Methods for Forecasting Extreme Events with Machine Learning and Extreme Value Statistics

Pasche, Olivier Colin

Methods for Forecasting Extreme Events with Machine Learning and Extreme Value Statistics

Méthodes de Prédiction d'Événements Extrêmes
par Apprentissage Automatique et
Statistique des Valeurs Extrêmes

Olivier C. Pasche

Thesis n° 163 submitted to the
Doctoral Program in Statistics,
Research Institute for Statistics and Information Science,
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of Ph.D. in Statistics

Jury:
Prof. Stefan Sperlich, University of Geneva (president)
Prof. Sebastian Engelke, University of Geneva (thesis advisor)
Prof. Hery Lam, Columbia University

February 2026

UNIVERSITÉ
DE GENÈVE

La Faculté d'économie et de management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par-là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 20 février 2026

Doyen
Salvatore DI FALCO

Impression d'après le manuscrit de l'auteur

À mes chers parents.

Acknowledgements

I would like to begin by expressing my deepest gratitude to my thesis advisor, Sebastian Engelke. Your trust, support, and kindness have been instrumental to everything I have achieved during this Ph.D.. These few years of collaboration have been truly amazing, both scientifically and personally. You were genuinely invested in my work and always available to offer thoughtful advice and support, even well beyond the scope of our research. At the same time, you trusted my ideas and choices, and gave me the freedom to pursue the ideas I found most compelling and meaningful, a freedom that contributed enormously to how much I enjoyed this journey. You also played a central role in building my research network, creating countless opportunities to present at international conferences, connect with leading collaborators, and establish partnerships with other universities. From early on, you entrusted me with real teaching responsibilities, such as course management, lecturing, designing a new Master's course, and supervising Master's theses, which helped me develop those teaching skills rapidly. Beyond all that, you helped foster a warm and stimulating work environment, with your genuine friendliness and with the many dinners, drinks, and outings in the mountains. I already look forward to our future collaborations!

My sincere thanks also go to Henry Lam, for inviting and welcoming me at Columbia University during my research stay, for believing in my research ideas, and for serving on my thesis jury. Our collaboration was both productive and genuinely enjoyable, and I am thrilled with what we have accomplished in under a year. Your great kindness made my time in New York City all the more memorable. In that regard, I extend my gratitude to all the friends I made there, in particular to Julien, Matías, Christian, Aapeli, and Darshan, for so warmly including me in your social circles and making New York feel like home.

I am deeply grateful to Anthony Davison and Valérie Chavez-Demoulin, for guiding and supporting my first steps in academic research, during and after my Master's thesis, and for introducing me to Sebastian: a recommendation that started this doctoral journey. I am also grateful to all my other collaborators, both for the work at the core of this thesis and for the parallel projects we pursued together: Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, Joel Zeder, Sebastian Sippel, Erich Fischer, Juraj Bodik, Gloria Buriticá, Manuel Hentschel, and Frank Röttger. It was a genuine pleasure working with each of you.

I would also like to warmly thank my friends and colleagues at our institute. In particular, thanks to my officemate Manuel, for the fun we shared at the office, out in the city, and up on the mountains. And thanks to everyone in our fifth-floor group: Alban, Alice, Anastasia, Anna, Beatrice, Cesare, Christophe, Dongmei, Edossa, Florian, Frank, Gloria, Manon, Marco, Marco, Nicola, Romain, Simon, and Zhongwei, as well as to Rim and Jérémy. You are in great part why I was looking forward to the days at the office, and I am grateful for the fun times we have together, whether that's skiing, going for a night out (or in Gland...), playing games, or chatting around the office. Your company made these years all the richer. Close to that environment, thanks also to

everyone having contributed to the CUSO and other Swiss researcher events being so enjoyable, including Jon, Tim, Sonia, Mario, Leonardo, Ho, Alessia, Kartik, Paul, Fabien, Ilia, Theresa, Charles, Jake, Patrick, Philip, Antoine, Louis, Ejub, and many others. I would also like to thank the professors of our statistics institute for our many pleasant interactions, at our various faculty events, and our collaborations in teaching. In particular, thanks to Stefan Sperlich for presiding over my thesis jury.

Furthermore, I am grateful to the extreme-value community. To the professors and senior researchers who have worked to build such a welcoming and inclusive environment for young researchers: thank you for your generosity, your consideration, and the many opportunities that you provided for presenting my work and creating an international network. For the many invitations to speak at conferences and workshops, I would like to thank, in particular, Olivier Wintenberger, Philippe Naveau, Valérie Chavez-Demoulin, Ioannis Papastathopoulos, Clément Dombry, Miguel de Carvalho, Claudia Klüppelberg, Raphaël Hüser, Jean-Marc Bardet, Surya Tokdar, and the Swiss Federal Statistical Office. And to the many friends made along the way, at conferences or during visiting periods, Jon, Cristina, Michaël, Silvia, Karla, Lambert, Pauline, Jordan, Viviana, Jakob, and many more, it has been a joy getting to know you, and meeting you here and there again over the years and across many countries.

J'aimerais également remercier tous mes amis et amies de longue date, en dehors du monde académique, qui ont été présents d'une manière ou d'une autre tout au long de ces dernières années, et avec qui j'ai partagé tant de moments mémorables. Vous êtes trop nombreux pour être toutes et tous nommés ici, mais, en particulier, merci à Cyril, Camille, Emil, Margaux, Robin, Audrey, Alej, Elliott, Dana, Cécile, Marco, Victor, Edouard, Andréa, et bien d'autres encore. Certaines de nos amitiés remontent à l'enfance, d'autres au collège, à nos années à l'EPFL, ou à d'autres occasions, mais toutes me sont profondément précieuses. Ces moments de joie que nous partageons et les aventures que nous vivons ensemble sont une de mes plus grandes sources de bonheur. Je suis très reconnaissant pour nos amitiés impérissables, et je me sens chanceux de vous avoir à mes côtés et de toujours pouvoir compter sur vous.

Et bien-sûr, merci à toi Ségo. Merci pour ton soutien sans faille, pour ton amour, et d'être toujours là pour moi, aussi bien pour les choix de vie complexes que pour les petites choses de tous les jours. Merci d'être qui tu es, de grandir avec moi, d'être ma meilleure motivatrice, et de m'encourager à devenir la meilleure version de moi-même. Tu rends les beaux jours encore plus rayonnants, et les jours difficiles bien plus légers. Merci pour tous ces moments de bonnheur que nous partageons ensemble, en toutes circonstances.

Enfin, et surtout, je souhaite remercier ma famille. Merci à mes parents, Claude et Cristina, pour votre amour et votre soutien inconditionnels. Merci de m'encourager à poursuivre mes rêves, mais aussi à profiter de la vie et de ses beaux moments. Je suis profondément reconnaissant pour tous vos efforts et vos sacrifices pour m'offrir les meilleures opportunités et le meilleur environnement possibles. Votre sagesse et vos conseils m'ont toujours guidé à travers chacune des étapes et décisions importantes de ma vie. Merci pour tout ce que vous avez investi en moi, en temps, en soutien et en amour, jusque dans les plus petits détails du quotidien. Merci aussi à toi Cornelia, pour ta présence et ton soutien continus. Une pensée va également à mes grand-parents et aux beaux souvenirs qu'ils m'ont laissés.

— O. C. P.

Abstract

Extreme events such as natural disasters, financial crashes, and overloaded infrastructures or services collapsing cause severe harm and have catastrophic lasting consequences, especially when they strike by surprise. Providing reliable forecasts and risk estimates is crucial for early warnings, disaster preparedness, and adaptation. They help policymakers make informed decisions, financial investors promptly mitigate losses, emergency services and communities prepare, and insurers anticipate sudden increases in claims. This can, in turn, save lives and ecosystems, and prevent economic recessions. In particular, with their increasing frequency and intensity under climate change, environmental extremes such as floods, heatwaves, wildfires, and hurricanes, are especially critical to predict accurately. However, foreseeing extreme events is statistically challenging, as they are, by nature, unprecedented or scarce in historical records, and have complex drivers. Existing methods generally either cannot extrapolate or are not designed for accurate forecasting.

In that light, this thesis develops novel methodologies for accurately forecasting the conditional risk of extreme events and for understanding their drivers, by combining the extrapolation capabilities of extreme value statistics with the predictive flexibility and versatility of machine learning and with the insightfulness of causal inference. The first contribution introduces a method providing accurate extreme quantile predictions when the dependence on predictors is complex or acts between observations, by combining neural networks with extreme value statistics. The model can also forecast other risk metrics, such as high-threshold exceedance probabilities or expected shortfalls, as the entire conditional tail of the response variable is modelled. The second contribution provides an additional type of forecasts: prediction intervals. Our extreme conformal procedure predicts informative and adaptive high-confidence intervals of likely values for the response variable, when the required confidence level is too high for classical conformal methods to be applicable. The third contribution proposes a permutation test for causal discovery in extreme regimes, and a way to mitigate confounding effects detrimental to the extremal causal analysis. The fourth studies the performance of state-of-the-art deep-learning global weather prediction models, during real extreme events, highlighting differences from operational physics-based systems. The new methods introduced in this thesis, and their implementation, aim to provide practical tools for risk assessment and forecasting, that are applicable to a wide range of domains.

Keywords: extreme events, prediction, forecast, risk, extreme value theory, generalized Pareto distribution, extreme value statistics, machine learning, recurrent neural network, deep learning, quantile regression, conformal prediction, prediction intervals, high confidence, causation, causal inference, confounding, natural disasters, flood, heatwave, forecast assessment, actuarial science.

Résumé

Les événements extrêmes tels que les catastrophes naturelles, les crashs financiers et l'effondrement d'infrastructures ou de services surchargés causent de graves dommages et ont des conséquences catastrophiques durables, en particulier lorsqu'ils surviennent par surprise. Fournir des prédictions et des estimations de risque fiables est crucial pour mettre en place des alertes précoces, la préparation aux catastrophes et l'adaptation. Elles aident les gouvernements à prendre des décisions éclairées, les investisseurs financiers à atténuer leurs pertes, les services d'urgence et les communautés à se préparer, et les assureurs à anticiper les hausses soudaines de sinistres. Cela peut, par conséquent, sauver des vies et des écosystèmes, et prévenir des récessions économiques. En particulier, compte tenu de leur fréquence et intensité croissantes dues au changement climatique, les extrêmes environnementaux, tels que les inondations, les canicules, les incendies de forêt et les ouragans, sont particulièrement critiques à prédire avec précision. Cependant, la prévision d'événements extrêmes est statistiquement difficile, car ils sont, par nature, sans précédent ou rares dans les archives, et ont des mécanismes complexes. Les méthodes existantes ne peuvent généralement pas extrapoler ou ne sont pas conçues pour des prédictions précises.
Dans cette optique, cette thèse développe de nouvelles méthodologies pour prédire le risque conditionnel d'événements extrêmes avec précision et pour comprendre leurs mécanismes, en combinant les capacités d'extrapolation de la statistique des valeurs extrêmes avec la flexibilité prédictive et la polyvalence de l'apprentissage automatique, et avec la sagacité de l'inférence causale. La première contribution introduit une méthode permettant des prédictions de quantiles extrêmes précises, lorsque la dépendance aux variables explicatives est complexe ou s'exerce entre les observations, en combinant les réseaux de neurones avec la statistique des valeurs extrêmes. Le modèle peut également prédire d'autres mesures de risque, telles que les probabilités de dépassement de seuils élevés ou les pertes attendues au-delà de ces seuils, puisque l'ensemble de la queue conditionnelle de la variable réponse est modélisé. La deuxième contribution fournit un type supplémentaire de prévisions : des intervalles de prédiction. Notre procédure conformelle extrême prédit des intervalles de haute confiance, informatifs et adaptatifs, pour les valeurs probables de la variable réponse, lorsque le niveau de confiance requis est trop élevé pour que les méthodes conformelles classiques soient applicables. La troisième contribution propose un test de permutation pour la découverte causale dans les régimes extrêmes, ainsi qu'une manière d'atténuer les facteurs de confusion néfastes à l'analyse causale des extrêmes. La quatrième étudie les performances de modèles récents de deep-learning pour la prévision météorologique globale, lors d'événements extrêmes réels, en mettant en évidence leurs différences avec les systèmes opérationnels traditionnels, basés sur la physique. Les nouvelles méthodes introduites dans cette thèse, ainsi que leurs implémentations, visent à fournir des outils pratiques pour l'évaluation et la prédiction de risque, applicables à un large éventail de domaines.

