# AI-Driven Data Discovery to Improve Earth System Predictability

Ranjeet Devarakonda[1], Jitendra Kumar[1], Dalton Lunga[1], Jong Choi[1], and Giri Prakash[1]

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

## Focal Area(s)

**(3)** Insight gleaned from complex data (both observed and simulated) using AI, big data analytics, and other advanced methods, including explainable AI and physics- or knowledge-guided AI.

## Science Challenge

The number of data sets, models, capabilities, and tools within Earth and Environmental Systems Sciences Division (EESSD) portfolios is proliferating, with a multitude of data formats, structures, designs, and languages. Given the sheer volume of information and fragmentation of data across multiple repositories, finding relevant data may not be a trivial task; conversely, scientists may also miss important data products or pre-trained models from other repositories that are critical for their research. The current status quo has led to many disparate model repositories and databases that are inaccessible to a wide community of scientific experts. A convergent AI-guided discovery framework can enable search across distributed repositories, provide seamless access through a consistent API, and provide tools for assimilating heterogeneous observations from remote sensing platforms and spatially sparse and distributed networks of sensors. Such a framework holds the key for reliable, accurate, and timely characterization of hydrometeorological conditions such as soil moisture and evapotranspiration fluxes; these are critical for many regional-scale applications, including numerical weather predictions, surface/subsurface hydrology, flood forecasting, drought monitoring, agricultural impacts and planning, and climate change studies.

## Rationale

Evapotranspiration and soil moisture are critical components of the hydrological system, influencing groundwater storage and recharge, surface runoff and water availability, dynamic land-atmosphere processes, and the global climate system. These fluxes are influenced by a wide range of conditions ranging from soil, topography, vegetation, and snow to land use and regional to global patterns of climate and feedback. Reliable, accurate, timely characterization of soil moisture and evapotranspiration fluxes is critical for many regional-scale applications, including numerical weather predictions, surface/subsurface hydrology, flood forecasting, drought monitoring, agricultural impacts and planning, and climate change studies. [1][2] However, the ability to measure, estimate, and model these processes at regional to global scales remains limited because of the underlying challenge of data discovery and access. There is a clear need for a scalable and integrated one-stop-shop framework to "predict" user data search preferences and "link" the most relevant products in real time. Answers to the questions when, where, and how are considered the critical pieces of metadata and play an essential role in finding relevant and similar products. Most of the data projects within the EESSD portfolio already have established processes that capture these critical metadata elements; however, they are siloed so that (1) they are not seamlessly interconnected and (2) they do not allow for searching across the projects. [3] As noted in the science challenge section, given the ongoing proliferation of data from Internet of Things sensors, models, and tools, there is a clear need for EESSD to embrace modern AI-powered techniques to connect scientists with the most appropriate, relevant data. An AI-based data discovery process is motivated by multiple benefits over classical machine learning methods. Most important are efficiency, generalization capability and superiority in exploiting spatio-temporal dependencies in Earth systems data.

DOI: https://doi.org/10.2172/1769671

## Narrative

Regional- to global-scale remote sensing products from satellite-based platforms such as NASA SMAP, SMOS, and ASCAT have enabled the assimilation of these hydrometeorological fluxes to derive near-real time data products. Observations for soil moisture and evapotranspiration fluxes using ground-based sensors are steadily increasing across the globe and can be assimilated with remote sensing products for accurate estimation of fluxes. However, efforts to develop assimilation products often have been limited to small watersheds at regional scales. The potential for regional to global estimation of soil moisture and evapotranspiration fluxes is yet to be realized. Limiting the scientific progress is the fact that ground-based observations of soil moisture and evapotranspiration fluxes collected by different research groups and agencies often are archived in independent databases, each with its own protocols for data and metadata and accessibility, and often are not readily usable together. Heterogeneous data collection protocols and sensors add sources of potential uncertainty when such data are used together in an assimilation framework. Traditional ML models are being exploited to address these challenges. However, these models operate under the assumption of statistical regularity and the assumption that training and test data are drawn from the same distributions. The design of models proceeds in a trial-error fashion using limited training data. When shifts in data distributions are encountered, the process is developed from scratch, thus impeding efforts to achieve reproducibility and reusability of data and their models. These challenges create opportunities for new research undertakings, not only to mitigate the need to collect training samples to enhance data discovery for Earth system predictability, but also to build models to characterize the underlying data and model uncertainties. Such a data and model discovery framework stands to advance the pace of scientific discovery for many Earth sciences applications.

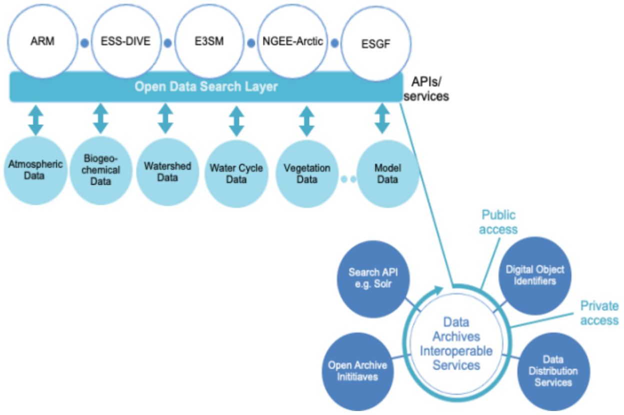

*Figure 1. EESSD project landscape—showing diverse data holdings and existing capabilities*

The current landscape of the Earth sciences encompasses complex, diverse, and extremely large data assets that pose a unique set of challenges, including the findability and linkability of data and the characterization of uncertainties in both data and models. Many archives hosting these diverse data are already FAIR [4] and follow established data management and data distribution practices (Figure 1). For example, The ARM data center archives distribute and provide end-to-end data management capabilities for over 2.8 PB of atmospheric data from instruments, value-added products, model outputs, field campaigns, and many other principal investigator–contributed data products. [5] Like many other data-intensive projects within EESSD, ARM provides robust data search capabilities via various methods that are already machine-readable per schema.org specifications. As an example, Apache Solr, one of the popular metadata search APIs, is currently being used by ARM and several other projects, including ESS-DIVE, NGEE Arctic, ESGF, and others. Using such existing machine-readable resources, a novel AI-based system can be built to search and connect data sets from these diverse domains to solve the science problem described above.

