Rapid assimilation and analysis of a suit of remote sensing data for predicting extreme events and their impact on ecological-human systems

Authors. N. Falco[1], B. J. Enquist[2], H. Wainwright[1], E. Anagnostou[3], M. Longo[4], X. Shen[3], E. Nikolopoulos[5], C. Hinojo[2]

[1]Lawrence Berkeley National Laboratory, [2]University of Arizona, [3]University of Connecticut, [4]Jet Propulsion Laboratory, [5]Florida Tech

Focal Areas. Insight gleaned from complex data using AI, and other advanced methods, including explainable AI and physics- or knowledge-guided AI. Predictive modeling through the use of AI techniques and AI-derived model components; the use of AI and other tools to design a prediction system comprising of a hierarchy of models.

Science Challenge. Ongoing and future remote sensing (RS) missions are expected to provide an unprecedented amount of data. The hope is that these data will improve our predictive understanding of the Earth System functioning and its response to extreme climate events. This 'data flood', however, poses several challenges: (1) how can we quickly synthesize the volume of data and identify emergent behavior of the ecosystems, particularly under extreme events; and (2) how can this knowledge be effectively incorporated in Earth system models (ESMs) and advance both applied and theoretical research? We envision the next-generation of ESMs will be deeply integrated with theory-informed RS-systems. However, we argue, that the most rapid pathway to get there is to develop a heuristic method based on theory/model-informed RS-based estimation will more rapidly advance scientific discoveries via AI capabilities. Such an approach will be able to identify how to more rapidly assimilate and analyze RS data. Developing such capabilities will be critical in capturing the key drivers of biosphere and ecosystem change - extreme events (e.g., storms, droughts) and associated hazards (e.g., fires, landslides, floods) - and quantifying their impact on both natural (biodiversity, environment) and human (infrastructure, energy, agriculture) systems.

Rationale. Biodiversity and the functioning of ecosystems are the key component of the Earth Systems, playing a central role in water, carbon and nutrient cycling (Lade et al., 2020, Norris et al., 2020, Ruckelshaus et al., 2020, Carpenter et al., 2009). The consequences of hydrological extremes – such as droughts and flooding – are measured by their impact on ecosystems such as permanent plant die-off events or species alternation (AghaKouchak et al., 2020, Trotsiuk et al., 2020, Steel et al., 2019, Ruthrof et al., 2018). To represent such events and their ecosystem impacts, ESMs rely on parametrization often supported by intensive spatially-limited in-situ observations. Models such as the Functionally Assembled Terrestrial Ecosystem Simulator (FATES), can forecast environmental variables (e.g., carbon/water fluxes, forest structure) from simulating plant dynamics of a single tree or plant communities. The outcome of models' prediction is, however, strictly connected to the quality, resolution, and scale, of the input parameters as well as the model parametrization. Computationally intensive models need to practice a trade-off between model complexity and spatio-temporal resolutions, making the identification of extreme patterns difficult to observe, and consequently producing high uncertainty responses to changes in ecosystem functioning when extreme events occur.

In recent years, RS capabilities for Earth Observations have increased in such a way that we are able to collect snapshots anywhere on our planet at multiple spatial and temporal scales and resolutions. The use of RS provides the opportunity to monitor and capture dynamics that are in general difficult to monitor and/or easy to miss. However, the challenge is to rapidly discover new patterns and processes in the Earth System but yet at the same time tame the data deluge from an increasing number of RS data sources. Several machine learning (ML) approaches, from the traditional supervised-learning to the more recent deep-learning networks, have been successfully applied to RS data to predict plant

DOI: https://doi.org/10.2172/1769770
DOE BER, AI4ESP 2021 #AI4ESP1040 (https://www.ai4esp.org)

community distributions, plant/leaf traits, vegetation structure, all which can be considered as important biodiversity aspects that need to be monitored to identify changes in ecosystem function. Although ML is a powerful tool, it is often used as a purely data-driven black box, where lack of domain knowledge affects both scalability and generalization capabilities. Learning complex scale-dependent relationships requires new theory-informed ML strategies able to not only identify correlations but also causality, allowing to cope with the gap between site-specific model-derived quantities and spatially extended estimations based on RS observations, while allowing in-situ data scarce regions (or problem) to leverage the knowledge from in-situ data rich regions (or problem).

Narrative. The science needed to understand and mitigate the impacts of global change on the biosphere will require both unprecedented access to diverse biological and environmental data across space, time, and scales and the synthesis and development of predictive theory (Dietze et al., 2018, Bush et al., 2017, Hampton et al., 2013). In this white paper, we argue that while environmental data from RS have been accumulating at a rapid pace, their broad scope generates major challenges for finding effective ways to discover, access, integrate, curate, and analyze the range and volume of relevant information. Second, to generalize and improve forecasts, there is an urgent need to harness big data and data synthesis with the vision and foresight of analytical and quantitative theory. We identify the key ML/AI capabilities to further enhance the predictability of ecosystem models: (1) enhanced connectivity from RS to model parameterization, (2) theory/model-informed RS-based estimation.

