P1.65. ENSEMBLE DATA ASSIMILATION AND INFORMATION THEORY

Dusanka Zupanski[1], Milija Zupanski[1], Mark DeMaria[2], Louis Grasso[1], Arthur Y. Hou[3], Sara Zhang[3], and Daniel Lindsey[1]
[1] Cooperative Institute for Research in the Atmosphere
Colorado State University
Fort Collins, Colorado, U. S. A.
[2] NOAA/NESDIS, Fort Collins, Colorado, U. S. A.
[3] NASA Goddard Space Flight Center
Global Modeling and Assimilation Office
Greenbelt, Maryland, U. S. A.

1. INTRODUCTION

Some of the most demanding goals of Numerical Weather Prediction (NWP) and new observing missions are (*i*) to determine analysis and forecast uncertainties, and (*ii*) to estimate information content of new observations. Achieving these goals within a unified methodology is desirable since these two goals are not independent: information content of new observations is dependent on the prior knowledge about the atmospheric state, which is often defined in terms of analysis and forecast uncertainties.

Ensemble based data assimilation approaches (Evensen 1994; Houtekamer and Mitchell 1998; Hamill and Snyder 2000; Keppenne 2000; Mitchell and Houtekamer 2000; Anderson 2001; Bishop et al. 2001; van Leeuwen 2001; Reichle et al. 2002a,b; Whitaker and Hamill 2002; Tippett et al. 2003; Zhang et al. 2004; Ott et al. 2005; Szunyogh et al. 2005; Zupanski 2005; Zupanski and Zupanski 2005) have a capability to update forecast error covariance, thus having a potential for appropriately addressing the first goal. Information theory (e.g., Shannon and Weaver 1949; Rodgers 2000) provides a theoretical framework for addressing the second goal. In this study we examine a unified approach employing both ensemble data assimilation and information theory.

2. METHODOLOGY

An ensemble-based method entitled Maximum Likelihood Ensemble Filter (MLEF, Zupanski 2005; Zupanski and Zupanski 2005) is used in this study as a data assimilation component of this general approach. Information measures defined in terms of Degrees of Freedom (DOF) for signal and entropy reduction (e.g., Rodgers 2000) are the components representing information theory within this general approach. This methodology is shortly described here.

The MLEF seeks a maximum likelihood state solution employing an iterative minimization of a cost function. The solution for an augmented state vector $\boldsymbol{x}$ (including initial conditions, model error, and empirical parameters), of dimension N_{state}, is obtained by minimizing a cost function J defined as

$$J(\boldsymbol{x}) = \frac{1}{2}[\boldsymbol{x}-\boldsymbol{x}_b]^T \boldsymbol{P}_f^{-1}[\boldsymbol{x}-\boldsymbol{x}_b] + \frac{1}{2}[\boldsymbol{y}-H(\boldsymbol{x})]^T \boldsymbol{R}^{-1}[\boldsymbol{y}-H(\boldsymbol{x})] \quad , (1)$$

where $\boldsymbol{y}$ is an observation vector of dimension equal to the number of observations (N_{obs}) and H is a nonlinear observation operator. Subscript b denotes a background (i.e., prior) estimate of $\boldsymbol{x}$, and superscript T denotes a transpose. The $N_{obs} \times N_{obs}$ matrix $\boldsymbol{R}$ is a prescribed observation error covariance. The matrix $\boldsymbol{P}_f$ of dimension $N_{state} \times N_{ens}$ is the forecast error covariance (N_{ens} being the ensemble size).

Uncertainties of the optimal estimate of the state $\boldsymbol{x}$ are also calculated by the MLEF. The uncertainties are defined as square roots of the analysis error covariance ($\boldsymbol{P}_a^{\frac{1}{2}}$) and the forecast error covariance ($\boldsymbol{P}_f^{\frac{1}{2}}$), both defined in terms of ensemble perturbations. The square root of the analysis error covariance is obtained as

$$\boldsymbol{P}_a^{\frac{1}{2}} = \begin{bmatrix} \boldsymbol{p}_a^1 & \boldsymbol{p}_a^2 & \dots & \boldsymbol{p}_a^{N_{ens}} \end{bmatrix} = \boldsymbol{P}_f^{\frac{1}{2}}(\boldsymbol{I}_{ens}+\boldsymbol{C})^{-\frac{1}{2}} \quad , (2)$$

