A library of AI-assisted FAIR water cycle and related disturbance datasets to enable model training, parameterization and validation

Authors

Robert Crystal-Ornelas[1], Charuleka Varadharajan[1], Danielle Christianson[2], Joan Damerow[1], Helen Weierbach[1], Emily Robles[1], Lavanya Ramakrishnan[2], Boris Faybishenko[1], Gilberto Pastorello[2]

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

[2] Computing 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. Here we describe how FAIR (Findable, Accessible, Reusable, Interoperable) datasets related to water cycle extremes are essential for successful implementation of ML in Earth System and other models. We also describe how AI can be used to acquire and integrate water cycle data related to extreme events to create a library of FAIR datasets for training and evaluating algorithms.

Science Challenge

To create FAIR water cycle datasets on extreme water cycle events, data collected from sensors must be: a) acquired across agencies, b) screened for QA/QC, and c) processed and augmented with sufficient metadata to identify periods of extreme water cycle events.

The 10 year vision is an openly accessible library with dozens of AI-ready extreme water cycle datasets that would lift the data curation burden from researchers.

Rationale

Research needs & gaps

Curated data products or databases are a critical component to a cyberinfrastructure that accelerates ML/AI and other models in Earth and Environmental Sciences (ESS; Maskey et al., 2020). In particular, benchmark datasets such as ImageNet and MNIST are responsible for the AI revolution (J. Deng et al., 2009; L. Deng, 2012), by providing a means to test and compare model performance. These datasets have inspired competitions such as the ImageNet Large-Scale Visual Recognition Challenge, which triggered major breakthroughs in AI and ML models. ML models not only need data hosted in a centralized place with scalable access protocols, but also must be curated with adequate metadata. For example, ImageNet contains over 14 million images, indexed by keywords and associated with metadata for training and validating models.

Similar datasets for Earth observations are limited, but are a key type of innovative data for model evaluation under the Data-Model Integration Grand Challenge (U.S. DOE, 2018, p. 19, Associated Research Goal 2). The FAIR data products that do exist tend to be heavily used for AI/ML. For example, FLUXNET2015 is a popular dataset used heavily to model carbon-water-energy fluxes (Pastorello et al., 2020). Similarly the CAMELS dataset containing

DOI: https://doi.org/10.2172/1769646
DOE BER, AI4ESP 2021 #AI4ESP1030 (https://www.ai4esp.org)

streamflow from the USGS NWIS system, meteorology from Daymet, NLDAS and Maurer datasets, and basin attributes from GAGES-II has been widely used in hydrological ML (Feng et al., 2020; Kratzert et al., 2018).

Because observations of extreme water cycle events are very limited, there are specifically few curated, labeled disturbance-related water cycle datasets. Thus, modelers are tasked with: a) gathering difficult to locate data across multiple organizations for an extreme water cycle event and b) classifying periods of extreme disturbance events so they can train their model accurately. Because these datasets do not exist, modelers who have used AI/ML to predict extreme events (e.g., hurricanes) sometimes train their algorithms to non-hydrological datasets (Kim et al., 2019). This leads to inefficiencies if researchers spend more time labeling data than addressing scientific questions.

Justification and benefits

To enable efficient modeling, scientists need benchmark datasets that are curated (e.g., pre-classified extreme water cycle events) and include standardized meteorological and hydrological parameters (e.g., precipitation, temperature, relative humidity, soil moisture; Maskey et al., 2020). Data curation for extreme disturbance events can be a massive time investment (Stevens et al., 2020, p. 119), and so a library of benchmark extreme water cycle datasets eliminates the burden of data curation. When researchers are not hindered by manual classification and data cleaning, they can better leverage modern exascale computing power with AI-ready data.

A library of real-world and simulated openly accessible datasets for extreme water cycle events, with complete metadata and data (i.e., FAIR data), would allow modelers to quickly locate curated datasets for disturbances that have been measured across the US for all the years and relevant variables that have been measured. Such a library would reduce bespoke model training and support model intercomparison.

