Hybrid Variational Ensemble Data Assimilation with Initial Condition and Model Physics Uncertainty

Luke T. Peffers
Luke.peffers@gmail.com
Department of Earth, Ocean, and Atmospheric Sciences, Florida State University

1. INTRODUCTION

The objective of this research is to evaluate the performance of the hybrid ensemble transform Kalman filter three-dimensional variational (ETKF-3DVAR) data assimilation scheme (Wang et al. 2008a,b) that was developed for the Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) and to explore potential improvements in the hybrid scheme. I employ realistic NWP experiments with various ensemble formulations to provide flow-dependent error covariances to the hybrid cost function that fit the observations. I conduct experiments of the hybrid system using ensembles that include ETKF initial perturbations, as well as physical parameterization diversity to account for model error.

2. DESCRIPTION OF THE HYBRID SCHEME

The WRF hybrid ETKF-3DVAR system adapts the extended control variable methodology (Lorenc 2003) to include estimates of ensemble covariance in the standard 3DVAR scalar cost function. The hybrid incremental cost function is defined as

$$J(\mathbf{x}_1', \alpha) = \frac{1}{2}\beta_1 (\mathbf{x}_1')^{\mathbf{T}} \mathbf{B}^{-1} (\mathbf{x}_1') + \frac{1}{2}\beta_2 \alpha^{\mathbf{T}} \mathbf{A}^{-1} \alpha + \frac{1}{2}(\mathbf{y}^{o'} - \mathbf{H}\mathbf{x}')^{\mathbf{T}} \mathbf{R}^{-1} (\mathbf{y}^{o'} - \mathbf{H}\mathbf{x}'). \quad (1)$$

Thus, the hybrid analysis increment, defined as

$$\mathbf{x}' = \mathbf{x}_1' + \sum (\alpha_k \circ \mathbf{x}_k^e), \quad (2)$$

the sum of the 3DVAR analysis increment and the increment associated with the ensemble covariance. In (2), $\mathbf{x}_1'$ is the 3DVAR analysis increment, ($\mathbf{x_1} - \mathbf{x}^b$), where $\mathbf{x}^b$ is the background or first-guess vector. The vector $\mathbf{x}^e$ contains the ensemble perturbations about the mean normalized by $(K\text{-}1)^{1/2}$, where K is the ensemble size. The extended control variables, α_k are vectors that multiply $\mathbf{x}^e$, where $\circ$ is the Schur product (element by element product). The vector α is formed by concatenating K vectors α_k, k = 1,…,K. The vector $\mathbf{y}^o$ is the observation vector, and $\mathbf{y}^{o'} = \mathbf{y}^o - H\mathbf{x}^b$ is the innovation vector (observation minus the background transformed to observation space). The block diagonal matrix **A**, with K blocks containing prescribed correlation matrices **S,** defines the spatial correlation of α, thereby defining the scale of the ensemble error covariance localization. A full derivation of these equations is given in Lorenc (2003) and references therein. Terms **H**, **B** and **R** in (1) are the standard observation operator and background and observation error covariance matrices, respectively.

Adjustable terms β_1 and β_2 determine the weights of the 3DVAR background-error covariances and the ensemble covariance. The restriction that $1/\beta_1 + 1/\beta_2 = 1$ ensures that the total background error covariances are conserved (Wang et al. 2008a). Thus, when $1/\beta_1 = 1$, the analysis is determined solely by the 3DVAR static background error covariances, while $1/\beta_2 = 1$ means that the analysis is determined solely by the ensemble covariances.

The ensemble transform Kalman filter (ETKF; Bishop et al. 2001; Wang and Bishop 2003) is used to transform forecast perturbations into analysis perturbations in a manner that is consistent with the EnKF analysis update equation. The analysis perturbations then are added to the hybrid analysis field to produce a new set of initial conditions for the next forecast cycle. ETKF is considered sub-optimal since it does not include covariance localization causing it to rely heavily on covariance inflation. Thus, an alternative to ETKF that is completely contained within and consistent with the WRF hybrid cost function minimization is explored in this study.

