at the bottom of the layer. The mass of the layer and amount of mixing at its base are chosen so that the layer simulates the diurnal cycle of SST.

The GEOS-5 configuration used here runs MOM4 and CICE on a tripolar grid with uniform 1/2° zonal resolution, variable meridional resolution ranging from 1/6° in the Tropics to 1/2° in the high latitudes, and has 40 vertical levels. The dimension of the grid is 720×410×40 (zonal×meridional×vertical). The AGCM has uniform 5/4° zonal resolution, 1° meridional resolution and 72 vertical levels (288×181×72 grid dimension). The system is run with a 15-minute time step. A faster, 15-second barotropic time step is also used to run MOM4.

The CGCM system has close to 100-million prognostic state variables in the configuration used herein. The ocean data assimilation modifies oceanic fields of temperature (T), salinity (S), zonal current (U), meridional current (V) and sea surface height (SSH) as well as sea ice concentration and thickness.

4. GEOS iODAS

The GEOS integrated ocean data assimilation system (iODAS) system has evolved from the first generation GMAO ocean data assimilation system introduced originally in Keppenne and Rienecker [2003] and discussed in more detail in *Keppenne et al.* [2005, 2008]. iODAS is implemented as an ESMF [*Hill et al.,* 2004] gridded component. The communications between the ocean and sea ice models and iODAS are managed by ESMF. The system is used routinely to produce the GMAO production ocean analysis [*Vernieres et al.,* 2012]. What follows is a summary of the main differences between the analysis algorithms used here and in *Keppenne et al.* [2008].

As is customary in EnKF applications to large-scale atmospheric or ocean models, the background-error covariances are localized to address the degree-of-freedom limitations encountered when the sample size is much less than that of the model state vector. The error-covariance localization is flow adaptive (following neutral density surfaces) and an iterative procedure is used to individually optimize the covariance localization scales involved in the processing of each observation. Incremental analysis updating [IAU: *Bloom et al.*, 1996] is used to apply them gradually over the assimilation interval.

The ocean assimilation is applied to the full CGCM, with the atmospheric component constrained during the integration by "replaying" the NASA Modern Era Retrospective-analysis for Research and Applications [MERRA: e.g., *Rienecker et al.*, 2008] into the GEOS-5 AGCM. The procedure involves integrating the AGCM to the next synoptic analysis time, reading the MERRA analysis fields and calculating analysis increments by taking their difference from the background atmospheric fields, rewinding the AGCM and, finally, integrating the full CGCM while incrementally applying both the atmospheric analysis increments thus computed and the ocean analysis increments produced by iODAS. The implementation is designed to facilitate consistent atmosphere-ocean states to initialize GEOS-5 seasonal climate forecasts. More details about the replay procedure are available in *Vernieres et al.* [2012].

5. Experiments and Results

To test the ERKF, four experiments were run spanning March-June 2006 and using 16 model trajectories. The initial condition comes from integrating a single instance of the CGCM while constraining its AGCM component by replaying the MERRA reanalysis. From this initial condition, the ensemble is first spun up during March 2006 by perturbing the analysis increments of the atmospheric analysis replay procedure and also applying daily perturbations to the OGCM T and salinity S fields. The ensemble configuration as of March 31 2006 is then used to initialize each ensemble run.

CE-16 is a control ensemble run without assimilation. ER-16 only applies the ensemble recentering step, EnKF-16 only applies the EnKF analysis step and ERKF-16 applies both the recentering and EnKF analysis steps. Starting on April 1, 2006, Argo T profiles are assimilated daily in EnKF-16 and in ERKF-16. These same Argo T profiles are also used to select $\boldsymbol{x}_p$ in the recentering steps of ER-16 and ERKF-16. The average wallclock run times per simulation day on 960 2.8 GHz Intel Altix Sandy Bridge cores are 687 seconds in CE-16, 691 seconds in ER-16, 744 seconds in EnKF-16 and 749 seconds in ERKF-16, illustrating the minimal cost of the recentering step. The performance of the data assimilation and of the ensemble recentering is assessed from how closely each run can reproduce the assimilated Argo T profiles and the unassimilated Argo S profiles.

