Probabilistic Machine Learning and Data Assimilation

Vishwas Rao (ANL), Sandeep Madireddy (ANL), Carlo Graziani (ANL),
Pengfei Xue (Michigan Tech), Romit Maulik (ANL)

Focal Area

This white paper responds to Focal Area 1. The associated portfolio of research activities is well-suited to DOE's asset mix of HPC platforms, climate expertise, climate simulation codes, and AI expertise, which creates an opportunity to use manifold-finding probabilistic AI methods to create more powerful data assimilation techniques that increase the fidelity and forecasting skill of Earth System Prediction.

Science Challenge

Data Assimilation (DA) is the problem of finding a map from a succession of weather observations to initial states for predictive Earth System Model (ESM) simulations. DA is of fundamental importance to numerical weather prediction (NWP), to model parameter estimation, and to model verification and validation (V&V). DA limitations are an important cause of the decay of forecasting skill of NWP forecasts beyond 10-14 days [19] - time-scales at which the chaotic effects are not prominent. Modern probabilistic machine learning methods offer opportunities to loosen this restriction, extending the skill window of NWP to the sub-seasonal and seasonal regimes, and enabling novel methods for optimization of data acquisition.

Rationale

Predictive errors in ESMs are compounded from two distinct but interrelated sources: model errors, and data assimilation errors. This whitepaper is concerned with the latter type of errors. ESMs are simulated using a discretized version of some dynamical system $\mathcal{M}$ that evolves a state $\mathbf{x}_n$ at time t_n to a new state $\mathbf{x}_{n+1} = \mathcal{M}(\mathbf{x}_n)$. The state $\mathbf{x}_n$ is a very large vector, representing many variables over a large mesh. The evolution must be started at some state $\mathbf{x}_0$ at time t_0. If the evolution is to be predictive of future observable conditions, the state $\mathbf{x}_0$ must be consistent with current observations. Given sparse observations $\mathbf{y}_1 = \mathcal{H}(\mathbf{x}_1), \ldots, \mathbf{y}_n = \mathcal{H}(\mathbf{x}_n)$ at times $t_n > t_{n-1}, \ldots, t_1 > t_0$ (where $\mathcal{H}(\cdot)$ is an *observation operator*), DA generally proceeds by inferring a state $\mathbf{x}_0$ that could plausibly have given rise to the subsequent observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ [3].

The main problem that arises is that this inverse problem is under-determined, and there is an infinitude of states $\mathbf{x}_0$ that are consistent with the observations. DA-based forecasts address this problem by first calculating an "optimal" $\mathbf{x}_0$ by using model simulations and observations in a loop, then deriving an ensemble of initial states from $\mathbf{x}_0$ by some (typically Gaussian) perturbation process, and evolving the ensemble forwards to obtain a sampling of possible predictions, which are then converted to a probabilistic forecast at time $t > t_n$ by a smoothing process.

The difficulty with this approach is that the resulting perturbed ensemble of initial conditions *does not sample the actual distribution of states conditioned on the observations*. That is, there exists an ideal conditional distribution $P(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_n)$, in principle ascertainable from data, representing initial conditions weighted by their probability in light of observations. This is the distribution that ought be sampled, in order for the evolved ensemble at time t to represent a sample from the

DOI: https://doi.org/10.2172/1769766
DOE BER, AI4ESP 2021 #AI4ESP1110 (https://www.ai4esp.org)

forecast distribution $P(\mathbf{x}(t)|\mathbf{y}_1, \ldots, \mathbf{y}_n)$. DA-derived perturbed ensembles do not have this property, because inferring the distribution $P(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_n)$ in the high-dimensional state space to which $\mathbf{x}_0$ belongs has been considered prohibitively difficult. As a consequence, probabilistic forecasts suffer large DA-sourced errors in probabilistic calibration. Other activities that use DA, such as model parameter estimation and model verification and validation (V&V) also suffer analogous errors.

At present, the most popular approaches to solve the DA problem are variational approaches (4D-Var), inspired by control theory, and ensemble-based approaches, inspired by statistical estimation theory. Variational methods rely on Gaussian modeling assumptions, which are oversimplifications [14, 20]. Ensemble-based methods such as the Ensemble Kalman Filter (EnKF) also embed Gaussian modeling assumptions, and are afflicted by variance underestimation and by spurious long-range correlations, addressed by means of covariance inflation [2], and localization [11]. Both families of approaches sample ensembles from Gaussian distributions, which, as discussed above, induces forecast distortions.

