4.2. CRITICAL ISSUES OF ENSEMBLE DATA ASSIMILATION IN APPLICATION TO GOES-R RISK REDUCTION PROGRAM

Dusanka Zupanski[1], Milija Zupanski[1], Mark DeMaria[2] and Louis Grasso[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.

1. INTRODUCTION

Ensemble data assimilation (EnsDA) approaches offer a great potential for addressing state-of-the-art problems of data assimilation and ensemble forecasting (e. g., 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. 2002; Whitaker and Hamill 2002; Ott et al. 2004; Tippett et al. 2003; Zupanski 2004; Zupanski and Zupanski 2004). The research in this area is, however, mostly limited to idealized cases, employing simplified models and a relatively small number of observations.

The emergence of next generation GOES satellites, beginning with GOES-R, poses a serious challenge to data assimilation in general. One of the most critical issues to be resolved in EnsDA approaches is how to handle many degrees of freedom of numerous observations and complex atmospheric models, while keeping the ensemble size relatively small. This paper employs the estimation theory (e. g., Jazwinski 1970) in order to define a balance between the number of observations and the ensemble size.

2. METODOLOGY

According to the estimation theory, the information content of the observations $\boldsymbol{y}$ is defined by the *observability* or *information* matrix $\boldsymbol{C}$, defined as

$$\boldsymbol{C} = \boldsymbol{M}^T \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H} \boldsymbol{M} , \qquad (1)$$

where $\boldsymbol{M}$ and $\boldsymbol{H}$ are linearizations of a forecast model M and an observation operator H, respectively, and $\boldsymbol{R}$ is an observation error covariance matrix.

Motivated by definition (1) we use a symmetric N_{ens} x N_{ens} matrix $\boldsymbol{A}$ (N_{ens} being ensemble size), as a measure of the information content of the observations in the ensemble subspace. This matrix is defined as

$$\boldsymbol{A} = (\boldsymbol{R}^{-1/2} \boldsymbol{H} \boldsymbol{M} \boldsymbol{P}_f^{1/2})^T (\boldsymbol{R}^{-1/2} \boldsymbol{H} \boldsymbol{M} \boldsymbol{P}_f^{1/2}) , \qquad (2)$$

where $\boldsymbol{P}_f^{1/2}$ is the square root of the forecast error covariance matrix, obtained via ensemble forecasting. Eq. (2) is applicable to both an ensemble filter and an ensemble smoother. In case of an ensemble filter, as in this study, $\boldsymbol{M}=\boldsymbol{I}$ ($\boldsymbol{I}$ is an identity matrix) is used in (2). Matrix $\boldsymbol{A}$ is also used to define the analysis error covariance $\boldsymbol{P}_a^{1/2}$ in EnsDA (e. g., Bishop et al. 2001; Zupanski 2004; Zupanski and Zupanski 2004):

$$\boldsymbol{P}_a^{1/2} = \boldsymbol{P}_f^{1/2} (\boldsymbol{I} + \boldsymbol{A})^{-1/2} . \qquad (3)$$

According to (2) and (3) the observations, with potentially many degrees of freedom, are being projected onto the ensemble sub-space with N_{ens} degrees of freedom. This could be a serious problem when dealing with satellite data, since the number of observations can outnumber the ensemble size by several orders of magnitude. In such cases a strategy for dealing with numerous observations is needed.

The approach taken in this study is to divide the observation vector $\boldsymbol{y}$ into groups $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$ and process each group of observations $\boldsymbol{y}_i$ separately. The analysis obtained as a result of the assimilation of current observations ($\boldsymbol{y}_i$) is used as a guess for the next group of observations ($\boldsymbol{y}_{i+1}$).

The ideas of processing observations in separate groups, or processing one observation at a time, have also been used in the previous studies (e. g., Anderson 2001; Bishop et al. 2001; Ott et

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

al. 2004). The new aspect proposed here is to use the information matrix ***A***, defined in ensemble subspace, to determine the appropriate balance between the observation number and the ensemble size.

3. EXPERIMENTAL RESULTS

EnsDA algorithm used in this study is the Maximum Likelihood Ensemble Filter (MLEF), developed at Colorado State University (Zupanski 2004; Zupanski and Zupanski 2004). The MLEF is applied to a non-hydrostatic atmospheric model: Regional Atmospheric Modeling System (RAMS). The case chosen for this experiment is hurricane Lili, which occurred from 21 September 2002 to 04 October 2002. Data assimilation results, valid at 08 UTC 01 October 2002, are examined here.

