AI-Driven Cross-Domain Knowledge Discovery and Hypotheses Generation for Enhanced Earth System Predictability

Authors

Svitlana Volkova, Chief Scientist, National Security Directorate, PNNL
Nathan Hodas, Chief Scientist, National Security Directorate, PNNL
Tim Scheibe, Lab Fellow, Earth & Biological Sciences Directorate, PNNL

Focal Area(s)

1. Data acquisition, multimodal data fusion and data-to-knowledge and knowledge-to-code transformations enabled by scaling AI and unsupervised deep learning.
2. Insights (cross-domain knowledge) extracted from complex data (both observational and simulated) using AI-driven approaches and large-scale data analytics.

Science Challenge

Human knowledge about Earth systems is extremely fragmented and incomplete. Therefore, when scientists, who are the domain experts, encode their incomplete fragmented knowledge into the Earth System Models (ESMs), these models have significant uncertainty and their predictive power is both limited and poorly understood (NASEM, 2020). Moreover, process-based models inherently embody our biases – and those are difficult to identify given the high level of other sources of uncertainty (e.g. parametric). These limitations are especially severe for hydrologic models of Earth's water cycle extremes and their impacts (Nearing et al., 2021). Data-driven knowledge acquisition could help to identify and reduce these model biases (structural errors) thus improving the model predictability. Earth scientists have abundant physics-based experimental data (both observational and simulated), but they do not yet know how to discover and extract predictive knowledge from these large heterogenous dynamic data streams. They also do not know how to fully encode cross-domain knowledge e.g., about multi-scale biological and geochemical processes into the model. Thus, we need fundamentally new ways to discover new cross-domain knowledge looking across disciplines and going beyond physics-based data (both observational and simulated) to enable AI-driven hypotheses generation and optimization of experimental design for ESMs, which will in turn improve ESM predictability (Abeliuk et al., 2020).

Rationale

Recently emerged AI techniques in representation learning, predictive modeling and domain-aware AI can help addressing scientific gaps in ESM predictability. These include but are not limited to:

- **Modeling algorithms:** Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and physics-based ML to improve ESM predictability.
- **Data to knowledge transformation (aka representation learning):** Transformer models (Vaswani et al., 2017), Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), and Convolutional Neural Networks (Szegedy et al., 2015) to learn data representations and autoencoders (Vincent et al., 2010) to learn observational and simulated multi-scale representations.

However, the above techniques are necessary but not sufficient for a paradigm shift needed to improve ESM predictability and reduce uncertainty. Instead, operating directly in the multi-disciplinary **knowledge space** and the software **code space** in addition to observational and simulated **data space** (where ESMs are operating now) and focusing on three S&T challenges below will allow to effectively bridge observational designs and useful predictions (Gleick, 2011).

DOI: https://doi.org/10.2172/1769670
DOE BER, AI4ESP 2021 #AI4ESP1139 (https://www.ai4esp.org)

AI-Driven Cross-Domain Knowledge Discovery and Hypotheses Generation for Enhanced Earth System Predictability

[1] How to extract "predictive" (aka useful) knowledge from multimodal observational and simulated data across scales?
[2] How to discover new multidisciplinary knowledge (in addition to physics-based observational and simulated data) from open source data e.g., scholarly outputs – publications, patents, Wikipedia, code bases etc.?
[3] How to encode this cross-domain "comprehensive" knowledge into ESMs to improve their predictive accuracy?

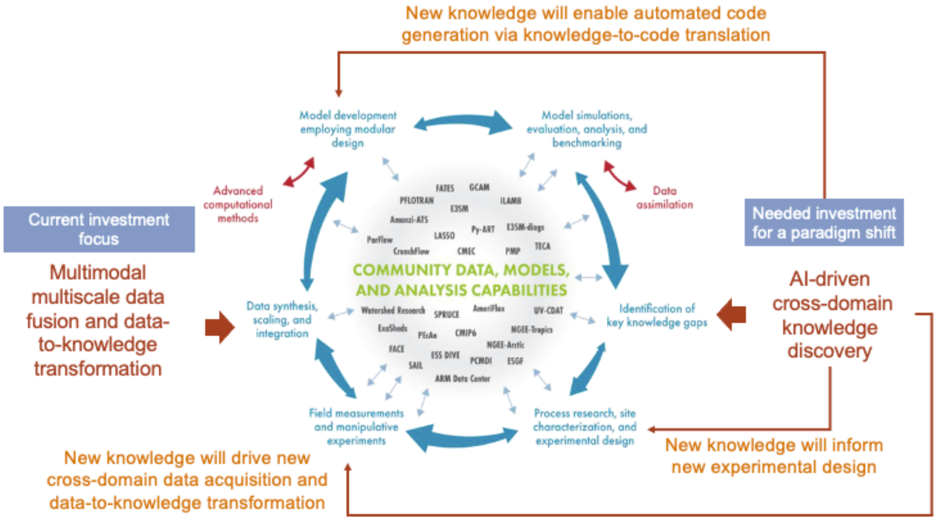
New knowledge will enable automated code generation via knowledge-to-code translation
Model development employing modular design
Model simulations, evaluation, analysis, and benchmarking
Advanced computational methods
Data assimilation
Current investment focus
Multimodal multiscale data fusion and data-to-knowledge transformation
Needed investment for a paradigm shift
AI-driven cross-domain knowledge discovery
FATES
GCAM
PFLOTRAN
E3SM
ILAMB
Amanzi-ATS
Py-ART
E3SM-diags
Parflow
LASSO
CrunchFlow
CMEC
PMP
TECA
COMMUNITY DATA, MODELS, AND ANALYSIS CAPABILITIES
Watershed Research
SPRUCE
AmeriFlux
UV-CDAT
ExaSheds
PEcAn
CMIP6
NGEE-Tropics
FACE
NGEE-Arctic
ESS DIVE
PCMDI
ESGF
SAIL
ARM Data Center
Data synthesis, scaling, and integration
Identification of key knowledge gaps
Field measurements and manipulative experiments
Process research, site characterization, and experimental design
New knowledge will drive new cross-domain data acquisition and data-to-knowledge transformation
New knowledge will inform new experimental design

