A Self-Evolution Data Fusion Platform for Large-Scale Water Models

Xinya Li (Data Scientist), Chris Vernon (Data Scientist), Min Chen (Earth Scientist), Heng Wang (Software Engineer), Jason Hou (Data Scientist)

Focal Area:

Data acquisition and assimilation enabled by machine learning (ML), artificial intelligence (AI), and advanced methods including experimental/network design/optimization, unsupervised learning (deep learning), leveraging advanced hardware (e.g., edge computing). This development is to enable solving the science questions regarding the human and climatic factors that interact with and drive global water scarcity.

Science Challenge:

Data-model integration in the context of complex high dimensional hierarchical nonlinear structure. A self-evolution, comprehensive data fusion platform; building data fusion approaches from the ground up into a web-based, globally accessible platform.

Rationale:

AI's superpower is fueled by data. The successes of applying AI for building integrative predictive models rely on data accessibility, adequacy, and quality, which are generally promoted for data managed in adherence to FAIR (findable, accessible, interoperable, and reusable) guiding principles. Even considering the ability to employ FAIR data, it is still non-trivial to integrate data for scientific discovery (data fusion) given the usually high-dimensional and hierarchical nonlinear cross-dependence structure nature of some domains (e.g., complex large-scale water modeling systems). Recent advances in computational resources, observation and monitoring methods, and sensing tools, bring unprecedented opportunities in data-model integration or data fusion but introduce new challenges that require a systematic approach to provide heuristic value.

In this white paper, we identify a method to address the importance of automated transferable data fusion for large scale, high-dimensional water cycling observational and simulation data, with the focus on adaptivity to various data formats, the ability to characterize temporal and spatial heterogeneity, and represent complex structures that can provide value to the scientific community.

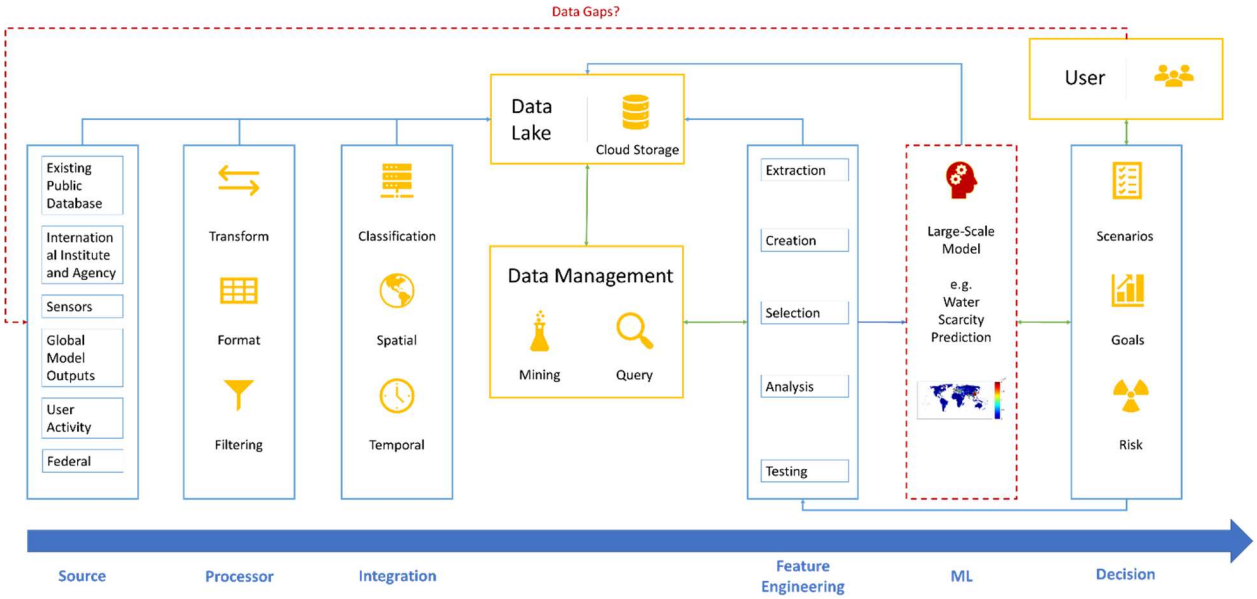
Narrative:

Conservation of Earth's water balance and integrative and associated water cycle extremes are a pointed interest from a socio-economic perspective and are among the most highly-research topics within the Earth sciences. A global-scale study related to water scarcity and its regional socioeconomic implications can serve as a testbed for the data fusion development. For example, the effort to collect and harmonize data to support global-scale water scarcity studies which have been conducted using GCAM (Global Change Analysis Model) and the suite of models that

