DOE BER Earth and Environmental Systems Science Division (EESSD) Call for White Papers to Advance an Integrative Artificial Intelligence Framework for Earth System Predictability: AI4ESP

Title – Integrating Applied Energy and BER Smart Data Capabilities to Develop a DOE Data Fabric for Energy-Water R&D

Authors/Affiliations: Kelly Rose[1] (lead); Jennifer Bauer[1], Tom Feeley[1], Paige Morkner[1,2], Chad Rowan[1,2], Mike Sabbatino[1,2], Chung Shih[1,2], and Dirk VanEssendelft[1]

[1] National Energy Technology Laboratory; [2] NETL Support Contractors

Focal Area(s): 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).

Science Challenge: DOE R&D, including DOE's Basic Energy Research (BER)'s Environmental Systems Science Division (EESSD) program and DOE's applied energy research (AER) programs (EERE, FE, and NE) are producers and consumers of Earth systems datasets. This white paper focuses on the first topic area from the call in relation to how crosscutting resources and innovations from DOE's EESSD and AER can be brought to bear to mutual benefit and more efficient energy-water, Earth system data resources through improved.

The overarching challenge posed by this call focuses on how DOE can directly leverage artificial intelligence (AI) to engineer a substantial (paradigm-changing) improvement in Earth System Predictability? While stemming from DOE BER's EESSD program, this is a challenge that is faced and also being addressed by DOE's AER programs. Over the past decade plus, FE, EERE, and NE programs have made important strides towards addressing this need. These strides are in many ways highly complementary to EESSD's MODEX efforts. Energy water systems spanning metocean to groundwater to surface water systems all are data driven whether for basic energy or applied energy.

These are remote, multi-variate, complex natural, and in many cases engineered, systems. Key needs and challenges of both EESSD and AER include developing data-focused tools to enhance data search and discovery to fill in knowledge gaps (address sparse data challenge), and rapidly transform datasets, including disparate and multi-source data. Leveraging DOE on-premise computing (HPC, exascale) infrastructure supports the computing-intensive algorithms required to execute these data acquisition and transformation processes to derive enriched knowledge and data, driving AI/ML and big data analytics for these systems. The opportunity lies in combining BER and AER efforts to provide a more robust, advanced, efficient and complete computing data fabric to address energy-water data acquisition and assimilation needs which currently pose significant impediments to AI/ML predictions and research.

Rationale: Data accessibility, discovery, and transformation resources to meet the need of each individual AI/ML research Earth system predictability project, remains one of the primary challenges facing DOE EESSD and AER affiliated researchers alike. These needs are well documented in the July 2020 release of the AI for Science technical report, but are also described in various news and technical publications (e.g. Gibney and Van Noorden 2013; Mons, 2020; Wilkinson et al., 2016). Studies estimate that data science projects, R&D or analytical, spend up to 80% of project time finding, gathering, and transforming information, and only 20% on modeling and analytics itself (Crowdflower). There have been inroads made by DOE EESSD and AER, including improved data curation and collection capabilities through DOE's FAIR data platforms, such as EDX®, OpenEI, ESS-DIVE. However, DOE affiliated AI/ML researchers still face significant challenges finding, exploring, cleaning, and efficiently transforming data to meet their AI/ML and data science goals. There are significant opportunities to leverage DOE's advanced computing resources with AI/ML and deep learning approaches to address the impediments researchers currently face, to flip the 80:20 challenge AI/ML projects currently face and mutually benefit EESSD and AER energy-water research.

DOI: https://doi.org/10.2172/1769726

EESSD and AER energy-water and Earth predictability R&D require the same foundation of data and knowledge. Likewise, both EESSD and AER programs have strengths and capabilities to support and address more systematic, "smart" development of this data foundation. AER programs have been tackling aspects of these challenges, using AI/ML and deep learning tools to develop custom solutions to these challenges that offer significant potential to benefit EESSD's MODEX ecosystem, and if partnerships with EESSD performers and capabilities can be scaled and further matured to offer mutual benefit and address the 80:20 challenge.