Data assimilation experiments presented here are performed over 139 consecutive sub-cycles. In each sub-cycle 50 *observations* are used, all valid at 08 UTC 01 October 2002. This time corresponds to the end of a 1-hour data assimilation interval. *Observation*s are defined by imposing random perturbations on the true state, obtained by RAMS integration. Each group of 50 "observations" includes each component of the wind vector (u, v, w), perturbation Exner function (π), ice-liquid water potential temperature (θ), and total-water mixing ratio (*r_total*), and they are all evenly distributed over the integration domain. The observation error magnitude varies with each variable and each model level. The integration domain is centered over the Gulf of Mexico and includes 75x55x35 grid points, with the horizontal grid distance of 60 km. The control variable of the data assimilation problem includes the same six components as the *observations* (u, v, w, π, θ, and *r_total*). The size of the control variable is 866250 (=75x55x35x6) and the ensemble size is N_{ens}=50. Data assimilation experiments are performed assuming a perfect model; that is, the model error is neglected. Preliminary experimental results are given in Figures 1 and 2.

The eigenvalues λ_i of the matrix $(\boldsymbol{I}+\boldsymbol{A})^{-1/2}$ are shown in Figure 1 for sub-cycles 1, 2, 15 and 122. Also shown are the eigenvalues of an experiment with approximately 7000 observations and 50 ensemble members. One can immediately notice that the experiment with 7000 observations has a flat eigenvalue spectrum. This is an indication that the number of ensembles is not sufficient to describe all degrees of freedom present in the observations. As a result, many more ensemble members are necessary to cover the entire interval [0,1] of possible values of λ_i. On the other hand, the eigenvalues obtained in the experiment with 50 observations in the first sub-cycle are distributed over almost the entire interval [0,1]. This indicates that the ensemble size is appropriate. Further, as the data assimilation sub-cycles repeat, the eigenvalue spectrum changes; in particular, the number of eigenvalues close to the value of 1 increases. As a consequence, those eigenvalues have a negligible effect. For example, in the sub-cycle 1, all 50 eigenvalues are effectively used, while only 16-17 eigenvalues are effectively used in sub-cycle 15, and 11-12 in sub-cycle 122.

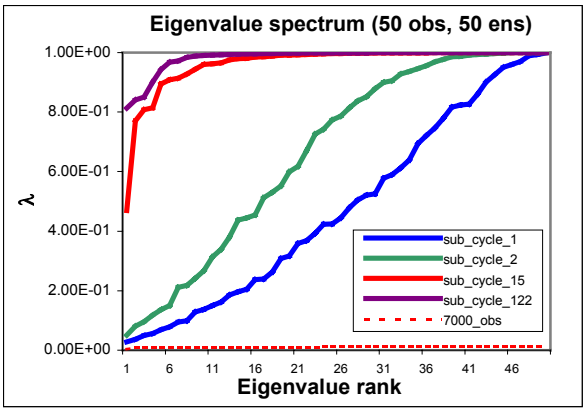

Figure 1: Eigenvalue spectrum of $(\boldsymbol{I}+\boldsymbol{A})^{-1/2}$ calculated in the MLEF experiments with 50 observations and 50 ensemble members in sub-cycles 1, 2, 15 and 122. The eigenvalues from the experiment with 7000 observations and 50 ensemble members obtained in first data assimilation cycle are also shown.

Based on the eigenspectrum of $(\boldsymbol{I}+\boldsymbol{A})^{-1/2}$ a strategy for balancing the ensemble size with the observation number can be developed; for example, data groups that do not bring *significant* amount of new information (e. g., sub-cycle 122) could either be excluded, or kept. Those that are kept need to be augmented with more observations. The strategy adopted in this

study is to keep groups of observations carrying *significant* new information. The observations are considered *significant* if

$$\gamma = \frac{\sum_{i=1}^{N_{ens}} \lambda_i}{N_{ens}} \le \gamma_c \qquad (4)$$

holds, where γ is the information content parameter and γ_c is a critical value (γ_c=0.97 is used in the experiments presented).