To date, the application of AI methods to DA has been confined to reduced-order phenomenological models with limited degrees of freedom, and has featured strictly deterministic (as opposed to probabilistic) neural network architectures, [5, 6], harnessed to the existing DA methodologies summarized above. These methods attempt to correct resolution filtering and physics model errors, but do not address the specific statistical errors embedded in their underlying DA methodologies.

The Opportunity: A new opportunity to address these errors has arisen in consequence of the advent of Exascale HPC platforms, and of probabilistic manifold-finding machine-learning methods that can efficiently learn data distributions in high-dimensional spaces. By training systems such as variational autoencoders [12] or information-conserving dimension-reduction systems such as variational information bottleneck (VIB) machines [1, 15, 17, 21], possibly supplemented by distribution-learning methods such as normalizing flows [13] on large datasets from ESMs, the required distribution $P(\mathbf{x}_0|\mathbf{y}_1, \ldots, \mathbf{y}_n)$ can be *learned*, either directly (in a discriminative setting) or from the joint distribution $P(\mathbf{x}_0, \mathbf{y}_1, \ldots, \mathbf{y}_n)$ (in a generative setting). The decoder part of these AI systems allows these learned distributions to be sampled, providing precisely the state ensemble required for proper predictive calibration. By comparing the evolution of such ensembles to observational data we can strengthen V&V activities such as parameter estimation and model validation.

The reason to believe that this program can be successful is that variational autoencoders and VIB-type systems have been demonstrated capable of providing substantial dimensional reduction of input data, in effect finding low(er)-dimensional representations of the data that preserves information concerning quantities of interest [15], such as (in this case) the initialization state $\mathbf{x}_0$. A successful implementation of this research program would have many benefits. At the most basic level, AI-based distribution-learning can free DA from the Tyranny of Gaussianism, allowing more general and expressive distributional properties that are better-suited to non-linear dynamics. A consequent benefit is the prospect of reducing predictive errors in NWP. Since NWP errors are compounds of model error and DA error, abating the DA error term is necessary in order to extend NWP prediction skill to the sub-seasonal and seasonal timescales, and to improve parameter estimation and V&V for weather and climate codes.

Narrative

Great Lakes Basin Supply: As an application of the proposed DA approach in a regional hydroclimate system, consider the Laurentian Great Lakes, which hold significant environmental, cultural, and economic value for both the region and the nation. The surface water elevations of

the Great Lakes are an ideal metric for understanding climate impacts on large hydrologic systems and assessing adaption measures [9]. Over the last two decades, the Great Lakes' water levels have swung from record lows to extreme highs, associated with increases in total and extreme regional precipitation, increased spring runoff, reduced over-lake evaporation, and fluctuating inter-lake flows. However, the contribution of these factors varied significantly across spatiotemporal scales [8, 10]. Higher water levels, along with more intense storms due to the hydrologic intensification that accompanies climatic warming, are also associated with flooding, coastal erosion, damage to infrastructure, and ecosystem impacts. Accurate predictions of these drivers, lake water levels and associated hydrological features in the coupled atmosphere-land-lake regional earth system are highly desirable. Assimilation of precipitation, runoff, and evaporation observations are key to the predictive modeling of these drivers.

AI/ML: There are a number of ML-based DA approaches possible, differing in ML architecture and in data selection. A reference implementation might work as follows: Training data is generated by an ESM configured with a meshed box covering the region of interest, and run for many simulation years, generating mesh data at some cadence. The data is used to create a large corpus of vectors $(\phi_t, \mathbf{x}_0, \mathbf{y}_1, \ldots \mathbf{y}_n)$, where ϕ_t is yearly phase (i.e. "day of year"), and where the simulated observations $\mathbf{y}_l$ are obtained by applying the observation operator $\mathcal{H}[\cdot]$ to the state $\mathbf{x}_l$. This data is then fed to a VIB system — an information-theoretic approach that learns an encoding of the inputs and outputs that is maximally expressive about outputs $\mathbf{x}_0$ (modeled using a decoder network) while being maximally compressive about inputs $\{\phi_t, \mathbf{y}_1, \ldots, \mathbf{y}_n\}$ (modeled using an encoder network). VIB learns a low(er)-dimensional representation of the data distribution $P(\mathbf{x}_0|\phi_t, \mathbf{y}_1, \ldots, \mathbf{y}_n)$. A key research question is how much training data is required to successfully approximate this distribution. Since the predicted observations are interpolated on mesh, both the encoder and decoder networks can be modeled using the convolutional neural networks (CNN) that take the spatial correlations into account. The convolution operation ensures stationarity and thus translational invariance. The latent space representations can be further enhanced by the use of normalizing flows which enable the modeling of complex posterior distributions [15].

