# AI-Based Upgrades to Observational Data Centers to Facilitate Data Interoperability

Giri Prakash[1], Nicki Hickmon[2], Adam Theisen[2], Debjani Singh[1], Cory Stuart[1], Ranjeet Devarakonda[1], and Jitu Kumar[1]

*[1]Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA; [2]Environmental Science Division, Argonne National Laboratory, Lemont, IL, USA*

## Focal Areas

**(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). Focal areas 2 and 3 have critical dependencies to the modernization described.

Key benefits to the focal areas: (1) Modernized observatory framework capable of agile adaptive observation, (2) Advanced instrument and data tagging supporting AI data acquisition for assimilation or validation, and (3) Widespread data interoperability bridging Earth system prediction scales

## Science Challenge

The integrated water cycle is composed of many processes. From fine scale processes such as evapotranspiration, entrainment and mixing to emergent behavior such as deep convection, sea ice formation and stratocumulus decks, each phenomena needs to be observed, modelled, understood and integrated into ESMs in order to improve earth system predictability. And, as outlined in the white paper call, MODEX is a key concept to achieve continuous improvement of numerical simulations of the earth system across spatial and temporal scales. To achieve the MODEX concept, it is critical to assess current observatory pipelines and their data management components and modernize them to support AI-based research missions. The community model, data, and analysis capabilities listed on the EESD's MODEX diagram represent the diverse nature of resources available to scientists and highlight the need for a robust data integration strategy to serve across the MODEX enterprise. This white paper proposes to address these challenges by evaluating the current data pipeline and improving it from data collection to distribution to meet the AI mission. The white paper's scope includes identifying specific data services components of ARM, applying them for similar sensor-based measurement projects, and developing and adapting community-based standards to enable data interoperability to support AI projects and broaden the scope of the datasets.

## Rationale

Data integration is the fundamental component for conducting interdisciplinary research for integrative and associated water cycle extremes. Many of the EESSD repositories provide highly relevant but domain specific datasets for this research. To enable edge computing–based, AI-driven sampling techniques and on-demand data access for AI-driven dynamic data assimilation for Earth system predictability, common standards and preferred data access methods must be established. Modernizing the data collection pipeline with a modular architecture for real-time data access with plug-and-play AI models will be extremely beneficial for the proposed MODEX enterprise. If this architecture is established, various EESSD field data collection projects, such as

DOI: https://doi.org/10.2172/1769667
DOE BER, AI4ESP 2021 #AI4ESP1107 (https://www.ai4esp.org)

ARM, NGEE, and Spruce, will be able to leverage and upgrade their data collection pipeline to support the MODEX concept.

The benefits of the proposed approach will include improved AI-friendly data collection, discovery, and distribution infrastructure that could be applied across the EESSD and other data-intensive projects. Specific benefits will include new standards and protocols for real-time data access from instruments, real-time data quality analysis and data reductions, modular data flow architecture to support various AI models, and community-based data discovery web services for providing a direct data access pipeline from EESSD data repositories to AI models.

The proposed approach coupled with advanced 5G network capability (Beckman et al., 2020) and edge-computing offers new opportunities for near real-time data analysis and data collection configurations.

## Narrative

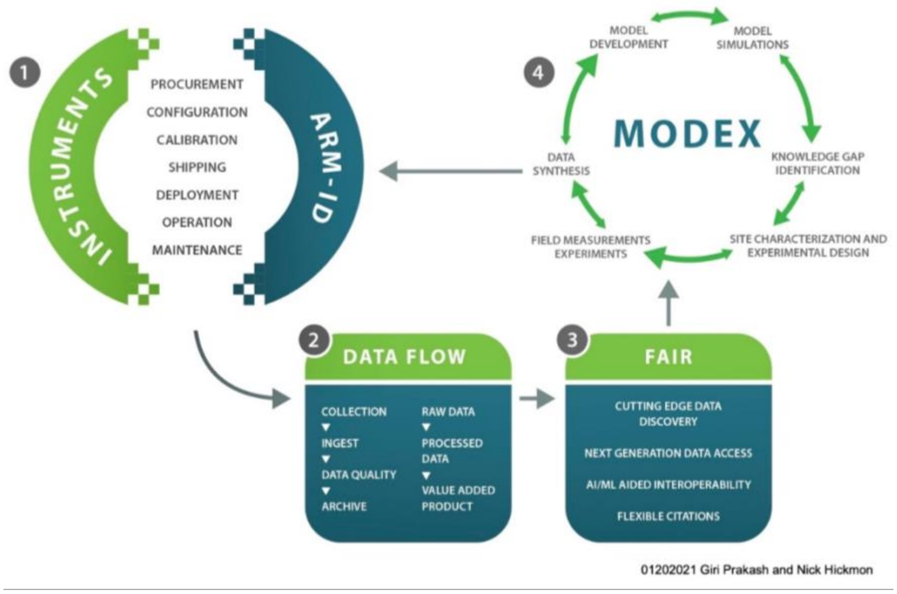
Our methodology follows four activities critical to advancing the AI science mission:

***Activity 1:*** *Evaluating the data pipeline components from collection to distribution by utilizing "digital twin" functionality for instrument interoperability to enhance MODEX design*

Current field sensor projects do not have strong and seamless connections to MODEX activities. To achieve such connections, a new metadata methodology describing the instrumentation and the data resulting from operation must be developed. The goal is to define the output of an instrument and the data products built off the output to enable the emulation of multiple data stream analysis and its use in AI techniques. Prepared datasets for use in an AI-driven dynamic assimilation method or instrument/measurement emulator, require data and information not only be efficiently discoverable and available but also must include all pertinent information on which to base selective decisions. This information would need to be processed at the speed that the specific use requires. Some information, such as instrument model and configuration, is somewhat straightforward to capture. But new formatting methods are required to flow calibration information necessary for uncertainty quantification with a format that can be used in "on-the-fly" processing using multi-instrument/observation data. A relevant, as opposed to realistic "digital twin" level of information regarding instruments and observations is required. We are looking for accurate representation of instrument measurements with all pertinent dependencies defined and accurately represented, either for actual operation or for adjustment in emulator settings. The result will be defined instruments/measurements available for AI-driven assimilation, emulation, parameterizations, and model algorithms.

The intent is to apply this methodology from an observatory perspective, such as ARM, which already adheres to FAIR (Findable, Accessible, Interoperable, and Reusable) principles ensuring they are maintained. This methodology will increase the interoperability of datasets by defining differences within datasets so they can be overcome, thus broadening the accessibility of DOE archived data to AI-driven techniques.

As illustrated in Figure 1, the team will develop a database to capture the details of instruments and their statuses, enabling a digital twin for field instruments that will be critical for instrument design and operation. This concept will also be applied to streamline the data pipeline used for processing and archiving petabytes of sensor data from remotely deployed instruments.



01202021 Giri Prakash and Nick Hickmon

*Figure 1. Digital twin concept to manage data sources and connect to MODEX.*

From MODEX to observations, the data simulators and emulators will help feed the observational network design and operation. Additionally, this and activity 2, supports edge computing with 5G for real-time data access, data quality analysis, and configuration of instruments that are deployed in remote locations.

***Activity 2:*** *Real-time instrument data quality analysis and processing using edge computing*

To support the real-time data flow required by activity 1, development and deployment of a modular edge computing framework at a site, such as the ARM Southern Great Plains central facility, would allow for the real-time quality analysis of instruments to be available. This real-time analysis would enable detection and flagging of problematic data by cross-analyzing datasets before their use in a dynamic selection process. The use of multiple variables for quality control is similar to efforts by the radar community to quality control radar data using neural networks with multiple polarimetric variables (Lakshmanan et al. 2014). Deep neural network has been shown to improve retrievals of temperature and humidity profiles from a microwave radiometer (Yan, et al. 2020) or a fully convolutional network to calculate cloud masks from lidar data (Cromwell and Flynn 2019). Coupling machine learning based retrievals with real-time high-quality data would allow for the production of advanced data products with lower uncertainties in a more efficient manner as compared to current products which can take up to 6-months or more to produce a final product. Combining real-time data with edge computing capabilities would also provide an avenue for real-time modifications to the measurement strategy to better capture events of scientific interest. This could include modification of radar or lidar scan strategies to focus on a single cloud or it could include modify the temporal frequency in which data are collected such as shifting from 1-minute averages to 1-second sampling intervals when atmospheric events of interest traverse over the domain. All of these capabilities—real-time data flow, quality control, data product generation, and adaptive sampling driven by edge computing and ML—will greatly benefit the scientific community and help accelerate the science.

