PhD Offer : ENS Lyon - France - more info http://perso.ens-lyon.fr/patrice.abry/ClimateLearningPhDSubject.pdf

Title : Studying and forecasting climate extreme and rare events from a machine learning based integration of observational and numerical model data

Early : April 23rd, 2021

Requirement : EU (or CH) Citizenship

Application : email (CV+motivation letter+Recommendation letter) patrice.abry@ens-lyon.fr

Abstract: The overall impact of climate change mostly stems from the rarest and most extreme events. Understanding the statistics and the spatiotemporal dynamics of extreme events thus constitute a major stake for human adaptation to climate changes. However, studying rare events constitutes an old and long-lasting significant scientific challenge, because of both the limited availability of observational data, and simplifications, biases and costs in global climate model numerical simulations.

The present works aims to investigate the use of machine learning and artificial intelligence to integrate observational and numerical model data in extreme event studies, so as to take advantage of both worlds: Observational data are scarce but produced by actual physical mechanisms; conversely, model data might be biased but are plentiful.

Forecasting heatwave occurrences and waiting times in the North hemisphere will be the specific goal of this work. Yet, the contributions aim to be of methodological values, with transfer learning to marry observational and model data, and to match the intrinsically nested nature of extreme events, with rare event simulation genetic algorithms to address resampling issues, with deep learning architecture design and data preprocessing driven by physical mechanisms and equations.

Context: Massive impacts of rare and extreme events. The impacts of climate change on biodiversity, human fatalities, societies and economy are likely to be dominated by the rarest and most extreme events, especially heat waves, drought, and extreme precipitations. For instance, recent decades witnessed a number of exceptionally warm summers and record-breaking heatwaves [1]. At Northern hemisphere mid-latitudes, relevant such examples were observed over France and Western Europe during summer 2003, with a death toll of about 70,000 people [2], or over Russia during summer 2010 [3], [4], with long lasting (several weeks) periods of anomalous heat. These events were unprecedented in the past 500 years and were outliers in the probability distribution of temperature [5]. Climate projections expect heat waves to become more severe toward the end of the 21st century [6].

Therefore, a difficult challenge for the scientific community, with critical stakes for the society (international agencies, governments and policy makers), consists in characterizing (the statistics of) extreme events and in predicting their occurrences, notably for events that were never observed before.

Scientific challenges: Studying the statistics of extreme events and predicting their occurrence. Studies of extreme climate events aim to assess the statistics and spatiotemporal dynamics of extreme events. Notably, the focus are: i) on the relations between intensity (return levels) and probability (return times or return period), a static property, and ii) on the clustering structures of extreme events, a dynamical property in time and space, that may stem from Earth & Ocean nonlinear longterm couplings. Notably, the way such intensity-return relations or clustering structures change through periods of time may convey significant information related to global climate change speed and strength. Assessing such relations and structures may also reveal physical features that can be used as precursors in extreme events forecasting.

Bottlenecks: lack of empirical data, model limitations and sampling issues. Current methodologies for studying extreme event statistics and predicting long return times and spatiotemporal clustering patterns suffer from significant limitations, stemming from the fact that extreme events are … rare events. This implies that huge amounts of data need to be analyzed for a relevant assessment of rare event properties.

From observational data, only a few units of extreme events of moderate to large intensities might be observed, and it may even be the case that events of most extreme intensity were never observed. Observational data will thus hardly be available in large enough quantity to permit relevant statistical analysis. Observational data also only provide partial information (sampling of the ground temperature, geopotential heights only for a finite and limited levels of pressure levels,…) as the quantity of information that can be measured is by nature limited.

Instead, Global Climate Models have been massively used to produce large amounts of numerically synthetized data. Thanks to their being based and constrained by the laws of physics, such model data permit to sample correctly the tail of extreme event distributions, and thus to estimate accurately return times or the evolution of their records [7]. However, relevant statistics require to simulate thousands of years of climate and are only thus achieved at massive storage and memory costs. Further, Global Climate Models need to account for intricate nonlinear multicomponent and longterm interactions and couplings and thus may require approximations, thus implying potential biases, whose impact on extreme event assessment is difficult to evaluate but could be large.

Beyond the direct use of observational data as initial conditions to global climate model solvers, there is thus a significant need to define methodologies aiming to combine the use of observational and numerical model data in extreme event studies.

Goal and contributions: Machine learning and artificial intelligence to integrate observational and numerical model data in extreme event studies. The overarching goal of this PhD thesis work is to contribute to devising artificial intelligence based methodologies aiming to combine observational and numerical model data to improve extreme event statistics assessment and forecasting.

