XAI4Extremes: An explainable AI framework for understanding extreme-weather precursors

Jiawen Wei[a], **Aniruddha Bora**[b], **Vivek Oommen**[b], **Chenyu Dong**[a], **Juntao Yang**[c], **Jeff Adie**[c], **Chen Chen**[d], **Simon See**[c], **George Karniadakis**[b], **Gianmarco Mengaldo**[a]

[a] *National University of Singapore, Singapore* (jiawenw, chenyu.dong)@u.nus.edu, mpegim@nus.edu.sg

[b] *Brown University, US* (aniruddha_bora, vivek_oommen, george_karniadakis)@brown.edu

[c] *NVIDIA AI Technology Centre, Singapore* (yjuntao, jadie, ssee)@nvidia.com

[d] *Centre for Climate Research Singapore, Singapore* chen_chen@nea.gov.sg

Extreme weather events are increasing in frequency and intensity due to climate change [1, 2, 3, 4, 5]. This, in turn, is exacting a significant toll in communities worldwide. While prediction skills are increasing with advances in numerical weather prediction and artificial intelligence tools, extreme weather still present challenges. More specifically, identifying the precursors of such extreme weather events and how these precursors may evolve under climate change remain unclear. In this paper, we propose to use post-hoc interpretability methods to construct relevance weather maps that show the key extreme-weather precursors identified by deep learning models. We then compare this machine view with existing domain knowledge to understand whether deep learning models identified patterns in data that may enrich our understanding of extreme-weather precursors. We finally bin these relevant maps into different multi-year time periods to understand the role that climate change is having on these precursors. The experiments are carried out on Indochina heatwaves, but the methodology can be readily extended to other extreme weather events worldwide.

Predictability drivers for heatwaves vary across different regions and involve several physical mechanisms. Better understanding them could help forecasting heatwaves and issuing early warnings [6, 7, 8, 9]. These predictability drivers, also referred to as precursors, are typically the result of human-expert knowledge, or briefly the "*human view*". In this work, we look at these precursors through the lenses of interpretable machine learning (ML), thereby providing a possibly complementary "*machine view*". The latter is obtained by identifying what data the machine deemed important to the onset of heatwaves, and it is used to understand (by working with human domain experts) whether it may be helpful in enriching our understanding of precursors – see also [10] for the use of explainable artificial intelligence (XAI) for scientific knowledge discovery. Without losing generality in the methodology proposed, we focus on tropical heatwaves in the Indochina peninsula, and attempt to answer two questions via interpretable ML: (i) What are the key precursors of these events? (ii) Is climate change influencing these precursors?

To outline our approach, we focus on dry-season (February-March-April-May) heatwaves in the Indochina peninsula (the latter depicted in Appendix A Figure A1). The key idea is to look at these dry-season heatwaves, using interpretable ML; more specifically post-hoc interpretability methods applied to a binary time series classification deep learning (DL) framework. This approach allows producing relevance maps, that highlight what input data the DL framework deemed important for the prediction it made. The binary DL time series classification framework is setup as follows. As input data, we consider the spatial (i.e., geographical) maps of 23 variables for the 7 days prior of a heatwave striking the Indochina peninsula. The 23 input variables characterize the large majority of dry-season heatwave precursors, and the 7 days time window provides a relevant time frame to capture the underlying pathways leading to these extremes. We then assume that the DL framework is able to identify patterns in the data that are causal to heatwaves; in other words, we assume that it could capture systematically the precursors to heatwaves. Indeed, we consider only true positive samples, such that the data deemed important by the DL framework is only associated to correctly classified heatwaves. The binary labels for the classification task are (1) heatwave and (0) non-heatwave, where the heatwaves are identified as outlined in Appendix A.