Figure 2 shows the Root Mean Square (RMS) errors of the analysis obtained in 45 successive sub-cycles, selected from 139 initial data groups according to (4), using 50 ensemble members and 50 observations in each sub-cycle. The RMS errors for the *u*-wind and *v*-wind components are shown in Figure 2a; the RMS errors for the total-water mixing ratio (*r_total*) are plotted in Figure 2b. The RMS errors are calculated with respect to the "truth", obtained by running the same model (RAMS) from different ("true") initial conditions. All model grid points are used in the RMS error calculations in each sub-cycle. The maximum magnitudes of the observation errors for *u* and *v* are denoted by *u_obs_err_max* and *v_obs_err_max*, respectively. Likewise the minimum error magnitudes for *u* and *v* could be denoted by *u_obs_err_min* and *v_obs_err_min*, respectively, however, both are identical and are denoted by *u_v_obs_err_min* (Figure 2a). Similarly *r_obs_err_max*, and *r_obs_err_min* are used for *r_total* (Figure 2b). As Figures 2a and 2b indicate, assimilation of observations in successive groups with observation numbers comparable to the ensemble size (50) is quite effective in reducing analysis error. One can also observe that the magnitude of the error reduction decreases with the increasing number of sub-cycles. This indicates that the system is learning about the true state and is approaching an asymptotic level of errors. This is also in agreement with the results presented in Figure 1; that is, the information content of the data decreases with repeating sub-cycles. One should be aware, however, that a perfect model is used in the experiments presented. In the case of an erroneous model, the learning process could be slower, or might not occur.

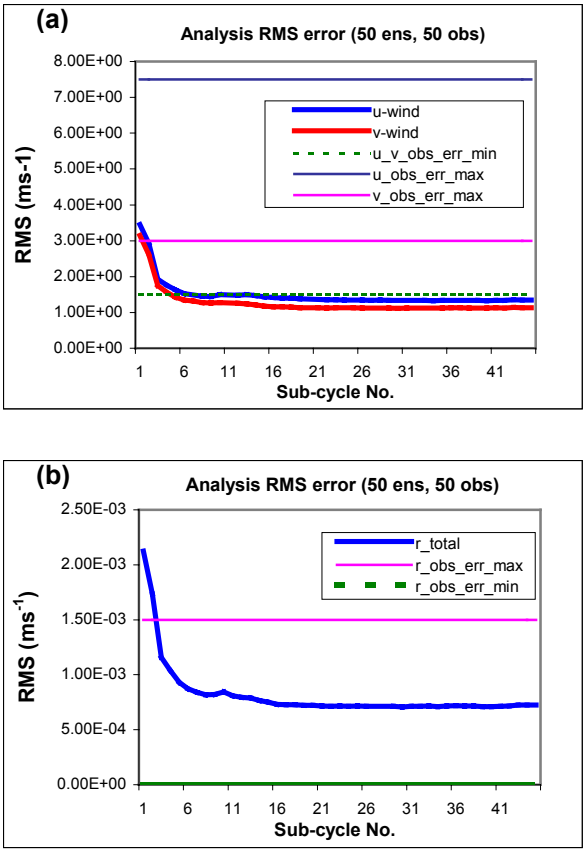

Figure 2: RMS errors of the analyses produced in the MLEF experiments over 45 sub-cycles. In each sub-cycle 50 observations and 50 ensemble members are used. The RMS errors for *u*- and *v*- wind, maximum (*u_obs_err_max*, *v_obs_err_max*) and minimum (*u_v_obs_err_min*) observation errors are given in (a). The RMS errors for *r_total*, along with maximum (*r_obs_err_max*) and minimum (*r_obs_err_min*) observation errors are given in (b).

4. SUMMARY

Preliminary results presented in this study indicate that it is possible to effectively assimilate a relatively large number of observations even though the ensemble size is relatively small. The information content of the observations, calculated in the ensemble sub-space, is a useful measure for defining an appropriate balance between the number of observations and the ensemble size. 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, and erroneous atmospheric models.

Acknowledgements

This research was partially supported by the GOES-R Risk Reduction Project under NOAA Grant NA17RJ1228. 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.

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.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

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, 2004: A local ensemble Kalman filter for atmospheric data assimilation. Posted http://arXiv.org/abs/physics/0203058. Submitted to *Mon. Wea. Rev.*

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

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.

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, 2004: Model error estimation employing ensemble data assimilation approach. Submitted to *Mon. Wea. Rev*. (also available at ftp://ftp.cira.colostate.edu/Zupanski/manuscripts/MLEF_model_err.revised.pdf).

Zupanski, M., 2004: Maximum likelihood ensemble filter: Theoretical aspects. Submitted to *Mon. Wea. Rev.* (also available at ftp://ftp.cira.colostate.edu/Zupanski/manuscripts/MLEF_MWR.pdf).