Enhanced connectivity from RS to model parameterization. RS data together with data collected from other resources such as drones, radar, weather stations, geographical information, and hydrology information raise a unique challenge for ML to deal with multi-view and multi-modal data and predict/assess extreme events. With multi-view and multi-modal data, an ML pipeline to align, fuse different views and different modalities of data is in great need to perform co-learning and decision making. For example*:*

If some observations are missing in a certain modality of data, by aligning and co-learning it with other modalities of data, we can mitigate this issue and still provide a reliable prediction/decision. Further, large high-resolution RS-based datasets used as input to ML algorithms can be used to enhance parameterizations of unresolved processes in current ESMs at high spatial and temporal scales.

Resolving convection in regional/global atmospheric models using ML and high-resolution datasets could lead to improvements in extreme event simulations. Furthermore, ML can relate extreme events and hazards with impacts such as agriculture, infrastructure, energy, transportation, etc. (Lazin et al., 2021, Cerrai et al., 2019, Nikolopoulos et al., 2018).

In addition, *ML can be used to forecast future conditions*. ML and deep learning methods (e.g., fully convolutional networks) can be trained to capture the regular spatial (in different horizontal and vertical resolutions) and temporal patterns in RS-related datasets to predict normal future situations and detect extreme events when the actual future scenario deviates from the prediction (Zhang et al., 2019).

Theory/model-informed RS-based estimation. In recent years, ML techniques (especially deep learning) have achieved great success in various domains, e.g., computer vision, speech recognition, and natural language processing. However, there are several substantial barriers for directly applying ML methods to RS data to facilitate the understanding of the Earth System functioning and its response to extreme events. First, most existing ML-based systems are data-driven. In Earth systems, however, domain knowledge in the forms of rules, theories, or physical constraints play an important role. For instance, it

has been shown that by combining ML with mechanistic theory applied to RS data it can be made more effective and provide insight into physiological and physical mechanisms.

Here we advocate that an important methodological development in application of ML methods is to examine and contrast results from theory-informed approaches with those that are theory-free. Theory informed choices of feature space hold the promise of greatly improving the convergence time, accuracy and inference of ML algorithms. This approach can inform our understanding, or lack thereof, of the underlying biological and physical processes, and potentially elucidate causal relationships. We point to three specific examples.

First, ML methods can be used to more rapidly identify mechanisms and advance theory by applying ML to theoretically informed feature spaces to leverage all available information and technology (Brummer et al., 2021). To search for patterns, ML algorithms are often applied to the full set of untransformed, standardized raw data. This is done because (i) in the absence of a prior theory, it is the most straightforward approach; and (ii) some practitioners of ML prefer to have a model- or theory-agnostic method arguably free of bias. For example, given a query RS observation, if we retrieve similar patterns in the historical observations (assuming most historical data are normal) using ML, then the query RS observation will also be determined as normal and we can use the retrieved patterns to interpret the current situation (e.g., geographic, hydrological, weather, biological etc.). If we cannot identify a similar pattern, then the query RS observation could be an extreme event we never encountered and need special attention. However, also applying ML to theoretically informed feature spaces can better inform mechanism and advance theory.

Second, examining and contrasting results from theory-informed approaches with those that are theory-free enables for "diagnosis" of impacts of extreme events. Assuming we have a database of extreme events, with the methods described above, after identifying an extreme event, we can retrieve similar RS observations in the extreme events database and leverage them to characterize the current observation (extreme event). While the raw data represent one feature space, there are always infinitely more choices of feature spaces based on specific combinations, subsets, mathematical operations (e.g. logarithms or ratios), or other transformations of the raw data.

Third, applications of ML methods and increased model complexity can help improve classification based on raw data and using feature spaces based on theory as well. Therefore, it is essential to leverage the domain knowledge to enhance the capability of ML-based prediction systems. Second, most existing ML methods (especially deep neural networks (DNNs)) lack explainability. For this purpose, we propose to design self-explanatory models (e.g., attention, exemplar) or post-hoc explanation methods (in which DNNs are treated as a black box and employ explainable models such as a decision tree to fit the input and output and perform explanation) to make the decision interpretable. ML-methods that can address the above two issues will provide enhanced prediction and classification results with a more transparent decision process.

Next generation of ESMs. We envision the next-generation of ESMs will be deeply integrated with theory-informed RS-systems. Advanced ESMs/RS integration would provide capabilities of constraining parameters and reducing parameter uncertainty to improve the process modelling, as well as improve the predicting capabilities of ecosystem functioning at high resolutions over large extents, in particular where in-situ data are not available. In support of this, a better integration with global sampling networks and ecological databases is also necessary. Such integration would allow a guided site selection based on representativeness to improve the extrapolation from plot-scale to larger areas.