Narrative: AER R&D programs capabilities, such as the National Risk Assessment Partnership (NRAP), Science-informed Machine Learning to Accelerate Real-Time Decisions (SMART) Initiative, Offshore Unconventional Portfolio, the Onshore Unconventional Portfolio, Carbon Storage program, and Computing Science and Engineering have data discovery, parsing, and simulation efficiency needs that strongly correspond to the MODEX EESSD Earth predictability needs to drive future AI/ML R&D breakthroughs. For this workshop, we propose building off AER data acquisition, assimilation, and computing technologies that have been in development under NETL's programs over the past five years, and capitalizing on NETL's relationships and collaborations with key performers (e.g. OS laboratories, academic, and agencies) who also align to EESSD's mission space, to address mutual AER, BER data acquisition and assimilation needs.

EDX SmartSearch is a massively parallel parsing infrastructure combined with natural language processing (NLP) and machine learning (ML) combined with a custom recommendation engine to facilitate automated data discovery. SmartSearch automates data discovery by analyzing training content such as images, documents, publication, search terms or phrases, and finding related content via worldwide web, local, enterprise datastores, and using advanced NLP parsing to evaluate relevance of data returns relative to the training corpus. SmartSearch leverages containerized services to support ingestion, analysis, discovery, and recommendations. Key features of tool include infinitely scalable (automated) data discovery capable of analyzing millions of files and generate comparison metrics, and generation of topic models, categorization in the desktop, cluster, or cloud environments. SmartSearch treats geospatial data (shapefiles, geodatabases) like a document, automatically extracting text, comparing textual versus geospatial data to identify relevancy. It can search for meta tags within HTML body of discovered web sites, and analyze archive files –even archives within archives (zips within zips, etc). SmartSearch v1 was run on NETL's cluster computing infrastructure to support development of the award-winning Global Oil and Gas Infrastructure (GOGI) database in a 4-month period of performance (Rose et al., 2018). Subsequent enhancements and the ongoing adaptation for the GCP environment, with support from DOE's OCIO, afford opportunities to support data discovery needs for a wide range of EESSD and AER energy-water data needs.

Complementary to the at-scale, deep learning capabilities of SmartSearch, SmartParse is a tool to help understand the natural language used in scientific papers and to help classify what topics were discussed within the corpus. This tool can be used to automatically pull and process information from papers and documents, providing a multifaceted analysis of a corpus of scientific documents. Utilizing NETL's Watt machine learning cluster, SmartParse can quantify and extract data from the collected papers to help with new scientific data discovery and data collection. When combined with the SmartSearch tool, researchers have a full suite of tools for scientific data discovery, data analysis, data extraction and data curation.

SmartSearch and SmartParse have been used in published and in press studies for AER R&D (e.g. Baker, et al., 2016; Morkner et al., in review; Rose et al., 2018). NETL smart-data quality and uncertainty quantification capabilities are also documented in (Bauer & Rose, 2016; Wenzlick et al., 2020). These capabilities, particularly SmartSearch and SmartParse, address key elements of the science challenge, "Developing data tools to improve predictions, fill in knowledge gaps and drive AI/ML and big data analytics for these systems."

SmartSearch and SmartParse are tools that address needs for data acquisition and assimilation both, through flexible, AI/ML enhanced approaches. Thus, for this workshop, the goal will be to

leverage these tools that have been tested and validated for AER projects and identify how best to expand and adapt them to meet EESSD/MODEX needs and workflows to reduce the significant burden data discovery and transformation currently weighs on AER and BER AI/ML analytics and research. We anticipate significant opportunities to enhance existing collaborations between NETL and BER entities, particularly in relation to addressing computing enhancements to further mature the scalability and flexibility of these tools which requires significant computing power. Thus, collaborations, such as NETL's ongoing discussions with BNL's National Nuclear Data Center researchers around NETL's tools and BNL's computing expertise and capabilities, would be leveraged to benefit EESSD Earth Science predictability needs, by more efficiently finding and transforming data at scale to feed AI/ML analytical and modeling.

As discussed below, NETL has HPC expertise but is also working with next-gen AI hardware accelerator such as Cerebras. Combining these competencies with BER exascale computing and expertise, could support the goals for data acquisition and assimilation at scale.

