

Fusing Information from Models and Measurements at the Exascale

A. Sandu*

*Computer Science Department, Virginia Tech. E-mail: sandu@cs.vt.edu. Phone: 540-231-2193.

Recognizing the need for a more detailed scientific understanding of the human impact on climate, significant advancements have been made in our ability to both measure and model the Earth system. A close integration of observations and models is essential for advancing our understanding of this system. Data assimilation is the process to dynamically integrate information from measurements into models. Sensor network configuration is the process of using model results to dynamically steer the measurement process. The exascale era opens exciting opportunities for using highly complex models in data assimilation, for utilizing huge volumes of data from a myriad of sensors, and for configuring complex observation networks such as to maximize the informational benefit.

In a variational approach the data assimilation problem is posed as an inverse problem, where parameters are adjusted such that the model predictions best fit the measurements. Sensor network configuration is realized in an optimization framework and relies on high order sensitivities. Current variational approaches face considerable challenges at extreme scales. The optimization approach taken in four dimensional variational (4D-Var) data assimilation does not parallelize well, due to the iterative and synchronous nature of traditional numerical optimization schemes. Adjoint models employed in gradient based optimization require to checkpoint enormous amounts of data - full state snapshots of the forward simulation. Characterizing uncertainty in the data and in the models is complex, and misspecification of these error covariances greatly affects the quality of the results.

We discuss several key directions that offer great opportunities for mathematical innovation, and where fundamental advances are needed in order to enable variational data assimilation to harness the power of future exascale systems.

Adjoint models. Adjoint models are an essential ingredient for the solution of inverse problems. They provide the gradients of the cost function with respect to model state and parameters, to be used in optimization and a posteriori error analyses, among other. The construction of adjoints for complex exascale models will become increasingly challenging due to the multiphysics nature of the applications, and to the large data that needs to be check-pointed in a distributed manner. On the positive side adjoint models inherit the parallel structure of the forward models, and . The use of automatic differentiation tools [15, 17] is expected to become essential in the future. Also, the use of second order adjoints to provide Hessian-vector products valuable for sensitivity analysis, optimization, and observation impact studies, will become important.

Weakly constrained 4D-Var. The perfect model assumption, on which current variational data assimilation systems rely, breaks down when the length of the assimilation window increases. Since the model is inexact, the inference adjusts the model parameters such as to compensate for model inaccuracies. Therefore long window assimilation results, such as carbon flux estimates, can be heavily corrupted by the errors in the underlying model. To address this problem the focus will shift to the weak constraint 4D-Var framework which accounts for all sources of error that impact data assimilation: prior, data, and model. The long assimilation window, and the weak constraint 4D-Var approach, bring a number of significant computational challenges. Estimates of model errors are needed in this approach, but they are difficult to obtain; moreover, it may not be possible to disambiguate model and data errors. The optimization problem increases in size, and becomes more ill-conditioned [21, 22]. The cost per iteration increases (at least) linearly with the window length.

Scalable 4D-Var formulations. New mathematical ideas need to be explored to make weak constraint 4D-Var feasible at the exascale; they include dual formulations in observation space, preconditioning, parallelization, and the use of approximate and reduced order models. Dual formulation of 4D-Var [6]) defines the control variables in the observation space [24] and leads to smaller optimization problems. The saddle point formulation trades the size of systems to be solved for more opportunities for parallelism. Long interval assimilation can be reformulated by shifting the assimilation window forward in time [21]. Suboptimal approaches employ simplified adjoints, e.g., based on reduced order models [1, 9]). New opportunities for parallelization in time are opened by the weak constraint formulation. During optimization the model, tangent linear, and adjoint solutions on different sub-windows can proceed in parallel. To reduce the number of optimization iterations adequate preconditioners need to be studied, such as approximate inverse Hessians [12, 13], limited memory approximations [23], and temporal multigrid [4].

Model errors. The statistical description of model error is one of the main challenges in data assimilation. To carry out long window 4D-Var assimilation one needs to estimate both the model bias and the model error covariance. Biases are expected values of the systematic and slowly growing model errors, while covariances characterize small random errors which evolve on time scales shorter than the assimilation window [21]. Pragmatic assumptions on the statistical distribution of the errors are necessary to develop computationally feasible covariance models. Homogeneous and isotropic correlation functions have been used for modeling spatial error correlations [14, 16] and the Kronecker (tensor) product allows the construction of high-dimensional spatio-temporal statistical models [5, 7]. The parameters of these models can be found from and tendency differences between ensembles of model runs.

Biases can corrupt the model, observations, and observation operators [11]. The variational approach can naturally estimate biases by including them as additional model parameters [10, 18, 21]. Bias and parameter estimation are essentially the same problem. It is difficult to disambiguate biases coming from different sources (e.g., model versus observation biases). Variational bias correction works best in situations where there is sufficient redundancy in the data, or where the model biases are small [11]. Robust inversion algorithms are needed where the inference results are correct regardless of whether one can fully apportion biases to sources.