Figure 1. Our holistic AI-driven scientific discovery approach and its benefits to improve the MODEX approach with various DOE data, models and ana analysis capabilities.

As shown in Figure 1, our holistic approach will create artificial intelligence to identify current key knowledge gaps in ESMs and generate hypotheses about new knowledge and experimental designs, drive additional cross-disciplinary data collection enable data and knowledge fusion, and knowledge to code "translation" (inspired by recent advances in large-scale language modeling[1]).

Narrative

We propose a transformational AI-driven approach for cross-domain knowledge discovery[2] and hypothesis generation[3] (e.g., about the effect of atmosphere, ocean, land, ice, and biosphere on climate change) to improve predictability of ESMs (Kuhn, 2012). Similar transformational approaches for data-driven knowledge discovery and AI-driven hypothesis generation have been used for biology, material science and chemistry to enable new drug discovery (Mak and Pichika, 2019), new material discovery (Tshitoyan et al., 2019) and automation of chemical experimentation (Burger et al., 2020).
Our holistic approach (Figure 2) will use a novel neural network architecture and recent successes in scaling AI to (1) learn additional cross-domain knowledge about biological and chemical processes from open knowledge bases (including but not limited to scientific literature, patents, Wikipedia, Google datasets[4] etc.), (2) combine this knowledge with observational and simulated data (e.g., from ARM,

[1] https://analyticsindiamag.com/open-ai-gpt-3-code-generator-app-building/
[2] https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf?la=en-GB&hash=5240F21B56364A00053538A0BC29FF5F
[3] https://www.weforum.org/agenda/2020/11/scientific-discovery-must-be-redefined-quantum-and-ai-can-help/
[4] https://datasetsearch.research.google.com/

CMIPs), and code developed to run ESMs using joint representations (aka embeddings), and (3) recommend additional data to collect, experiments to run and generate code to improve ESMs.

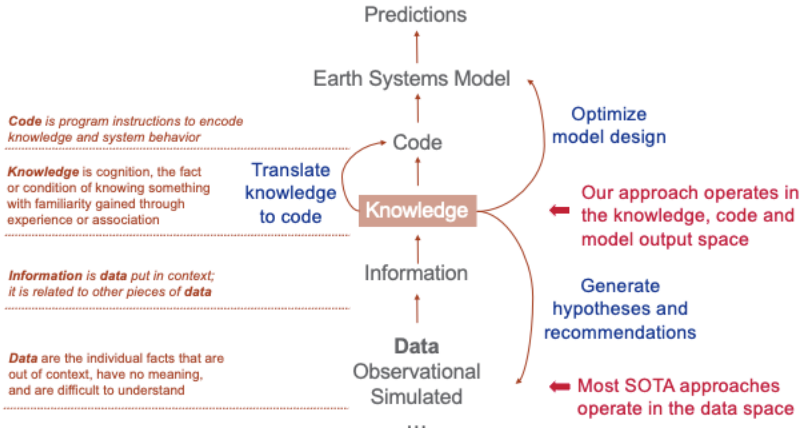
Predictions
Earth Systems Model
Code
Knowledge
Information
Data
Observational
Simulated
...
Code is program instructions to encode knowledge and system behavior
Knowledge is cognition, the fact or condition of knowing something with familiarity gained through experience or association
Information is data put in context; it is related to other pieces of data
Data are the individual facts that are out of context, have no meaning, and are difficult to understand
Translate knowledge to code
Optimize model design
Our approach operates in the knowledge, code and model output space
Generate hypotheses and recommendations
Most SOTA approaches operate in the data space

Figure 2. The differences between data vs. information vs. knowledge – inputs to ESMs.