The contributions intend to be of methodological values and thus to be of usable for the study and forecast of extreme events of different natures in climate studies and, beyond climate studies, in other application fields. These methodological contributions are organized along several lines:

- Data structuration to make joint use of physical quantities different in nature and units,
- Transfer learning and nested extreme events (most extreme events are included in less extreme ones),
- Transfer learning with pre-training on numerical model data and final training on observational data,
- Resampling strategies to address the intrinsic unbalanced class sizes between extreme and non-extreme events, with physically-driven resampling strategies,
- Design of artificial intelligence architectures design driven by physical mechanisms and equations.

The first focus will however be on studying extreme heatwaves over the north hemisphere.

Research program:

The goal is to devise artificial intelligence numerical tools that can actually be used data to forecast the probability of occurrences or the waiting time till the next extreme heatwaves of given intensities, from a limited set of real-observational data, with a limited number of computational resources and in a time that is compatible with actual societal reactivity and action.

The empirical data will be the ERA5, ERAClim and NCEP reanalysis data sets, which extend over several decades. The need for and availability of other empirical data will be investigated in the course of the work.

The research tracks will be organized along different lines.

Resampling strategies. By nature of the problem addressed, there is an intrinsic class-size unbalance between extreme and non-extreme events. Such an unbalance is well documented to constitute a severe issue in designing learning procedures and in achieving relevant prediction or estimation performance. Several strategies will be investigated. The simplest approach consists in downsampling the largest class. Alternatively, existing upsampling strategies for the smallest class will be tested. However, we also intend to explore two original approaches.

First, deep learning architectures, in the spirit of Discrete-Convolutional Generative adversarial Networks (DC-GAN), will be used to synthesize artificial climate time series containing extreme events, with spatiotemporal dynamics that reproduce those of observational climate data. This would thus enlarge the available sets of data with extreme events. The training of such DC-GAN will make use of Global Climate Model and observational data.

Second, and in attempt to marry physics (thus expert knowledge) and artificial intelligence, rare event algorithms will be used. They have been developed in the fields of statistical physics and molecular dynamics in order to compute trajectories that are too rare to be computed by direct numerical simulations. The general principle is to work with ensemble simulations. The trajectory ensemble is then modified akin to population dynamics (genetic algorithms) with some selection rules. Some trajectories are selected and reproduced such that the rare events become more common. The algorithm gives a rule to compute the probability of the new ensemble. The algorithm then produces extremely efficiently rare events that could not be observed in a reasonable computational time and estimate those event probabilities. Using such a rare event algorithm, we were able to evaluate return times for events that cannot be observed otherwise in comprehensive Global Climate Models [8]. Moreover, the number of observed heatwaves for a return time of 100 years is multiplied by a factor about 100, which improves the statistics of extremes [9].

Transfer learning. Transfer learning strategies, which consists in applying what has been trained under certain conditions to (slightly) different situations, will be investigated in two different instances.

First, the idea is to exploit the intrinsically nested nature of extreme events: most extreme events are included in less extreme ones. Training will hence first be performed on the less extreme events, and the learned architecture will be used to seed and initialize the training for more extreme events, iteratively up to the most extreme and hence rarest events.

Second, transfer learning will be used to combine global climate model and observational data. The pre-training of artificial intelligence architectures will be first based on global climate model data. The achieved architectures will be used to initialize the fine training on observational data. The idea is that global climate model data are plenty and thus can yield efficient pre-training but may contain model biases and approximations. The final training from observational data may correct for such biases and avoid that the training locks on model defects, even if mild, rather than on feature of real physical significance, a classical pitfall in Deep Learning architecture training.

Data structuration. Climate data are heterogeneous in nature and in physical units (temperature, pressure, humidity). While is it has often been claimed that artificial intelligence will make the best of raw data to preprocess them and extract relevant information, it is now commonly accepted that data preprocessing prior to feeding deep learning architectures has positive impact of performances. A classical case was made by the team of S. Mallat, promoting the scattering transform outputs as first non-trainable layers for deep learning architectures [10]. In the present work, we will seek to investigate how data preprocessing based on physical mechanisms and equations may improve data assimilation.

Deep learning architecture design. In any artificial intelligence tool, the question of the choice or design of the architecture needs to be addressed. Because climate data are governed by coupled partial differential equations inducing spatiotemporal dynamics and patterns, convolutional neuronal networks appear as natural candidate architectures. Tailoring their complexity, e.g., using Vapnik-Chervonenkis Dimension [11,12,13], to the complexity of the task potentially quantified by the rareness of the extreme events (or more technically by the likelihood of their statistics), will be addressed.

Further, following interesting propositions very recently reported in [14], we will investigate how to take advantage of the knowledge of the partial differential equations, governing climate dynamics, and of the analysis of their iterative resolution schemes, to better guide the choice and design of deep learning architectures. This will be conducted in the spirit of recent work conducted by Nelly Pustelnik, a member of the Signal, Systems and Physics team, aiming to use convex minimization iteration schemes to design deep learning layer design [15].

