A HPC Theory-Guided Machine Learning Cyberinfrastructure for Communicating Hydrometeorological Data Across Scales

Haowen Xu[1], Melissa R. Dumas[1], Anne Berres[1], Kuldeep R. Kurte[1], Yan Liu[1], Guannan Zhang[1], and Jibonananda Sanyal[1]

[1]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*

Focal Area(s)

(2) Predictive modeling through the use of AI techniques and AI-derived model components. In this white paper, we propose the design and development of a high-performance computing (HPC) –powered cyberinfrastructure to provide a computation-efficient downscaling solution that can interpolate high- resolution predictions of hydrometeorological variables (e.g., precipitation and flood depth grid) using low-resolution raster-based simulation outputs. The cyberinfrastructure adopts a theory-guided machine learning approach to enable timely and high-resolution hydrometeorological predictions for supporting decisions in hydropower operations and mitigation of multiple water-related hazards (e.g., flood, sedimentation, stream bank instability).

Science Challenge

What are the empirical connections among hydrometeorology data across different spatial scales? What role does the process-related parameter (e.g., land use land cover, soil type, and topography) play in this connection? How can we interpolate high-resolution hydrometeorological data (e.g., precipitation and flood depth grid) from low-resolution simulation outputs without a tremendous increase in the simulation run time and computation cost? How can the generated high-resolution hydrometeorology data be validated?

Rationale

High-resolution predictions of hydrometeorological variables are critical for supporting hydropower generation decisions and flood control at hydroelectric power plants. Traditional climate and hydrologic models rely on the numerical simulation of detailed physical processes. Therefore, running these simulations is time-, labor-, and computation-intensive. Improving the spatial and temporal resolution in these modeling outputs could lead to cubic increases in both the simulation time and computational demands, rendering high-resolution hydrometeorological predictions expensive and impractical. Many past studies (Rodrigues et al., 2018; Shi et al., 2016; Chang et al., 2018) apply the super resolution (SR) technique (Park et al., 2003) to downscale climate models using deep learners. However, deep learners are deemed "black-boxes," as their derivation processes from low-resolution outputs to high-resolution outputs are often hidden. Their results are difficult for domain scientists to interpret and validate. Thus, there is a need for an exploratory machine learning approach that can partially integrate domain-specific theory and knowledge into the data-driven mapping process between simulation outputs of different spatial scales. The domain-specific theory and knowledge can be incorporated into the data model through an inductive approach (Olden et al., 2012; Wagener et al., 2010, 2007) in which process-related environmental variables are used and analyzed as key drivers (i.e., environmental surrogates) to reflect the complex physical processes (Auerbach et al., 2015). Many of these variables, such as land use land cover, soil types, topography, digital elevation, air temperature, and various watershed characteristics, can be directly measured through sensors or remote sensing techniques. Additionally, SR applications that can downscale hydrological and hydrodynamics models to efficiently produce high-resolution (1 m) flood depth grids are still rare. Since the flood depth

DOI: https://doi.org/10.2172/1769644
DOE BER, AI4ESP 2021 #AI4ESP1148 (https://www.ai4esp.org)

grid can be used to support critical decisions for flood control operation at hydroelectric power plants, it is crucial to enable an SR-based capability for interpolating high-resolution flood inundation maps.

Narrative

We propose an HPC-powered cyberinfrastructure based on the theory-guided machine learning approach (Karpatne et al., 2017). The approach is designed to leverage domain knowledge and scientific models to improve the effectiveness of data analytics models in optimizing the performance of traditional physical- based models (Shi et al., 2016; Bar-Sinai et al., 2018). The cyberinfrastructure aims to provide an exploratory problem-solving environment that allow researchers and hydropower managers to accomplish the following:

1. Upload and share hydrometeorology data, such as precipitation and flood depth grids.
2. Explore the empirical relationships between data of different spatial resolution scales derived from a combined workflow of machine learning techniques, with the consideration of environmental- process variables defined based on the domain knowledge.
3. Enable computation- and time-efficient interpolation of high-resolution results from low-resolution hydrometeorology data and simulation outputs defined by users.

This proposal's deliverable is a cyberinfrastructure with the capability to interpolate high-resolution simulation results as a web service. This capability will be generic and adaptive so that it can be applied to interpolate raster-based data/simulation outputs that characterize various types of hydrometeorological variables. In the prototyping stage, we target the interpolation of precipitation outputs from the WRF- Hydro Modeling System and flood inundation depth grids from the National Water Model and the Height Above Nearest Drainage tool.