3. PROPOSED ALTERNATIVE TO ETKF

The proposed alternative to ETKF is termed the hybrid Lanczos ensemble filter (HLEF). The hybrid analysis increment $\mathbf{x}'$ was defined previously in (2) as the local (Schur product) linear combination of ensemble perturbations. The ETKF formulation does not include a contribution from the 3DVAR background error covariance. Therefore, neglecting for now the contribution from the increment associated with 3DVAR, the analysis increment (2) is defined as

$$\mathbf{x}' = \sum_{k=1}^{K} (\alpha_k \circ \mathbf{x}_k^e)$$

Thus, if the analysis increment is determined solely by ensemble perturbations (i.e., $1/\beta_2=1$), the analysis-error covariance in state space $\mathbf{P_a}$ is defined as in the Maximum Likelihood Ensemble Filter (MLEF: Zupanski et al. 2008; Zupanski 2005), by

$$\mathbf{P}_a = \mathbf{P}_f^{1/2} \mathbf{K}^{-1} \mathbf{P}_f^{\mathbf{T}/2}, \quad \text{where}$$

$$\mathbf{K}^{-1} = \left(\nabla^2 J(\mathbf{x}')\right)^{-1}$$

represents the inverse Hessian of the cost function, or an appropriate equivalent, at the optimal point in the sub-space spanned by the ensemble perturbations. The ETKF transform used in the current study provides an equivalent eigenvalue decomposition of the generalized form $\mathbf{K}^{-1}$ (Zupanski et al. 2008). Here, instead of approximating $\mathbf{K}^{-1}$ with ETKF that describes the analysis-error covariance corresponding to EnKF, the estimation of **K** from the Lanczos minimization algorithm is exploited to approximate **K** with $\mathbf{L}_\mathrm{n}$, where

$$\mathbf{L}_\mathrm{n} = \mathbf{Q}_n(\mathbf{L}_n\mathbf{D}_n\mathbf{L}_n^{\mathbf{T}})\mathbf{Q}^{\mathbf{T}}$$

represents the approximation of **K** after *n* iterations of the cost function minimization. Here, $\mathbf{Q}_\mathrm{n}$ contains the *n* Lanczos vectors, and $\mathbf{D}_\mathrm{n}$ and $\mathbf{L}_\mathrm{n}$ are diagonal and lower bi-diagonal matrices (see Golub and Van Loan (1996) for details). As the number of iterations *n* increases, $\mathbf{L}_\mathrm{n}$ becomes a better approximation of **K**. Therefore, the ability of $\mathbf{L}_\mathrm{n}$ to approximate **K** depends on the degrees of freedom available for minimization. For a *K*-member ensemble, one can expect at most that the $n = K$ largest eigenvalues of $\mathbf{L}_\mathrm{n}$ will emerge from the minimization process. The eigenvalue decomposition of $\mathbf{L}^{-1/2}$ is used to construct a symmetric square-root matrix representing $\mathbf{K}^{-1/2}$, similar to both ETKF and MLEF.

By prescribing the correlation matrix **A** in (1), as in Wang et al. 2008ab, the effect is to localize control variables with a Gaussian function that reduces the coefficients' influence to a user-determined horizontal distance surrounding an assimilated observation. This has the effect of increasing the degrees of freedom that are available to the DA system (Lorenc 2003). It is hypothesized that the additional degrees of freedom and removal of spurious correlations at distant grid points through localization will improve the estimate, thereby providing more skillful initial ensemble perturbations than ETKF. Furthermore, HLEF also includes the effect of hybridization as governed by the adjustable terms β_1 and β_2, also not possible with ETKF.

4. METHODOLOGY

The study domain was centered over East Asia using version 3.2 of the Advanced Research WRF (ARW). The experiments used a 45-km horizontal grid spacing with 160 × 160 grid points, 35 vertical levels, and a model top at 50 hPa. The simulations were run for a month period (1 March through 31 March 2010) to ensure that the filter had time to converge to a steady analysis cycle. The initial ensembles at the start of the assimilation cycle, and the LBC ensembles throughout the cycles, were generated by adding random perturbations to the NCEP FNL analysis. The random perturbations were taken from a normal distribution with zero mean and with the same covariance as the WRF 3DVAR background-error covariance (Torn et al. 2006). Model error is represented by constructing a 20-member ensemble with varying configurations of physical parameterizations. The results, including implicit model error, will be compared to those from a single model configuration. Then, the proposed HLEF ensemble generation algorithm is tested and compared to the ETKF cycling ensemble.

