atmosphere

Editorial

Efficient Formulation and Implementation of Data Assimilation Methods

Elias D. Nino-Ruiz [1,*], Adrian Sandu [2] and Haiyan Cheng [3]

[1] Applied Math and Computational Science Laboratory, Department of Computer Science, Universidad del Norte, Barranquilla 080001, Colombia
[2] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA; asandu7@vt.edu
[3] Department of Computer Science, Willamette University, 900 State Street, Salem, OR 97301, USA; hcheng@willamette.edu
* Correspondence: enino@uninorte.edu.co; Tel.: +57-5-3509268

Received: 3 July 2018; Accepted: 4 July 2018; Published: 6 July 2018

Abstract: This Special Issue presents efficient formulations and implementations of sequential and variational data assimilation methods. The methods address three important issues in the context of operational data assimilation: efficient implementation of localization methods, sampling methods for approaching posterior ensembles under non-linear model errors, and adjoint-free formulations of four dimensional variational methods.

Keywords: ensemble Kalman filter; posterior ensemble; modified Cholesky decomposition; sampling methods; empirical orthogonal functions; Gaussian mixture models

1. Efficient Formulation and Implementation of Data Assimilation Methods

Data Assimilation is the process by which imperfect numerical forecasts and sparse observational networks are fused in order to estimate the state $\mathbf{x}^* \in \mathbb{R}^{n \times 1}$ of a system [1,2] which (approximately) evolves according to some model operator,

$$\mathbf{x}_p^* = \mathcal{M}_{t_{p-1} \rightarrow t_p}\left(\mathbf{x}_{p-1}^*\right), \text{ for } 1 \leq p \leq M, \tag{1}$$

where, for instance, $\mathcal{M} : \mathbb{R}^{n \times 1} \rightarrow \mathbb{R}^{n \times 1}$ is a numerical model which mimics the ocean and/or the atmosphere dynamics, n is the number of model components, M is the number of observations (which form the assimilation window), and p denotes time index at time t_p. Sequential and smoothing methods are commonly utilized in order to perform the estimation process [3–5]. In the context of sequential data assimilation, when Gaussian assumptions are done over background and observational errors, based on Bayes rule, the posterior mode of the error distribution can be computed as follows:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{A} \cdot \mathbf{H}^T \cdot \mathbf{R}^{-1} \cdot \left[\mathbf{y} - \mathcal{H}\left(\mathbf{x}^b\right)\right] \in \mathbb{R}^{n \times 1}, \tag{2a}$$

where $\mathbf{x}^a \in \mathbb{R}^{n \times 1}$ is known as the analysis state. The analysis covariance matrix reads as

$$\mathbf{A} = \left[\mathbf{B}^{-1} + \mathbf{H}^T \cdot \mathbf{R}^{-1} \cdot \mathbf{H}\right]^{-1} \in \mathbb{R}^{n \times n}, \tag{2b}$$

where m is the number of observed components from the model domain, $\mathcal{H} : \mathbb{R}^{n \times 1} \rightarrow \mathbb{R}^{m \times 1}$ is the observation operator, $\mathbf{B} \in \mathbb{R}^{n \times n}$ is the unknown background error covariance matrix, and $\mathbf{R} \in \mathbb{R}^{m \times m}$ stands for the data error covariance matrix. Likewise, $\mathcal{H}'(\mathbf{x}) \approx \mathbf{H}^T \in \mathbb{R}^{n \times m}$ is a linearized observation operator (with the linearization performed about the background state). Typically, the moments of the prior distribution,

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}^b, \mathbf{B}\right) , \tag{3}$$

can be estimated based on an ensemble of model realizations [6]. However, since ensemble members come at high computational costs owing to current operational data assimilation settings (i.e., numerical grid resolutions), ensemble sizes are bounded by the hundreds, while their underlying error distributions range on the order of billions [7]. Consequently, sampling errors can impact the quality of the analysis state [8]. In practice, localization methods can be utilized in order to mitigate the impact of sampling errors during the assimilation steps [9]. However, the implementation of localization methods is not immediate, and further analyses are needed before they can be implemented. Besides, for highly non-linear observation operators, ensemble-based methods such as the ensemble Kalman filter (EnKF) can fail to obtain reasonable estimates of posterior moments. Another important issue is that prior errors might not follow a normal distribution (as is commonly assumed), and therefore non-Gaussian models should be chosen to describe prior error distributions. For instance, Gaussian mixture models (GMMs) are an option in this context. This Special Issue addresses all of these important concerns in the context of sequential and variational data assimilation:

1. In the EnKF implementation based on a modified Cholesky decomposition (EnKF-MC) [10,11], the covariance matrix estimator proposed by Bickel and Levina in [12] and the conditional independence of model components regarding their spatial distances are exploited in order to obtain sparse Cholesky factors of the precision background error covariance matrix. This is done in order to reduce the computational cost of the analysis step, and to mitigate the impact of spurious correlations during the assimilation of observations. Given the relation between $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ in (2b) and by using the Bickel and Levina estimator, Nino-Ruiz proposes a "*A Matrix-Free Posterior Ensemble Kalman Filter Implementation Based on a Modified Cholesky Decomposition*" [13].
2. Non-linear observation operators can be commonly found in the context of observations mapped from satellite radiances. Consequently, posterior kernels of error distributions are no longer Gaussian. Thus, alternatives to EnKF formulations are a must under such circumstances, and therefore, sampling methods based on Markov chain Monte Carlo (MCMC) methods can be exploited to successfully sample from posterior error distributions. In "*Cluster Sampling Filters for Non-Gaussian Data Assimilation*" [14], Attia et al. propose filters which account for non-Gaussian errors in prior and observations. Furthermore, the convergence of MCMC is sped up by using Verlet integrators. On the other hand, Nino-Ruiz et al. [15] proposes "*A Robust Non-Gaussian Data Assimilation Method for Highly Non-Linear Models*" wherein prior errors are modeled by fitting GMMs while gradient approximations of the three-dimensional variational cost function are exploited for accelerating its convergence towards posterior modes.
3. The application to actual scenarios of operational data assimilation methods are widely discussed by Soldatenko et al. in [16] and by Kou et al. in [17].

We hope that you enjoy reading this Special Issue about efficient formulations and implementations of ensemble-based methods.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lorenc, A.C. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **1986**, *112*, 1177–1194. [CrossRef]
2. Sakov, P.; Bertino, L. Relation between two common localisation methods for the EnKF. *Computat. Geosci.* **2011**, *15*, 225–237. [CrossRef]

3. Buehner, M.; Houtekamer, P.; Charette, C.; Mitchell, H.L.; He, B. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. *Mon. Weather Rev.* **2010**, *138*, 1550–1566. [CrossRef]
4. Buehner, M.; Houtekamer, P.; Charette, C.; Mitchell, H.L.; He, B. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Weather Rev.* **2010**, *138*, 1567–1586. [CrossRef]
5. Caya, A.; Sun, J.; Snyder, C. A comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation. *Mon. Weather Rev.* **2005**, *133*, 3081–3094. [CrossRef]
6. Nino-Ruiz, E.D.; Sandu, A. Ensemble Kalman filter implementations based on shrinkage covariance matrix estimation. *Ocean Dyn.* **2015**, *65*, 1423–1439. [CrossRef]
7. Anderson, J.L. Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation. *Mon. Weather Rev.* **2012**, *140*, 2359–2371. [CrossRef]
8. Jonathan, P.; Fuqing, Z.; Weng, Y. The Effects of Sampling Errors on the EnKF Assimilation of Inner-Core Hurricane Observations. *Mon. Weather Rev.* **2014**, *142*, 1609–1630.
9. Buehner, M. Ensemble-derived Stationary and Flow-dependent Background-error Covariances: Evaluation in a Quasi-operational NWP Setting. *Q. J. R. Meteorol. Soc.* **2005**, *131*, 1013–1043. [CrossRef]
10. Nino-Ruiz, E.D.; Sandu, A.; Deng, X. A parallel ensemble Kalman filter implementation based on modified Cholesky decomposition. In Proceedings of the 6th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, Austin, TX, USA, 15 November 2015; ACM: New York, NY, USA, 2015; p. 4.
11. Nino-Ruiz, E.D.; Sandu, A.; Deng, X. A parallel implementation of the ensemble Kalman filter based on modified Cholesky decomposition. *J. Comput. Sci.* **2017**, in press. [CrossRef]
12. Bickel, P.J.; Levina, E. Regularized estimation of large covariance matrices. *Ann. Stat.* **2008**, *36*, 199–227. [CrossRef]
13. Nino-Ruiz, E.D. A Matrix-Free Posterior Ensemble Kalman Filter Implementation Based on a Modified Cholesky Decomposition. *Atmosphere* **2017**, *8*, 125. [CrossRef]
14. Attia, A.; Moosavi, A.; Sandu, A. Cluster Sampling Filters for Non-Gaussian Data Assimilation. *Atmosphere* **2018**, *9*, 213. [CrossRef]
15. Nino-Ruiz, E.D.; Cheng, H.; Beltran, R. A Robust Non-Gaussian Data Assimilation Method for Highly Non-Linear Models. *Atmosphere* **2018**, *9*, 126. [CrossRef]
16. Soldatenko, S.; Tingwell, C.; Steinle, P.; Kelly-Gerreyn, B.A. Assessing the Impact of Surface and Upper-Air Observations on the Forecast Skill of the ACCESS Numerical Weather Prediction Model over Australia. *Atmosphere* **2018**, *9*, 23. [CrossRef]
17. Kou, X.; Huang, Z.; Liu, H.; Zhang, M.; Shen, S.; Peng, Z. Evaluating the Role of the EOF Analysis in 4DEnVar Methods. *Atmosphere* **2017**, *8*, 146. [CrossRef]