Integrating Models with Real-time Field Data for Extreme Events: From Field Sensors to Models and Back with AI in the Loop

Shreyas Cholia[1], Charuleka Varadharajan[2], Gilberto Pastorello[1]

[1] Computational Research Division, Lawrence Berkeley National Laboratory
[2] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory

Focal Area(s)

This whitepaper is responsive to 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). We discuss Artificial Intelligence and Machine Learning (AI/ML) enabled integration of real-time data into the extreme event modeling workflow to improve the predictive capabilities of these models, and deliver real-time feedback to remote sensors, including software and data engineering challenges.

Science Challenge

The DOE's ModEx approach involves combining data acquisition and model development iteratively, with the objective of improving model predictability. However, most models are not capable of assimilating real-time data, in part due to challenges with obtaining, processing, and integrating relevant data in near real-time from field observations. This, in particular, affects models that seek to predict water cycle extremes and their impacts, as the location, timing, and scale of disturbances are not known a priori, and because we are increasingly seeing disturbances with no historical precedent. This impacts the quality of the models themselves, and in-turn the predictions that can be fed back to field sensors during an actual event. How can models make more effective use of data from the field? And how can we enable a tight feedback loop from these models back to the sensors?

BER EESSD's data-model integration grand challenge [1] calls for the development of "a broad range of interconnected infrastructure capabilities and tools that support the integration and management of models, experiments, and observations across a hierarchy of scales and complexity to address EESSD scientific grand challenges." Specifically called out is cyberinfrastucture that enables the two-way feedback between observational data collection and model simulations. Tools and data-model pipelines are often fragmented and bespoke for each application/dataset – this approach will not scale across the range of possible water cycle extreme events, with data from diverse sensors and experiments.

Rationale

To meet the needs of rapidly expanding field data from DOE Earth Systems sensors and instruments, it will be critical to enable data acquisition, transformation, and analytics workflows that can integrate real-time data streams with predictive models, particularly for extreme events. Live data-streams from sensors can then be used to train models using a hybrid of AI/ML and deterministic techniques, which can, in-turn, generate real-time predictions and feedback for sensors out in the field. The two-way connectivity between models and data can also enable adaptive sampling, tailored to measuring the perturbation events of interest, and collecting the most optimal data streams when and where it is necessary.

DOI: https://doi.org/10.2172/1769727
DOE BER, AI4ESP 2021 #AI4ESP1025 (https://www.ai4esp.org)

Integrating Models with Real-time Field Data for Extreme Events: From Field Sensors to Models and Back with AI in the Loop

Narrative

The next-generation of scientific advances in the Earth and environmental sciences will be predicated on a rapid growth in experimental and observational data from large sensor networks and remote instruments. It will be critical to embrace AI/ML and hybrid techniques to utilize this data and harness its predictive capabilities. There needs to be a coordinated investment in the cyberinfrastructure required to support these advances – otherwise we risk piecemeal and ad-hoc approaches to data engineering that will cause each individual effort to reinvent the wheel.

While we are collecting increasing volumes of data from field sensors, there is a disconnect with extreme event models, many of which don't use real-time data. AI has the power to transform this field, by enabling an integrated approach where real-time data from field sensors can be fed into hybrid predictive models. These models can be trained, calibrated, and improved against live data, the results of which can then be fed back to devices in the field. Instruments can be targeted to collect specific types of data, or re-calibrated on the fly, thus creating a real-time feedback system. There is also an opportunity to combine multi-modal data at different scales into a set of common pipelines, where models can be informed by new diverse data sources.

This will require a multi-disciplinary approach: computer systems engineers that work on real-time data streaming, platform integration, and edge computing challenges; AI/ML experts that implement methods to train and refine AI driven models using real-time data; and, most importantly, domain scientists that build the core predictive models and combine them with physics-informed AI/ML methods.

This is important for disturbances because we cannot pre-suppose what data are needed ahead of time during an extreme event – for instance, we can't know a priori the scale, timing, and geographical locations impacted. Since each disturbance manifests with a unique set of circumstances, modelers need to be able to query a diverse set of data types on the fly for the regions impacted (see Varadharajan et al. AI4ESP whitepaper [2]). Manually curating data for every single disturbance would be labor-intensive, if not prohibitive, and might result in insufficient representation of different types of disturbances in the models. AI/ML can help accelerate this process of harmonizing multiple data sources, and generating data products in real-time to address science questions that are most pertinent to the disturbance. Key here is the ability to fuse data from across different datasets and perform transformations in real-time.

The highly distributed nature of sensor networks, and the volumes of data collected, will necessitate a data "backbone" for AI/ML based scientific workflows. Such a platform would support:

1. data ingestion from multiple sources across multiple domains;
2. data analysis and creation of hybrid AI/ML models, through synthesis and fusion of datasets;
3. data feedback loop for sensors enabling real-time decision making for optimal data acquisition.

The integrated platform should provide core capabilities allowing users to store, process, analyze, integrate, curate, search, and disseminate data. A toolbox of data analysis algorithms and data processing

pipeline software should be maintained and available to the data producers. The ultimate aim would be to enable AI/ML driven workflows where researchers across domains can perform advanced analyses across multiple datasets and generate predictive models that can then be used directly by field sensors to target optimal, real-time data acquisition. AI/ML also provides the opportunity to perform quality assurance and control (QA/QC) against real-time streams on the fly, including metadata extraction, data validation, and sensor error checking.

Inherent in the design of such a platform, is the need to make data broadly available to algorithms, instruments and researchers. Any data products or models delivered by this system should adhere to the FAIR principles [3] to ensure that they are findable, accessible, interoperable, and reusable.

Our vision is a data ecosystem that crosses agency boundaries, with the ability to create these data model pipelines across multiple sources that are currently silo-ed. We can use AI/ML to synthesize and harmonize water cycle and related disturbance data across these sources. In the event of a disturbance there will be a need to pull together data from multiple federal agencies – having a common pipeline that can enable data fusion and hybrid AI/ML driven models across these will be critical in terms of generating useful and coordinated predictions.

Suggested Partners/Experts

- Matei Zaharia – Stanford University, Databricks (Data Infrastructure for AI/ML)
- DOE Facilities Representatives – ESnet (Edge Computing), NERSC (Superfacility Project)
- Partners from SFA field observatories

References

1. U.S. DOE. 2018. “Climate and Environmental Sciences Division Strategic Plan 2018–2023”, DOE/SC–0192, U.S. Department of Energy Office of Science. https://science.osti.gov/-/media/ber/pdf/workshop-reports/2018_CESD_Strategic_Plan.pdf
2. Charuleka Varadharajan, Yuxin Wu, Andrew Wiedlea, Kolby Jardine, Haruko Wainwright, Robert Crystal-Ornelas, Joan Damerow, Helen Weierbach, Taylor Groves, Gilberto Pastorello, Lavanya Ramakrishnan, Juliane Mueller, Danielle Christianson. “Observational Capabilities to Capture Water Cycle Event Dynamics and Impacts in the Age of AI”, AI4ESP Whitepaper.
3. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* “The FAIR Guiding Principles for scientific data management and stewardship”, *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18