The final heatwave binary classification dataset consists of 720 samples with an approximate ratio of (1) heatwave vs (0) non-heatwave being 1:5. We split the dataset into training, validation, and testing sets with a ratio of [0.6:0.2:0.2], and then train the Transformer model for heatwave classification. We apply four different post-hoc interpretability methods, namely Integrated Gradients [11], DeepLIFT [12], DeepSHAP [13], and GradSHAP [13], to the trained Transformer model. To guarantee that we obtain the most accurate and robust relevance maps, we adopt the interpretability evaluation frameworks in [14] and [15]. Integrated Gradients performs the best among four post-hoc methods according to the evaluation results; thereby we use the relevance maps it generates for analysis. The overall approach, that

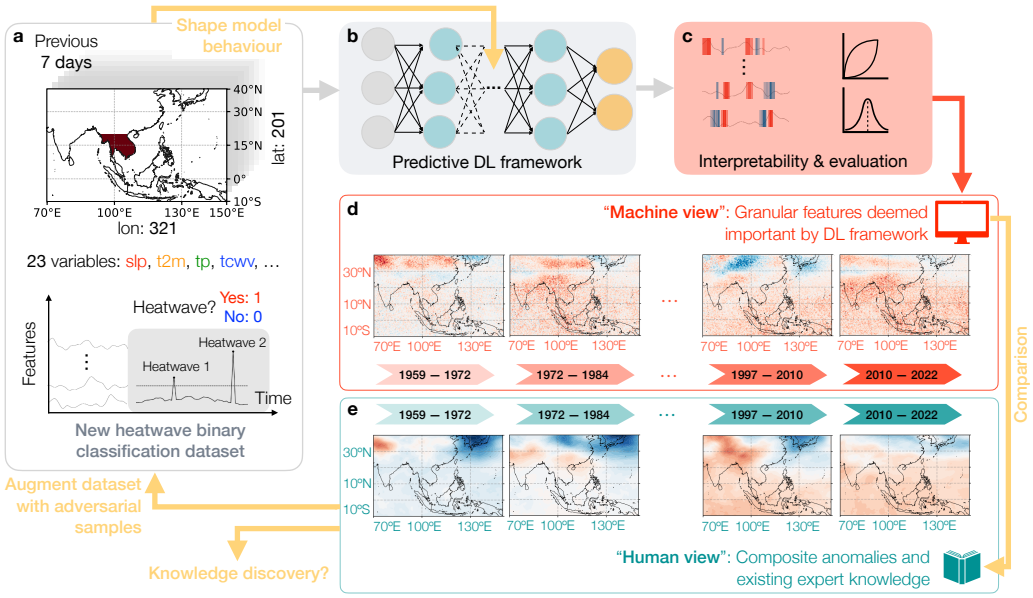

Fig. 1: The XAI4Extremes framework proposed, composed of a novel extreme weather dataset (a), a DL predictive model (b), an interpretability block along with its evaluation (c), that produces relevance maps, or what we called the "*machine view*" (d). The latter (d) is then compared with existing human expert knowledge (e) for knowledge discovery or for augmenting the dataset with e.g., adversarial samples that can shape and improve model behavior.

we name XAI4Extremes, is depicted in Figure 1: we propose a new dataset for weather extremes – heatwaves in this particular case (panel a, in gray), that is used by a predictive DL framework (panel b, in blue), to which we apply post-hoc interpretability and its evaluation (panel c, in red). The relevance maps produced by the post-hoc interpretability method, what we also refer to as "*machine view*" (panel d, in red), are then compared against human expert knowledge, what we also refer to as "*human view*" (panel e, in green). This comparison may lead to knowledge discovery in terms of heatwave precursors and role of climate change in heatwave precursors. This may be the case when the machine view enriches human expert knowledge, by providing a scientifically plausible use of data that was unknown to human domain experts, but that domain experts can explain. Indeed, it is responsibility of human domain experts to respond to the question why the interpretable ML framework deemed important a specific set of input data. The relevance maps can also be used to generate adversarial samples to augment the dataset and shape model behavior, thereby improving the performance of the predictive DL framework. We remark that the approach outlined in this section can readily be applied to other types of weather extremes in different regions worldwide.