Résumé

Mots-clés : événements extrêmes, prédiction, prévision, risque, théorie des valeurs extrêmes, distribution de Pareto généralisée, statistique des valeurs extrêmes, apprentissage automatique, réseau de neurones récurrent, apprentissage profond, régression quantile, prédiction conformelle, intervalles de prédiction, haute confiance, causalité, inférence causale, facteur de confusion, catastrophes naturelles, inondation, canicule, évaluation de prédictions, science actuarielle.

Contents

Introduction

Motivation and aims

The timing and magnitude of extreme events often seem unpredictable, and when they strike, they often have disastrous consequences. Essential infrastructures and services such as electrical grids, hospitals, insurances, urban buildings and structures, and internet services can collapse under exceptional demand, load or wear. Financial crashes can cripple markets and trigger recessions. Notable examples include the Great Depression following the Wall Street crash of 1929, and the 2008 financial crisis and Great Recession. Industrial accidents can have severe effects on health and communities. Floods, heatwaves, wildfires, droughts, earthquakes, and hurricanes are environmental extremes that kill, injure, destroy infrastructures and ecosystems, and cause large economic losses. For example, the 2005 floods in Switzerland caused around 3 billion CHF in damages, the loss of lives, and countless injuries. The left-hand photograph in Figure 1 shows a scene from that flooding event in Bern, the Swiss capital, where, despite the seemingly lighthearted picture, the situation was particularly severe. Heatwaves are another frequent and widespread example. The 2021 Pacific Northwest heat dome produced record-shattering temperatures and has been associated with around 1400 excess deaths. During the 2023 humid heatwave in South Asia, the dangerous combination of extreme heat and high humidity led to even more fatalities. Furthermore, a byproduct of extreme temperatures is the increased likelihood of wildfires. Appearing on the front-page of newspapers all around the world after the Australian 'Black Summer' fires, the right-hand photograph of Figure 1 captures, all in a single frame, their disastrous impact on wildlife, homes, and vegetation. As the Intergovernmental Panel on Climate Change (IPCC) documents, extreme environmental events are expected to increase in both frequency and intensity due to climate change, resulting in greater and broader impacts (Seneviratne et al., 2021). Moreover, the impacts of extreme events often fall disproportionately on the most vulnerable communities and ecosystems.

In all contexts impacted by extreme events, accurate forecasts and reliable risk quantification are crucial to reduce these harms, save lives and ecosystems, prevent economic recessions, and reduce financial losses. Clear risk metrics and timely early warnings let policymakers, emergency services, investors, insurers, and communities prepare, allocate resources effectively, and act. The most damaging disasters were often so costly, in part, because their risk was underestimated or not foreseen well enough in advance to allow those mitigating measures. This was for example the case for the 2005 floods in Switzerland (Bezzola and Hegg, 2007).


Figure 1: Photographies. Left: A woman pulls her daughter in a makeshift boat, with the help of a firefighter, through a flooded street in Bern, Switzerland, on 22 August 2005 (credits: Edi Engeler, Keystone). Right: A kangaroo fleeing a wildfire, in front of a burning home, on New Year's Eve 2020, in New South Wales, Australia (credits: Matthew Abbott, New York Times).

Extreme events are inherently hard to foresee. By definition, they are rare and, often, have never yet occurred with the same magnitude in historical records. This scarcity in data requires statistical extrapolation, from observed levels of magnitude, to larger unobserved levels. Moreover, the mechanisms leading to extreme events are often complex. They typically involve many interacting factors, sometimes in time and space, are multivariate, likely non-linear, and often non-stationary or high-dimensional. These properties pose significant statistical challenges for modelling and prediction.

Extreme value statistics provides principled tools for extrapolation to unprecedented events. Its asymptotic theory comprehensively characterizes the behaviour of distribution tails, in a general context with mild assumptions, even with limited data. In particular, for a given probability level or return period of interest (e.g., a one in 100-year event), extreme quantiles or return levels give a concrete estimate of the extreme-event magnitude reached with that frequency (e.g., the height exceeded by the water level, on average, only once every 100 years, or, in other words, exceeded with a 1% probability during any given year). In finance, that estimate is often called the value-at-risk. For a more detailed introduction to extreme value statistics, one can refer to Coles (2001). However, classical extreme value methods are not sufficient to capture complex high-dimensional patterns and make accurate multi-factor risk forecasts.

Machine learning excels at learning complex relationships in high-dimensional data settings. Given many examples of a response variable Y (e.g., a river's water level) together with a vector of predictor variables $\mathbf{X}$ (e.g., the recent upstream rainfall and glacier melt), supervised learning methods can learn the potentially complex relationship between $\mathbf{X}$ and Y to produce accurate conditional predictions for Y when new values of $\mathbf{X}$ are observed (knowing the recent rainfall and glacier melt, how high is the river's water level likely to be?). For a statistical introduction to machine learning, one can refer to James et al. (2021). But standard machine learning methods are not built to extrapolate beyond the range of training data and typically lack reliable uncertainty quantification. When an event is more extreme than the examples seen during training (e.g., the rainfall or the water flow is larger than the examples previously shown to the model), machine learning predictions become unreliable.

In that light, the main aim of this Ph.D. thesis is to create and assess novel methodologies for *forecasting extremes*, as opposed to the traditional use of extreme value theory for *estimating* risk metrics. In practice, classical estimation typically provides a fixed, static estimate of a risk metric and of its uncertainty, based on historical data: for instance, the 100-year return level described previously. Those estimates usually do not vary with current conditions of the potential extreme-event drivers, nor with time, or only vary according to a simplistic dependence model, with inference as an aim rather than making accurate predictions. They are, thus, only suitable for fixed long-term planning, and only if the risk is stationary over time, for example to plan the appropriate height of river banks to avoid most floods, to know the potential loss (value-at-risk) of a given portfolio of financial investments over a fixed time period, or to set insurance premiums. In contrast, the methods developed in this thesis aim to provide accurate conditional predictions of risk metrics, by capturing their potentially complex dependencies on other predictor variables and over time. This is enabled by building novel intersections between extreme value statistics and machine learning, to combine the extrapolation power to unprecedented or low-probability events of the former with the predictive flexibility and accuracy of the latter. On a day following heavy rainfall, the height expected to be reached with 1% conditional probability by water-levels, and the conditional probability that water-levels overflow river banks, are both much higher than after dry weather. Similarly, given the current state of the market, economic indicators, and changes in the portfolio's composition, the value-at-risk and expected shortfall of potential losses might vary significantly. The likelihood of cumulated insurance claims reaching a critical amount may also vary drastically, for example, depending on recent weather, public holidays, the time of year, and their interactions. Accurately forecasting such conditional risks opens the door, for instance, to dynamic short-term risk assessment, early warning systems, and informed adaptive decision-making, contrary to static estimates. This perspective, in our view, initiates a new, modern era of extreme-value methods, that is concerned with prediction and machine learning problems to address extreme event prevention challenges with more accuracy, flexibility, and adaptivity.

To this aim, Chapter 1 introduces a new method, combining the predictive flexibility and architectural versatility of neural networks with the extrapolation capabilities of extreme value statistics, to produce accurate and reliable extreme quantile predictions, when the drivers of extreme events are complex and dependent in time. In fact, the method delivers forecasts for the entire conditional tail distribution of the response Y, allowing the prediction of additional risk metrics such as the probability of exceeding a high threshold (e.g., of a river critically overflowing its banks), or the expected shortfall beyond that threshold. This provides a practical tool for early warning systems and risk management, widely applicable to various situations ranging from meteorology to finance. We show, for instance, that our method's on-day-ahead forecasts would have been able to trigger a warning for the 2005 flood in Bern, contrary to the operational methods used at the time, despite relying on less information.

Stemming from the same aim and motivation, the work presented in Chapter 2 proposes a different type of predictions, namely conformal prediction intervals. Contrary to classical (mean) or quantile regression, that yield point predictions for the conditional mean or quantiles of Y given $\mathbf{X}$, conformal methods provide interval predictions: a range of likely values of Y, given $\mathbf{X}$. Given a pre-specified confidence level, these prediction intervals satisfy finite-sample coverage guarantees,

with remarkably weak assumptions on the data; that is, they contain the true value of Y with the desired confidence probability. As a concrete example, given it rained a lot, the water level is predicted to be, say, between 3.1 and 3.6 meters with 95% confidence. On a dry day, that 95%-confidence prediction interval might be between 2.2 and 2.3 meters. A notable strength of those conformal methods is their versatility, as they can be built on top of any black-box predictor, to provide uncertainty quantification, no matter its complexity. The most adaptive rely on quantile predictions. For those intervals to cover the rare extreme events motivating this thesis, the confidence levels must be correspondingly very high. However, when the target confidence level is too high relative to the sample size, classical conformal methods fail to provide informative intervals. To solve this limitation, we propose a novel conformal procedure, relying on extreme value statistics, to provide informative prediction intervals at those high confidence levels for which classical approaches are unapplicable. This method in part relies on extreme quantile predictions, such as the one introduced in Chapter 1, as base predictors. While the latter aims at accurate extreme quantile prediction, the conformalized quantiles provide predictive regions with coverage guarantees, that are typically more conservative.

A second, closely-related, aim of this thesis, is providing tools for understanding the causal drivers of extreme events; this is the focus of Chapter 3. Although being harder to identify, knowing the causal relationships between variables of interest enables a deeper and more robust understanding of the system at hand and its underlying mechanisms, compared to only knowing their statistical association. Linking to the first predictive aim of this thesis, predictive models built on causal relationships provide more robust and accurate predictions under changing conditions. In our river-flow prediction example, changing conditions may come from climate change, or human interventions such as dams or land-use changes. Causal models also enable counterfactual reasoning, such as assessing the impact of hypothetical interventions. Such a question could be: how would flood risk change if upstream deforestation occurred? In extreme regimes, causal mechanisms sometimes behave differently. Some may even only arise in those regimes, making causal discovery methods specifically designed for extremes essential. Examples could include the causal effects of extreme temperatures on the likelihood of wildfires and mortality rates, of extreme rainfall on landslide occurrences, or of a financial crash on currency valuation. Moreover, some complex causal mechanisms simplify in the tails of distributions, making causal discovery, in some regards, easier during extreme events. The novel methodology presented in Chapter 3 contributes to the intersection between extreme value statistics and causal inference. It proposes a way to mitigate the unwanted effect of confounders, a common challenge in causal studies, on the extreme causal analysis, as well as a permutation test for detecting causal links, tailored to heavy-tailed variables and extreme regimes.

In meteorology, large deep-learning models for global weather forecasting, such as those described in Chapter 4, are becoming increasingly accurate, often outperforming traditional physics-based operational methods in benchmark studies. As their use increases in popularity, they might replace or complement the traditional methods in the future. These large-scale models are typically trained to take as input the state of the atmosphere (i.e., the values of many meteorological variables at every location on a grid covering the entire surface of the earth at different altitudes) and to output predictions of that state a few hours later. These models are then used recursively for longer lead-time predictions. Their accuracy is typically benchmarked based on average performance

metrics. However, as afore argued, extremes are the most critical weather events to predict accurately, which is not ensured by a good average performance. For extreme-event forecasting, they are unlikely to extrapolate well. However, their global scale might potentially be a strength: an unprecedented event at a given location might resemble past extreme events elsewhere in the world, which the model has already learned from. Yet, validation of their reliability and uncertainty for extreme events is challenging, due to the scarcity and fundamental variety of environmental extremes. The study presented in Chapter 4 discusses and assesses deep-learning weather forecasts during three concrete extreme weather events of different nature. We uncover different error patterns, depending on lead-time, between deep-learning and physics-based predictions, as well as limitations to the applicability of the considered deep-learning weather models.

Contributions and thesis structure

Collectively, the scientific contributions of the four articles presented in this thesis range from novel statistical methodologies, applicable to a wide range of domains, to concrete applications to extreme weather event forecasting and flood risk mitigation. They resulted in publications in both statistical and atmospheric-science journals. Moreover, open-source software implementations of the proposed methods have been developed, with the aim of facilitating their practical use, their applicability, and reproducibility.

Parallel to the work selected as part of this Ph.D. thesis, I also contributed to other closely related research projects. These include the identification of biases existing in extreme event attribution studies (Zeder et al., 2023), highlighting the complexity of accurate risk estimation, and motivating the need for conservative approaches, such as the profile-likelihood based uncertainty quantification discussed in Chapter 2. They also include the development of a new framework and mathematical theory for Granger-type extreme causal discovery for time series, in the presence of confounders (Bodik and Pasche, 2024), which is, in aim, an extension of the work presented in Chapter 3 to lagged causal effects. Finally, as part of our competitive data-driven approaches to modelling extreme events (Buriticá et al., 2025), the method described in Chapter 1 achieved the most accurate extreme quantile predictions in the EVA (2023) Conference Data Challenge.