Feasibility: Freddy Bouchet already conducted preliminary work of the use of machine learning for climate prediction [16]. Patrice Abry already published articles related to deep learning in International conferences [17, 18, 19], one of which has recently been selected to be transformed into a journal paper [18]. Freddy Bouchet and Patrice Abry have (in collaboration with Pierre Borgnat) already started preliminary investigations related to such topics via the joint supervision of internships. These several months of preliminary efforts lead to the current writing of a first paper (close to be submitted and available upon request) [20], as well as to the structuration of this PhD research program.

Research Unit support: This project received formal approbation from the Director of LPENSL. Please see supporting letter appended below.

[1] IPCC.Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[2] R. Garcıa-Herrera, J. Dıaz, R.M. Trigo, J. Luterbacher and E.M. Fischer. A Review of the European Summer Heat Wave of 2003. Critical Reviews in Environmental Science and Technology, 40(4):267–306, 2010.

3] D. Barriopedro, E.M. Fischer, J. Luterbacher, R.M. Trigo, and R. Garcia-Herrera. Redrawing the temperature record map of Europe. Science, 332:220–224, 2011.

[4] F.Otto, N. Massey, Geert Jan Van Oldenborgh, R. Jones, and M. Allen. Reconciling two approaches to attribution of the 2010 Russian heat wave. Geophysical Research Letters, 39:4702–, 02 2012.

[5] Schaer, C., Vidale, P., Luthi, D., Frei, C., Haberli, C., Liniger, M., & Appenzeller, C. (2004). The role of increasing temperature variability in European summer heatwaves. *Nature*, *427*(6972), 332–336.
[6] Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. In C. B. Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, et al. (Eds.), *A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC SREX Report)*. Cambridge: Cambridge University Press.
[7] Bador, M., Terray, L., Boe, J., Somot, S., Alias, A., Gibelin, A.-L., & Dubuisson, B. (2017). Future summer mega-heatwave and record-breaking temperatures in a warmer France climate. Environmental Research Letters, 12(7), 074025.
[8] Wouters, J., & Bouchet, F. (2016). Rare event computation in deterministic chaotic systems using genealogical particle analysis. *Journal of Physics A: Mathematical and Theoretical*, *49*(37), 374002.
[9] F. Ragone, J. Wouters and F. Bouchet, 2018, Computation of extreme heat waves in climate models using a large deviation algorithm, Proceedings of the National Academy of Sciences, vol 115, no 1, pages 24-29, [pdf].
[10] J. Bruna, S. Mallat, (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1872-1886.
[11] V. N. Vapnik and A. Y. Chervonenkis, (2015). On the uniform convergence of relative frequencies of events to their probabilities, in Measures of complexity, pp. 11–30. Springer.

[12] E. B. Baum and D. Haussler (1989). "What size net gives valid generalization?" in Advances in neural information processing systems, pp. 81–90.

[13] G. Friedland and M. Krell, (2017). A capacity scaling law for artificial neural networks," arXiv preprint arXiv:1708.06019.

[14] O. Pannekoucke and R. Fablet (2020). PDE-NetGen 1.0: from symbolic PDE representations of physical processes to trainable neural network representations. Preprint: arxiv.org/pdf/2002.01029.pdf

[15] M. Jiu and N. Pustelnik (2020). A deep primal-dual proximal network for image restoration, preprint.

[16] D. Lucente, S. Duffner, C. Herbert, J. Rolland and F. Bouchet, 2019, Machine learning of committor functions for predicting high impact climate events, *Climate Informatics CI2019 proceedings*, *arXiv:1910.11736*, [pdf].
[17] Liotet, P., Abry, P., Leonarduzzi, R., Senneret, M., Jaffrès, L., Perrin, G. (2020, May). Deep Learning Abilities to Classify Intricate Variations in Temporal Dynamics of Multivariate Time Series. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3857-3861).

[18] Mauduit, V., Abry, P., Leonarduzzi, R., Roux, S. G., Quemener, E. (2020, September). Dcgan for the Synthesis of Multivariate Multifractal Textures: How do We Know it Works?. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE

[19] B. Pascal, V.Mauduit, P. Abry, and N. Pustelnik *Scale-free texture segmentation: Expert feature-based versus Deep Learning strategies,* European Signal Processing Conference (EUSIPCO), The Netherlands, Amsterdam, January 18 - 22, 2021 (pdf)

[20] Valerian Jacques-Dumas, Freddy Bouchet, Pierre Borgnat and Patrice Abry (2021). Deep Learning based Extreme Heatwave Forecast. In preparation. Current version available upon request.