Multi-scale Multi-physics Scientific Machine Learning for Water Cycle Extreme Events Identification, Labelling, Representation, and Characterization

Authors

Jason Hou, Huiying Ren, Arun Veeramany, Larry Berg, Tim Scheibe, Ruby Leung (PNNL)
Alex Sun (UT Austin), Roger Ghanem (USC), Ty Ferre (U Arizona)

Focal Area(s)

Insight gleaned from complex data (both observed and simulated) using AI, big data analytics, and other advanced methods, including explainable AI and physics- or knowledge-guided AI

Science Challenge

Impacts of climate are usually felt through extreme events such as droughts, floods, thunderstorms, windstorms, wildfires, and so on, that are intimately tied to the water cycle. Predicting the frequency and severity of extreme events under climate change remains a significant challenge; meanwhile, the mechanisms and impacts of these extremes are far from well understood. There are several major science challenges: (1) Lack of labelled extreme events data and missing standards in defining extremes; (2) Computational demand of high-resolution ensemble climate modeling; (3) Modeling the multiscale multi-physics hierarchical structure of compound extremes; (4) Lack of understanding of mechanisms of extreme events; (5) Large uncertainty in extreme events impacts on infrastructure; (6) Subjective assessment of weather-related risk from seasonal to multi-decadal time scales and lack of metrics for risk assessment and mitigation control.

We identify the following high-priority research needs that artificial intelligence (AI), machine learning (ML) and deep learning (DL) may enable transformational breakthroughs, by integrating general purpose GPU, cloud, and edge computing as well as database management:

- Water cycle extreme events identification, labelling, characterization, imputation, and augmentation, towards benchmark and FAIR (findable, accessible, interoperable, reusable) datasets to be integrated with high-dimensional multivariate predictors;
- Multi-scale multi-physics hierarchical structural representation of the multivariate joint extremes system;
- Development of standard and quantitative impact and risk metrics.

Rationale

Extreme events are featured by extreme values of irregular and rare occurrences, e.g., extreme weather conditions, outages and cascading events in power grids, system oscillations and instabilities; they occur infrequently but usually impede the normal functioning of the system with high-impact consequences. Effective ML of the mechanisms and predictive understanding of the extremes is usually limited by the inadequacy of training data for extreme events, which becomes even more challenging if the events are not correctly or confidently labelled (classified). In addition to the relatively small samples of extremes, complementary numerical simulations are also not adequate, given the computational demand and great uncertainty associated with the large degrees of freedom of key variables in the spatiotemporal modeling domain. Furthermore, few observations are available for each location and time point and strong

DOI: https://doi.org/10.2172/1769751
DOE BER, AI4ESP 2021 #AI4ESP1062 (https://www.ai4esp.org)

assumptions that the shape and scale parameters of the extreme-value distributions are constant across the spatiotemporal domain also confound our ability to understand and predict changes in extreme events in the future. As a result, novel ML methods are needed to produce data that accurately represents the extreme values/regions associated with adverse events of interest. These methods must be stable during optimization for training while rendering data samples that fully capture diverse process distribution at the extremes via open datasets and architects.

Extremes, such as droughts, wildfires, floods, and storms, also exhibit complex spatial cross-dependence structure across temporal scales. This makes the characterization and modeling of extremes difficult. Novel approaches are needed to estimate extreme-value temporal pattern such as seasonality and trend, but also the distribution of extreme events in the future. Graphical models (e.g., graphical neural networks) have great potential to capture highly structured dependencies among the parameters of extreme-value distributions. Few-short learning (learning with limited data), together with graphical models and generative models can be potentially integrated towards comprehensive spatiotemporal representations of compound extreme events.

Defining and quantifying risks associated with extreme events under climate change is also challenging as it requires comprehensive uncertainty quantification and coupled earth, environmental, and energy system modeling, for example, to quantify the impacts of extreme events type, magnitude, and duration on hydrologic and electricity infrastructures. ML and DL approaches are needed for data augmentation and system representation, but they need to be modified to provide decision making models with quantified uncertainty. In addition, a set of standards and protocols are needed through coordinated effort to make sure the risk measures and mitigation strategies are consistent across scales, regions, and countries.