Hybrid variational-ensemble algorithms. Hybrid algorithms, combining the benefits of sampling/ensemble based and variational methods, constitute a highly promising approach to assimilation. New methods need to be developed to fully solve the parameter estimation problem and provide a quantitative representation of posterior uncertainty in state and parameter optimal estimates [2]. An important new idea is the randomized implicit sampling [19]. Ensembles can represent complex posterior distributions, e.g., multimodal or non-symmetric; moreover, it captures the correlations between uncertainties in state and in parameters, as well as cross-correlations between uncertainties in different parameters. Weighted ensemble averages provide unbiased minimum variance estimators of the state and parameters, in contrast to the biased maximum likelihood estimators obtained by 4D-Var.

Sensor network configuration. Quantification of the data impact on inference results is required in order to address important issues such as assessing the contribution of each individual observation to the forecast error reduction, extracting the maximum of information from a sparse data set, thinning massive data sets, and configuring sensor networks. Promising algorithms involve high order derivatives, and singular value decompositions to identify the most important components in both the model and the data spaces. Novel data analysis techniques are necessary to quantify the uncertainty in assessing the value of various observing system components [8], [20]. Information theoretic concepts to quantify the contribution of different sensors, and to define optimal configurations, are expected to become increasingly important in the future.

References

- [1] A. C. Antoulas. *Approximation of large-scale dynamical systems*. SIAM, Philadelphia, 2005.
- [2] Ethan Atkins, Matthias Morzfeld, and Alexandre J. Chorin. Implicit particle methods and their connection with variational data assimilation. *Monthly Weather Review*, 2012/11/27 2012.
- [3] J. Barkmeijer, R. Buizza, and T.N. Palmer. 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125:2333–2351, 1999.
- [4] A. Cioaca, A. Sandu, and E. de Sturler. Efficient methods for computing observation impact in 4D-Var data assimilation. *Journal of Computational Physics*, submitted, 2012.
- [5] E.M. Constantinescu, A. Sandu, T. Chai, and G.R. Carmichael. Autoregressive models of background errors for chemical data assimilation. *Journal of Geophysical Research*, 112(D12309), 2007.
- [6] P. Courtier. Dual formulation of four-dimensional variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 123:2449–2461, 1997.
- [7] N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics, 2011.
- [8] D.N. Daescu. On the deterministic observation impact guidance: a geometrical perspective. *Monthly Weather Review*, 137:3567–3574, 2009.
- [9] D.N. Daescu and I.M. Navon. Efficiency of a POD-based reduced second order adjoint model in 4D-Var data assimilation. *International Journal for Numerical Methods in Fluids*, 53:985–1004, 2007.
- [10] D. Dee. On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, 123:1128–1145, 2005.
- [11] Dick Dee. Bias correction in data assimilation. ECMWF Meteorological Training Course, April 2012.
- [12] G. Desroziers and L. Berre. Accelerating and parallelizing minimizations in ensemble and deterministic variational assimilations. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1599–1610, 2012.
- [13] Amal El Akkraoui, Yannick Trémolet, and Ricardo Todling. Preconditioning of variational data assimilation and the use of a bi-conjugate gradient method. *Quarterly Journal of the Royal Meteorological Society*, pages n/a–n/a, 2012.
- [14] G. Gaspari and S.E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125:723–757, 1999.
- [15] R. Giering and T. Kaminski. Recipes for adjoint code construction. *ACM Trans. Math. Soft.*, 24(4):437–474, 1998.
- [16] T. Gneiting. Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, 125:2449–2464, 1999.
- [17] A. Griewank. *Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation*. Frontiers in Applied Mathematics. SIAM, Philadelphia, 2000.

- [18] J. F. Lamarque, B. V. Khattatov, V. Yudin, D. P. Edwards, L. K. J. C. Gille Emmons, M. N. Deeter, J. Warner, D. C. Ziskin, G. L. Francis, S. Ho, D. Mao, J. Chen, and J. R. Drummond. Application of a bias estimator for the improved assimilation of Measurements of Pollution in the Troposphere (MOPITT) carbon monoxide retrievals. *JGR*, 109(D16):D16304, 2004.
- [19] Matthias Morzfeld, Xuemin Tu, Ethan Atkins, and Alexandre J. Chorin. A random map implementation of implicit filters. *Journal of Computational Physics*, 231(4):2049 – 2066, 2012.
- [20] R.D. Torn and G.J. Hakim. Ensemble-based sensitivity analysis. *Monthly Weather Review*, 136:663–677, 2008.
- [21] Yannick Tr’emolet. Accounting for an imperfect model in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2483–2504, 2006.
- [22] Yannick Trémolet. Model-error estimation in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1267–1280, 2007.
- [23] J. Tshimanga, S. Gratton, A. T. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 134(632):751–769, 2008.
- [24] Liang Xu, Thomas Rosmond, James Goerss, and Boon Chua. Toward a weak constraint operational 4D-Var system: application to the burgers’ equation. *Meteorologische Zeitschrift*, 16(6):741753, 12 2007.