As a use-case scenario, a user would upload the simulated flood inundation map (as raster-based depth grids within the same spatial extents) at different spatial resolutions (e.g., 1 m, 10 m, and 30 m) to train the theory-guided machine learning model behind the proposed cyberinfrastructure. The machine learner, consisting of multivariate clustering and classification algorithms, would create a regionalization based on the relevant process variables (e.g., slope, elevation, and soil permeability) and establish the empirical communication between the simulated hydrologic data of different resolutions within individual regions. The process variables should be selected based on domain-specific knowledge and should be readily available through public sources. In this use case scenario, the watershed slope and elevation data can be retrieved from Terrain Analysis Using Digital Elevation Models (TauDEM), and the soil-related information is available from Soil Survey Geographic Database (SSURGO). Once reliable communication is established, the user can upload a coarse-resolution depth grid (e.g., 30 m) of a new spatial extent (e.g., a catchment) to the cyberinfrastructure, and then expect in return an interpolated 1m flood inundation depth grid. Meanwhile, interactive visual interfaces are provided to allow users to explore the derived across-scale data communication (as multivariate relationships between data of different resolutions and process variables). In this setting, the user can justify the derived results using domain-specific knowledge. The derived across-scale data communication can also provide users with additional data-driven insights that can benefit hydropower operation and flood mitigation.

We will employ a unsupervised deep self-organizing map (SOM) (Sakkari and Zaied, 2020) to map high- dimensional input data—consisting of multi-resolution hydrometeorological data and various process- related variables—to lower-dimensional map space, establishing physically

interpretable linkages among different resolutions of data and their associated process variables. The map space can readily be visualized through the SOM hexagonal layout and other multivariate visualizations to help domain scientists explore and understand the linkages, as well as to justify the interpolated high-resolution results. The overall design of the cyberinfrastructure and the SOM-based data analytics is illustrated in Figure 1. For the cyberinfrastructure implementation, we propose the adoption of the Tensorflow-based SOM package (Khacef et al., 2020) in this project, as the package can be readily deployed on the Summit supercomputer at ORNL.