DOI: https://doi.org/10.2172/1769652

provide data to, and downscale from, GCAM [1-5] have exposed the value of this type of analysis and have invoked a flurry of community data requests for those seeking to conduct their own analysis. For global-scale water models, large data sources range from observations or simulations such as, population (density), country/region distribution, water basin and geopolitical region distribution, irrigated areas, livestock density, land area, soil moisture, average temperature, precipitation, and sectoral water withdrawals. Each of the data structures requires a data fusion strategy to be useful from the perspective of reproducibility, consistency, and accuracy. The fused and validated data provide higher confidence and large spatiotemporal coverage, as well as keep an ML-based up-to-date roadmap for the data story to alleviate the information gap introduced into the subsequent predictive modeling. Data fusion outcomes reflected by model adoption is also expected to reveal important physical controls/dynamics by testing hypothesized drivers of water cycling.

A self-evolution, comprehensive data fusion platform will be substantially beneficial to multiple exascale water models, which will include generic and domain-specific approaches to solve fusion challenges of database storage, data structures, redundancy and deficiency, downscaling and upscaling, assumption compatibility, projection and mapping and joint-calibration. The fused datasets will also provide the flexibility to evolve through additional ML methods. Also, building data fusion approaches from the ground up into a web-based, globally accessible platform is essential to encourage contribution by the scientific community, decision makers, and public for maximum impact. This web application will also take advantage of the most recent applications using cloud data fusion that leverage a convenient user interface for storing and building data pipelines. Adopting Bayesian vision for dependence on use of data, this platform of data fusion will allow incremental assimilation and propagation of updates/inferences across the complete data stack and improve the predictability of global water cycle and human-Earth system dynamics.



An initial design is illustrated in the above figure from the raw data collection to the final decision of application in a ML model. Datasets from various sources are collected, processed, and integrated before storage into cloud-based data lake. Management tools through querying and mining are applied to build interface to this data lake, which is for the purpose of exchanging information to the following feature engineering. Feature engineering connects data from platform and ML model to be integrated. The data user will be responsible for controlling the platform to prepare feature datasets according to ML goals and scenarios. The output dataset of feature engineering can be used for the ML model, which is separately developed. Also, the feature datasets and ML model outputs will recharge to the data lake for future use and fill possible data gaps. As a result, this platform will achieve a self-evolution through ML model developments and data-model integration by community users. Due to the large-scale data storage, cloud infrastructure (such as Cloud DataLab, Cloud Dataprep, BigQuery) is suggested to help facilitate the involved data processing steps.

Suggested Partners/Experts:

University of Wisconsin
University of Maryland

References:

1. Graham N.T., M.I. Hejazi, M. Chen, E. Davies, J.A. Edmonds, S.H. Kim, and S. Turner, et al. 2020. "Humans drive future water scarcity changes across all Shared Socioeconomic Pathways." Environmental Research Letters 15, no. 1: Article No. 014007. PNNL-SA-151297. doi:10.1088/1748-9326/ab639b
2. Vernon C.R., M.I. Hejazi, S. Turner, Y. Liu, C.J. Braun, X. Li, and R.P. Link. 2019. "A Global Hydrologic Framework to Accelerate Scientific Discovery." Journal of Open Research Software 7, no. 1: Article No. 1. PNNL-SA-137438. doi:10.5334/jors.245
3. Huang Z., M.I. Hejazi, X. Li, Q. Tang, C.R. Vernon, G. Leng, and Y. Liu, et al. 2018. "Reconstruction of global gridded monthly sectoral water withdrawals for 1971-2010 and analysis of their spatiotemporal patterns." Hydrology and Earth System Sciences 22, no. 4:2117-2133. PNNL-SA-129126. doi:10.5194/hess-22-2117-2018
4. Li X., C.R. Vernon, M.I. Hejazi, R.P. Link, Z. Huang, L. Liu, and L. Feng. 2018. "Tethys - A Python Package for Spatial and Temporal Downscaling of Global Water Withdrawals." Journal of Open Research Software 6, no. 1:9. PNNL-SA-129638. doi:10.5334/jors.197
5. Chen, M., Vernon, C.R., Graham, N.T. et al. “Global land use for 2015–2100 at 0.05° resolution under diverse socioeconomic and climate scenarios.” Sci. Data, 7, 320 (2020). https://doi.org/10.1038/s41597-020-00669-x