Machine Learning to Enable Efficient Uncertainty Quantification, Data Assimilation, and Informed Data Acquisition

Authors

Juliane Mueller[1], Charuleka Varadharajan[2], Yuxin Wu[2], Erica Siirila-Woodburn[2]

[1] Lawrence Berkeley National Laboratory, Computational Research Division

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

Focal Area

This whitepaper addresses the following focal area: (1) Data acquisition and assimilation enabled by machine learning, AI, and advanced methods including experimental/network design/optimization, unsupervised learning (including deep learning), and hardware-related efforts involving AI (e.g., edge computing)

Science Challenge

10-year vision statement: To develop and deploy an ML framework for fast and efficient uncertainty quantification (UQ) of extreme water cycles, whose outcomes will be used for informing adaptive data acquisition for reducing uncertainty and data collection biases, and for improving data assimilation.

Process-based models are usually deterministic, i.e., if executed twice with the same parameters and initial conditions, the outcomes are identical. In contrast, machine learning models exhibit stochasticity for several reasons that include randomness introduced during training and random dropout masks used during predictions. Uncertainties arise in all model projections from structural considerations as well data considerations (e.g., noisy data that are used during model calibration, poor or under-constrained model parameters, or changes in or addition of data). This is particularly problematic when modeling climate extremes and their impacts on hydrology and biogeochemistry since dynamic changes that occur during disturbance events have large measurement uncertainties and data biases, as well as model uncertainties arising from how disturbances are represented in the model structure. Moreover, it can be expected that the UQ outcomes are highly influenced by data collection biases (including location, frequency, and the type of data collected) and data changes (due to data updates and addition of new data). For example, many types of sensor-based observations are sparse and their site locations can be biased due to resource and logistical constraints, rather than being driven by uncertainty reduction targets guided by models. This potentially leads to an underestimation of the large-scale model uncertainty due to local-scale regimes that have been over-sampled, or hot spots which outweigh contribution but are under-sampled. However, being able to identify these collection biases and to analyze the sensitivities of the model uncertainties to data changes offers a unique opportunity for informing adaptive data acquisition and improved data assimilation.

DOI: https://doi.org/10.2172/1769743
DOE BER, AI4ESP 2021 #AI4ESP1097 (https://www.ai4esp.org)

Rationale

In order to quantify the uncertainty that is associated with the predictions made by simulations that model, for example, unsaturated flow, reactive transport, or coupled process interactions, the computational burden of running the simulation must be alleviated. Off-the-shelf UQ methods would have to query the simulation thousands of times, which is computationally intractable. For example, when quantifying the uncertainty of evapotranspiration dynamics over time and space, both a time dependent component and spatial correlations must be taken into account, i.e., the simulation uncertainties for one cell at time *T* will impact the uncertainties in the next cell at the previous and the following time steps. This is particularly important when modeling extreme perturbations, particularly such as hurricanes or wildfires where the state space can be highly dynamic and will evolve with the disturbance. ML models have been shown to be able to approximate complex responses accurately and are therefore good contenders for being used as surrogate models in place of the simulations during UQ. ML models can be used as approximations of individual time-consuming simulation model parts (or modules), especially those that are not based on process understanding and that are approximations and parametrizations.

Being able to efficiently quantify model uncertainty through the use of ML models also enables us to derive optimal data collection strategies. Here the goal may be to select measurement locations and frequencies such that the global model prediction uncertainty is minimized, which also leads to a reduction of selection biases. Using ML in such an optimization formulation allows us to directly take into account constraints, such as those imposed by the terrain and locations of already installed measurement devices and ML enables us to solve this optimization problem efficiently without having to repeatedly query the costly simulation. However, the mechanistic simulation models are still needed, in particular for interpreting and explaining the UQ outcomes especially in regard to the effects of including or excluding certain datasets.

Narrative

ML can help to improve mechanistic models and our understanding of them in the following ways: (1) Basing UQ on ML model approximations of (parts of) the mechanistic models will make UQ computationally tractable as the expensive-to-evaluate simulation does not need to be queried repeatedly; (2) The outcomes of (1) will enable us to adaptively design new data acquisition strategies to reduce the global simulation model uncertainty and data collection biases; (3) Together with our mechanistic understanding, ML will enable us to estimate and interpret the importance and impacts of data changes on UQ outcomes.