As a summary, the thesis is organized as follows. The first chapter introduces our novel extreme quantile regression method, combining extreme value statistics and neural networks for flexible and accurate risk prediction. It is a postprint of an article published in *The Annals of Applied Statistics* (Pasche and Engelke, 2024). The second chapter presents our extreme conformal method for high-confidence interval predictions. It is a preprint of an invited paper in press for publication in the special issue 'Bridging Heavy Tails and Artificial Intelligence' of *Extremes* (Pasche et al., 2025a). The third chapter details our methodology for mitigating confounding effects in extreme causal discovery and testing for direct causal relationships in extreme regimes. It is a postprint of an article published in *Extremes* (Pasche et al., 2023). The fourth chapter reports our validation study of deep-learning weather model predictions during real extreme events. It is a postprint of an article published in *Artificial Intelligence for the Earth Systems* (Pasche et al., 2025b). Each chapter is self-contained and can be read independently. Some parts, thus, overlap. The appendices and supplements to each article are provided in a second part, as appendix chapters.

Research Articles

1 Neural networks for extreme quantile regression with an application to forecasting of flood risk

Olivier C. Pasche[1], Sebastian Engelke[1]

[1]*Research Institute for Statistics and Information Science, University of Geneva, Switzerland*

This chapter is a postprint of the homonymous article published in *The Annals of Applied Statistics* (Pasche and Engelke, 2024), with doi:10.1214/24-AOAS1907.

Abstract

Risk assessment for extreme events requires accurate estimation of high quantiles that go beyond the range of historical observations. When the risk depends on the values of observed predictors, regression techniques are used to interpolate in the predictor space. We propose the EQRN model that combines tools from neural networks and extreme value theory into a method capable of extrapolation in the presence of complex predictor dependence. Neural networks can naturally incorporate additional structure in the data. We develop a recurrent version of EQRN that is able to capture complex sequential dependence in time series. We apply this method to forecast flood risk in the Swiss Aare catchment. It exploits information from multiple covariates in space and time to provide one-day-ahead predictions of return levels and exceedance probabilities. This output complements the static return level from a traditional extreme value analysis, and the predictions are able to adapt to distributional shifts as experienced in a changing climate. Our model can help authorities to manage flooding more effectively and to minimize their disastrous impacts through early warning systems.

Keywords: extreme value theory, generalized Pareto distribution, machine learning, prediction, recurrent neural network.

1.1 Introduction

Risk assessment is concerned with the analysis of rare events, which have small occurrence probabilities but carry the potential of serious impacts on our health, the environment, or the

economy. Examples of such extreme events are floods in hydrology, crises in the financial system, or heatwaves in a changing climate. In these applications, the quantity of interest is typically a univariate response variable Y representing the random risk factor. The goal is to estimate a quantile $Q(\tau) = F_Y^{-1}(\tau)$ at level $\tau \in [0,1]$, where we denote by F_Y^{-1} the generalized inverse of the distribution function of Y.

Since for risk quantification the level τ is usually very close to 1 so that the quantile $Q(\tau)$ goes beyond the range of the data, the classical approach is to model the tail of the distribution of Y using extrapolation results from extreme value theory. Two main approaches exist. When Y represents, say, a daily quantity, then the generalized Pareto distribution (GPD) can be used to approximate the tail above a high threshold u by (Balkema and de Haan, 1974; Pickands, 1975)

$$\mathbb{P}(Y > y) \approx \mathbb{P}(Y > u)\left(1 + \xi\frac{y-u}{\sigma(u)}\right)_+^{-1/\xi}, \quad y \geq u, \tag{1.1}$$

where $\xi \in \mathbb{R}$ and $\sigma(u) > 0$ are the shape and scale parameters, and the second factor on the right-hand side is the GPD approximation to $\mathbb{P}(Y > y \mid Y > u)$. On the other hand, if Y represents an annual maximum, then the generalized extreme value distribution provides a good fit (Fisher and Tippett, 1928). In hydrology and climate science, risk is often assessed as the T-year return level Q^T, that is, the size of an event that is exceeded on average once every T years. If Y represent a quantity with n_Y independent recordings per year (e.g., $n_Y = 365$ for daily data and $n_Y = 1$ for annual maxima), then the T-year return level is the quantile $Q^T = Q(1 - 1/(n_Y T))$.

Figure 1.1 shows the river catchment of a gauging station on the Aare river in Bern, Switzerland. It is part of the Aare–Rhein basin, where flooding is a major economic and safety concern (Andres et al., 2021). The Swiss Federal Office for the Environment (FOEN) provides recordings of daily average discharges throughout the country. For risk assessment they report the 100-year return levels using the GEV method for annual maxima and the GPD for large daily discharges. The horizontal lines in Figure 1.2 show estimates of this return level at the Bernese station based on data from the years 1930–1958. Both methods give very similar results.

The disadvantage of such an unconditional approach is twofold. First, the return level is static and unable to reflect changes in the size of extreme floods over time, which can occur, for instance, due to climate change. For the Bernese station, for instance, a structural break has been observed in the nineties without a clearly defined cause[1], making classical extreme value modelling challenging. Second, while the return level Q^T is relevant for the construction of long-term flood infrastructure, it can not be used to assess the risk of flooding on a given day. Such forecasting of extreme events is crucial for early warning systems. Indeed, the probability of exceeding on a particular day a given high threshold, say the (constant) 100-year return level, depends on many covariates $\mathbf{X}$ such as the river flows upstream and precipitation in the catchments during the preceding days and weeks.

In this paper, we, therefore, advocate a conditional version of return levels defined as the

[1] See flood report of the FOEN at `https://www.hydrodaten.admin.ch/en/2135.html`.

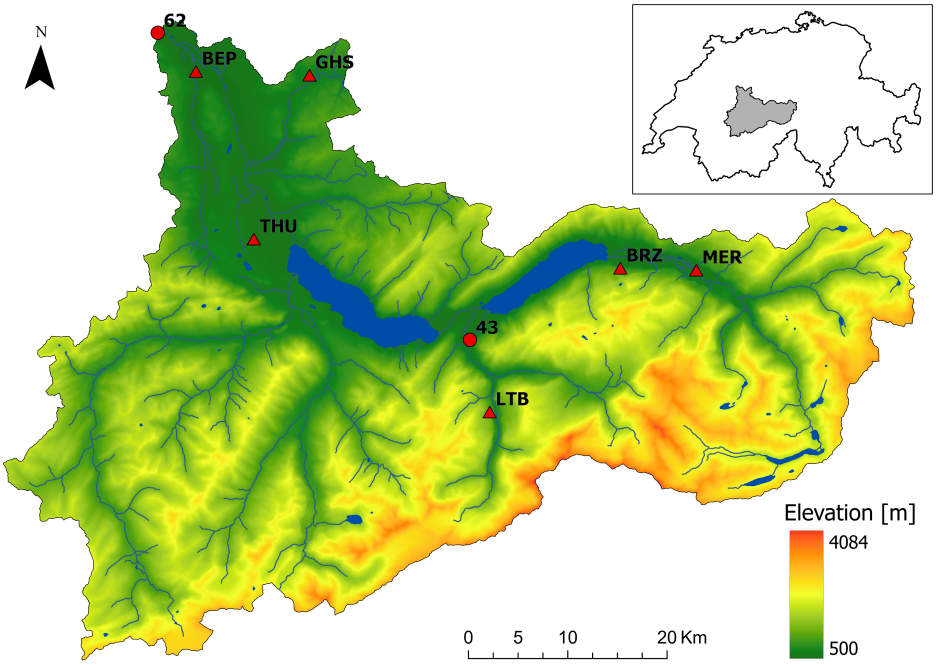

Figure 1.1: Topographic map of water catchment of the gauging station in Bern–Schönau (62) on the Aare in Switzerland. Another gauging station upstream in Gsteig (43) on the Lütschine river and six meteorological stations with precipitation measurements (triangles) are also shown.

conditional quantile of Y given a vector of observed covariates $\mathbf{X} = \mathbf{x}$, that is,

$$Q_{\mathbf{x}}(\tau) = F^{-1}_{Y|\mathbf{X}=\mathbf{x}}(\tau). \tag{1.2}$$

The interpretation of such a conditional T-year return level $Q^T_{\mathbf{x}} = Q_{\mathbf{x}}(1 - 1/(n_Y T))$ is different from the unconditional return level Q^T. Since $Q^T_{\mathbf{x}}$ depends on the exact configuration of the covariates $\mathbf{x}$, one can see a conditional T-year return level as the size of an event that is exceeded in average once every T years, if the covariate vector $\mathbf{X}$ of all observations of Y had the same value $\mathbf{x}$. A more precise interpretation is to see $Q^T_{\mathbf{x}}$ as the value that is exceeded in the next time step with probability $1/(n_Y T)$, and we refer to it as the conditional quantile with return period T years; for a comprehensive discussion of return levels and quantiles, see Bücher and Zhou (2021).

The top panel of Figure 1.2 shows one-day-ahead forecasts of such conditional 100-year quantiles $Q^{100}_{\mathbf{x}}$ for the Bernese river data from the method of this paper, fitted to the years 1930–1958. We see that, as opposed to the unconditional return level Q^{100}, the conditional quantile changes from day to day, depending on past precipitation and river flows. In fact, on August 21, the day before the first exceedance of the unconditional Q^{100} during the serious flood of August 2005 in Bern, the size of the conditional 100-year event predicted for the next day was much higher than on other days, and the forecasted conditional probability of such an exceedance (bottom panel of Figure 1.2) was 920 times larger than the static 100-year probability. Both outputs can be used as triggers for early warnings and additional flood management measures. In particular, the days when an exceedance above Q^{100} is likely can be effectively pinpointed thanks to their temporal sparsity in the probability forecast.

Forecasting extreme events is notoriously difficult due to the low occurrence probabilities involved.

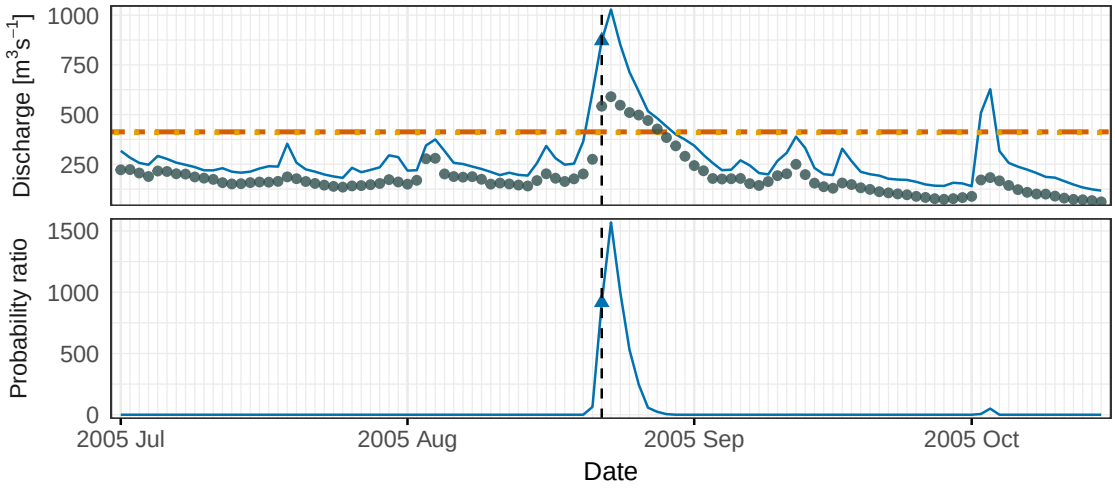

Figure 1.2: Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead EQRN forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead EQRN forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability. The vertical line indicates August 22, the day of the first exceedance.

This is particularly true for hydrological models, which are typically used by national agencies and which do not use explicit tail extrapolation. In the aftermath of the 2005 flood, the FOEN published a detailed analysis of the internal forecasting procedures during this event (Bezzola and Hegg, 2007). The report showed that too-late warnings lead to more severe consequences since forecasters did not trust the predicted precipitation amounts during this extreme scenario and underestimated the flood risk. This gives another motivation why our statistical model and its output in Figure 1.2 could have been a helpful tool during this flood.