Narrative

Scientific and technical description of opportunities and approach

The benchmark datasets will focus on including data during, for example, extreme flood and drought events (or other such disturbances) to eliminate gaps in water cycle extremes which limit accurate model predictions.

Barriers to advancing the science

There are five challenges related to curating model datasets on extreme water cycle disturbances:

1) Data diversity - Existing data holdings are multimodal (e.g., timeseries, camera and remote sensing imagery, spatial data layers). The data types will range across scientific disciplines including hydrology, climate, biogeochemistry, ecology, land use, and related human activities. These diverse data need to be standardized for use in models.

2) Data discovery - Although there are a few centralized databases, including within the DOE (e.g., ESS-DIVE, ARM, and AmeriFlux), data needed to comprehensively model water cycle perturbations will be spread out across a range of federal, local, state, academic, and industry databases. Each data source will have different levels of metadata describing their products, and different means of querying, subsetting, and accessing the data. Thus a substantial amount of time will be spent in pulling together different datasets for data mining, pattern recognition, and modeling.

3) Data integration - Different data sources also tend to use a variety of formats, units, and variable naming conventions that make it challenging to create harmonized datasets that can easily be fed into machine learning, deterministic, or hybrid models.

4) Data assimilation - Ideally, models should be able to assimilate new data to incorporate the latest information available into their parameter or state space. This would only be possible if datasets are expanded in near real-time to include additional data as they become available.

5) Data Quality - A typical QA/QC problem is distinguishing between the real extreme events associated with droughts and floods and bad or anomalous data collected in the field due to malfunctioning of monitoring sensors or data acquisition systems, or due to processing mistakes. QA/QC can take a significant amount of time and effort as there are no automated, scaleable tools that enable easy detection, flagging, and removal of suspicious data.

ML-related activities to advance the science

While modelers will benefit from a library of FAIR water cycle datasets, we can also leverage ML/AI clustering algorithms to identify datasets that have broad spatial and temporal coverage (Stevens et al., 2020, p. 118). ML/AI can enable QA/QC through anomaly detection when creating and updating water cycle benchmark datasets. Automating QA/QC checks has the twofold benefit of decreasing the data curation burden when updating datasets and ensuring that scientists have the most recent data available.

Scientific impact

The dataset library would be dynamic and developed with real-time data assimilation so that newly collected data gets QA/QC and then integrated into the benchmark dataset. Examples of curated FAIR DOE datasets that would emerge in the next 10 years include those from hydrobiogeochemical research being conducted in the East River testbed of LBNL's Watershed Function SFA in conjunction with measurement of atmospheric processes from the ARM's mobile facility SAIL campaign, and water flux measurements from the AmeriFlux network (which include NSF's NEON flux observation sites).

Suggested Partners/Experts

Kevin Booth, Radiant Earth Foundation
Hannah Kerner, University of Maryland, NASA Harvest
Gregory Dusek, NOAA AI Executive Committee
Manil Maskey, NASA
Karianna Bergen, Brown University

References

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). ieeexplore.ieee.org.

Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long‐short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9). https://doi.org/10.1029/2019wr026793

Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S. E., Kashinath, K., & Prabhat, M. (2019). Deep-Hurricane-Tracker: Tracking and Forecasting Extreme Climate Events. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1761–1769).

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall--runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022.

Maskey, M., Alemohammad, H., Murphy, K. J., & Ramachandran, R. (2020). Advancing AI for Earth Science: A data systems perspective. *Eos* , *101*. https://doi.org/10.1029/2020EO151245

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225.

Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D. (2020). *AI for Science Report*. Department of Energy.

U.S. DOE. (2018). *Climate and Environmental Sciences Division Strategic Plan 2018–2023* (No. DOE/SC–0192). U.S. Department of Energy Office of Science.