5. SUMMARY OF RESULTS

4a. Ensemble spread

- The ETKF perturbations used as an ensemble-generation scheme were able to maintain the ensemble spread that is appropriate for the WRF hybrid scheme. However, rank histograms of the ensemble spread indicate that the ensembles are systematically under-dispersive. Conversely, the inflation coefficient that measure the ration of innovation variance over ensemble variance converge to unity, indicating that the ensembles' spread is adequate. Since these diagnostic tools rely on imperfect observations the true degree of skill cannot be known for sure.
- The multi-physics ETKF ensemble improved the characteristics of ensemble spread compared to the single-physics ensemble by allowing the ensemble spread to 1) reduce dependence on covariance inflation, 2) sample a larger range of innovation variance, and 3) maintain variance more evenly in the available directions of ensemble sub-space.
- The multi-physics ETKF ensemble was characterized by larger (smaller) error growth (reduction) during the model integration than the single-physics ensemble as measured by ensemble spread. Ensemble error reduction is an area of concern since ETKF perturbations should excite fast-growing errors and cause ensemble spread to increase, not decrease. Including physical-parameterization diversity appears to mitigate this issue.
- Use of the ensemble mean as the first guess in the 3DVAR cost function significantly improved the skill of the analyses.
- Tuning the static 3DVAR background error covariances using the ETKF ensemble perturbations instead of time-lagged perturbations improved the skill of the deterministic and ensemble 3DVAR analyses as measured by 12- through 48-h deterministic forecast skill.

4b. Hybrid data assimilation

- Incorporating ensemble-based flow-dependent error covariances from limited 20-member ensembles into the hybrid cost function added significant skill to the analyses. This added skill was in addition to the skill achieved by using the ensemble mean as the first guess and using the tuned background error covariances. See Fig. 1.
- The greatest improvements in analysis skill were observed when a multi-physics ensemble was used to supply error covariances to the hybrid cost function. See Fig. 1.
- Vertical localization adds skill (although not statistically significant) to the analyses of wind components but mostly at longer lead times and when the localization length scale is less restrictive. See Fig. 2.

4c. Proposed alternative to ETKF

- The proposed HLEF ensemble generation scheme was shown to be equivalent to the ETKF scheme (as theory suggests) when no inflation was applied and the HELF perturbations did not include the effect of covariance localizations or hybridization.
- Both vertical and horizontal covariance localization in the HLEF perturbations ameliorated the under estimation of analysis uncertainty.
- 10-day cycling experiments with adaptively-inflated and localized HLEF perturbations required less than 30% of the inflation required by ETKF. See Fig. 3.
- Experiments addressed the possibility of producing analysis perturbations that are consistent with the hybrid variational cost function by including the effect of covariance localization and hybridization of the hybrid cost

function thereby reducing the dependence of covariance inflation. Further work is needed to fully explore this approach.

REFERENCES

Bishop C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.

Golub,G. H. and C. F.Van Loan,1996: Matrix Computations, John Hopkins University Press, Baltimore

Lorenc A. C., 2003: The potential of the ensemble Kalman filter for NWP—A comparison with 4Dvar. *Quart. J. Roy. Meteor. Soc.*, **129**, 3183–3203.

Skamarock W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR.

Torn R. D., G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.*, **134**, 2490–2502.

Wang, X., D.M. Barker, C. Snyder, and T.M. Hamill, 2008: A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part I: Observing system simulation Experiment. *Mon. Wea. Rev.*, **136**, 5116–5131.

Wang X., D. M. Barker, C. Snyder, and T. M. Hamill, 2008: A hybrid ETKF–3DVAR data assimilation scheme for the WRF Model. Part II: Real observation experiments. *Mon. Wea. Rev.*, **136**, 5132–5147.

Wang X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158.

Zupanski, M, 2005: Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Wea. Rev.*, **133**, 1710–1726.

Zupanski, M., D. Zupanski, D. F. Parrish, E. Rogers, G. DiMego, 2002: Four-dimensional variational data assimilation for the blizzard of 2000. *Mon. Wea. Rev.*, 130, 1967–1988.