Conditional quantiles (1.2) are studied in the field of quantile regression, where many flexible methods exist (Meinshausen, 2006; Cannon, 2011; Zhang et al., 2019; Athey et al., 2019). While they work well for quantile levels within the data range, they break down for extreme values of τ close to 1. Such extreme quantile regression relies on extrapolation results as in (1.1), where extreme value parameters, such as the scale $\sigma(\mathbf{x})$ and shape $\xi(\mathbf{x})$, may be modelled as functions of the covariates through linear models (Wang et al., 2012; Li and Wang, 2019), generalized additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), or kernel methods (Daouia et al., 2011; Gardes and Stupfler, 2019; Velthoen et al., 2019). To overcome limitations of additive and kernel-based methods in higher dimensions, more recently, flexible tree-based methods have been combined with the GPD extrapolation for predicting extreme conditional quantiles (Velthoen et al., 2023; Gnecco et al., 2024) or predictive tail distributions (Koh, 2023) on complex data. Tree-based methods have the advantage of requiring little tuning for good prediction performance. However, they can not incorporate additional structure of the data as encountered in time series or spatial applications.

The goal of our work is to combine ideas from extreme value theory and machine learning to propose an extreme quantile regression model that has the ability to extrapolate in the direction of the response Y and to model complex covariate dependencies in the predictors $\mathbf{X}$. We propose an extreme quantile regression network (EQRN) that models covariate-dependent GPD parameters

$\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ as outputs of a neural network. Conditional quantile estimates at the desired extreme level are then readily derived from the estimated conditional tail distribution. Neural networks are known for their ability to model complex dependencies and to approximate any measurable function arbitrarily well (Hornik, 1991). The second advantage is versatility. The deep learning literature is rich in network architectures, activation functions, and regularization methods. In particular, convolutions produce shift-invariant models for covariates with spatial dependencies, and recurrent architectures provide models for sequentially dependent observations such as time series. In our application of flood forecasting with time-dependent data, recurrent neural networks (Werbos, 1988; Elman, 1990) are of particular interest.

The main output of our EQRN model for sequentially dependent data is the one-day-ahead risk forecast, either in terms of conditional quantiles or exceedance probabilities. Thanks to the GPD approximation, the method can extrapolate beyond the range of the data, as illustrated in Figure 1.2. The strength of our recurrent model lies in the ability to exploit information from multiple covariates and capture complex time dependence. It is, therefore, an effective early warning tool even for unprecedented, record-shattering events, like the 2005 floods. This is of particular importance in a nonstationary system, where climate change makes extreme events increasingly likely (Fischer et al., 2021).

The main contributions of the paper are threefold. First, our EQRN method is the first to model the conditional GPD parameters through neural networks. Thanks to the large number of neural network architectures, this expands the range of possible applications for extreme quantile regression to new areas. While we concentrate on sequentially dependent data, our method can be used, for instance, in combination with convolutional or graphical neural networks for spatial covariates. Second, a major technical contribution is to make neural networks applicable in the extreme value context. To stabilize the prediction of extreme quantiles in these highly flexible regression methods, we propose to use an orthogonal reparametrization of the GPD deviance, a suitable choice of activation functions, and the use of the intermediate quantiles as additional covariates; the latter is a new idea that seems to also improve other extreme quantile methods. Finally, a main novelty is the application to flood forecasting and the notion of conditional return levels that can be used as early warnings. With this perspective our method enables applications in many other areas, such as the one-day-ahead forecasting of the value-at-risk or of the expected shortfall in financial time series.

The paper is organized as follows. In Section 1.2 we provide background on quantile regression, extreme value theory, and neural networks. We propose our EQRN model for both independent and sequentially dependent data in Section 1.3. Section 1.4 contains a simulation study to assess the performance of our approach in comparison to existing methods. In Section 1.5 we describe the Swiss river data, apply our methodology to forecast flood risk and discuss the implications of the results. Section 1.6 concludes with a brief discussion.

1.2 Background

1.2.1 Quantile regression

In the classical quantile regression setup, we observe an independent and identically distributed sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of the random vector $(\mathbf{X}, Y)$, where Y is the real-valued response variable and $\mathbf{X}$ is a vector of p covariates (or predictors). One aims at predicting the conditional quantile $Q_{\mathbf{x}}(\tau)$ defined in (1.2) of Y, given $\mathbf{X} = \mathbf{x}$, for some predictor value $\mathbf{x} \in \mathbb{R}^p$ and probability level $\tau \in (0, 1)$ of interest.

Analogously to regression that minimizes the mean squared error, quantile regression minimizes the quantile loss,

$$Q_{\mathbf{x}}(\tau) = \arg\min_q \mathbb{E}[\rho_\tau(Y - q) \mid \mathbf{X} = \mathbf{x}], \tag{1.3}$$

where $\rho_\tau(t) := t(\tau - \mathbb{1}_{\{t<0\}})$ is the quantile check function (Koenker and Bassett, 1978). Many parametric and nonparametric quantile regression models exist, including linear models (Chernozhukov, 2005), random forests (Athey et al., 2019), and neural networks (Cannon, 2011; Zhang et al., 2019). They yield a conditional quantile estimate by minimizing the empirical quantile loss over the training sample $\mathcal{D}$, that is,

$$\hat{Q}_{\mathbf{x}}(\tau) = \arg\min_{q_\tau \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - q_\tau(\mathbf{x}_i)), \tag{1.4}$$

where $\mathcal{M}$ is the set of possible quantile functions $q_\tau(\cdot)$ characterized by the model.

Classical methods for quantile regression that rely on the quantile loss (1.3) perform well for "moderate" probability levels τ. To define what that means exactly, we typically let τ_n depend on the sample size n. The expected number of exceedances of y_i over the respective conditional quantile $Q_{\mathbf{x}_i}(\tau_n)$, $i = 1, \ldots, n$, is given by $n(1 - \tau_n)$. With moderately extreme, or intermediate, we refer to a sequence $\tau_n \to 1$ with $n(1 - \tau_n) \to \infty$, meaning that the quantile goes to the upper endpoint of the distribution, but there are more and more exceedances with growing sample size n. On the other hand, we call a quantile level $\tau_n \to 1$ extreme if $n(1 - \tau_n) \to c \in [0, \infty)$; that is, there are finitely many, or possibly zero, exceedances over $Q_{\mathbf{x}_i}(\tau_n)$ in the sample. In this situation, classical quantile regression methods do not perform well due to the scarcity of observations in the tail of the response.

The left panel of Figure 1.3 illustrates this issue for a sample $y_1, \ldots, y_n$ with $n = 1000$ and no covariates. The dashed line shows for different quantile levels τ_n the empirical quantile estimates, obtained by solving the respective quantile loss function without covariates. It can be seen that as soon as the number of exceedances $n(1 - \tau_n) < 1$, that is, $\tau_n > 99.9\%$, there is a significant bias compared to the true quantiles (solid line). The reason is that the empirical estimates can not predict higher than the largest observation. When covariates are present, this issue persists, since quantile regression relies on solving the empirical quantile loss (1.4).

In the sequel we omit the dependence on n and write τ instead of τ_n. Intermediate quantile levels will be denoted by τ_0.

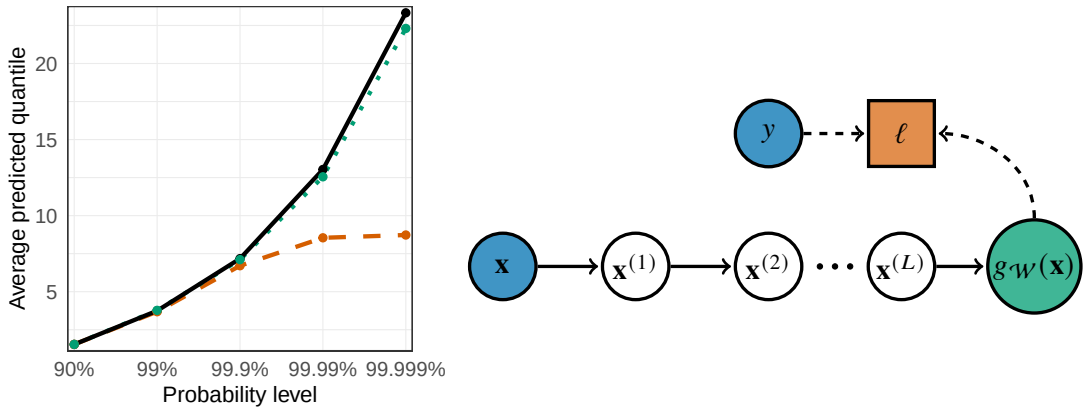

Figure 1.3: Left: True τ-quantiles (solid line) compared to empirical estimates (dashed line) and GPD based estimates (dotted line) for moderate to extreme probability levels (log-scale) for sample size 1,000. Estimates are averages over 100 trials. Right: Multilayer perceptron flowchart from input $\mathbf{x}$ to output $g_{\mathcal{W}}(\mathbf{x})$, with loss function ℓ and corresponding response y.

1.2.2 Generalized Pareto distribution

In order to predict well on extreme quantiles, a method should rely on asymptotic results from extreme value theory for accurate extrapolation beyond the sample. In particular, we rely on the generalized Pareto distribution (GPD). In the presence of covariates, it arises as an approximation of the tail of the distribution of $Y \mid \mathbf{X} = \mathbf{x}$. More precisely, we use a conditional version of (1.1),

$$\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x}) \approx (1 - \tau_0) \left(1 + \xi(\mathbf{x}) \frac{y - u(\mathbf{x})}{\sigma(\mathbf{x})}\right)_+^{-1/\xi(\mathbf{x})}, \quad y > u(\mathbf{x}), \tag{1.5}$$

where the threshold $u(\mathbf{x})$ is chosen as an intermediate quantile $Q_{\mathbf{x}}(\tau_0)$ at level $\tau_0 \in (0,1)$ close to 1, and the shape $\xi(\mathbf{x}) \in \mathbb{R}$ and scale $\sigma(\mathbf{x}) > 0$ depend on the covariates; here we omit the dependence of $\sigma(\mathbf{x})$ on the intermediate level τ_0 in the notation. This approximation holds under weak conditions on the tail of $Y \mid \mathbf{X} = \mathbf{x}$; see Balkema and de Haan (1974); Pickands (1975) for the precise statement. This condition is of univariate nature, and it can, therefore, be verified even in more complex situations, for instance, where $\mathbf{X}$ represents the history of a multivariate time series; see Section 1.4.1 for details. The shape parameter $\xi(\mathbf{x})$ is important since it encodes the tail heaviness of the response: if it is positive, the response has a heavy-tailed distribution such as Pareto or Student-t; if it is zero, the response is light-tailed such as a Gaussian or exponential; if it is negative, then the response has a finite upper endpoint.

In order to predict an extreme quantile at level $\tau > \tau_0$ from approximation (1.5), we can invert this expression to find

$$Q_{\mathbf{x}}(\tau) := Q_{\mathbf{x}}(\tau_0) + \frac{\sigma(\mathbf{x})}{\xi(\mathbf{x})} \left[\left(\frac{1 - \tau_0}{1 - \tau}\right)^{\xi(\mathbf{x})} - 1\right]. \tag{1.6}$$

This shows that an estimate $\hat{Q}_{\mathbf{x}}(\tau)$ of an extreme quantile requires estimates of the intermediate quantile $\hat{Q}_{\mathbf{x}}(\tau_0)$ and of the conditional GPD parameters $\hat{\xi}(\mathbf{x})$ and $\hat{\sigma}(\mathbf{x})$ as functions of the predictor vector. For the intermediate quantile function, we can use any of the existing methods for quantile regression since they work well for this moderate quantile level, as discussed above. Estimation

of the GPD parameters can be done by specifying a parametric or nonparametric model. We will use neural networks for this purpose, which are introduced in the next section.

The green line in Figure 1.3 shows estimates $\hat{Q}(\tau)$ for different quantile levels τ, using the approximation (1.6) without covariate dependence, with empirical intermediate quantile at $\tau_0 = 90\%$ and GPD parameters estimated with maximum likelihood. It can be seen that the extrapolation solves the bias issue of empirical methods.

1.2.3 Neural networks and conditional density estimation

The literature on neural networks is vast, and existing methods are being improved constantly. We concentrate in this section on well-established techniques that are most relevant for our purpose of modelling extreme quantiles.

A multilayer perceptron (MLP) or fully-connected feed-forward neural network model is a parametric family of nonlinear functions $g_{\mathcal{W}} : \mathbb{R}^p \to \mathbb{R}^q$ that map a p-dimensional input $\mathbf{x}$ to a q-dimensional output by

$$\mathbf{x} \mapsto \mathbf{x}^{(L+1)}, \text{ with } \mathbf{x}^{(l)} = \sigma^l \left(W^l \mathbf{x}^{(l-1)} + b^l \right) \forall l = 1, \dots, L+1, \tag{1.7}$$

where $\mathbf{x}^{(0)} = \mathbf{x}$. The number of hidden layers $L \in \mathbb{N}$, the hidden layer dimensions $h_1, \dots, h_L \in \mathbb{N}$, and the choice of activation functions $\sigma^l : \mathbb{R}^{h_l} \to \mathbb{R}^{h_l}, l = 1, \dots, L+1$ (with $h_0 = p$ and $h_{L+1} = q$) are hyperparameters that need to be chosen, for instance, by cross-validation. The set of trainable parameters to be inferred from data contains all weights and bias terms of the network, that is, $\mathcal{W} = \left\{ (W^l, b^l); l = 1, \dots, L+1 \right\}$, with $W^l \in \mathbb{R}^{h_l \times h_{l-1}}$ and $b^l \in \mathbb{R}^{h_l}$. Figure 1.3 shows a schematic illustration of the transformations inside the MLP.