The time mean April-June 2006 global RMS OMF differences of each run with the Argo T and S data are listed in table 1. They show that applying the recentering step without following it with an EnKF analysis has only a small impact on the RMS T OMF (7% reduction from CE-16 to ER-16) and that there is no noticeable benefit for T in applying the recentering step on top of the EnKF analysis step (1% RMS T OMF increase from EnKF-16 to ERKF-16). However, the results are markedly different in terms of the unassimilated S variable. The RMS S OMF from

ER-16 are 24% lower than those from CE-16 and those of ERKF-16 are 24% lower than those from EnKF-16. In contrast with ER-16 and ERKF-16, while EnKF-16 performs best for T, it does not significantly improve over CE-16 for S.

	CE-16	ER-16	EnKF-16	ERKF-16
RMS T OMF (°C)	1.62	1.51	0.91	0.92
RMS S OMF (PSU)	0.51	0.39	0.49	0.37

Table 1. April-June 2006 global mean RMS OMF differences with the assimilated Argo T data and unassimilated Argo S data in each of the CE-16, ER-16, EnKF-16 and ERKF-16 runs.

Figures 2 and 3 expand upon the information provided by Table 1. Figure 2 shows the vertical average of the RMS OMF for the assimilated Argo T data averaged over the 3-month data assimilation period and binned to 3°×3° horizontal boxes. Figure 3 shows the corresponding binned RMS OMF horizontal distributions for the unassimilated Argo S data. For the T data, Figures 2c (EnKF-16) and 2d (ERKF-16) show very similar distributions, while the distribution of RMS T OMF in ER-16 (Figure 2b) resembles closely that of CE-16 (Figure 2a). For the unassimilated S data, there are strong similarities between the distributions of RMS OMF of CE-16 (Figure 3a) and EnKF-16 (Figure 3c) and of ER-16 (Figure 3b) and ERKF-16 (Figure 3d).

To better understand the effect of the recentering step on the unassimilated S variable, Figures 4 and 5 show the differences of the RMS S OMF of CE-16 from those of EnKF-16 and ER-16. Figure 4 corresponds to the upper 200 meters. Figure 5 is for the 200-2000 meter depth range. Warm (cold) colors indicate that the analysis is closer to (further away from) the unassimilated Argo S data than the control. Clearly, the better performance in terms of RMS S OMF of the recentering step from ER-16 over the EnKF analysis from EnKF-16 is mostly due the recentering's effectiveness in the upper 200 meters, where the RMS S OMF of EnKF-16 are not significantly different from those of CE-16 (Figure 4). Both the recentering and the EnKF analysis are effective below 200 meters, but while ER-16 is generally closer than both EnKF-16 and CE-16 to Argo S in the tropics where there are more Argo profiles than in the extratropics, it is less effective south of 45°S where the observations are sparse. However, when the recentering is applied first and then followed by the EnKF analysis, as is the case in ERKF-16, the poor performance of the recentering in the data-sparse high southern latitudes is compensated by the EnKF analysis ability to use its background error covariance estimates to propagate information from data rich regions to data poor regions.

6. Conclusions

The ensemble recentering Kalman filter (ERKF) first applies a recentering step to adjust the distribution of an ensemble of model states so that the ensemble mean is shifted to the prior position of the most likely sample without altering the ensemble statistics, after which an EnKF analysis step is applied. In our experiments, the ERKF is better able than the EnKF to improve model estimates of the unassimilated S variable when Argo T profiles are assimilated into the GEOS-5 CGCM. The effectiveness of the recentering step is attributed to its ability to gradually correct model biases over time and to the fact that it replaces the ensemble mean with a balanced state. Although the recentering is less effective in poorly observed regions, the EnKF analysis compensates for this fact by propagating information from data rich areas to data poor areas using its background error covariance estimates. In contrast to particle filters, which must

integrate a very large number of model trajectories to be competitive with the EnKF, the ERKF can improve upon the performance of the EnKF without requiring a larger ensemble.

Acknowledgement

This work is supported by NASA's Modeling Analysis and Prediction Program under WBS 802678.02.17.01.25. The infrastructure for the runs is provided by the NASA Center for Climate Simulation (NCCS). Yuri Vikhliaev, Max Suarez and Bin Zhao helped configure the GEOS-5 modeling system. Guillaume Vernieres helped configure the data assimilation system and Robin Kovach assisted with the observation preprocessing and with plotting the results. Michele Rienecker provided critical insights that helped refocus the initial draft manuscript.

