Semi-automated Design of Artificial Intelligence Earth Systems Models

Philipe Dias[1,2], Henry Medeiros[1], Dalton Lunga[2], Nagendra Singh[2], and Ranjeet Devarakonda[2]

[1]Marquette University, Milwaukee, WI, USA; [2]Oak Ridge National Laboratory, Oak Ridge, TN, USA

Focal areas

(1) Predictive modeling using AI methods. **(2)** Extracting insights from complex data using AI.

Scientific challenge

Prediction and observation of water cycles at various scales involve not only patterns isolated in space and time, but also modeling of complex spatio-temporal relationships across multiple domains. For instance, evapotranspiration (ET) and leaf area indexes (LAI) are two parameters that are needed to accurately model and understand land-atmosphere processes. Accurate assessments of ET and LAI are critical for understanding hydrological processes, deforestation, crop yield, and irrigation impacts. However, ET estimates for global simulations are available at very coarse spatial resolution. They are usually derived from satellite data based on broad plant functional types (PFTs), which fail to capture fine-scale variations because of changes in vegetation type across the globe. Similarly LAI estimates have typically been derived from vegetation indices at global scales or estimated locally using physical models, both of which suffer from a range of uncertainties [1] that impact model sensitivity. The new era of AI model development for Earth systems (ES) calls for data-driven methods that provide domain scientists with uncertainty-aware estimations of biophysical parameters such as ET and LAI in a generalizable, interpretable, and discoverable manner.

Rationale

As climate change continues to advance and the probability of extreme events increases, relationships among humans, water cycles, and their environmental and socioeconomic impacts become increasingly volatile. In tandem with the quest to understand fundamental drivers of the water cycle, a deluge of data related to the ES is swiftly accumulating, presenting invaluable potential for scientific breakthroughs.

We present potential key research directions for AI in ES: (1) estimate LAI at tree-species level to contribute finer estimates of PFT by leveraging high-resolution satellite imagery and (2) uncover relevant features for ET analysis, as well as promote data fusion from different sources and locations. Such strategies would increase accuracy and reduce uncertainty for ET and LAI estimations, yet the extraction of actionable information and insights from these vast and heterogeneous data sources is marked by the following challenges.

First, the large *volume* of data requires *autonomous or semi-autonomous approaches for sample selection, annotation and processing*. Since traditional machine learning (ML) approaches rely on extensive human labor for these tasks, it is impractical to scale such approaches for large and complex datasets. This issue is particularly pertinent to ES domains, where many relevant patterns can only be reliably identified by experts.

Although some relationships between subsystems governing Earth system models (ESMs) are known by domain scientists, many other potential connections require automation. For instance, an observation-based study over the West Sahel region [2] identified information transfer between

DOI: https://doi.org/10.2172/1769777
DOE BER, AI4ESP 2021 #AI4ESP1034 (https://www.ai4esp.org)

precipitation and LAI, but highlighted monsoons, dust, and greenhouse gases as other potentially important interactions for investigation. Combined with a growing *variety* of the available data—stemming from resolution, dimensionality, domain, and quality—a major challenge thus remains to *develop data-driven autonomous systems that effectively explore hypotheses spaces at a global scale to uncover relevant features and relationships*.

In addition to requiring large amounts of accurately annotated data, existing AI models provide limited *interpretability*, which compromises reliability assessments, as well as opportunities for gathering novel insights. Given that physical understanding of models is often central for scientific discovery, model predictions need to be related to domain knowledge for possibly uncovering previously unknown relationships.

Moreover, predictions generated by such models are often *poorly calibrated and lack explicit measures of uncertainty*. Quantification of a model's uncertainty is fundamental to many aspects of scientific data analysis, including its combination with prior knowledge and additional models, and informs a model's potential performance in new domains. In particular, *the need for domain adaptation* within ES applications is motivated by the fact that water cycle and remote-sensing observation techniques often rely on specialized models that fail to generalize to domains with no ground-truth data.

Finally, there is a need to *simplify the process of design and adjustment of model architectures for new applications and specific domains*. Currently, this is commonly done by means of expert-guided trial and error which, combined with the limited standardization of available frameworks, imposes significant learning curves for either domain scientists to specialize in ML techniques, or ML scientists to acquire domain-specific knowledge. While the Earth science domain provides ample opportunities for model sharing among related tasks, effective models must not require large amounts of data and computation for every task; like humans, they should instead adapt to novel tasks by building upon prior knowledge.

Narrative

As a new research agenda, we advocate for the design of frameworks that are interpretable, are generalizable, and reduce the burden on domain scientists by means of data-driven semi-automated approaches for (1) data selection and annotation, (2) exploration and optimization of model architectures, (3) quantification of different uncertainty sources, and (4) enabling of task-driven model discoverability. These challenges encapsulate key data-model integration research goals highlighted in the EESSD Strategic Plan, and we argue that recent advances in the topics of active learning, domain adaptation, neural architecture search, and uncertainty quantification represent promising and realistic avenues toward these goals.