In the general setting, p is the number of features or covariates considered in the model, and q depends on the task at hand. In order to train a model, a loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is required that maps a tuple $(y, g_{\mathcal{W}}(\mathbf{x}))$ of response and prediction to a positive number quantifying their discrepancy. Common tasks include mean regression with $q = 1$ and squared error loss, quantile regression with $q = 1$ and quantile loss, and classification with q equal to the number of possible classes and cross-entropy as loss.

For conditional density estimation, or distribution regression, we suppose that Y follows a distribution with parametric probability density $f_Y(\cdot; \theta)$ and parameter $\theta = \theta(\mathbf{x})$ depending on the vector $\mathbf{X} = \mathbf{x}$. Conditional density estimation networks are neural networks that aim at outputting conditional estimates $g_{\mathcal{W}}(\mathbf{x}) = \theta(\mathbf{x})$ based on realizations of $\mathbf{X} = \mathbf{x}$ as input (e.g., Cannon, 2012). In this setting, p is the dimension of $\mathbf{X}$, and q is the dimension of θ. The loss function is the deviance or negative log-likelihood loss $\ell(y, \theta(\mathbf{x})) = -\log f_Y(y; \theta(\mathbf{x}))$.

To train a neural network on the training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we find the optimal parameter values minimizing the average empirical loss, that is,

$$\hat{\mathcal{W}} \in \arg\min_{\mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, g_{\mathcal{W}}(\mathbf{x}_i)). \tag{1.8}$$

This is generally achieved via backpropagation using mini-batches and optimization algorithms, such as the well-performing gradient descent variants (Kingma and Ba, 2014; Tieleman and Hinton, 2012; Duchi et al., 2011). Since neural networks are typically overparameterized, overfitting has to be prevented with regularization methods, such as L_2 weight penalties for narrow networks and dropout (Srivastava et al., 2014) for deeper architectures. As the optimization problem (1.8) is often nonconvex, local-minima convergence is an issue. Restricting training to a subset of $\mathcal{D}$ and keeping the rest to track the validation loss at the end of each epoch helps to avoid local minima by learning rate decay. Restarting training with different initializations and keeping the best fit in terms of validation loss often leads to lower minima. Early stopping based on the validation loss is another measure against overfitting. The final validation loss is used for model selection and the choice of optimal hyperparameters; for more details on the fitting of neural networks, see Goodfellow et al. (2016).

When observations are dependent in space or time, generalizations of the MLP exist to account for these particular structures. Convolutional and graph neural networks exploit neighbourhood information with parsimonious architectures that are effective for images, graphs, or spatial observations (LeCun et al., 2015; Scarselli et al., 2009). We concentrate here on methods for sequential dependence that arises typically in time series $\{(\mathbf{X}_t, Y_t)\}_{t=1}^T$. For this type of data, recurrent architectures of the network allow capturing dependence between observations. A simple recurrent neural network (RNN) layer (Werbos, 1988; Elman, 1990) takes as input a vector $\mathbf{x}_t$ and outputs the hidden recurrent state vector

$$\mathbf{h}_t = \tanh(W_{\mathrm{xh}}\mathbf{x}_t + W_{\mathrm{hh}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{h}}),$$

depending both on $\mathbf{x}_t$ and the hidden state $\mathbf{h}_{t-1}$ recursively resulting from the previous inputs $\mathbf{x}_{t-1}$ and $\mathbf{h}_{t-2}$ in the sequence. Here and in the sequel, the bias vectors $b_\cdot$ and weight matrices $W_{\cdot\cdot}$, indexed by the input and output variables, are the trainable parameters. This model has then been improved by the addition of a gating cell state $\mathbf{c}_t$, to avoid vanishing gradient issues, as well as a forget gate $\mathbf{f}_t$, an input gate $\mathbf{i}_t$, and output gate $\mathbf{o}_t$, to control both short- and long-term dependencies in the sequence. This yields the long short-term memory (LSTM) layer (Hochreiter and Schmidhuber, 1997; Gers et al., 2000, 2003; Jozefowicz et al., 2015)

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_{\mathrm{xi}}\mathbf{x}_t + W_{\mathrm{hi}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{i}}), & \mathbf{f}_t &= \sigma(W_{\mathrm{xf}}\mathbf{x}_t + W_{\mathrm{hf}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{f}}),\\
\mathbf{g}_t &= \tanh(W_{\mathrm{xg}}\mathbf{x}_t + W_{\mathrm{hg}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{g}}), & \mathbf{o}_t &= \sigma(W_{\mathrm{xo}}\mathbf{x}_t + W_{\mathrm{ho}}\mathbf{h}_{t-1} + \mathbf{b}_{\mathrm{o}}),\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, & \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t),
\end{aligned} \tag{1.9}$$

where $\sigma(z) = 1/(1+\exp(-z))$ is the sigmoid activation and $\odot$ is the Hadamard (or componentwise) product; see Figure 1.4 for an illustration. The common dimension of the vectors defined in (1.9), is a hyperparameter of the layer. The input of this layer $\tilde{\mathbf{x}}_t := (\mathbf{x}_{t-s}, \ldots, \mathbf{x}_{t-1})$ can include predictors from the past to model longer dependencies, where $s \in \mathbb{N}$ determines the time horizon.

The LSTM model has been simplified by Cho et al. (2014) into the gated recurrent unit (GRU) layer, which has become a popular alternative. A multilayer recurrent network is obtained by considering the $\mathbf{h}_t$ as a sequence of inputs for the following recurrent layers; see Figure A.1 in Supplementary Material A.1. Usually, a fully connected layer as in (1.7) is used to map the hidden state of the final recurrent layer to the network output $\tilde{g}_{\mathcal{W}}(\tilde{\mathbf{x}}_t)$.

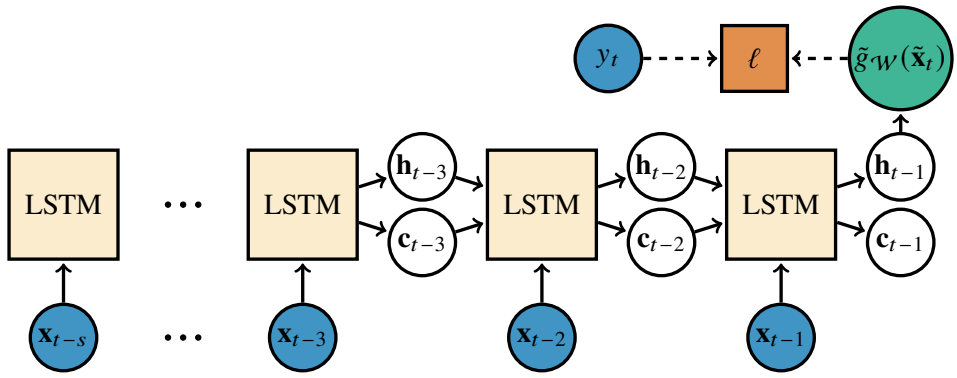

Figure 1.4: Single-layer LSTM network flowchart from input $\tilde{\mathbf{x}}_t := (\mathbf{x}_{t-s}, \ldots, \mathbf{x}_{t-1})$ to output $\tilde{g}_{\mathcal{W}}(\tilde{\mathbf{x}}_t)$, with loss evaluation. The LSTM cells represent the transformation in (1.9).

1.3 Extreme quantile regression neural networks

In this section we propose a new methodology that combines the extrapolation power of the GPD model with the high-dimensional predictor space capabilities and flexibility of neural networks to obtain accurate estimates for quantile functions $Q_{\mathbf{x}}(\tau)$ at extreme levels τ. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training dataset. Estimation of conditional extreme quantiles $\hat{Q}_{\mathbf{x}}(\tau)$, using (1.6), requires estimators for the intermediate quantile function $\hat{Q}_{\mathbf{x}}(\tau_0)$ with $\tau_0 < \tau$ and the GPD parameter $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$. It is customary to proceed in two steps.

First, we model the intermediate quantile at level τ_0 using classical quantile regression methods. We then define the conditional exceedances

$$z_i := y_i - \hat{Q}_{\mathbf{x}_i}(\tau_0), \quad i \in \mathcal{I} := \{i = 1, \ldots, n : \ y_i > \hat{Q}_{\mathbf{x}_i}(\tau_0)\}.$$

The intermediate probability τ_0 should be chosen low enough to allow for stable estimation of $Q_{\mathbf{x}}(\tau_0)$ with classical empirical methods, but high enough for the approximation in (1.5) to be accurate, so that the exceedances z_i are approximate samples of a GPD. However, it is not a classical tuning parameter, since different values for τ_0 yield different subsets of exceedances $\mathcal{I}$. Comparison of the loss function (1.11) on these datasets would, therefore, not be meaningful. Instead, the threshold is, in the univariate case, usually selected in terms of stability plots and sensitivity analyses.

In the second step, we estimate the GPD parameters $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ based on the set of exceedances z_i, $i \in \mathcal{I}$. Modelling these parameters directly in the extrapolation formula (1.6) may lead to strong dependence between the estimates and numerical instabilities. We, therefore, rely on an orthogonal reparametrization that has a diagonal Fisher information matrix. As for the standard asymptotic GPD likelihood properties, the latter is well-defined for the GPD model, when $\xi(\mathbf{x}) > -0.5$, and the reparametrization

$$(\sigma(\mathbf{x}), \xi(\mathbf{x})) \mapsto (\nu(\mathbf{x}), \xi(\mathbf{x})), \qquad \nu(\mathbf{x}) := \sigma(\mathbf{x})(\xi(\mathbf{x}) + 1)$$

yields the desired orthogonality (Cox and Reid, 1987; Chavez-Demoulin and Davison, 2005). In our experiments this reparametrization significantly improves stability and convergence in every

considered setting.

In this section we propose a flexible neural network model for the orthogonalized GPD parameters $\nu(\mathbf{x};\mathcal{W})$ and $\xi(\mathbf{x};\mathcal{W})$, where $\mathcal{W}$ denotes the collection of all model parameters. This can be seen as conditional density estimation with output dimension $q = 2$, where the parametric family is the GPD model with parameters $\theta = (\nu, \xi)$ depending on the covariate $\mathbf{X} = \mathbf{x}$. In general, an estimate of the model parameters $\hat{\mathcal{W}}$ is thus obtained as a minimizer of the GPD deviance loss over the training exceedances

$$\hat{\mathcal{W}} \in \arg\min_{\mathcal{W}} \sum_{i \in \mathcal{I}} \ell_{\mathrm{OGPD}}\{z_i; \hat{\nu}(\mathbf{x}_i;\mathcal{W}), \hat{\xi}(\mathbf{x}_i;\mathcal{W})\}, \tag{1.10}$$

where the deviance or negative log-likelihood of the GPD in terms of the orthogonal reparametrization is

$$\ell_{\mathrm{OGPD}}(z;\nu,\xi) = \left(1 + \frac{1}{\xi}\right) \log\left\{1 + \xi\frac{(\xi+1)z}{\nu}\right\} + \log(\nu) - \log(\xi+1). \tag{1.11}$$

This yields a flexible model for the conditional tail distribution of $Y \mid \mathbf{X} = \mathbf{x}$ with which not only $\hat{Q}_{\mathbf{x}}(\tau)$ can be regressed but also conditional exceedance probabilities over a high threshold or conditional expected shortfalls, for example.

In the next two subsections, we discuss the details of the model for independent observations and for time series data with sequential dependence, respectively.

1.3.1 Independent observations

We first consider the case where the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is a set of independent, identically distributed observations of $(\mathbf{X}, Y)$. The goal in this case is the estimation of the conditional quantile $Q_{\mathbf{x}}(\tau)$ for a predictor value $\mathbf{X} = \mathbf{x}$ at an extreme level $\tau > 0$. For the first step of estimating the intermediate quantile function, in principle, any classical quantile regression method can be used. To avoid overfitting and obtain unbiased generalization error estimates from the training set $\mathcal{D}$, the predicted $\hat{Q}_{\mathbf{x}_i}(\tau_0)$, $i = 1, \ldots, n$, should be constructed out of training sample. This is achievable in two ways. Using bagging methods, such as generalized random forests (Athey et al., 2019), is a convenient choice since they allow for out-of-bag predictions where only a single fit on $\mathcal{D}$ is required. For other methods such as quantile regression neural networks (Cannon, 2011), out of training sample predictions can be obtained in a foldwise manner similar to cross-validation. In the sequel we assume that the intermediate quantile, and thus the exceedances, are given.