We present our preliminary results in Appendix B) due to the page limit. Results (Figure A2, panel a) show the temperature field at 200 hPa (i.e., the temperature in the upper troposphere between approximately 11 and 12 km altitude), is deemed more important by the machine for heatwaves in Indochina in more recent decades, with a clear upward trend. If we compare the interpretability results (i.e., the relevance maps or machine view) with something more understandable by humans, i.e., composite anomalies, we note that there is indeed a warming of the upper troposphere that is associated to heatwaves in Indochina (Figure A2, panel b). This indicates that the temperature at 200 hPa is becoming a key precursor of Indochina heatwaves, especially in recent decades, aspect that may indicate the fingerprint of climate change. The result points to a human-understandable explanation where higher 200 hPa temperature can suppress convection and increase subsidence, thereby leading reduced cloud cover that amplifies surface heating, potentially leading to heatwaves.

The overarching explainable AI framework we propose in this work, namely XAI4Extremes, aims to better understand weather extremes and their evolution under climate change. We propose to couple a predictive DL framework with interpretability methods, in order to understand ***what*** data the machine deemed important for its predictive performance of true positive samples (i.e., correctly identified heatwaves), something we refer to as "*machine view*". We finally propose to compare this machine view to existing human expert knowledge (what we call "*human view*"), to respond the question ***why*** the machine used those data. The latter aspect may lead to knowledge discovery, or it can be used to shape model behavior by e.g., generating ad-hoc adversarial samples based on the machine view. We note that there are still several, yet stimulating, open challenges to be overcome [16]. We believe that these limitations are open opportunities for the AI and broader scientific research communities that can be tackled over the next few years.

Acknowledgments

J.W. and G.M. acknowledge support from MOE Tier 2 grant T2EP50221-0017, and from MOE Tier 1 grant 22-4900-A0001-0.

References

[1] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Y Chen, L Goldfarb, MI Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021.

[2] Chenyu Dong, Robin Noyelle, Gabriele Messori, Adriano Gualandi, Lucas Fery, Pascal Yiou, Mathieu Vrac, Fabio D'andrea, Suzana J Camargo, Erika Coppola, Gianpaolo Balsamo, Chen Chen, Davide Faranda, and Gianmarco Mengaldo. Indo-pacific regional extremes aggravated by changes in tropical weather patterns. *Nature Geoscience*, pages 1–8, 2024.

[3] Davide Faranda, Gabriele Messori, Aglae Jezequel, Mathieu Vrac, and Pascal Yiou. Atmospheric circulation compounds anthropogenic warming and impacts of climate extremes in europe. *Proceedings of the National Academy of Sciences*, 120(13):e2214525120, 2023.

[4] SE Perkins-Kirkpatrick and SC Lewis. Increasing trends in regional heatwaves. *Nature communications*, 11(1):3357, 2020.

[5] Markus G Donat, Andrew L Lowry, Lisa V Alexander, Paul A O'Gorman, and Nicola Maher. More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, 6(5):508–513, 2016.

[6] Erin Coughlan De Perez, Maarten Van Aalst, Konstantinos Bischiniotis, Simon Mason, Hannah Nissan, Florian Pappenberger, Elisabeth Stephens, Ervin Zsoter, and Bart Van Den Hurk. Global predictability of temperature extremes. *Environmental Research Letters*, 13(5):054017, 2018.

[7] Lisa-Ann Kautz, Olivia Martius, Stephan Pfahl, Joaquim G. Pinto, Alexandre M. Ramos, Pedro M. Sousa, and Tim Woollings. Atmospheric blocking and weather extremes over the euro-atlantic sector–a review. *Weather and Climate Dynamics*, 2022.

[8] Wenju Cai, Simon Borlace, Matthieu Lengaigne, Peter Van Rensch, Mat Collins, Gabriel Vecchi, Axel Timmermann, Agus Santoso, Michael J McPhaden, Lixin Wu, et al. Increasing frequency of extreme el niño events due to greenhouse warming. *Nature climate change*, 4(2):111–116, 2014.

