Machine Learning for Adaptive Model Refinement to Bridge Scales

Authors

Juliane Mueller[1], Charuleka Varadharajan[2], Erica Siirila-Woodburn[2], Charles Koven[2]

[1] Lawrence Berkeley National Laboratory, Computational Research Division

[2] Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area

Focal Area

This whitepaper is responsive to focal area (2) Predictive modeling through the use of AI techniques and AI-derived model components: the use of AI and other tools to design a prediction system comprising of a hierarchy of models (e.g., AI driven model/component/parameterization selection). Here we describe scale-aware ML models for adaptive model refinement that allow us to bridge the spatial and/or temporal scales in simulation models and observation data for capturing and predicting extreme water cycles.

Science Challenge

10-year vision statement: To develop and exploit ML models that enable an adaptive model refinement for seamlessly bridging the scales between different types of simulation models and data on the fly, with the goal to improve the prediction and understanding of disturbance events and their impacts.

Modeling of the water cycle is done on vastly different scales with very different types of models that range from low-accuracy (low-fidelity) empirical models to detailed highly accurate (high-fidelity) process-based models. Usually, different types of models are developed for a certain use: e.g., low-fidelity models may be employed when large spatial and temporal scales must be covered for which detailed high-fidelity models are computationally too expensive. On the other hand, high-fidelity models are needed for studying sudden short-term disturbances such as floods or wildfires [1]. For some processes such as large-scale river biogeochemistry, mechanistic models do not exist and statistical or semi-empirical models with approximations and parametrizations are relied on. Models that were built for a certain scale usually do not easily transfer to other scales due to differences in relevant features, fundamentally different physics at the different scales, and input datasets required. It is currently not possible to adaptively switch between these different types of models on the fly. ML methods could allow us to address this challenge, and thus bridge the gap between different-scale models and exploit the information and advantages that individual models have to offer.

Similar to models at different scales, data is collected at different scales (see, e.g., [3]), leading to questions about how to use data collected at one scale for informing a model at a different scale. For example, using observation averages when upscaling highly resolved data may lead to an inaccurate representation of the overall system behavior, as averages are highly sensitive to outliers such as extreme events, yet these outliers should not be removed as they capture the important disturbances we seek to better understand. Averaging leads to information loss, introduction of errors, and a significant increase in the prediction uncertainty. For example, a coarse-grid of atmospheric forcing in mountainous terrain is known to lead to systematic biases in variables such as temperature, precipitation phase partitioning, and radiative effects due to a “smearing” of

DOI: https://doi.org/10.2172/1769741

orographic effects. In the subsurface, models are parameterized to represent "effective" flow and transport behavior, so that the average behavior of a solute represented in the model at coarse scales may in reality move much faster (given fracture flow) or much slower (given matrix flow).

Rationale

Currently, there are no methods that are able to bridge the spatial and/or temporal scales of simulations and observations for predicting the onset and the impacts of extreme water cycle events. It is challenging to assimilate data collected at different scales, to adaptively switch between different-scale models on the fly, and to take into account different durations of system memories and the effects that disturbances have.

Disturbances can occur over a large range of spatial and temporal scales. For large scale disturbances such as prolonged droughts, highly accurate high-fidelity models designed for approximating short-term disturbances are not practical due to their computational demand, and thus fast-to-evaluate models, which are less accurate, are needed. On the other hand, for short-term disturbances such as floods or wildfires, high-fidelity models that capture detailed dynamics are needed. In contrast to adaptive mesh refinement methods, whose purpose is to dynamically increase the accuracy of solutions in turbulent regions of the simulations, different types of disturbances are described by entirely different types of models that capture different types of physics at different scales. Thus, it is not possible to model different types of disturbance events with the same model by simply refining or coarsening mesh sizes and time steps. The question then becomes when to use which model? Ideally, one could adaptively switch between different-scale models, i.e., seamlessly combine and make use of different-scale models as needed for predicting disturbance events and their impacts. Thus, low-fidelity models could be used in regimes when prediction accuracy is less important ("business as usual") and crude approximations are sufficient. However, at the onset of a disturbance event such as a flood or a wildfire, one wants to switch to the detailed high-accuracy or high-resolution model.