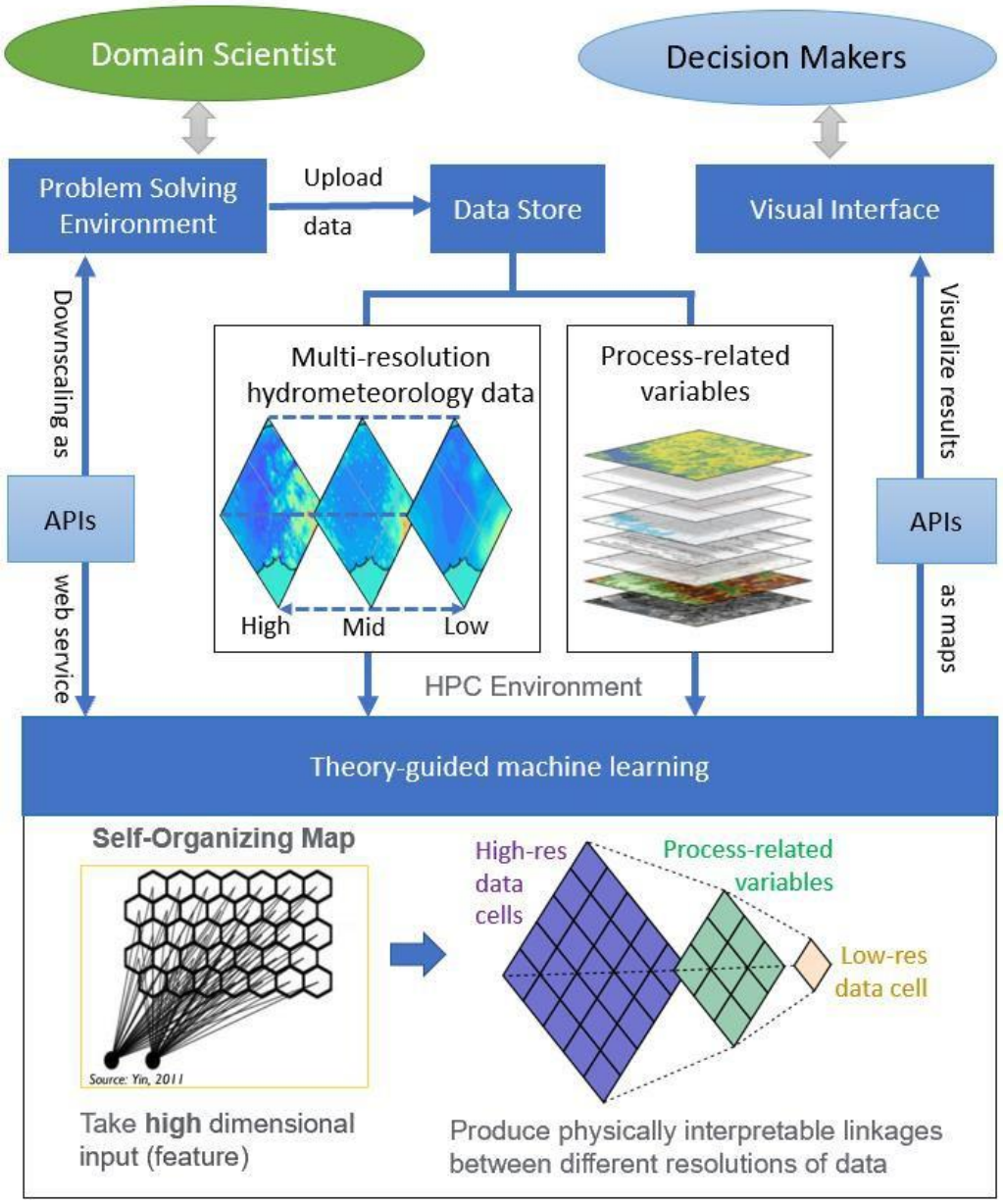

Figure 1. Overall design and data analytics pipeline of the cyberinfrastructure.

Suggested Partners/Experts

Prof. Dr. Marian Muste, Research Engineer, IIHR Hydroscience and Engineering (https://www.engineering.uiowa.edu/faculty-staff/marian-muste) Expertise: Cyberinfrastructure, Hydroinformatics, River Engineering

Caglar Koylu, Assistant Professor, (https://clas.uiowa.edu/geography/people/caglar-koylu) Expertise: Machine Learning, Self-organizing Map

References

Auerbach, D. A., Buchanan, B. P., Alexiades, A. V., Anderson, E. P., Encalada, A. C., Larson, E. I., and Flecker, A. S. 2016. "Towards catchment classification in data-scarce regions", Ecohydrology. 9 (7): 1235-1247.

Bar-Sinai, Y., Brenner, M., Getreuer, P., Hickey, J., Hoyer, S., and Milanfar, P. 2018. Using image super-resolution techniques as a coarse-graining method for physical systems. In *APS March Meeting Abstracts* (Vol. 2018, pp. F54-007).

Chang, Y. C., Acierto, R., Itaya, T., Akiyuki, K., and Tung, C. P. 2018. A deep learning approach to downscaling precipitation and temperature over Myanmar. EGUGA, 4120.

Karpatne, A., et al. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on knowledge and data engineering, 29(10), 2318-2331.

Khacef, L., Gripon, V., and Miramond, B. 2020. GPU-based Self-Organizing Maps for post-labeled few-shot unsupervised learning. In International Conference on Neural Information Processing (pp. 404- 416). Springer, Cham.

Olden, J., Kennard, M., and Pusey, B. 2012. A framework for hydrologic classification with a review of methodologies and applications in ecoyhydrology. Ecohydrology 5. doi:10.1002/eco.251

Park, S. C., Park, M. K., and Kang, M. G. 2003. Super-resolution image reconstruction: a technical overview. IEEE signal processing magazine, 20(3), 21-36.

Rodrigues, E. R., Oliveira, I., Cunha, R., and Netto, M. 2018. DeepDownscale: A deep learning strategy for high-resolution weather forecast. In 2018 IEEE 14th International Conference on e-Science (e- Science) (pp. 415-422). IEEE.

Sakkari, M., and Zaied, M. 2020. A Convolutional Deep Self-Organizing Map Feature extraction for machine learning. Multimedia Tools and Applications, 1-20.

Shi, W., et al. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1874-1883).

Wagener, T., Sivapalan, M., Troch, P., and Woods, R. 2007. Catchment classification and hydrologic similarity. Geography Compass 1, 901 – 931. doi:10.1111/j.1749-8198.2007.00039.x

Wagener, T., et al. 2010. The future of hydrology: An evolving science for a changing world. Water Resources Research 46. doi:10.1029/2009WR008906