References

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

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

Gordon, N. J.; D.J. Salmond, and A.F.M. Smith (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEEE Proceedings on Radar and Signal Processing*, **140 (2)**, 107–113.

Hill, C., C. Deluca, V. Balaji, M. Suarez and A. Da Silva (2004), The architecture of the earth system modeling framework, *Computing in Science and Engineering*, **6 (1)**, 18-28.

Hunke, E.C., W.H. Lipscomb, A.K. Turner, N. Jeffery, S. Elliott (2013), CICE: the Los Alamos Sea Ice Model Documentation and Software User's Manual Version 5.0, LA-CC-06-012, Los Alamos National Laboratory, Los Alamos NM.

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

Keppenne, C.L., and M.M. Rienecker (2003), Assimilation of temperature into an isopycnal ocean general circulation model using a parallel ensemble Kalman filter, *J. Mar. Sys.*, **40-41**, 363-380.

Keppenne, C.L., M.M. Rienecker, N.P. Kurkowski, and D.A. Adamec (2005), Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction, *Nonlinear Processes in Geophysics*, **12**, 491-503.

Keppenne, C.L., M.M. Rienecker, J.P. Jacob, and R.M. Kovach (2008), Error covariance modeling in the GMAO ocean ensemble Kalman filter, *Mon. Wea. Rev*., **136**, 2964-2982.

Keppenne, C.L., M.M. Rienecker, R.M. Kovach and G. Vernieres (2013), Background error covariance estimation using information from a single model trajectory with application to ocean data assimilation, *Ocean Modelling*, submitted

Oke, P.R., G.B. Brassington, D.A. Griffin, and A. Schiller (2010), Ocean data assimilation: a case for ensemble optimal interpolation, *Australian Meteorological and Oceanographic Journal*, **59**, 67-76.

Rienecker, M.M., M.J. Suarez, R. Todling, J. Bacmeister, L. Takacs, H.-C. Liu, W. Gu, M. Sienkiewicz, R.D. Koster, R. Gelaro, I. Stajner, and J.E. Nielsen (2008), The GEOS-5 data assimilation system. Documentation Versions 5.0.1, 5.1.0, and 5.20, *NASA Tech. Rep. Series on Global Modeling and Data Assimilation* **27**, NASA/TM-2008-104606, 1–118.

311 Vernieres, G., C.L. Keppenne, M.M. Rienecker, J.P. Jacob and R.N. Kovach (2012), The
312 GEOS-ODAS description and evaluation. *NASA Tech. Rep. Series on Global Modeling and*
313 *Data Assimilation* **30**, NASA/TM-2012-104606.
314 Weerts, A.H., and G.Y.H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for
315 state updating with hydrological conceptual rainfall-runoff models, *Water Resources*
316 *Research*, **42 (9)**, W09043, 17pp.
317

Figure Captions

Figure 1. Schematic representation of the recentering step of the ERKF showing the sample mean, $\bar{\boldsymbol{x}}$ (doubly outlined small open circle), the most likely sample $\boldsymbol{x}_p$ (small open circle), the control observations, $\boldsymbol{y}$ (large open circle) and the recentering increment, $\boldsymbol{\Delta}_p$ (arrow connecting $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}_p$). The filled circles represent particles other than $\boldsymbol{x}_p$. The cluster of circles shown to the left represents the prior sample forecast. The cluster to the right represents the recentered sample.

Figure 2. RMS OMF differences with respect to the assimilated Argo T data averaged vertically and binned to 3°×3° boxes over April-June 2006 in (a) CE-16, (b) ER-16 (c) EnKF-16 and (d) ERKF-16.

Figure 3. Same as Figure 2 for the unassimilated Argo S data.

Figure 4. Difference of CE-16 RMS OMF for Argo S from those of (a) EnKF-16 and (b) ERKF-16 averaged over the upper 200 meters for the unassimilated Argo S data. Warm (cold) colors indicate areas where the analysis is closer to (further from) the observations than CE-16.

Figure 5. Same as Figure 4 for the 200-2000 meter depth range.