NETL and Cerebras executed a research agreement in 2020 to begin developing a cognitive in-the-loop simulation capability on the new CS1 hardware platform. The CS1 is the first hardware device that can be thought of as a single chip supercomputer. It consists of the world's largest silicon device with 400,000 cores and a multi-petaflop peak performance. NETL and Cerebras demonstrated the world's fastest linear solver for computational fluid dynamics on this hardware. The solver was found to be over 200 times faster than current generation supercomputers can be at scales up to half a billion cells. While this is certainly impressive, it is not yet a practical tool for scientific simulations. NETL and Cerebras are developing a set of libraries to enable scientific computing on this platform. This work will have a significant impact on all scientific modeling and provide an incredible boost in speed and productivity and can be leveraged by the EESSD. The CS1 is ideal for hybrid AI/HPC applications as it can do both equally well. This has benefits for AI assisted scientific modeling because there is no need to move data on inference and slow down the simulation progress. In addition, the same hardware can be used for training AI models, even while simulations are running. Further, the CS1 is only the first generation of this new single chip supercomputer, and the capabilities will expand and grow in the near term.

These novel approaches to data discovery, exploration, and assimilation will provide key AI/ML driven data infrastructure needed to address Earth system research at a more integrated, and efficient scale. For example, data acquisition and assimilation is needed to support rapid, distributed analytics and modeling for groundwater, fluvial, coastal and offshore systems; examine linkages across the highly coupled energy, water, and land systems, to help assess co-evolution in relation to both short- and long-term perturbations. Authoritative, timely data and simulation capabilities are also required to support modeling of these complex systems, and their interactions and how they evolve/respond to various perturbations. This includes a focus on coastal and offshore systems, due to increase in extreme weather events and their effect on the terrestrial-aquatic interface.

These are complex, multivariate, multi-source, multi-scale Earth systems that require a holistic acquisition and assimilation of data and knowledge. Incorporating the SmartSearch, SmartParse at-scale AI/ML approaches with Cerebras and other advanced computing approaches described above would significantly aid in rapid assimilation and initial processing of data. A few examples of how we envision this benefitting EESSD: i) SmartSearch for data assimilation (they suggest non-linear data assimilation, and I'd agree with proposing this if we can, as well as support for distributed data management/repositories); ii) SmartParse to help amass additional data and contextual insight on data sets to make recommendation as towards data use, quality, etc.); iii) Relate data and information to help researchers quickly identify the information needed from distributed big-data volumes, iv) Generate critical information to help researchers determine high-level trends and patterns in data (including spatial and temporal patterns, similarities); and v) During assimilation and aggregation use novel AI/ML scripts to determine data quality and recommend applications/usability of data for models

Suggested Partners/Experts (Optional):

NETL has contracting and collaborative relationships with a wide array of government research, commercial, academic, and non-governmental research entities that could be engaged in support of the workshop and its goals. Below are relationships we highlight as strategically aligned to the workshop, whitepaper and AI4ESP goals in general:

- DOE's applied energy research laboratories, NETL, NREL and INL have an MOU and existing collaborative R&D efforts that could be leveraged to engage not just NETL personnel in this workshop, but key collaborators from NREL and INL including:
    - INL's Digital Innovation Center of Excellence (DICE)
        - **Potential speaker**: Chris Ritter, DICE Director, INL
    - NETL's Science-based AI/ML Institute (SAMI)
        - **Potential speaker:** D. Vic Baker, computational researcher and lead developer of SmartSearch/SmartParse, MATRIC
    - NETL's Energy Data eXchange® (EDX)
    - NETL's Geologic and Environmental Sciences (GES) R&D Programs, NRAP, SMART, Carbon Storage, Onshore and Offshore
        - **Potential speaker:** Grant Bromhal, Senior Fellow for GES at NETL
    - NETL-Cerebras
        - **Potential speaker**: Dirk VanEssendelft, Energy and Geo-Environmental Engineering and Computational Scientist, and Cerebras lead at NETL
    - NREL's OpenEI data curation and virtualization platform & team
        - **Potential speaker**: Debbie Brodt-Giles, Data Group Manager, NREL
- Brookhaven National Laboratory's, National Nuclear Data Center – BNL's NNDC and NETL have ongoing discussions regarding advanced, AI/ML search and parsing capabilities, and complementary efforts between the two entities, including BNL computing engineering competencies.
    - **Potential speaker:**
        - Research physicist **Adam Hayes**, BNL's National Nuclear Data Center (NNDC) or,
        - Director of the NNDC, **Alejandro Sonzogni**, BNL's National Nuclear Data Center (NNDC)

NETL has active collaborations or conversations related to the energy-water data acquisition and assimilation goals with many researchers and labs associated with EESSD's mission and efforts, including ANL, BNL, LANL, LBL, LLNL ORNL, PNNL, SNL. For the workshop, key personnel aligned to ESSDD's MODEX may be engaged to support the workshop and AI4ESP goals from these institutions.

