Surrogate multi-fidelity data and model fusion for scientific discovery and uncertainty quantification in Earth System Models

Romit Maulik, Virendra P. Ghate, William Pringle, Yan Feng, Vishwas Rao, Julie Bessac, Bethany Lusch

Focal Area

This white paper responds to Focal Area 2. We seek to develop a fast multi-fidelity model and data fusion framework with inductive biases and explainable AI components for cost-effective input uncertainty quantification (UQ) and correlation extraction between drivers and responses in Earth System Models (ESMs).

Science Challenge

This whitepaper addresses the Earth and Environmental Systems Sciences Division (EESSD)'s predictability challenges in modeling the integrated water cycle and data-model integration. Specifically, it focuses on reducing and characterizing the uncertainty in the representation of process models for unresolved physics, either due to model resolution or limited by the physical understanding or computational efficiency, and the use of observational data for in-situ process parameter optimization within ESM. The described methods may also be used to determine the nature of responses (e.g. strength and direction), and hence to identify critical processes that drive the overall ESM responses to perturbation in the forcing

Rationale

Traditionally in multiple-model assessments of climate simulations such as the Coupled Model Intercomparison Project (CMIP6) framework, the uncertainty in the predicted components of Earth's water cycle (e.g., clouds, precipitation and sea ice) in a warmer climate are assessed by calculating the spread in their forecasted values by different ESMs for the same climate change scenario. However, such an approach overlooks the inherent uncertainty in the predicted values of these quantities in a particular model, which is crucial not only for generating an accurate forecast but also for informing future model development efforts. The UQ of predicted variables by a particular ESM is assessed through following approaches, i) by making short-term global simulations in a hindcast mode [10], ii) by making single column model (SCM) simulations of the parameterization under consideration for selected weather condition, and comparing it to observations or high-resolution model output [16], iii) by performing perturbed parameter ensemble (PPE) short-term simulations [11]. Although widely used, these parametric models may only be evaluated for a limited set of weather conditions, and for limited time-periods as they are based

DOI: https://doi.org/10.2172/1769781
DOE BER, AI4ESP 2021 #AI4ESP1090 (https://www.ai4esp.org)

on short-term simulations. This is further complicated by the fact that the impact of uncertainty in one of the model variables on the predicted climate might be offset by that of another variable. As such, this necessitates potentially thousands of model simulations covering a sufficient time-span for the UQ of ESM predicted components of the water cycle. Currently, simulations for all such parameterizations is computationally infeasible for fully assessing the inherent uncertainty in the predicted climate. This also prevents the understanding of how parameters affect posterior distributions of the quantities of interest. Here we propose an interpretable multi-fidelity Artificial Intelligence (AI) framework that could be used for addressing this critical research gap. The framework and approach may be applied to two vital components of the Earth's water cycle: the clouds and the ocean surface, that reside under different components of the ESM. Oceans encompass about 70% of the Earth's surface and clouds blanket 70% of the ocean surface, thereby making marine clouds the most vital components of the Earth's energy budget. Additionally, atmospheric and ocean surface dynamics over the high-gradient coastal regions are one of the under-represented areas in the current ESMs where UQ of the cloud and ocean surface processes are particularly critical. Hence, we highlight the framework for these two vital components of the water cycle.

Clouds: Clouds are a major component of the atmospheric water cycle that vary at a variety of spatial and temporal scales. They are intimately coupled to the turbulence in the atmosphere that is known to change from millimeter to kilometer scales and are susceptible to aerosol direct and indirect effects, thereby affecting precipitation. Their properties change at spatial and temporal scales finer than ESM resolutions, and their effects are parameterized using resolved scale variables. In contrast, cloud droplets and interactions with aerosols are fully resolved in high-resolution Large Eddy Simulation (LES) or Direct Numerical Simulation (DNS) models [8]. In addition, detailed observations of aerosol, cloud and turbulence properties are routinely made at the Atmospheric Radiation Measurement (ARM) Climate Research Facility (ACRF). ESM cloud parameterizations have several modules, each of which have O(10) tuning parameters. LES or DNS do not simulate all of these tuning parameters, and many of them cannot be directly observed. However, a subset of the tuning parameters may be simulated and can be observed with sufficient accuracy. These observations could then be used to regularize the physics-informed emulators built from high-resolution models for predicting cloud microphysical properties.

Ocean surface: The surface of the ocean is representative of the ocean water cycle and how it interacts with the atmosphere and land, and thus improving the understanding of these interactions will contribute to the predictability of the integrated water cycle. Ocean surface components such as sea ice, sea-surface temperature (SST), and wave conditions are known to change substantially in a warmer climate. Although recent progress has been made to include wind-waves in DOE's Energy Exascale Earth System Model, waves and their interactions with sea-ice, the atmosphere, and coastal regions have not yet generally been included in ESMs. This is at least partially because so-called 3rd generation spectral-wave models are computationally expensive as they are formulated in 5-dimensional space (wave direction, wave frequency, two spatial dimensions, and time), a necessary trade-off given the implausibility of phase-resolving all surface waves of periods O(sec) in the ocean. Moreover, spectral-wave models combine complicated parametrizations of phenomena such as white-capping, and depth-limited breaking; and one of the key underlying assumptions of nonlinear energy transfer has been shown to result in inconsistent spectral evolution [1]. Thus, including these processes to describe the ocean surface would dramatically increase the computational expense and complexity of the ESM, which would further decrease the opportunity for ensemble forecasts to explore the parametererization space. Fortunately, the ocean surface is the

most widely observable part through remote sensing and floating in-situ instrumentation, and thus lends itself to data-driven ML methods for emulation. Recent research by our team has shown that SST emulated from sparse sensor networks can be predicted weeks in advance with comparable accuracy to high-resolution 3D ocean-ice forecast models [6, 7]. Hence, there is the opportunity for this geophysical emulation framework to be extended to ensemble forecasting of sea-ice coverage, sea-surface heights, and significant wave heights and mean wave periods, in addition to SST.