* *Corresponding author address*: Dusanka Zupanski, Colorado State University/CIRA, Foothills Campus, Fort Collins, CO 80523 (e-mail: Zupanski@cira.colstate.edu)

where $\boldsymbol{I}_{ens}$ is a diagonal identity matrix of dimension $N_{ens}\times N_{ens}$, and $\boldsymbol{p}_a^i$ are column vectors representing analysis perturbations in ensemble subspace. Matrix ***C*** of dimension $N_{ens}\times N_{ens}$ is defined as

$$\boldsymbol{C} = \boldsymbol{Z}^T \boldsymbol{Z} \quad ; \quad \boldsymbol{z}^i = \boldsymbol{R}^{-\frac{1}{2}} H(\boldsymbol{x} + \boldsymbol{p}_f^i) - \boldsymbol{R}^{-\frac{1}{2}} H(\boldsymbol{x}) \text{ ,(3)}$$

where vectors $\boldsymbol{z}^i$ are columns of the matrix ***Z*** of dimension $N_{obs}\times N_{ens}$. Note that, when calculating $\boldsymbol{z}^i$, a nonlinear operator *H* is applied to perturbed and unperturbed states ***x***. Vectors $\boldsymbol{p}_f^i$ are the columns of the square root of the forecast error covariance matrix obtained via ensemble forecasting employing a nonlinear dynamical model *M* (e.g., an NWP model)

$$\boldsymbol{P}_f^{\frac{1}{2}} = \left[\boldsymbol{p}_f^1 \quad \boldsymbol{p}_f^2 \quad \cdots \quad \boldsymbol{p}_f^{N_{ens}} \right] \quad ;$$
$$\boldsymbol{p}_f^i = M(\boldsymbol{x}_a + \boldsymbol{p}_a^i) - M(\boldsymbol{x}_a) \ , \qquad (4)$$

where $\boldsymbol{x}_a$ is the optimal solution for the model state (analysis).

Equations (1)-(3), referred to as analysis equations, are solved iteratively in each data assimilation cycle, while equation (4), referred to as a forecast equation, is used to advance the columns of the forecast error covariance matrix $\boldsymbol{P}_f^{\frac{1}{2}}$ from one cycle to another.

Measures of information content of observations referred to as DOF for signal and entropy reduction, denoted d_s and *h*, respectively, are often used in information theory (e.g., Rodgers 2000). In data assimilation applications, d_s and *h* are commonly defined in terms of analysis and forecast error covariances, $\boldsymbol{P}_a$ and $\boldsymbol{P}_f$, (e.g., Wahba 1985; Purser and Huang 1993; Wahba et al. 1995; Rodgers 2000; Rabier et al. 2002; Fisher 2003; Johnson 2003; Engelen and Stephens 2004). These information measures can also be calculated employing the eigenvalues λ_i^2 of thr matrix ***C***, defined in (3), that we also refer to as *the information matrix in ensemble subspace*. Thus, the following formulas for DOF for signal d_s and entropy reduction *h* can be used:

$$d_s = \sum_i \frac{\lambda_i^2}{(1+\lambda_i^2)} \ ; \ h = \frac{1}{2} \sum_i \ln(1+\lambda_i^2) \ , \qquad (5)$$

which are essentially the same formulas as in Rodgers (2000). The difference is that the eignevalues of the information matrix defined in ensemble subspace (***C***) are used in our formulation, while in the formulation of Rodgers (2000), the eigenvalues of the information matrix, defined either in the model space or in the observation space, are used. The advantage of the information matrix defined in ensemble subspace is that it is commonly a small matrix (of dimensions $N_{ens}\times N_{ens}$), so it is possible to evaluate the full eigenvalue spectrum of it, even when using complex NWP models and numerous observations. A potential disadvantage is that a small ensemble size might be insufficient to accurately determine the information measures. The experimental results examining the impact of ensemble size on the information measures will be presented and discussed in this paper.

3. EXPERIMENTAL RESULTS

3.1 Experiments with GEOS-5 single column model

Experiments examining the impact of ensemble size on the information content measures are performed using a single column version of the Goddard Earth Observing System (GEOS-5) Atmospheric General Circulation Model (AGCM). In Fig.1, experimental results obtained using simulated observations of temperature and humidity are shown. The location chosen for the experiments is a Tropical Western Pacific site (130E, 15N). Experimental results over a 10-day period from May 7 to May 17, 1998 are shown in Fig. 1. Information measures d_s, calculated in data assimilation experiments employing 10, 20, and 40 ensemble members, are plotted as functions of analysis cycles. It can be seen that values of d_s are smaller in the experiments with smaller number of ensembles. However, the relative changes in the information content from one cycle to another (i.e., trends of increase or decrease) are similar in all experiments. These results indicate that even insufficient ensemble sizes (10 and 20 in this case) could still provide useful comparisons of the information content measures, providing the comparisons are done only within the experiments with the same number of ensemble members. More detailed information content analysis employing GEOS-5 single column model can be found in Zupanski et al. 2005).