Beyond these relationships for energy-water R&D specifically, NETL has a strong and important suite of relationships with DOE EIA, EM, LM, OCIO, OSTI, OTT, Chief Counsel, AITO in relation to data management, curation and best practices to ensuring FAIR data practices are compliant with DOE policies and procedures, while supportive and aligned to the mission and goas of DOE AER science and research elements. Most recently, NETL's EDX was named as the agency's priority geospatial data repository by the OCIO's Geospatial PMO. While EDX is the first, there is the expectation that other key data repositories from EIA, OpenEI, and EESSD will soon be nominated for addition to this more integrated, findable, and vetted cohort of data resources to support the Department's goals and mission.

In addition to key relationships and collaborations from within the DOE itself, NETL has key partnerships and collaborations with other federal agencies key to energy-water systems. These include the Department of Defense (DOD) (USCG, NRL, NGA, DITRA, etc), Department of Interior (DOI) (USGS, BSEE, BOEM), and Department of Commerce (DOC) (EPA, NOAA, etc). Suggested speakers below would support workshop goals in identifying authoritative

energy-water-Earth system data sources and gaps that the "smart" data fabric approach described above could address.

- U.S. Geological Survey (USGS) – NETL has two Memorandums of Agreement (MOAs) between the USGS and NETL, one focused on energy water systems specifically, and one related to data and sample sharing agreement as it relates to rare earth element and critical mineral systems.
    - **Potential speakers**:
        - USGS, Timothy D Oden, Associate Director for Data Programs, Pennsylvania Water Science Center
        - USGS, Warren Day, Earth MRI Science Coordinator/Research Geologist, Mineral Resources Program
- Offshore Analysis of Seafloor Instability and Sediments (OASIS) is a multidisciplinary, multi-agency collaboration applying state-of-the-art science to study the timing and location of Mississippi River Delta Front seafloor instability. Participants include BOEM, BSEE, NETL, NOAA, USGS, Naval Research Lab, Naval Oceanographic Office, and the National Geospatial Intelligence Agency. This is a metocean-centric effort that NETL is a partner within, and may afford opportunities to engage DOD, DOC, and additional DOI data management and Earth system experts for the workshop.

References (Optional)

Baker, D.V., Kelly, R., Bauer, J., Rager, D., 2016. Computational Advances and Data Analytics to Reduce Subsurface Uncertainty. Presented at the 50th U.S. Rock Mechanics/Geomechanics Symposium, American Rock Mechanics Association.

Bauer, J. R., and Rose, K., 2015, Variable Grid Method: an Intuitive Approach for Simultaneously Quantifying and Visualizing Spatial Data and Uncertainty, *Transactions in GIS*. 19(3), p. 377-397. https://doi.org/10.1111/tgis.12158

Gibney, Elizabeth, and Richard Van Noorden. "Scientists losing data at a rapid rate." *Nature News* (2013).

Mons, Barend. "Invest 5% of research funds in ensuring data are reusable." *Nature* 578.7796 (2020): 491.

Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P., Greenburg, R., Walker, S., Yeates, D., Rose, K. *in review.* Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences***.**

Rose, K., Bauer, J., Baker, V., Bean, A., DiGiulio, J., Jones, K., Justman, D., Miller, R.M., Romeo, L., Sabbatino, M., Tong, A., 2018. Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase. https://doi.org/10.18141/1427573

Wenzlick, M., Bauer, J.R., Rose, K., Hawk, J., Devanathan, R., 2020. Data Assessment Method to Support the Development of Creep-Resistant Alloys. Integr Mater Manuf Innov 9, 89–102. https://doi.org/10.1007/s40192-020-00167-3

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.