*Technical Implementation:* We propose developing a deep-learning–based interface, convolutional neural network (CNN) API aimed at providing multiple data source aggregation, metadata training, and connecting AI-powered analytical engines. The CNN API performs unsupervised learning, data filtering, and building clustering models to inform the results. Finally, the engine will further devise a relational model that leverages the relative ordering uncertainty estimates on different CNN models with respect to multiple datasets to make inferences regarding model parameters and the underlying complexities in the data.

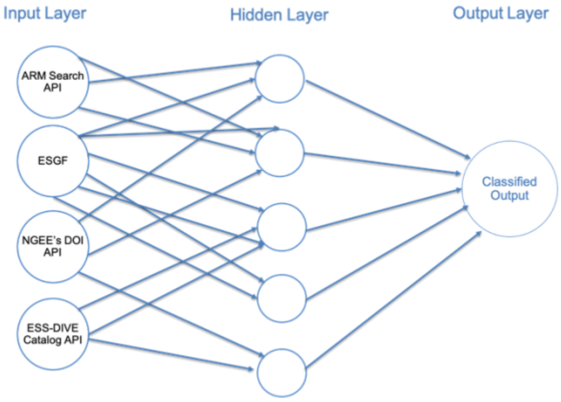

*Figure 2. CNN API to build a trained metadata two-vector dataset.*

This primary purpose of the CNN API (Figure 2) is to interconnect, provide uncertainty estimates, and aggregate data sources from various data archives. The input layer for CNN will be the existing "heterogenic" metadata search API endpoints, such as Solr or ElasticSearch. The hidden layers will apply continuous bag-of-words [6] and skip-gram [7] algorithms to classify relevant metadata from heterogeneous sources to build a metadata vector model. The output of that process will be a well-classified trained metadata and uncertainty estimates database that can be readily leveraged for information or critical metadata properties such as measurement, location, type of data, and spatial information of the data set. To build the "uncertainty-aware" AI-powered data search system, we will filter, and exploit Bayesian-based [8] methods to cluster the metadata using the vectored metadata dataset. Content-based filtering and collaborative filtering are commonly used techniques for filtering data. We can apply a hybrid approach that uses both techniques to leverage critical metadata properties from the trained database with content-based properties and the collaborative recommendations provided by our domain experts.

After the data have been retrieved and filtered, we can use them to create an unsupervised learning model. Clustering is one of the most popular analytical techniques for grouping similar objects. The goal is to segregate groups with similar traits and assign them to clusters. There are multiple ways to approach the clustering task, and the importance of the technique depends on whether the clustering criterion is associated with the phenomenon under study. We propose to use a hierarchical clustering algorithm, [9] a simple and straightforward form, to segment the data into similar groups and build a predictive model for each group by computing the distances based on the derived metadata vector. Once the distance is computed, we will cluster the results and determine an appropriate tree size to map the results accurately.

*Future studies and conclusion:* The proposed outcome of this work will be (1) a fully trained dataset that contains "vectored" discovery-level metadata from data/models/tools (meta2vec) and (2) an AI-powered data search system that can connect with relevant datasets in real time. The meta2vec can also be used independently in non-DOE projects to transform the unlabeled raw corpus of scientific metadata into well-classified scientific data inventory. There will be no need for preprocessing and thus little to no need for memory. The AI-powered data search system can also work independently and can be used by any external science repository to search across the EESSD portfolio while hiding the coding complexities.

## References

[1] Brocca, L., Melone, F., Moramarco, T., and Morbidelli, R. Spatial-temporal variability of soil moisture and its estimation across scales, Water Resour. Res., 46, W02516 (2010). doi:10.1029/2009WR008016.

[2] Naz, B. S., et al. A 3 km spatially and temporally consistent European daily soil moisture reanalysis from 2000 to 2015. Sci Data 7, 111 (2020). https://doi.org/10.1038/s41597-020-0450-6.

[3] Geernaert, Gary, et al. Earth and Environmental Systems Sciences Division Strategic Plan - Data-Model Integration Scientific Grand Challenge: 2018–2023. No. DOE/SC–0192. US DOE Office of Science (United States) (2018).

[4] Wilkinson, M., 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

[5] Devarakonda, R., et al. “Big Federal Data Centers Implementing FAIR Data Principles: ARM Data Center Example.” 2019 IEEE International Conference on Big Data (Big Data). IEEE (2019).

[6] Zhang, Y., Jin, R., and Zhou, Z.-H. “Understanding bag-of-words model: A statistical framework.” International Journal of Machine Learning and Cybernetics 1.1-4 (2010): 43–52.

[7] Guthrie, D., et al. “A closer look at skip-gram modelling.” LREC. Vol. 6, (2006).

[8] Heller, K. A., and Ghahramani, Z. “Bayesian hierarchical clustering.” In International Conference on Machine Learning 2005, (2005).

[9] Devarakonda, R., Kumar, J., and Prakash, G. “Clustering-Based Predictive Analytics to Improve Scientific Data Discovery. In 2020 IEEE International Conference on Big Data (Big Data), (2020).