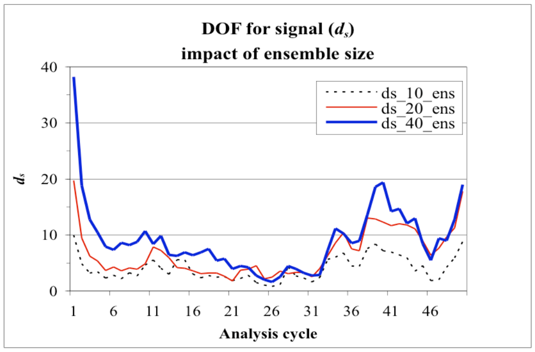

Fig 1. Values of d_s plotted as functions of analysis cycles, calculated in the experiments with 10 (*ds_10_ens*), 20 (*ds_20_ens*), and 40 (*ds_40_ens*) ensemble members. Simulated observations of temperature and humidity (total of 80 observations) are assimilated in each data assimilation cycle. Note that the values are generally larger for larger ensemble sizes. The essential trends (changes from one cycle to another) are similar in all experiments.

3.2 Experiments with RAMS model

The information content analysis is also performed employing a more complex atmospheric model, the Colorado State University Regional Atmospheric Modeling System (RAMS, Pielke et al., 1992; Cotton et al., 2003). This is a non-hydrostatic model that integrates predictive equations for wind components, Exner function, ice-liquid water potential temperature, and total water mixing ratio on a vertically stretched Arakawa C-grid. This research is performed with the aim to develop a methodology for information content analysis of future GOES-R observations.

We have performed two different types of information content analysis: (*i*) *conditional information content analysis*, and (*ii*) *unconditional information content analysis*. Conditional information content analysis estimates information content of new observations taking into account information content of observations assimilated previously, and it is dependent on the order of observations. Unconditional information content analysis takes into account information from new observations without considering previously assimilated observations, and it is not dependent on the order of observations.

Examples of conditional and unconditional information content analysis are shown in Fig. 2. These experiments are performed for Hurricane Lili case, which occurred from 21 September 2002 to 04 October 2002. Simulated observations, grouped in 24 groups, including wind (***u***, ***v***, and ***w*** component), perturbation Exner function (***p***), ice-liquid water potential temperature (***th***) and total water mixing ratio (***r***) observations, are used in the experiments. As the figure indicates, the conditional information content measure DOF for signal (d_s) is dependent on the order of assimilation of various observation groups (forward or reverse order). Unconditional information content is not dependent on the order of assimilation. The values of d_s obtained in the experiment using unconditional information content analysis are larger than the values obtained in the two experiments using conditional information content analysis. This is because the information content of each data group is calculated independently of other groups in the unconditional information content analysis.

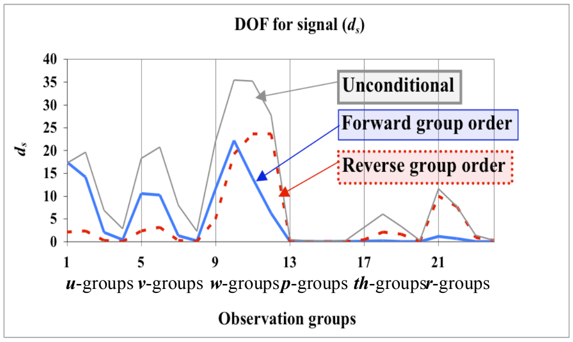

Fig. 2. DOF for signal (d_s) for various groups of observations in data assimilation experiments with the conditional information content calculations, using (i) forward group order and (ii) reverse group order. (iii) The unconditional information content calculation results are also shown. In groups numbered 1 through 4 ***u***-observations are assimilated, in groups 5 through 8 ***v***-observations are assimilated and so on, as indicated on the horizontal axis. Note that ***w***-groups of observations carry most information in all experiments. Experimental results employing 50 ensemble members are shown. The size of the control vector ***x*** is 54000. In each group of data, 1200 observations are used.

4. SUMMARY

Preliminary results presented in this study indicate that it is possible to effectively calculate information content measures of complex atmospheric models with large-scale vectors of state variables, using numerous observations. This is of special importance for assimilation of current and future satellite observations. Further studies are planned in the future in applications to real observations.

Acknowledgements

This research was supported by the GOES-R Risk Reduction Project under NOAA Grant NA17RJ1228 and by NASA grants 621-15-45-78, NAG5-12105 and NNG04GI25G. The views, opinions, and findings in this report are those of the authors and should not be construed as an official NOAA and or U. S. Government position, policy, or decision.