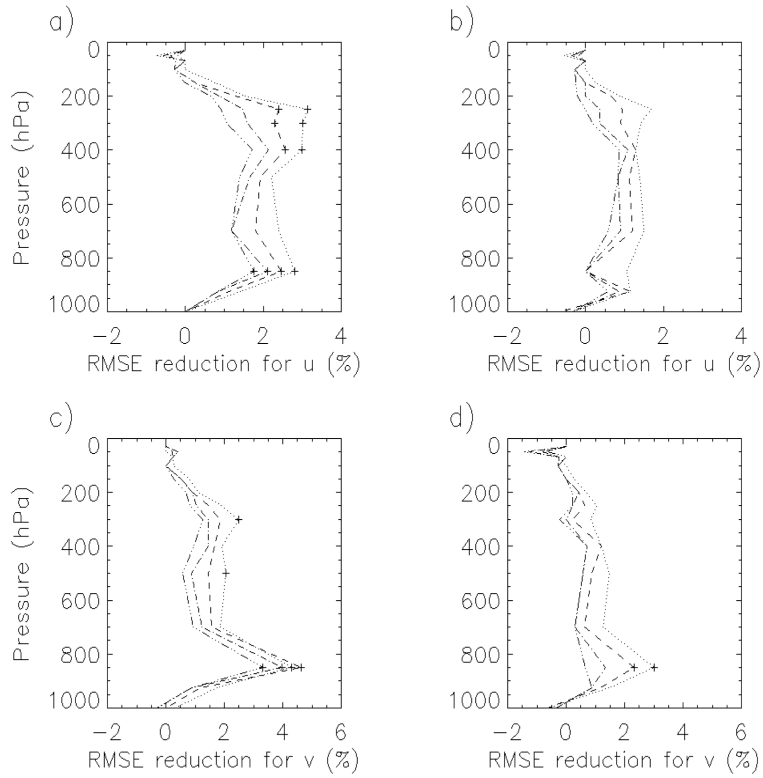

Fig. 1. Percent reduction of time averaged wind component rmse scores for hybrid experiments using $1/\beta_2 = 0.2$ (i.e., 20% contribution from ensemble-based covariances) compared to ensemble *3DVAR*. Results for four horizontal localization length scales (S) are shown: 250-km (dotted), 500-km (dashed), 1000-km (dot/dash), and 1500-km(3-dot/dash). Left panels are for multi-physics experiments compared to ensemble *3DVAR* with 3DVAR background error covariance using the ensemble method with a multi-physics ETKF ensemble. Right panels are for single-physics experiments compared to ensemble *3DVAR* with 3DVAR background error covariance using the ensemble method with a single-physics ETKF ensemble. Scores are shown for 12-h deterministic forecasts from each of the 59 analyses from 1200 UTC 1 March through 0000 UTC 31 March 2010. (+) symbols indicate statistically-significant differences at the 95% confidence interval. Values greater than zero indicate improvement, while negative values indicate degraded forecast skill.