***Activity 3:*** *Data tagging to identify benchmarking datasets*

AI-based models require accurate, clean, well-labeled, and well-prepared data and metadata to produce the desired output. With thousands of diverse data streams collected for domain-specific research, it is difficult to identify the most appropriate data for AI-based inter-domain research. As an example, the ARM Data Center repository (Prakash et al., 2018) currently holds over 2.8 petabytes of data covering more than 11,000 diverse data products. Of this, more than 50 different data streams contain cloud properties data at various spatiotemporal scales. Data tagging based on recommendations and suitability for use in broader research areas, such as the integrative water cycle, will help the AI models to readily access the most appropriate data. It will also help the data repositories prepare and provide training datasets for unsupervised learning. ML techniques can support the identification of the so called "master data," or the most relevant data. With techniques such as pattern extraction, information retrieval, and classification using genetic algorithms; support vector classifiers and k-nearest neighbors; and the use of input from subject matter experts, these data streams could be tagged and used beyond the originally anticipated use cases. These techniques will intelligently detect similar data by clustering similar columns, identifying likely matches, and recognizing patterns within data files. A combination of deterministic, heuristic, and probabilistic algorithms can be used to synthesize contextual attributes (e.g., when, what, where, who) from unstructured data for use in the matching process. Data tagging based on recommendations and suitability using probabilistic classifiers similar to supervised learning with Naïve-Bayes algorithms and MaxEnt (multinomial logistic regression), natural language processing, and Stanford's Named Entity Recognizer can be used in broader research areas, such as the integrative water cycle. This approach will help the AI models readily access the most appropriate data or master data and help data repositories prepare and provide training datasets for unsupervised learning.

***Activity 4:*** *Enabling data interoperability between data repositories and AI models, developing and extending community-based standards and protocols*

Although datasets may technically be interoperable, integration or communication among these datasets often fails because of a lack of cross-domain ontologies and standards, significantly impacting data sharing with inter-domain AI-based research activities such as Earth system prediction. Successful data interoperability can be achieved by creating or extending currently available data sharing standards and protocols (e.g., ISO 19115 , FGDC, web services) and establishing frameworks to facilitate dynamic data discovery (Devarakonda et al., 2020) and data transformation into AI analysis-ready forms. In addition, ontologies provide background knowledge that can be exploited in ML models with domain-specific keywords as training sets (Kulmanov et al., 2020). The white paper team proposes to work with user communities, EESSD's data repositories, and other data centers to develop/extend the standards, protocols, and cross-domain ontologies to achieve maximum data interoperability to support AI-based research activities including integrative water cycle and associated water cycle extremes.

### Suggested Partners/Experts

**Yaxing Wei** is the lead scientist of the ORNL Distributed Active Archive Center (DAAC). His research interests include data management, visualization, sharing, and analysis.

**Shaocheng Xie** is the group leader of the LLNL cloud processes research and modeling group. He is the project leader for the development of the DOE E3SM next generation of atmospheric physics. He also leads the ARM effort to bridge ARM data and climate model development. His research interests include climate modeling and evaluation, cloud and convection parameterizations, and ARM data integration, quality and uncertainty quantification, and objective analysis.

We plan to leverage components of the "AI-Driven Data Discovery to Improve Earth System Predictability" white paper submitted by Devarakonda et al. to implement the data interoperability and data sharing concepts explained in this white paper.

### References

Beckman, P., et al. 2020. 5G Enabled Energy Innovation: Advanced Wireless Networks for Science, Workshop Report. United States: N. p., 2020. Web. doi:10.2172/1606538.

Cromwell, E. and D. Flynn, "Lidar Cloud Detection With Fully Convolutional Networks," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 2019, pp. 619-627, doi: 10.1109/WACV.2019.00071.

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

Kulmanov, M., Smaili, F. Z., Gao, X., and Hoehndorf, R. 2020. Semantic Similarity and Machine Learning with Ontologies, *Briefings in Bioinformatics*, https://doi.org/10.1093/bib/bbaa199

Lakshmanan, V., Karstens, C., Krause, J., & Tang, L. (2014). Quality Control of Weather Radar Data Using Polarimetric Variables, Journal of Atmospheric and Oceanic Technology, 31(6), 1234-1249. Retrieved Feb 5, 2021. from https://journals.ametsoc.org/view/journals/atot/31/6/jtech-d-13-00073_1.xml

Prakash, G., Devarakonda, R., Records, R., and Dumas, K. K. 2018. "Modern Scientific Data Management Practices: ARM Data Center Example," vol. 2018.