In the second step, we propose to model the orthogonalized GPD parameters $\nu(\mathbf{x})$ and $\xi(\mathbf{x})$ by a fully-connected feed-forward neural network with parameter vector $\mathcal{W}$ and deviance loss function as in (1.10); see Section 1.2.3 for details. Choices of the network architecture such as the number of neurons, the number of layers and activation functions, are hyperparameters, denoted by Θ, to be selected. We provide sensible default values in our implementation, but one can also choose them in a data-driven way based on a validation set.

The only restrictions are on the output activation functions, since $\nu(\mathbf{x})$ should be strictly positive.

We find the exponential function or the SELU activation (Klambauer et al., 2017) shifted above zero to be good choices. Regarding the output activation for ξ, no strict restrictions apply and the identity would be a natural choice. However, standard likelihood regularity properties are not satisfied for the GPD model when $\xi \leq -0.5$, which very rarely occurs in practice. We observe that smoothly restricting the shape estimates, for example, between -0.5 and 0.7 with the activation $x \mapsto 0.6\tanh(x) + 0.1$, helps to improve training stability. This avoids aberrant ξ estimates in the early stages of the training and still covers almost all practical cases. In many situations it is reasonable to assume that only the scale $\nu(\mathbf{x})$ varies locally but $\xi(\mathbf{x}) \equiv \xi$ is constant (e.g., Kinsvater et al., 2016). This can be achieved by restricting the network so that the shape output only depends on a bias term.

Algorithm 1 summarizes our extreme quantile regression network (EQRN) for independent observations, which takes the intermediate quantiles and training data as input and outputs the extreme quantile at a desired test predictor value $\mathbf{x} \in \mathbb{R}^p$ and level $\tau > \tau_0$. Optionally, the conditional GPD parameters can also be obtained.

The conditional GPD estimation in the second step relies on the exceedances to ensure that only information from the tail is used for extrapolation. There may, however, be residual information in the moderately extreme observations that should not be discarded. We propose to use the intermediate quantiles $\hat{Q}_{\mathbf{x}_i}(\tau_0)$ as an additional feature in the conditional density estimation. This feature engineering seems to consistently and significantly improve the accuracy of the final prediction $\hat{Q}_{\mathbf{x}}(\tau)$ on test data in our simulations. This idea is, to some degree, related to stacked learning (Breiman, 1996; Wolpert, 1992) and is achieved by considering the $(p+1)$-dimensional vector $(\mathbf{x}_i, \hat{Q}_{\mathbf{x}_i}(\tau_0))$, $i = 1, \ldots, n$, instead of $\mathbf{x}_i$ as predictors, for the network input. To avoid overfitting, it is again important that the intermediate quantile estimates are constructed out of training sample. This new feature also improves other extreme quantile regression methods, such as the GBEX model (Velthoen et al., 2023), as observed in our simulations in Section 1.4.

Classical backpropagation and optimization procedures are performed to find $\hat{\mathcal{W}}$. To avoid overfitting and local minima convergence, the validation loss can be tracked as discussed in Section 1.2.3. The hyperparameters Θ of this network include choices of optimization algorithms and regularization, for instance. More details on the function calls in the algorithm can be found in Supplementary Material A.2.

Since the true quantile $Q_{\mathbf{x}}(\tau)$ is unknown in real-world data, we can not assess the performance of $\hat{Q}_{\mathbf{x}}(\tau)$ with metrics such as mean squared or absolute errors. As illustrated in Section 1.2.1, the quantile loss is also unreliable due to the data scarcity at extreme quantiles. We, therefore, choose the final validation loss based on the GPD deviance to compare different choices of hyperparameters, as it is the most reliable surrogate metric.

1.3.2 Sequential dependence

In many applications the observations are not independent but display sequential dependence, such as in time series. In this case we denote the training data by $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^T$, which are observed sequentially from a time series $\{(\mathbf{X}_t, Y_t)\}_{t=1}^T$. The goal here is different from the case of

Algorithm 1 EQRN for independent observations

The tuning parameters Θ for the conditional GPD density estimation network $g_{\mathcal{W}}$ and the intermediate quantile model $\hat{Q}_\cdot(\tau_0)$ capable of out-of-sample prediction are prespecified. The training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and test covariates $\mathbf{x}$ are observed. Let $\tau \in (\tau_0, 1)$ be the desired probability level.

procedure EQRN-Fit($\mathcal{D}$, Θ, $\hat{Q}_\cdot(\tau_0)$)
  $\mathcal{I} \leftarrow \{i = 1, \dots, n : y_i > \hat{Q}_{\mathbf{x}_i}(\tau_0)\}$
  $z_i \leftarrow y_i - \hat{Q}_{\mathbf{x}_i}(\tau_0) \ \ \forall i \in \mathcal{I}$
  $\mathcal{T}, \mathcal{V} \leftarrow$ RandomValidationSplit($\mathcal{I}$) ▷ If no validation: $\mathcal{T} = \mathcal{I}, \mathcal{V} = \emptyset$
  $\hat{\mathcal{W}} \leftarrow$ InitializeNetworkWeights(Θ)
  for $e = 1$ to maximum number of epochs E **do**
    for all $\mathcal{B} \in$ GetMiniBatches($\mathcal{T}$) **do**
      $\{(\hat{\nu}_i, \hat{\xi}_i)\}_{i \in \mathcal{B}} \leftarrow g_{\hat{\mathcal{W}}}(\mathbf{x}_\mathcal{B}, \hat{Q}_{\mathbf{x}_\mathcal{B}}(\tau_0))$
      $\ell \leftarrow \sum_{i \in \mathcal{B}} \ell_{\text{OGPD}}(z_i, \hat{\nu}_i, \hat{\xi}_i)/|\mathcal{B}|$
      $\hat{\mathcal{W}} \leftarrow$ BackPropUpdate(ℓ, $\hat{\mathcal{W}}$, $\mathbf{x}_\mathcal{B}$, $\hat{Q}_{\mathbf{x}_\mathcal{B}}(\tau_0)$, Θ)
    stop if $\mathcal{V} \neq \emptyset$ and LossNotImproving($\hat{\mathcal{W}}$, $\mathbf{x}_\mathcal{V}$, $\hat{Q}_{\mathbf{x}_\mathcal{V}}(\tau_0)$, $z_\mathcal{V}$)
  output $\hat{\mathcal{W}}$

procedure EQRN-Predict($\mathbf{x}$, τ, $\hat{\mathcal{W}}$, $\hat{Q}_\cdot(\tau_0)$)
  $\{\hat{\nu}(\mathbf{x}), \hat{\xi}(\mathbf{x})\} \leftarrow g_{\hat{\mathcal{W}}}(\mathbf{x}, \hat{Q}_\mathbf{x}(\tau_0))$
  $\hat{\sigma}(\mathbf{x}) \leftarrow \hat{\nu}(\mathbf{x})/\{\hat{\xi}(\mathbf{x}) + 1\}$
  compute $\hat{Q}_\mathbf{x}(\tau)$ w.r.t. $\hat{\sigma}(\mathbf{x})$, $\hat{\xi}(\mathbf{x})$, $\hat{Q}_\mathbf{x}(\tau_0)$, τ and τ_0 using equation (1.6)
  output $\hat{Q}_\mathbf{x}(\tau)$, and optionally $\{\hat{\sigma}(\mathbf{x}), \hat{\xi}(\mathbf{x})\}$

independent observations. Indeed, we would like to predict as well as possible high quantiles of the response Y_u at some time point u one step in the future based on all past information $\tilde{\mathbf{X}}_u := \{(\mathbf{X}_t, Y_t)\}_{t<u}$. Therefore, the target is

$$Q_{\tilde{\mathbf{x}}_u}(\tau) := F^{-1}_{Y_u|\tilde{\mathbf{X}}_u = \tilde{\mathbf{x}}_u}(\tau), \tag{1.12}$$

where $\tilde{\mathbf{x}}_u := \{(\mathbf{x}_t, y_t)\}_{t<u}$ are observations that are not necessarily part of the training set. In this section we propose a recurrent neural network to solve this task. Several principles are the same as in the case of independent observations, such as the choice of output activation functions. We thus focus on the differences to the independent case.

While in principle it is still possible to use any classical quantile regression method to model the intermediate conditional quantiles at level τ_0, we recommend using quantile regression neural networks (Cannon, 2011; Zhang et al., 2019) with a recurrent architecture. These models are specifically designed for sequential dependence and can easily adapt to varying sequence lengths of the input features $\tilde{\mathbf{X}}_t$. In our experiments, recurrent quantile regression neural networks consistently outperformed generalized random forests in the presence of sequential dependence. Although a varying sequence length of input features is possible, we restrict this length to a fixed horizon $s \ll T$ for computational efficiency. Thus, for any time point t, we define its past by $\tilde{\mathbf{x}}_t = \{(\mathbf{x}_j, y_j)\}_{j=t-s}^{t-1}$. For simplicity, we denote our augmented training set by $\tilde{\mathcal{D}} = \{(\tilde{\mathbf{x}}_t, y_t)\}_{t=s+1}^T$.

Algorithm 2 summarizes our EQRN for sequential data. For the estimation of the conditional GPD parameters, the main difference compared to the independent model is the use of a recurrent

Algorithm 2 EQRN for sequential observations

The tuning parameters Θ for the recurrent conditional GPD density estimation network $\tilde{g}_{\mathcal{W}}$, the intermediate quantile model $\hat{Q}_{\cdot}(\tau_0)$ capable of out of sample prediction and horizon s are prespecified. The training data $\tilde{\mathcal{D}} = \{(\tilde{\mathbf{x}}_t, y_t)\}_{t=s+1}^{T}$ and test covariates $\tilde{\mathbf{x}}_u$ are observed. Let $\tau \in (\tau_0, 1)$ be the desired probability level.

procedure EQRN-Fit($\tilde{\mathcal{D}}$, τ_0, s, Θ, $\hat{Q}_{\cdot}(\tau_0)$)
  $z_t \leftarrow y_t - \hat{Q}_{\tilde{\mathbf{x}}_t}(\tau_0) \;\; \forall t \in \mathcal{I} := \{t = s+1, \ldots, T : \; y_t > \hat{Q}_{\tilde{\mathbf{x}}_t}(\tau_0)\}$
  $\mathcal{T}, \mathcal{V} \leftarrow$ SequentialValidationSplit($\mathcal{I}$)
  $\hat{\mathcal{W}} \leftarrow$ InitializeRecurrentNetWeights(Θ)
  for $e = 1$ to maximum number of epochs E **do**
    for all $\mathcal{B} \in$ GetMiniBatches($\mathcal{T}$) **do**
      $\{(\hat{\nu}_t, \hat{\xi}_t)\}_{t \in \mathcal{B}} \leftarrow \tilde{g}_{\hat{\mathcal{W}}}(\tilde{\mathbf{x}}_{\mathcal{B}}, \hat{Q}_{\tilde{\mathbf{x}}_{\mathcal{B}}}(\tau_0))$
      $\ell \leftarrow \sum_{t \in \mathcal{B}} \ell_{\text{OGPD}}(z_t, \hat{\nu}_t, \hat{\xi}_t)/|\mathcal{B}|$
      $\hat{\mathcal{W}} \leftarrow$ BackPropUpdate(ℓ, $\hat{\mathcal{W}}$, $\tilde{\mathbf{x}}_{\mathcal{B}}$, $\hat{Q}_{\tilde{\mathbf{x}}_{\mathcal{B}}}(\tau_0)$, Θ)
    stop if $\mathcal{V} \neq \emptyset$ and LossNotImproving($\hat{\mathcal{W}}$, $\tilde{\mathbf{x}}_{\mathcal{V}}$, $\hat{Q}_{\tilde{\mathbf{x}}_{\mathcal{V}}}(\tau_0)$, $z_{\mathcal{V}}$)
  output $\hat{\mathcal{W}}$
procedure EQRN-Predict($\tilde{\mathbf{x}}_u$, τ, $\hat{\mathcal{W}}$, $\hat{Q}_{\cdot}(\tau_0)$)
  $\{\hat{\nu}(\tilde{\mathbf{x}}_u), \hat{\xi}(\tilde{\mathbf{x}}_u)\} \leftarrow \tilde{g}_{\hat{\mathcal{W}}}(\tilde{\mathbf{x}}_u, \hat{Q}_{\tilde{\mathbf{x}}_u}(\tau_0))$
  $\hat{\sigma}(\tilde{\mathbf{x}}_u) \leftarrow \hat{\nu}(\tilde{\mathbf{x}}_u)/\{\hat{\xi}(\tilde{\mathbf{x}}_u) + 1\}$
  compute $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau)$ w.r.t. $\hat{\sigma}(\tilde{\mathbf{x}}_u)$, $\hat{\xi}(\tilde{\mathbf{x}}_u)$, $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau_0)$, τ and τ_0 using equation (1.6)
  output $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau)$, and optionally $\{\hat{\sigma}(\tilde{\mathbf{x}}_u), \hat{\xi}(\tilde{\mathbf{x}}_u)\}$

architecture to capture the sequential nature of the data. If validation splitting is used during training, the split should preserve the sequential structure instead of being performed randomly. For the use of the intermediate quantile as a feature, two approaches seem relevant. The first one is to only use $\hat{Q}_{\tilde{\mathbf{x}}_t}(\tau_0)$ as a separate additional input to $\tilde{\mathbf{x}}_t$ in the network. The second approach is to also use past intermediate information by considering $\{(\mathbf{x}_j, y_j, \hat{Q}_{\tilde{\mathbf{x}}_j}(\tau_0))\}_{j=t-s}^{t-1}$, instead of $\tilde{\mathbf{x}}_t$, as input features to model the GPD parameters $\nu(\tilde{\mathbf{x}}_t)$ and $\xi(\tilde{\mathbf{x}}_t)$. We prefer the second approach, as it can pass more information to the tail model. More details on the function calls in the algorithm can be found in Supplementary Material A.2.