Rapid assimilation and analysis of a suit of remote sensing data for predicting extreme events and their impact on ecological-human systems

Suggested Partners/Experts. Partners would include team members in Watershed Function SFA, BioFi, Botanical Information and Ecology Network (BIEN), Coastal Observations Mechanisms and Predictions Across Systems and Scales - Field Measurements and Experiments (COMPASS-FME).

References

Lade, Steven J., Will Steffen, Wim De Vries, Stephen R. Carpenter, Jonathan F. Donges, Dieter Gerten, Holger Hoff, Tim Newbold, Katherine Richardson, and Johan Rockström, 2020. "Human impacts on planetary boundaries amplified by Earth system interactions." Nature Sustainability 3, no. 2: 119-128.

Norris, K., Terry, A., Hansford, J. P., & Turvey, S. T., 2020. Biodiversity conservation and the earth system: Mind the gap. Trends in Ecology & Evolution.

Ruckelshaus, Mary H., Stephen T. Jackson, Harold A. Mooney, Katharine L. Jacobs, Karim-Aly S. Kassam, Mary TK Arroyo, András Báldi et al., 2020. The IPBES global assessment: Pathways to action. Trends in ecology & evolution 35, no. 5: 407-414.

Carpenter, Stephen R., Harold A. Mooney, John Agard, Doris Capistrano, Ruth S. DeFries, Sandra Díaz, Thomas Dietz et al., 2009. Science for managing ecosystem services: Beyond the Millennium Ecosystem Assessment. Proceedings of the National Academy of Sciences 106, no. 5: 1305-1312.

AghaKouchak, Amir, Felicia Chiang, Laurie S. Huning, Charlotte A. Love, Iman Mallakpour, Omid Mazdiyasni, Hamed Moftakhari, Simon Michael Papalexiou, Elisa Ragno, and Mojtaba Sadegh., 2020. Climate extremes and compound hazards in a warming world. Annual Review of Earth and Planetary Sciences 48: 519-548.

Trotsiuk, V., Hartig, F., Cailleret, M., Babst, F., Forrester, D. I., Baltensweiler, A., ... & Schaub, M., 2020. Assessing the response of forest productivity to climate extremes in Switzerland using model–data fusion. Global change biology, 26(4), 2463-2476.

Steel, Emma J., Joseph B. Fontaine, Katinka X. Ruthrof, Treena I. Burgess, and Giles E. St J. Hardy, 2019. Changes in structure of over-and midstory tree species in a Mediterranean-type forest after an extreme drought-associated heatwave. Austral Ecology 44, no. 8: 1438-1450.

Ruthrof, K.X., Breshears, D.D., Fontaine, J.B., Froend, R.H., Matusick, G., Kala, J., Miller, B.P., Mitchell, P.J., Wilson, S.K., van Keulen, M. and Enright, N.J., 2018. Subcontinental heat wave triggers terrestrial and marine, multi-taxa responses. Scientific reports, 8(1), pp.1-9.

Dietze, M. C. et al. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. 2018. Proc. Natl. Acad. Sci. U. S. A. 115, 1424–1432.

Bush, A. et al., 2017. Connecting Earth observation to high-throughput biodiversity data. Nat Ecol Evol 1, 176.

Hampton, S. E. et al., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment vol. 11 156–162.

Lazin R., X. Shen and E. Anagnostou, 2021. Estimation of Flood-Damaged Cropland Area Using a Convolutional Neural Network. Environmental Research Letters, Article reference: ERL-110343.R1

Cerrai D., E.N. Anagnostou, et al., 2019. Improving Predictability of Storm Power Outages by Evaluating a New Representation of Weather and Vegetation in Non-parametric Modeling. IEEE-Access, DOI: 10.1109/ACCESS.2019.2902558.

Nikolopoulos, E., Elisa Destro, Md Abul Ehsan Bhuiyan, Marco Borga, and Emmanouil N. Anagnostou, 2018. Evaluation of predictive models for post-fire debris flows occurrence in the western United States. Nat. Hazards Earth Syst. Sci., 18, 2331-2343, 2018, https://doi.org/10.5194/nhess-18-2331-2018.

Zhang et al. 2019. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. AAAI 2019: 1409-1416.

Brummer, A. B., Lymperopoulos, P., Shen, J., Tekin, E., Bentley, L. P., Buzzard, V., Gray, A., Oliveras, I., Enquist, B. J., & Savage, V. M., 2021. Branching principles of animal and plant networks identified by combining extensive data, machine learning and modelling. Journal of The Royal Society Interface, 18(174), 20200624. https://doi.org/10.1098/rsif.2020.0624