Uncertainty: Uncertainty in simulation models is often related to using simplified process approximations, parametrizations, and nonlinear coupled processes. However, the data used for calibrating the model parameters also introduce uncertainties and sensitivities. For example, measurement data usually have noise, data collection may be biased, and datasets are often updated (new versions are published) or new data are added. Thus, when quantifying the uncertainty of mechanistic models, ideally both types of uncertainty (model and data) will be taken into account. When noisy data are used for constraining model parameters, full statistical inference should be used (rather than perturbation or deterministic methods). However, inference is computationally too expensive as the simulation model would need to be queried potentially thousands of times.
Recent developments in the ML/AI space have shown that, provided with the necessary data, these models can make highly accurate predictions and enable better understanding of complex phenomena [1,2]. Thus, ML models can be used as surrogates to replace a complete expensive-to-evaluate simulation. On the other hand, individual simulation modules can be replaced by ML models, especially those that are not based on mechanistic understanding but rather complex empirical approximations and parametrizations that have been developed over multiple years. Thus, during UQ, instead of repeatedly querying the computationally expensive simulation, the fast-to-evaluate ML model is queried and therefore the computational complexity of conducting

UQ will be significantly reduced. These computationally lightweight ML models also have the advantage of being deployable on the edge in the future, allowing us to autonomously adjust the data collection in order to reduce uncertainties, which are in particular high during extreme events during which a higher measurement frequency is required.

Observational data bias: The uncertainty associated with data used in calibration tasks also depends on data collection biases. For example, on a local scale, the model uncertainty can be expected to be significantly lower when oversampling took place. This local uncertainty does, however, not represent the global model uncertainty. Therefore, these data collection biases must be automatically detected and corrected for when determining the global model uncertainty. ML models can be used to automatically identify data collection biases through a variety of clustering and pattern recognition methods. Moreover, once identified, data collection biases can be accounted for during the training of ML models, for example, by defining additional input features that encode our knowledge of the presence of biases, which is not possible when using the simulation model in UQ directly. The lessons learned from discovering data collection biases in one river basin can also be transferred to inform better and less biased data collection in other basins.

Data Assimilation: ML models will also allow us to analyze the sensitivity of trained models and simulations with respect to assimilating new data. As increasingly more data are collected or new versions of old datasets are published, the question arises whether the ML model needs to be retrained to account for potentially new unseen behavior (e.g., from a 100-year flood that may occur more frequently with more intense and frequent hurricanes). If the ML model is an accurate representation of the simulation model and if it is insensitive to data changes, does this also hold true for the simulation model it approximates?

AI-guided observations: Being able to study the effects of data changes in a computationally efficient way will also allow us to decide whether a recalibration of the simulation to the new data is necessary. Insights into the sensitivity of ML and simulation models to the data will enable us to assess the need for additional data collection. If the models are highly sensitive to the inclusion or exclusion of certain data or features, it is an indication that additional data should be collected. Here, the ML models will help us to make adaptive sampling decisions (i.e., what to measure where and at which frequency). Criteria for making these sampling decisions could be to reduce the global uncertainty and sensitivity of the models, to maximize information gain, to better constrain simulation model parameters, or to optimize for some other metric of interest for the specific use case. For example, if the goal is to model extreme water cycle disturbances such as floods, oversampling is appropriate as more data needs to be collected over a short period of time in order to capture the fast variations. On the other hand, if no major changes take place, this kind of oversampling is not necessary.

The proposed ML-driven UQ framework lends itself well to the FAIR principles: It is an opportunity to deploy an open-source tool that is built with the goal to enable reproducibility, and whose data acquisition, data assimilation, and bias correction methods can be developed such that the data products 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] Han et al., Deep Learning with Long Short Term Memory Based Sequence-to-Sequence Model for Rainfall-Runoff Simulation, *Water*, 2021, *13*(4), 437; https://doi.org/10.3390/w13040437

[2] Reichstein et al., Deep learning and process understanding for data-driven Earth system science. *Nature*. 2019 Feb; 566(7743):195-204. doi: 10.1038/s41586-019-0912-1.