The training data consists of time points $\{1, \ldots, T\}$, while the test data uses information from time points $\{u-s, \ldots, u-1\}$ to predict at time u. These two intervals are typically disjoint when the model was fitted in the past and is applied for prediction in the present. The prediction model can, of course, be used to predict on the training data when $u \leq T$, but such predictions might be overly precise since y_u was used in the training procedure.

1.4 Simulation study

1.4.1 Setup

In this section we assess the accuracy of our EQRN model in predicting extreme conditional quantiles on simulated data and compare it to existing state-of-the-art methods. The aim is to study a simplified version of the application, which motivates the simulation setup and modelling

choices. We thus focus on the case of sequentially dependent data; a simulation study for independent data can be found in Supplementary Material A.3.

The main competitors from the extreme value literature in terms of flexibility are the generalized additive models (EGAM) (Youngman, 2019) and gradient boosting for extreme quantile regression (GBEX) (Velthoen et al., 2023), which both use conditional GPD modelling. For sequentially dependent data, we also consider the extreme quantile autoregression (EXQAR) (Li and Wang, 2019) as, although assuming linear quantile dependence, the model is designed for time series. As a benchmark we consider an unconditional GPD model that ignores the covariate dependence altogether and a semiconditional GPD model that uses the covariate only for the intermediate quantile. We include results for the generalized random forests for quantile regression (GRF) (Athey et al., 2019), which does not use extrapolation for high quantiles. The training data $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^{T}$, with $T = 7{,}000$, are sequentially generated from the time series

$$\begin{cases} Y_t = \sigma_t \left|\varepsilon_t^Y\right|, \quad X_t = 0.4 \cdot X_{t-1} + \left|\varepsilon_t^X\right|, \quad \varepsilon_t^Y, \varepsilon_t^X \sim \mathcal{N}(0,1), \\ \sigma_t^2 = 1 + 0.1 \cdot \{2Y_{t-1}^2 + Y_{t-2}^2 + Y_{t-3}^2 + Y_{t-4}^2 + Y_{t-5}^2\} + \\ \qquad + 0.1 \cdot \{3X_{t-1}^2 + 2X_{t-2}^2 + X_{t-3}^2 + X_{t-4}^2 + X_{t-5}^2\}. \end{cases} \tag{1.13}$$

Figure A.5 in Supplementary Material A.4 shows part of the simulated data. To have a fair comparison, all methods use the same covariate vectors $\tilde{\mathbf{x}}_t = \{(\mathbf{x}_j, y_j)\}_{j=t-s}^{t-1}$ with $s = 10$. This model admits a GPD approximation as in (1.5) since the conditional distribution $Y_t \mid \tilde{\mathbf{X}}_t = \tilde{\mathbf{x}}_t$ is a folded normal distribution, and its tail can, therefore, be approximated by a GPD with shape parameter $\xi(\tilde{\mathbf{x}}_t) = 0$. For the methods that use covariate-dependent intermediate quantiles, we use the same estimates $\hat{Q}_{\tilde{\mathbf{x}}_t}(\tau_0)$ with $\tau_0 = 80\%$ from a recurrent quantile regression neural network (QRN); for a sensitivity analysis of the choice of τ_0, see Supplementary Material A.4. The best QRN architecture and hyperparameters are chosen based on validation quantile loss. For the methods that use covariate-dependent GPD parameters, we also use $\hat{Q}_{\tilde{\mathbf{x}}_t}(\tau_0)$ as additional covariate. Although designed for univariate time series, we adapted EXQAR to accept several covariate sequences. Those two choices significantly improve the competitors' performances.

For the EQRN model, 2,000 observations are kept for validation tracking, and thus only the remaining 5,000 are effectively used for weight training. The best choices for EQRN hyperparameters are made based on validation loss by performing a grid search over a set of possible values and network architectures. All other models are fitted on the whole training dataset, as they do not use validation loss tracking. The best set of hyperparameters for GBEX (tree depths, learning rate and number of trees) are chosen using cross-validation, and the ground truth for whether the shape is constant is given to EGAM. For EXQAR we use $\delta_{2n} = n^{-0.9}$, as recommended by the authors, but set $\delta_{1n} = 1 - \tau_0$, thus increasing the number of quantile pseudo-observations used for inference and allowing better comparison with the other methods. The predictions of all models are evaluated by their root mean squared error (RMSE) compared to the true conditional quantiles on a newly generated test dataset that follows the same distribution as the training data in (1.13).

1.4.2 Results

For estimation of the conditional GPD parameters with our recurrent EQRN model, we consider both LSTM and GRU architectures with one to three recurrent layers and hidden dimensions between 32 and 256. As the networks are not too deep, L_2 penalty was chosen over dropout for regularization during training, with possible penalty $\lambda \in \{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. Both constant and covariate-dependent shape parameter outputs are considered. The model with minimum validation loss is a single LSTM layer with hidden dimension 128, followed by the usual fully connected output layer, with constant shape and $\lambda = 10^{-4}$. As a comparison we also retain predictions for the best unpenalized network with $\lambda = 0$ and fixed shape, which has two LSTM layers of hidden size 128.

The left panel of Figure 1.5 shows the RMSE of best penalized and unpenalized EQRN models, compared to the improved competitors, as a function of the quantile level τ. We observe that, for the lowest level $\tau = \tau_0 = 80\%$, all structured GPD models have the same performance since they use the same intermediate quantiles. GRF and the unconditional model have already higher errors since they are not able to capture the sequential dependence at the intermediate level sufficiently. For growing quantile levels τ, the errors of the covariate-dependent GPD models start to diverge. This is due to the differences in modelling flexibility in terms of the GPD parameters of each method. We observe a similar behaviour for EXQAR, as its linear quantile dependence is not flexible enough. Our EQRN method based on recurrent neural networks seems to be best at modelling sequential tail dependence.

Figure 1.5 also shows the predicted quantiles $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau)$ on the test data compared to the true $Q_{\tilde{\mathbf{x}}_u}(\tau)$ for a fixed $\tau = 99.95\%$ for the best penalized and unpenalized EQRN models. In general, both models seem to perform well in predicting the high conditional quantiles. The weight penalty seems to mainly affect the larger quantile predictions. Compared to the unpenalized model, we observe that the reduction in the variance of the predictions comes at the cost of a bias for larger quantile values. This bias-variance trade-off is typical with penalization. The poor performance of the semiconditional estimates highlights the added value of covariate dependence in the GPD parameters.

As discussed in Section 1.3, the choice of the intermediate level τ_0 generally has an impact on the prediction accuracy. In our covariate-dependent setting, where the model for the conditional GPD is a flexible regression model, this choice seems to have less importance. Indeed, even for a fairly low value of τ_0, the flexibility of the neural network model seems to be able to absorb some of the approximation bias; see Supplementary Material A.4 for details.

Additional results on the quantile R squared coefficient and bias-variance decomposition of the RMSEs presented in Figure 1.5 are also discussed in Supplementary Material A.4.

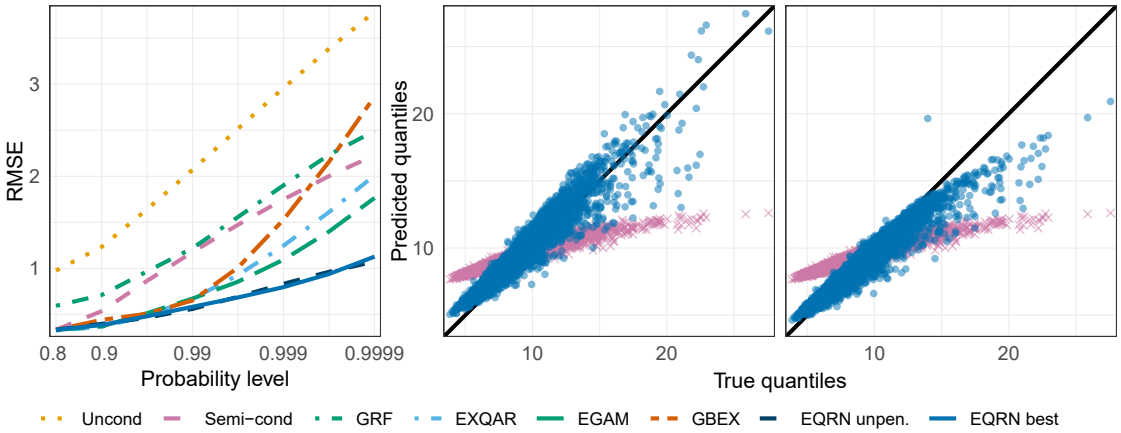

Figure 1.5: Left: Root mean squared error between predicted and true conditional quantiles at different probability levels τ (log-scale), for the selected EQRN models and the improved competitors. Centre-right: True vs. predicted quantiles at probability level $\tau = 99.95\%$ for the best unpenalized (middle) and penalized (right) EQRN models (dots), compared to the semiconditional estimates (crosses).

1.5 Application

1.5.1 Motivation

Flood risk is a major natural hazard in Europe, which causes huge economic damage and endangers human lives. There is a longstanding interest in statistical methods of extreme value theory for hydrology (e.g., Katz et al., 2002; Keef et al., 2009; Asadi et al., 2015; Engelke and Hitz, 2020), and national agencies commonly use them to assess the long-term risk of flooding in cities, at power plants, and other key locations. Return levels with long return periods can be estimated using the GEV distribution for annual maxima or the GPD model for daily threshold exceedances. The output then guides effective long-term flood management measures.

An example of an important location in Switzerland is the gauging station in Bern on the Aare river, which is shown within its water catchment in Figure 1.1. The Swiss Federal Office for the Environment (FOEN) monitors the Aare, and we use daily average discharges (in m^3s^{-1}) in Bern and another upstream station together with recordings of daily precipitation (in mm) at six locations in the Bern catchment; see Figure 1.1 for details. All time series are available in the period from 1930–2014 and can be obtained from the FOEN[2] (for discharges) and MeteoSwiss[3] (for precipitation). Figure 1.6 shows an excerpt for the two river stations and one precipitation gauge.

To illustrate possible drawbacks of a classical extreme value analysis, the left panel of Figure 1.7 shows the annual maxima of river discharges at the Bernese station on the Aare. The dashed line is the estimated 100-year return level based on the GEV approximation using the training period from 1930–1958. One can see that starting from the year 1999 there are several exceedances over this return level, somewhat contradicting the fact that it should only be exceeded on average once in 100 years. The solid line is the same return level based on data from 1930–y, where $y \in \{1959, \dots, 2014\}$ denotes the end of the training period. While the predictions are fairly

[2] `https://www.hydrodaten.admin.ch/`.
[3] `https://opendatadocs.meteoswiss.ch/`.