More specifically, our approach will focus on addressing three S&T gaps outlined above:

[1] How to extract "predictive" (aka useful) knowledge from multimodal observational and simulated data across scales? ➔ How: AI modeling algorithms e.g., transformer models, LSTMs, GCNs, GANs etc.
Inputs: measurement data from ARM and archived model outputs (e.g. CMIPs)
Outputs: useful knowledge representations (aka embeddings).

[2] How to discover new cross-domain knowledge (in addition to physics-based understanding encoded into ESMs) from open source data e.g., scholarly outputs (publications, patents, Wikipedia, code bases)? ➔ How: AI-driven knowledge discovery and hypotheses generation.
Inputs: human-generated knowledge from open data sources.
Outputs: additional cross-domain knowledge (aka joint embeddings), actionable hypotheses, optimized experimental design and knowledge-to-code translation.

[3] How to learn "comprehensive" knowledge representations including code developed to encode ESMs and automatically translate knowledge-to-code? ➔ How: AI-enabled representation learning e.g., large-scale language models, autoencoders and domain-aware ML methods.
Inputs: Observational and simulated data, cross-domain knowledge and code
Outputs: Cross-domain knowledge representations and recommendations to acquire new data, optimize experimental designs and regimes, and generate code (generative models).

Data-driven knowledge acquisition and fusion will help identify and reduce current ESM shortcomings and limitations (e.g., biases and structural errors), thus improving ESM predictability. Moreover, incorporating cross-domain knowledge that encodes interactions among the physical climate, the biosphere, and the chemical constituents of the atmosphere and ocean into ESMs in a novel previously unexplored way – via hypotheses generation, optimization of experimental design and knowledge-to-code translation – will drive a paradigm shift and lead to a cross-disciplinary research framework of the future.

AI-Driven Cross-Domain Knowledge Discovery and Hypotheses Generation for Enhanced Earth System Predictability

Suggested Partners/Experts (Optional)

PNNL has extensive expertise and publication track record in developing novel AI-driven approaches to predict real-word system behavior from multimodal heterogenous open data streams (Volkova et al., 2011; Glenski et al., 2018; Shrestha et al., 2019) and perform causal knowledge discovery from observational data (Saldanha et al., 2020).

Academic partners: Dr. Yulia Tsvetkov (CMU) will bring expertise in Natural Language Processing, Dr. William Wang (UCSB) will bring expertise in deep learning and knowledge graph reasoning, Dr. Andreas Zufle (George Mason University) will bring expertise in data mining, simulation and prescriptive analytics.

Industry partners: Emre Kiciman (Microsoft Research) will bring expertise in causal discovery and inference, Hannaneh Hajishirzi (AI2, University of Washington will bring expertise in multimodal representation learning and deep learning.

References (Optional)

1. Abeliuk, A., Huang, Z., Ferrara, E. and Lerman, K., 2020. Predictability limit of partially observed systems Scientific reports, 10(1), pp.1-10.
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
3. Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R. and Rankin, N., 2020. A mobile robotic chemist. Nature, 583(7815), pp.237-241.
4. Saldanha E., Cosbey R., E. Ayton, M. Glenski, J. Cottam, K. Shivaram, B. Jefferson, B. Hutchinson, D. Arendt, S. Volkova. Evaluation of Algorithm Selection and Ensemble Methods for Causal Discovery. NeurIPS Workshop on Causal Discovery and Causality-Inspired Machine Learning 2020.
5. Gleick, J., 2011. Chaos: Making a new science. Open Road Media.
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661.
7. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.
8. Kuhn, T.S., 2012. The structure of scientific revolutions. University of Chicago press.
9. Glenski M., Weninger T., and Volkova S. Improved Forecasting of Cryptocurrency Price using Social Signals. ArXiv. Featured by Bloomberg. 2018.
10. Mak, K.K. and Pichika, M.R., 2019. Artificial intelligence in drug development: present status and future prospects. Drug discovery today, 24(3), pp.773-780.
11. NASEM (National Academies of Sciences, Engineering, and Medicine), 2020. Earth System Predictability Research and Development: Proceedings of a Workshop in Brief. Washington, DC: The National Academies Press.
12. Nearing, G.S., F. Kratzert, A.K. Sampson, C.S.Pelissier, D. Klotz, J.M. Frame, C. Prieto, and H.V.Gupta, 2021. What role does hydrological science play in the age of machine learning? Water Resources Research, online publication in press, doi: 10.1029/2020WR028091

13. P. Shrestha, S. Maharjan, D. Arendt, and S. Volkova. Learning from Dynamic User Interaction Graphs to Forecast Diverse Social Behavior. CIKM'19.
14. S. Volkova, E. Ayton, K. Porterfield, and C. Corley. Forecasting Influenza-like Illness Dynamics for Military Populations using Neural Networks and Social Media. PLoS ONE 10(9). 2017.
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
16. Tshitoyan, Vahe, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571.7763 (2019): 95-98.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.
18. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A. and Bottou, L., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(12).