References

Anderson, J. L., 2001: An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.,* **129**, 2884–2903.

Bishop, C. H., B. J. Etherton, and S. Majumjar, 2001: Adaptive sampling with the ensemble Transform Kalman filter. Part 1: Theoretical aspects. *Mon. Wea. Rev.,* **129**, 420–436.

Cotton, W.R., R.A. Pielke, Sr., R.L. Walko, G.E. Liston, C. J. Tremback, H. Jiang, R. L. McAnelly, J. Y. Harrington, M.E. Nicholls, G. G. Carrió and J. P. Mc Fadden 2003: RAMS 2001: Current Status and future directions. *Meteor. Atmos. Phys.*, **82**, 5-29.

Engelen, R. J., and Stephens G. L., 2004: Information Content of Infrared Satellite Sounding Measurements with Respect to CO_2. *J. Appl. Meteor.* **43**, 373–378.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, (C5), 10143-
10162.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter/3D-variational analysis scheme. *Mon. Wea. Rev.,* **128**, 2905–2919.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126**, 796–811.

Johnson, C., 2003: Information content of observations in variational data assimilation. Ph.D. thesis, Department of Meteorology, University of Reading, 218 pp. [Available from University of Reading, Whiteknights, P.O. Box 220, Reading, RG6 2AX, United Kingdom.]

Keppenne, C., 2000: Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Wea. Rev.,* **128**, 1971–1981.

Mitchell, H. L, and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.,* **128**, 416–433.

Ott, Edward, B. R. Hunt, I. Szunyogh, A. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, and J. A. Yorke, 2005: A local ensemble kalman filter for atmospheric data assimilation. Posted http://arXiv.org/abs/physics/0203058. Submitted to *Tellus*

Pielke, R.A., W.R. Cotton, R.L. Walko, C.J. Tremback, W.A. Lyons, L.D. Grasso, M.E. Nicholls, M.D. Moran, D.A. Wesley, T.J. Lee, and J.H. Copeland, 1992: A comprehensive meteorological modeling system-RAMS. *Meteorol. Atmos. Phys.*, **49**, 69-91.

Purser, R.J. and Huang, H.-L. 1993: Estimating effective data density in a satellite retrieval or an objective analysis. *J. Appl. Meteorol.*, **32**, 1092–1107.

Rabier F., Fourrie N., Djalil C., and Prunet P., 2002: Channel selection methods for Infrared
Atmospheric Sounding Interferometer radiances. *Quart. J. Roy. Meteor. Soc.,* **128,** 1011–1027.

Reichle, R. H., D. B. McLaughlin, D. Entekhabi, 2002a: Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Wea. Rev.*, **130**, 103–114.

Reichle, R.H., J.P. Walker, R.D. Koster, and P.R. Houser, 2002b: Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometorology*, **3**, 728-740.

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, 238 pp.

Shannon, C. E., and Weaver W., 1949: *The Mathematical Theory of Communication*. University of Illinois Press, 144 pp.

Szunyogh, I., Kostelich E. J., Gyarmati G., Patil D. J., Hunt B. R., Kalnay E., Ott E., and Yorke J. A., 2005: Assessing a local ensemble Kalman filter: Perfect model experiments with the NCEP global model. Submitted to *Tellus*.

Tippett, M., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square-root filters. *Mon. Wea. Rev.*, 131, 1485–1490.

van Leeuwen, P. J., 2001: An ensemble smoother with error estimates. *Mon. Wea. Rev.,* **129**,
709–728.

Wahba, G., Johnson D. R., Gao F., and Gong J., 1995: Adaptive tuning of numerical weather prediction models: Randomized GCV in three- and four-dimensional data assimilation. *Mon. Wea. Rev.*, **123**, 3358–3370.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.,* **130**, 1913–1924.

Zupanski D. and M. Zupanski, 2005: Model error estimation employing ensemble data assimilation approach. Submitted to *Mon. Wea. Rev*. [also available at http://rammb.cira.colostate.edu/projects/goes_r/applications.asp].

Zupanski D., M. Zupanski, A.Y. Hou, S. Zhang, C.D. Kummerow, and S.H. Cheung, 2005: Information theory and ensemble data assimilation. Submitted to *J. Atmos. Sci.*
[also available at
ftp://ftp.cira.colostate.edu/Zupanski/manuscripts/Geos5_Information_JAS.pdf]

Zupanski, M., 2005: Maximum Likelihood Ensemble Filter: Theoretical Aspects. *Mon. Wea. Rev.*, **133**, 1710–1726.