The implementation of such an ***Adaptive Model Refinement*** raises many questions: ***When should one switch from a large-scale to a small-scale model and vice versa? How can the different time and spatial scales of disturbances modeled by inherently different types of simulations be connected (multi-year droughts versus a weeklong flood)? How can the different length memories that are inherent in ecosystems be aligned across scales? How can the different approaches to modeling and implementing the underlying physics be connected? How can the different datasets used for informing different-scale models be assimilated?***

Answers to these questions do not currently exist. Therefore, we envision the development of *scale-aware ML methods* that fill this research gap and that build the bridge between different-scale models. ML models can be exploited to decide when to switch from a large-scale to a small-scale model and when relying on predictions made by an ML model approximation is simply sufficient. ML models have the benefit of being highly flexible and, provided with the necessary data, they are able to approximate highly nonlinear responses [2]. With the recent advancements in ML, there are opportunities to develop new model types that can assimilate and be trained on data collected at different scales with different measurement devices to make predictions. The result will be a computationally efficient adaptive model refinement method for autonomously deciding when to use which model and for making accurate predictions across scales.

Narrative

In order to study and better predict hydrological processes, ecosystems dynamics, and fluctuations in water quality and quantity, both process-based simulation and ML models are being developed. Depending on the research goals, models at different scales are used. For example, for studying disturbance events such as floods, wildfires, or multi-year droughts, and their long-term effects on the water cycle, models that incorporate detailed physics are needed, yet these models operate on

different time and spatial scales and calibrating them requires different types of observation datasets. Some of the necessary data, especially data obtained from remote sensing, are plentiful, whereas other data such as subsurface heterogeneity are not easily measured.

Although there are established methods for up- and down-scaling, they rarely (if at all) quantify the uncertainty that arises from scaling. If, instead of a single "representative' value, a distribution of the observations was used, at least some statistics of the values could be retained, which would open the possibility for propagating the uncertainty arising from scaling through the models. However, simulation models are deterministic and only return point estimates, making calibration against data distributions difficult, and it cannot be expected that a simulation that uses one set of model parameters will be able to capture the whole spectrum of possible observations. Mean system behaviors are likely to be better approximated (because more data exists) than rare extreme events. On the other hand, when down-scaling data to a local scale, detailed information about impactful short-term disturbances such as floods may not be represented, yet these disturbances must be captured as they have long-term impacts on the water cycle. Taking the down-scaled data as truth-inputs for calibrating simulation models has a cascading and detrimental effect on prediction accuracy.

In order to address the shortcomings of current scaling approaches, we envision adaptive model refinement methods that allow an autonomous switch between different-scale models on the fly. To achieve this, ML models can be exploited to predict when and where to use which model (large-scale, small-scale, high-or low-fidelity, or an ML approximation) and for how long. For example, during average "business as usual" years, no rapid changes occur and coarser large-scale models will make adequate predictions. ML can be used to predict when and where an extreme event is likely to occur and how long it will last (e.g., flooding or wildfires after a prolonged drought) and in response switch to a higher-accuracy small-scale model. ML models are suitable tools for this task as they have the ability to approximate complex nonlinear functions. New types of ML models, whose architectures allow them to capture and predict different durations of system memories, can assimilate and learn from data (observation and/or simulation) that are recorded at different scales. These new models will be able to provide predictions in the form of probability distributions (rather than point estimates), allowing for efficient uncertainty quantification. Moreover, for processes where mechanistic understanding is lacking, ML models are good approximation candidates.

We envision that this research of developing and exploiting scale-aware ML/AI tools for adaptive model refinement to bridge the scales of simulation models and data will enable (1) an autonomous adaptive coupling of different-scale models at different sub-domains and times to obtain high-accuracy predictions at locations and times when they are needed (hot spots); (2) efficient uncertainty quantification of predictions when moving between scales, and thus enabling reliable predictions; (3) the reduction of information loss when moving between scales by relying on distributions rather than point estimates.

The envisioned scale-aware ML methods for bridging scales will adhere to the FAIR principles: It is an opportunity to develop, implement, and deploy an open-source tool that couples open-source simulation models and data, and that enables reproducibility. The methods would be implemented such that the data products will adhere to the FAIR principles.

Suggested Partners/Experts

Todd Munson and Alp Dener in the Mathematics and Computing Science Division at Argonne National Lab who develop mathematical optimization methods that make training ML models more efficient.

References

[1] Rodman et al., Studying wildfire behavior using FIRETEC, 2002 *International Journal of Wildland Fire* **11**, 233-246, https://doi.org/10.1071/WF02007

[2] Pathak et al., Using Machine Learning to Augment Coarse-Grid Computational Fluid Dynamics Simulations, arXiv:2010.00072

[3] Wainwright et al., A multiscale Bayesian data integration approach for mapping air dose rates around the Fukushima Daiichi Nuclear Power Plant, *Journal of Environmental Radioactivity*, 2016, **167**: 62-69