Narrative

Water cycle extreme events identification, labelling, characterization, imputation, and augmentation: Although extreme events database is not adequate to enable most of the advanced ML and DL techniques, there do exist plenty of historical extreme weather events data, atmospheric modeling outputs, forecast reanalysis data and remote sensing imagery information, which allow ML/DL-assisted labeling and augmentation, for example, by integrating unsupervised hybrid clustering, self-organization maps, supervised classification, generative networks, and few-shot learning (learning with limited data), to name a few. Such defined and augmented extremes labels could significantly expand the existing extreme events database to be FAIR (findable, accessible, interoperable, and reusable) for supervised, physics-informed, explainable scientific ML, for characterizing, modeling, and predicting extremes such as tornadoes, storms, droughts, floods and wildfires within present and future climate scenarios.

Multi-scale multi-physics hierarchical structural representation and modeling of compound extremes: Hierarchical graph representation is critical to characterizing and modeling the high-dimensionality and cross-dependence of features that is prevalent in complex dynamic systems. To support hierarchical graph representation, we can develop synthetic network generation methods as well as graph coarsening methods that will enable multi-level ML methods. In addition to representing (complex) interrelationships among input entities, graph neural networks (GNNs) that operate over structures represented as graphs are an important target, which enables ML, including label propagation, graph coarsening in support of multi-level methods for ML, and active learning.

Dimension reduction is a critical challenge in complex extreme events system modeling and simulation. The state-space explosion problem coupled with the curse of dimensionality (in terms of feature space) within such a large-scale complex system of extreme events can lead to intractable optimization outcomes required for actionable decision support insights. To address the dimension-reduction challenge, one can take advantage of both implicit (e.g., generative adversarial networks (GANs)) and explicit (e.g., variational autoencoders (VAEs)) methods and develop novel scalable hybrid variational autoencoder generative adversarial network (VAE-GAN) based non-parametric approaches to handle high-dimensional state space information for function approximation and sampling. The coupled network enables dealing with unbalanced training data by learning the underlying data distribution, where one can apply an encoder to learn complex data distribution and output the approximate posterior of the latent variable which are fed to the decoder to generate new samples to augment the sparse data and improve data adequacy, and match the original data distribution from generated samples through re-sampling the dataset from random noise and has a better performance of generating good samples but is challenged by convergency issue due to unstable training, mode collapsed generation, diminished gradient, overfitting, and sensitivity to hyper-parameters. Such a network uses learned feature representations in the discriminator as basis for reconstruction, capture data distribution while offering invariance towards translation, and can effectively generate unseen features and classes to characterize the unbalanced data of extremes. Methods based on diffusion maps (DMAP) can also help extract intrinsic structure from extremes data as they permit the localization of the statistical fluctuations to a lower dimensional manifold. Large samples can then be generated on these manifolds that are statistically consistent with the training dataset.

Extreme events impact and risk metrics: Quantification of the impacts of extreme events on infrastructure resilience to high impact low probability events is of growing concern, for instance to address the impacts of extreme weather on critical hydrological and electricity infrastructures worldwide. However, up to date, there is no clear methodology or set of metrics to quantify resilience in the context of infrastructure systems, in terms of both operational and infrastructure integrity. Therefore, we identify another high-priority research component for extreme events study that can be assisted by ML/DL, which is to identify a standard and comprehensive list of impact and risk metrics that are not only transferrable across temporal scales and spatial locations, but also sensitive to major operational parameters or factors to allow development of effective and automated ML (e.g., random forest, gradient boosting, artificial neural networks) decision-support models, e.g., based on integration of observational network and ensemble earth/environmental/energy system models.

The above research and development can leverage the large library of big data analytics, uncertainty quantification, and ML/DL algorithms that the BER community scientists (e.g., the white paper writing team) developed, tested, and implemented with GPU/cloud/edge computing, in previous and ongoing projects.

Suggested Partners/Experts (Optional)

- NOAA/USGS database teams
- E3SM modeling teams
- PNNL ARM/ESS-SFA
- DOE Research computing
- PCSD/NSD data science teams