Narrative

Data-driven emulator design for geophysical processes has seen explosive growth in the past few years with several algorithms being proposed to obtain cost-effective process models. As such, the use of machine learning in the geophysical sciences is hardly new [14]. In fact, there are several competing emulation strategies for various geophysical processes [2, 4, 12, 13, 15] with their associated pros and cons. An improvement on current methods to solve our science challenges is through development of a comprehensive model and data fusion framework where multiple partial differential equation (PDE), hybrid and fully data-driven emulators along with simulation and observation data may be used to identify and characterize driver processes and how their uncertainties propagate through to the fused outputs. This fusion may be performed by building linear [3] and nonlinear multi-fidelity process-models [9] which utilize Gaussian processes to build error relationships between different emulators, which are themselves functions of input parameters and predictions by various models. The overall framework may then be equipped with inductive biases (such as hard and soft physics constraints) and explainable AI components to effectively extract science from fused raw simulation and observational datasets. For example, explainability can be obtained through the use of game theoretic optimal credit allocation to identify the directional correlations of different free-parameters on model outputs [5]. This multi-fidelity paradigm is also useful for healthy estimates of parametric uncertainty by relying on large volumes of low-fidelity evaluations but introducing physics-based biases through relatively few medium and high fidelity assessments. The overwhelming advantage of using a model-fusion framework stems from the telescopic cancellation of errors across various model and observational fidelities/accuracies in the calculation of ensemble statistics, which in turn acts as an effective risk-mitigation strategy. Fast sensitivity analysis and information content, using the fused models and data, can be used in UQ, for optimally configuring sensor networks, and for model calibration. Using these techniques requires expensive operations such as solutions to forward model, adjoint model, tangent linear model, second order adjoints, and second order tangent linear model. Unfortunately, this is intractable for realistic complex climate models. The use of a differentiable multi-fidelity framework can enable the fast calculation of such quantities of interest thereby surpassing this bottleneck which in turn can help greatly improve predictability. Many applications require evaluation of integrals to obtain the statistics of the quantities that are functions of stochastic parameters. The most common means to achieve this is to sample from the posterior distribution using techniques similar to Markov-Chain Monte Carlo. However, this is computationally infeasible for high-dimensional problems. In contrast, AI-based surrogate models reduce sampling costs in such situations. Note also that the dangers of extrapolation, ever-present in purely data-driven models can be alleviated through multi-fidelity paradigms with model-form UQ. Finally, the construction of a hierarchy of models with varying computational costs and fidelities will allow for easy integration of novel developments in PDE-based or data-driven emulation strategies into practically deployed systems.

Suggested Partners/Experts

Paris Perdikaris (University of Pennsylvania)

References

[1] D. Ardag and D. T. Resio. Inconsistent spectral evolution in operational wave models due to inaccurate specification of nonlinear interactions. *Journal of Physical Oceanography*, 49(3):705–722, 2019.

[2] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh. Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001958, 2020.

[3] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[4] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins, et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*, 2016.

[5] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[6] R. Maulik, R. Egele, B. Lusch, and P. Balaprakash. Recurrent neural network architecture search for geophysical emulation. SC '20. IEEE Press, 2020.

[7] R. Maulik, K. Fukami, N. Ramachandra, K. Fukagata, and K. Taira. Probabilistic neural networks for fluid flow surrogate modeling and data recovery. *Physical Review Fluids*, 5(10):104401, 2020.

[8] J.-P. Mellado, C. Bretherton, B. Stevens, and M. Wyant. Dns and les for simulating stratocumulus: Better together. *Journal of Advances in Modeling Earth Systems*, 10(7):1421–1438, 2018.

[9] P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

[10] T. J. Phillips, G. L. Potter, D. L. Williamson, R. T. Cederwall, J. S. Boyle, M. Fiorino, J. J. Hnilo, J. G. Olson, S. Xie, and J. J. Yio. Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. *Bulletin of the American Meteorological Society*, 85(12):1903–1916, 2004.

[11] Y. Qian, H. Wan, B. Yang, J.-C. Golaz, B. Harrop, Z. Hou, V. E. Larson, L. R. Leung, G. Lin, W. Lin, et al. Parametric sensitivity and uncertainty quantification in the version 1 of e3sm atmosphere model based on short perturbed parameter ensemble simulations. *Journal of Geophysical Research: Atmospheres*, 123(23):13–046, 2018.

[12] S. Rasp, M. S. Pritchard, and P. Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018.

[13] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[14] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

[15] S. Scher. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45(22):12–616, 2018.

[16] X. Zheng, S. Klein, V. Ghate, S. Santos, J. McGibbon, P. Caldwell, P. Bogenschutz, W. Lin, and M. Cadeddu. Assessment of precipitating marine stratocumulus clouds in the e3smv1 atmosphere model: A case study from the arm magic field campaign. *Monthly Weather Review*, 148(8):3341–3359, 2020.