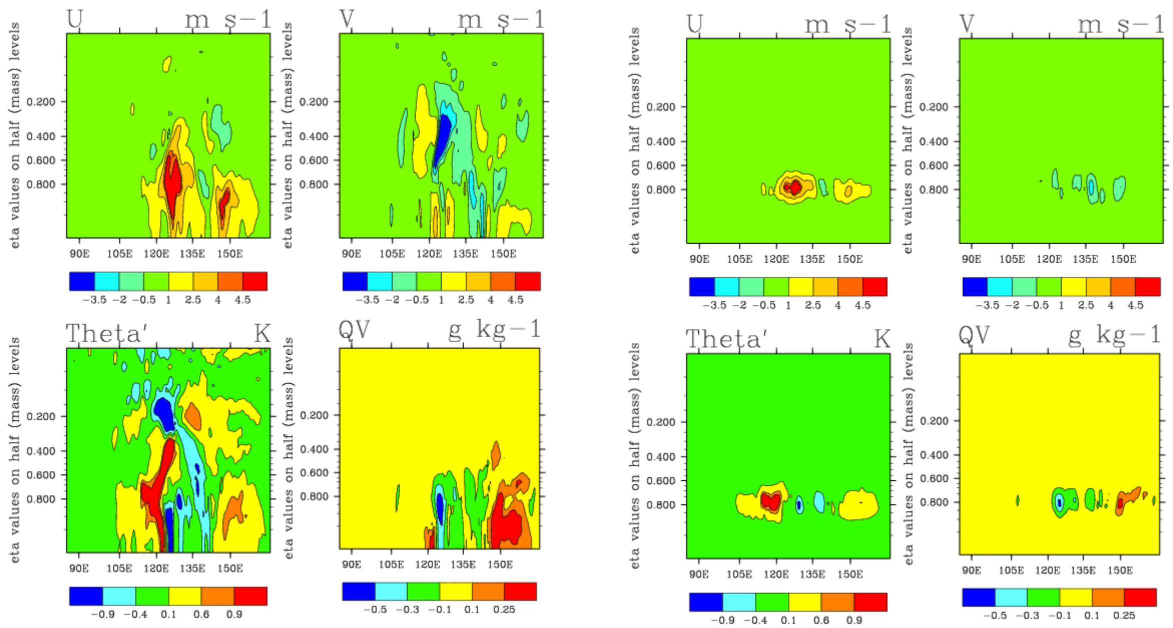

Fig. 2. Cross sections for pseudo observation experiment showing the analysis increments for constant length scale vertical covariance localization functions. Left panel shows analysis increments with 26 grid point vertical localization and right panel is shows two grid point vertical localization.

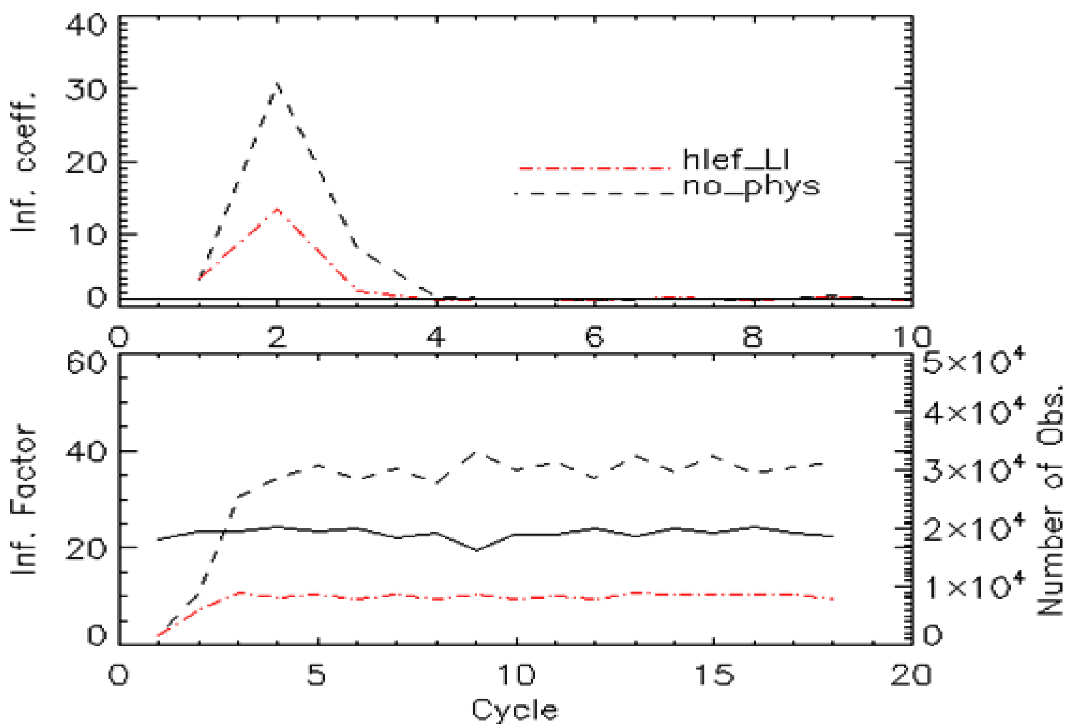

Fig. 31. Adaptive inflation coefficients (top) and inflation factors (bottom dashed) for the 10-day cycling experiment from 1-10 March 2010 for single-physics cycling ensemble simulations using ETKF ("no_phys" in the figure) with adaptive inflation factor and HLEF with horizontal localization (1000 km) and adaptive inflation factor. The number of observations (solid line) is shown in the bottom panel.