[9] Daniela IV Domeisen, Elfatih AB Eltahir, Erich M Fischer, Reto Knutti, Sarah E Perkins-Kirkpatrick, Christoph Schär, Sonia I Seneviratne, Antje Weisheimer, and Heini Wernli. Prediction and projection of heatwaves. *Nature Reviews Earth & Environment*, 4(1):36–50, 2023.

[10] Gianmarco Mengaldo. Explain the black box for the sake of science: the scientific method in the era of generative artificial intelligence. *arXiv preprint arXiv:2406.10557*, 2024.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[12] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[13] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[14] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023.

[15] Jiawen Wei, Hugues Turbé, and Gianmarco Mengaldo. Revisiting the robustness of post-hoc interpretability methods. *arXiv preprint arXiv:2407.19683*, 2024.

[16] Hugues Turbé, Mina Bjelogrlic, Gianmarco Mengaldo, and Christian Lovis. Protos-vit: Visual foundation models for sparse self-explainable classifications. *arXiv preprint arXiv:2406.10025*, 2024.

[17] SE Perkins, LV Alexander, and JR Nairn. Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, 39(20), 2012.

[18] Ming Luo and Ngar-Cheung Lau. Synoptic characteristics, atmospheric controls, and long-term changes of heat waves over the indochina peninsula. *Climate Dynamics*, 51:2707–2723, 2018.

Appendix A. Identification of heatwaves

Identifying heatwaves remains a significant challenge. Currently, there are numerous definitions of heatwaves in the research community, yet there is no consensus on a standard definition. This complexity arises from the varied spatial coverage and duration of heatwaves. In our study, we adopted a relatively

simple two-stage definition that combines index-based and event-based approaches, which have been widely used in other research.

We first define heatwaves on each individual grid point in the daily ERA5 reanalysis data from 1959 to 2022 using the heatwave index TX90pct [17]. The threshold for one day at one grid point is the calendar day 90th percentile of the daily maximum temperature, based on a centered 15-day window. A heatwave is defined as three or more consecutive days exceeding this threshold, and all days belonging to this heatwave are considered as heatwave days for that grid point. We note that we removed a grid point by grid point linear trend from the the data. This is because we want to maintain a relatively uniform distribution of heatwaves in the studied period.

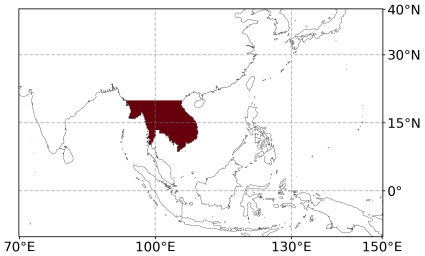

Fig. A1: Indochina region used to define heatwaves (dark red).

Based on this grid point by grid point definition of local heatwaves, we further define heatwave events in Indochina using the regional mask illustrated in figure A1. These events can be divided into heatwaves in the dry and in the wet seasons, whereby the precursors and onset mechanisms differ [18]. We focus on dry-season (FMAM) heatwaves without lacking generality on the methodology proposed here. For each region, one heatwave event is defined when a minimum number of grid points are identified as heatwaves. Specifically, this threshold is set at the 90th percentile of the number of grid points classified as heatwaves during the season of interest. We define the first day that exceeds the predefined threshold as the heatwave onset day. To avoid overlapping events, we stipulate that no day within the seven days preceding any onset day should exceed this threshold. For the onset days of non-extreme events, we randomly select days when the number of grid points falls below the predefined threshold within the same season, following specific criteria: We ensure that there are no heatwave onset days or other non-extreme events within a 7-day window before and after these selected days. We provide the dataset with a ratio of non-extreme events to extreme events set at 5:1. This is the maximum ratio achievable while following the selection strategy outlined above.