At deployment, a set of field observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are obtained — in the basin supply problem, these would include (at least) runoff, precipitation, and evaporation rates. The learned distribution $P(\mathbf{x}_0|\phi_t, \mathbf{y}_1, \ldots, \mathbf{y}_n)$ is then sampled to yield an ensemble of initializations $\mathbf{x}_0$, and this sample is then evolved to produce a sample from the forecast distribution. Key observables of interest such as temperature, precipitation, surface runoff, etc. are extracted from the state $\mathbf{x}(t)$ at forecast time t using appropriate observation operators, and their forecast distribution inferred.

Optimizing Data Acquisition: Optimal data acquisition is important in improving the seasonal or sub-seasonal predictability. Some of the existing methods that address this important problem requires repeated use of forward, adjoint, tangent linear model, and second order adjoint model evaluations. These computations are expensive and their application to real models is impractical, so that most of the existing research in these directions tackle small/toy problems. AI-based DA can help approach the data optimization problem in operational settings. Access to the distribution $P(\mathbf{x}_0|\phi_t, \mathbf{y}_1, \ldots, \mathbf{y}_n)$, and to the information-quantifying properties of VIB enables an approach in which the expected information (in bits) gained by a new observation (a new component of the $\mathbf{y}$) is estimated using current data, and the observation is chosen to maximize the expected information gain, which is equivalent to minimizing uncertainty [4, 7]. In this connection, since the distribution $P(\mathbf{x}_0|\phi_t, \mathbf{y}_1, \ldots, \mathbf{y}_n)$ needs to be updated every time new data is observed, a key research topic is the use of transfer learning [16] and preconditioning [18] to update the posteriors of the distribution, treating the previous posterior distribution as a prior for the updated problem.

Suggested Partners/Experts

Kayo Ide (University of Maryland, College Park)
Adrian Sandu (Virginia Tech)
Dacian Daescu (Portland State University)
Jeffrey Anderson (NCAR)

References

[1] A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

[2] J. L. Anderson and S. L. Anderson. A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127(12):2741–2758, 1999.

[3] D. Assimilation. Making sense of observations. *Eds. W. Lahoz, B. Khattatov, R. Ménard. Heildelberg: Springer*, 718, 2010.

[4] A. Attia, A. Alexanderian, and A. K. Saibaba. Goal-oriented optimal design of experiments for large-scale bayesian linear inverse problems. *Inverse Problems*, 34(9):095009, 2018.

[5] J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino. Combining data assimilation and machine learning to infer unresolved scale parametrisation. *arXiv e-prints*, pages arXiv–2009, 2020.

[6] A. Farchi, P. Laloyaux, M. Bonavita, and M. Bocquet. Using machine learning to correct model error in data assimilation and forecast applications. *arXiv preprint arXiv:2010.12605*, 2020.

[7] V. Fedorov. Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589, 2010.

[8] A. Gronewold, J. Bruxer, D. Durnford, J. Smith, A. Clites, F. Seglenieks, S. Qian, T. Hunter, and V. Fortin. Hydrological drivers of record-setting water level rise on earth's largest lake system. *Water Resources Research*, 52(5):4026–4042, 2016.

[9] A. D. Gronewold, V. Fortin, B. Lofgren, A. Clites, C. A. Stow, and F. Quinn. Coasts, water levels, and climate change: A great lakes perspective. *Climatic Change*, 120(4):697–711, 2013.

[10] A. D. Gronewold and R. B. Rood. Recent water level changes across earth's largest lake system and implications for future variability. *Journal of Great Lakes Research*, 45(1):1–3, 2019.

[11] P. L. Houtekamer and H. L. Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020.

[14] F.-X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography*, 38(2):97–110, 1986.

[15] S. Madireddy, N. Li, N. Ramachandra, J. Butler, P. Balaprakash, S. Habib, and K. Heitmann. A modular deep learning pipeline for galaxy-scale strong gravitational lens detection and modeling. *arXiv preprint arXiv:1911.03867*, 2019.

[16] S. Madireddy, J. H. Park, S. Lee, P. Balaprakash, S. Yoo, W.-k. Liao, C. D. Hauck, M. P. Laiu, and R. Archibald. In situ compression artifact removal in scientific data using deep transfer learning and experience replay. *Machine Learning: Science and Technology*, 2(2):025010, 2020.

[17] O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[18] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. A. Witte, and F. J. Herrmann. Preconditioned training of normalizing flows for variational inference in inverse problems. *arXiv preprint arXiv:2101.03709*, 2021.

[19] H. Stern and N. E. Davidson. Trends in the skill of weather prediction at lead times of 1-14 days. *Quarterly Journal of the Royal Meteorological Society*, 141(692):2726–2736, 2015.

[20] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

[21] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.