Data selection and annotation: The availability of high-quality ground truth is particularly limited in Earth science domains. For instance, the FLUXNET database provides robust ground-truth for many ESM parameters, but acquisition sites are limited and sparsely distributed. In other domains, efforts to minimize costs of data selection and annotation include the design of better sample selection strategies [3] and semi-automated annotation tools that exploit stochastic simulations [4]. Similarly, for Earth science data, there is greater potential to design semi-automated spatio-temporal based interpolation techniques for simplified ground-truth acquisition. Such a capability could leverage relationships and similarities in terms of land cover shared among neighboring regions, as well as combining multiple data sources with varying spatial, spectral and/or temporal resolutions.

Uncertainty quantification: Parametric uncertainty is a large source of predictive uncertainty in ESM and therefore is of critical importance for understanding controlling processes, as well as guiding model development and data acquisition [5]. For these models, parametric uncertainty is most commonly estimated through sensitivity analysis (SA), with studies of land surface models demonstrating that common subsets of sensitive parameters are shared among PFTs [5]. While uncertainty estimation and calibration are still open problems within the AI community, some recent work represents potential avenues to replace SA needs with uncertainty-aware AI models. This includes models that learn to distinguish between model and data uncertainty [6, 7], novel formulations that exploit the perspective of evidence accumulation [8, 9], and approaches that place priors over output probabilities to learn distributions rather than point estimates [10].

Model exploration and adaptation: In addition to the multimodal, multiangular, and multisensor nature of remote sensing data, many Earth systems processes, including ET and LAI, are themselves heterogeneous in space and time. This together with the limited availability of representative ground-truth impedes the generalizability of models. Thus, efforts to seek autonomous identification of relevant features, model architecture adjustment, and exploration of hypotheses spaces while leveraging unsupervised domain adaptation, will be critical. Emerging AI advances, including stochastic neural architecture search, [11] are effective at end-to-end simultaneous learning of neural operation parameters and architecture distribution parameters and could advance ES applications. In addition, techniques for the explicit quantification of distributional uncertainty [10] and domain separation networks [12] are emerging as techniques for explicitly modeling shared domain characteristics; therefore, they are positioned to extract domain-invariant features for relevant phenomenon.

Task-driven model discoverability: Existing AI formulations for ES applications are designed with limited scope for model reuse. This is demonstrated from NN algorithms that have shown remarkable progress on single task use-cases but poor generalization when confronted with a large and diverse set of tasks. We envision flexible frameworks that enable domain scientists to share and access repositories of models defined on previously collected data. To achieve this, learning models are not to require large amounts of data or compute for every task but rather, like humans, should be able to rapidly adapt to novel tasks by building upon experience from related tasks. Key to such a framework are methods that include task-based metadata in a standard format to facilitate their discovery and retrieval. Leveraging insights derived from model uncertainty estimates, new model descriptors could be designed and linked to new tasks to perform a constrained search on existing model repositories. Domain scientists could discover the relevance of existing models to new tasks, thereby accelerating the pace of scientific discoveries.

Allied with DOE's unique exascale computing facilities, such techniques offer opportunities for harnessing the available computational resources to relieve the burden typically imposed on domain experts for the design of AI systems. We envision an era where domain scientists will collaborate with machines to learn complex relationships among data sources while advancing their domain-specific knowledge.

References

[1] H. Fang, S. Wei, C. Jiang, and K. Scipal, "Theoretical uncertainty analysis of global MODIS, CYCLOPES, and GLOBCARBON LAI products using a triple collocation method," *Remote Sensing of Environment*, vol. 124, pp. 610–621, 2012.

[2] B. Y. Liu, Q. Zhu, W. J. Riley, L. Zhao, H. Ma, M. Van Gordon, and L. Larsen, "Using information theory to evaluate directional precipitation interactions over the west sahel region in observations and models," *Journal of Geophysical Research: Atmospheres*, vol. 124, no. 3, pp. 1463–1473, 2019.

[3] A. Siddhant and Z. C. Lipton, "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[4] P. A. Dias, Z. Shen, A. Tabb, and H. Medeiros, "Freelabel: A publicly available anno- tation tool based on freehand traces," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[5] D. Ricciuto, K. Sargsyan, and P. Thornton, "The impact of parametric uncertainties on biogeochemistry in the e3sm land model," *Journal of Advances in Modeling Earth Systems*, vol. 10, no. 2, pp. 297–319, 2018.

[6] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[7] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *arXiv preprint arXiv:2002.08791*, 2020.

[8] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, 2018.

[9] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *arXiv preprint arXiv:1910.02600*, 2019.

[10] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018.

[11] S. Xie, H. Zheng, C. Liu, and L. Lin, "SNAS: stochastic neural architecture search," *arXiv preprint arXiv:1812.09926*, 2018.

[12] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in neural information processing systems*, 2016.