Appendix B. Preliminary results

Figure A2 show the temperature field at 200 hPa, that is the temperature between approximately 11 and 12 km altitude (i.e., the temperature in the upper troposphere), for two different regions, region 1 and 2. Region 1 comprises the Indian Ocean, and India, while region 2 comprises the Maritime continent and part of the Pacific Ocean. In Figure A2, panel a, we show the mean trend of relevance for region 1 (top row), region 2 (middle row), and region 1 and 2 combined (bottom row). It is possible to see how the temperature in the upper troposphere is deemed more important by the machine for heatwaves in Indochina in more recent decades for both regions, with a clear upward trend. If we compare the interpretability results (i.e., the relevance maps or machine view) with something more understandable by humans, i.e., composite anomalies, we note that there is indeed a warming of the upper troposphere that is associated to heatwaves in Indochina (Figure A2, panel b). This indicates that the temperature at 200 hPa is becoming a key precursor of Indochina heatwaves, especially in recent decades (in agreement with composite anomalies), aspect that may indicate the fingerprint of climate change.

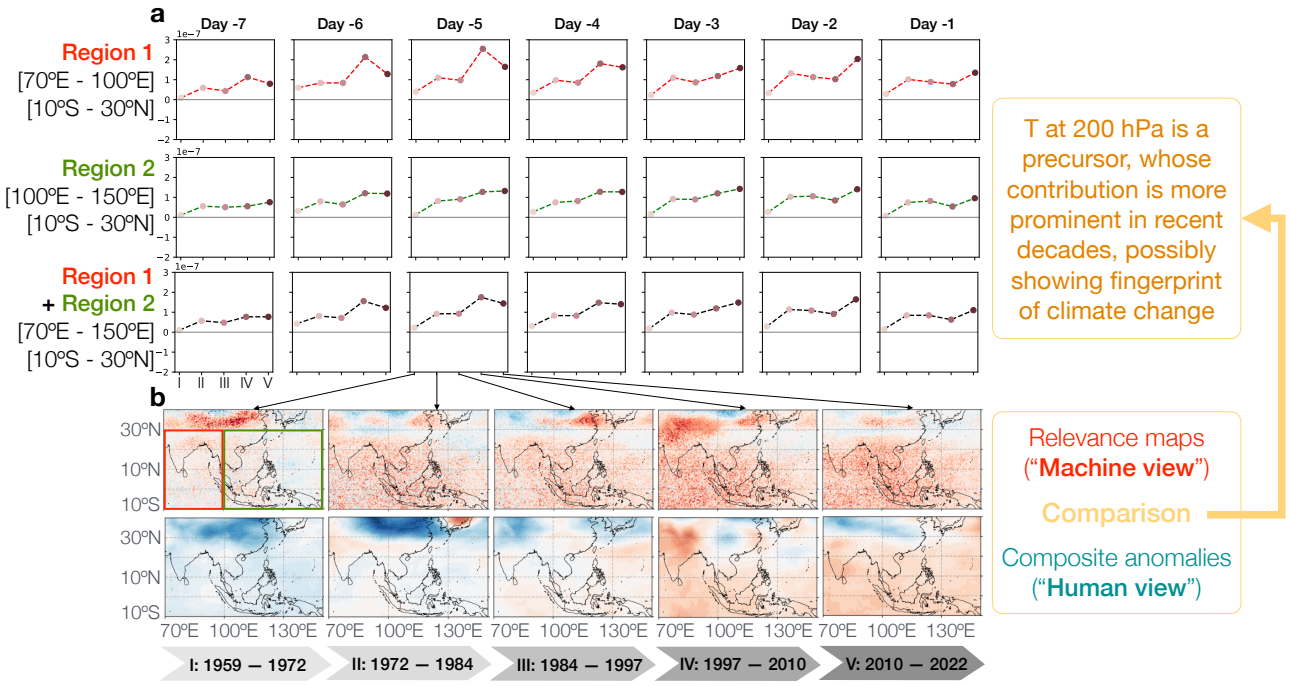

Fig. A2: Mean relevance of temperature at 200 hPa for the 5 historical time periods considered and on the 7 days prior to heatwaves in Indochina, for region 1 (a, top), region 2 (a, middle), region 1+2 (a, bottom), along with the corresponding relevance maps – i.e., "*machine view*" – associated to 7 days prior to heatwave, and composite anomalies – i.e., "*human view*" – (b).