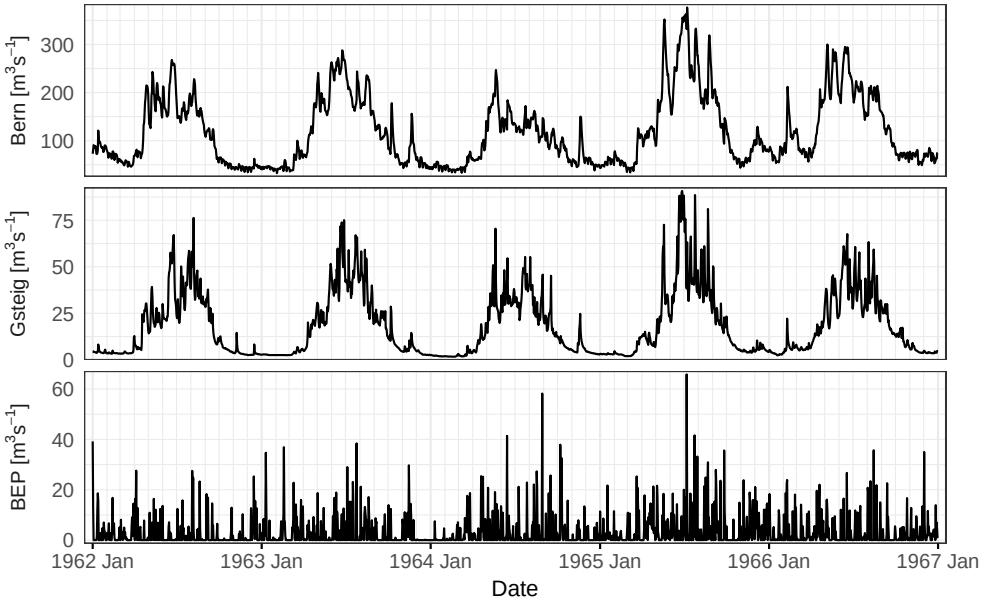

Figure 1.6: Daily average discharge observations at Bern–Schönau (62) and at the upstream station at Gsteig (42), and daily precipitation at the closest meteorological station to Bern (BEP), over five years; see Figure 1.1 for geographical locations of the gauging stations.

stable until 1999, an extreme value analysis performed after that year would yield much higher values for the 100-year return level. Conversely, historical estimates before such a break-point would severely underestimate the flood risk. In general, distributional shifts can be due to climate change, changes in the river system, or other structural breaks in factors influencing discharge at this location. For the Bernese station, the FOEN indeed reports a significant break-point in extreme discharges in the nineties but acknowledges that a clear cause can not be identified[4]. One factor may be a multidecadal variability of flood occurrence, as described in Schmocker-Fackel and Naef (2010).

The aim of our methodology is complementary to classical extreme value analysis and addresses this issue with static return levels. We apply our EQRN model to estimate one-day-ahead extreme quantiles of the river discharge conditionally on previous observations of discharge and precipitation in the catchment. This allows the forecasting of flood risk even in nonstationary systems, such as a changing climate. The strength of our approach lies in the ability to exploit information from multiple covariates and capture the complex time dependence. Even in situations where the causes for structural changes are unknown, our method implicitly accounts for them through their effects on the covariates. The output of the model can help practitioners and authorities to manage flooding more effectively and help to minimize their disastrous impacts by early warning systems.

To illustrate our methodology and show its effectiveness in comparison to classical forecasting approaches, we consider in more detail the flood in August 2005 in Switzerland. At the Bernese gauging station, it was the largest event since the beginning of the recordings, and it caused severe economic damage across large parts of the country and the loss of several lives.

[4]See flood report of the FOEN at `https://www.hydrodaten.admin.ch/en/2135.html`.

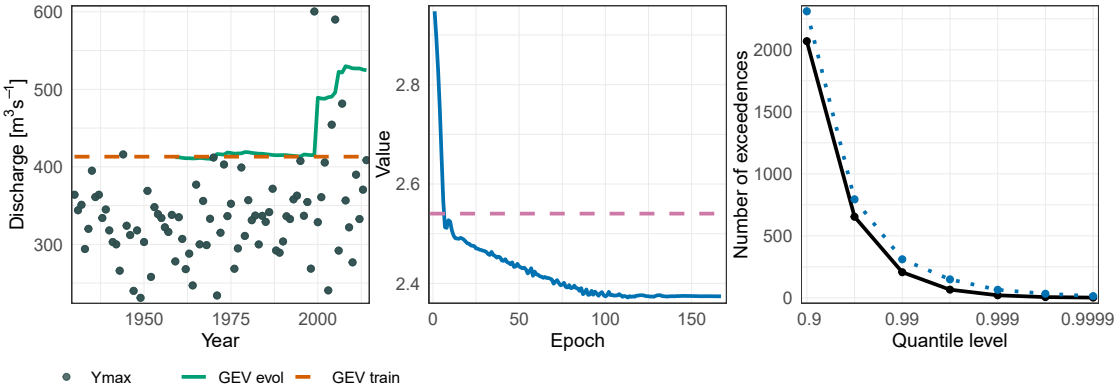

Figure 1.7: Left: Annual maxima of daily average discharges (points) at Bern–Schönau (62) together with the unconditional 100-year return level based on GEV fitted on the training data of 1930–1958 (dashed line), and the evolution of the same return level (solid line) using data from 1930–y, where $y \in \{1959, \dots, 2014\}$ denotes end of the period. Middle: Evolution of the validation loss (solid line) of the selected EQRN network for the river discharge data as a function of the training epoch; the dashed line shows the validation loss of the semiconditional model. Right: Number of observations exceeding the EQRN quantile predictions on the test set (dotted line) compared to the expected number of exceedances (solid line) for different probability levels (log-scale).

1.5.2 Model specification

The whole dataset consists of 31,046 daily observations $(\mathbf{x}_t, y_t)$ between 1930–2014. The response y_t is the daily average discharge at the Bernese gauging station on the Aare, and the covariates $\mathbf{x}_t \in \mathbb{R}^p$, $p = 7$, consist of discharge at another upstream station and daily precipitation measurements from six locations in the same catchment; see Figures 1.1 and 1.6. The discharges show significant seasonality, both in trend and variance, with the largest extremes only appearing in the summer. We do not reduce artificially the nonstationarity via classical approaches from times series analysis (e.g., Cleveland et al., 1990), as we believe the seasonality and other trends are captured through the covariates.

We split the data into training and test sets. The first $T = 10,349$ observations in the period between 1930–1958 are used to train the models, whereof the first three-quarters serve the parameter estimation, and the remaining quarter is a validation set to determine hyperparameters (sets $\mathcal{T}$ and $\mathcal{V}$ in Algorithm 2, respectively). The test set contains 20,697 observations from 1958–2014, which is used for neither fitting nor selection of parameters but only to evaluate the model performance on an independent time period. We choose this rather small proportion of training data to study the ability of the model to adapt to possible nonstationarity over time without refitting. In particular, the model weights are not updated with any information from data after 1958, even for forecasts in the study of the 2005 flood of interest. A large test set is also required to evaluate extreme properties of the data distribution. As augmented covariates at time t for the recurrent neural network models, we use the $s = 10$ preceding days and set $\tilde{\mathbf{x}}_t = \{(\mathbf{x}_j, y_j)\}_{j=t-s}^{t-1}$. The augmented training set is then $\tilde{\mathcal{D}} = \{(\tilde{\mathbf{x}}_t, y_t)\}_{t=s+1}^{T}$. Only observations of the $p + 1 = 8$ variables during the preceding 10 days are used to predict one day ahead for a new test time point.

As discussed in Section 1.3, we perform two steps for the estimation of the conditional tail model. First, we fit an intermediate quantile regression model to estimate $Q_{\tilde{\mathbf{x}}_t}(\tau_0)$. As in the simulations

with sequential dependence, we choose a recurrent QRN for this purpose and set $\tau_0 = 0.8$. For the second step, we include the intermediate quantile estimates $\hat{Q}_{\tilde{\mathbf{x}}_j}(\tau_0)$ during the same time horizon $j = t-s, \dots, t-1$ as additional covariates, and, slightly abusing notation, we denote $\tilde{\mathbf{x}}_t$ as the new covariate vector; see Section 1.3.2. A recurrent EQRN is then fitted to the exceedances for estimation of the conditional GPD parameters $\sigma(\tilde{\mathbf{x}}_t)$ and $\xi(\tilde{\mathbf{x}}_t)$; see Algorithm 2. Similarly to the study in Section 1.4.2, a grid search is performed on the training data to select the best hyperparameters and architectures for both recurrent neural network models based on validation losses.

The final model chosen to regress the intermediate quantiles is a QRN with two LSTM layers of dimension 256, followed by the usual fully connected layer, and L_2 weight penalty with parameter $\lambda = 10^{-6}$. The chosen EQRN model has two LSTM layers of dimension 16, followed by a fully connected layer, and L_2 weight penalty with parameter $\lambda = 10^{-6}$.

1.5.3 Results

The middle panel of Figure 1.7 shows the validation loss (solid line) of the selected EQRN model as a function of the training epoch. It can be seen that already after a few epochs, the method has a lower loss than the simple semiconditional model with constant GPD parameters σ and ξ (dashed line). This shows that the GPD distribution varies with the predictor values $\tilde{\mathbf{x}}_t$ and that a flexible model is beneficial.

The main output of the EQRN model are the extreme quantile estimates $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau)$ for a level τ and a time point u of interest, conditionally on the past covariates $\tilde{\mathbf{x}}_u$. These one-day-ahead risk forecasts are shown as a function of time on the test set in the top panel of Figure 1.2. We observe that the model is able to extrapolate beyond the range of the data since the event shown in the plot is unprecedented and the predictions still anticipate the first exceedance of the unconditional 100-year return level Q^{100}.

An unconditional τ-quantile is defined as the value that is exceeded by a proportion of $1-\tau$ of the data. An analogous property holds for conditional quantiles in data with sequential dependence, which yields a natural model assessment tool. On the population level, if $(\tilde{\mathbf{X}}_t, Y_t)_{t=1}^T$ is the random time series with augmented covariate vectors, then the expected number of exceedances over the true conditional τ-quantiles $Q_{\tilde{\mathbf{X}}_t}(\tau)$ is

$$\mathbb{E}\sum_{t=1}^{T} 1\{Y_t > Q_{\tilde{\mathbf{X}}_t}(\tau)\} = (1-\tau)T.$$

Consequently, plugging in the data $(\tilde{\mathbf{x}}_t, y_t)_{t=1}^T$ and estimates $\hat{Q}_{\tilde{\mathbf{x}}_t}(\tau)$ from a quantile regression method, the equation should approximately hold if the model is well-calibrated. Such a model assessment plot is shown in the right-hand panel of Figure 1.7 for our EQRN fit as a function of the quantile level τ. We observe that the model is fairly well-calibrated, with a slight bias toward more exceedances than expected.

An additional output are the corresponding GPD parameters $\sigma(\tilde{\mathbf{x}}_u)$ and $\xi(\tilde{\mathbf{x}}_u)$, which together

with the intermediate quantile $\hat{Q}_{\tilde{\mathbf{x}}_u}(\tau_0)$ specify the whole tail of the distribution of $Y_u \mid \tilde{\mathbf{X}}_u = \tilde{\mathbf{x}}_u$ according to (1.5). For a given threshold level of interest Q, we can plot the flood risk for the next day as the one-day-ahead forecast of the exceedance probability over Q, that is, an estimate of the function $u \mapsto \mathbb{P}(Y_u > Q \mid \tilde{\mathbf{X}}_u = \tilde{\mathbf{x}}_u)$. The bottom panel of Figure 1.2 shows the EQRN-based estimate of this function on the test set as a ratio to the unconditional $\mathbb{P}(Y_u > Q)$, where the threshold Q is chosen as the static 100-year return level Q^{100} based on the GEV distribution fitted on the training set; this threshold is relevant since it is often used to determine the height of dams for flood management. It results in a daily measure of how likely the exceedance on the next day is compared to what was expected unconditionally. Times with large predicted probability ratios are apparently times of imminent danger that can be used as triggers for early warning systems or additional flood management measures.

As an example, one may issue a warning when the forecasted conditional probability of exceeding Q^{100} is, say, a hundred times larger than the baseline unconditional probability $\mathbb{P}(Y_u > Q^{100})$. In the test data, there are four time clusters, which typically last several days, when Q^{100} is exceeded; see Figure 1.8 for one of these events. Applying this early warning system, in all four of these cases a timely warning would have been issued on the days preceding the first exceedance of the cluster. Such a decision rule would lead to an average of only 1.3 warnings for clusters of exceedances per year on the test set and is, therefore, not overly conservative. As the model does not need refitting on the test set, the daily forecast and possible warnings are obtained in less than one second of computation time, even on a CPU-only laptop computer[5]. The training time for the selected GPD network took less than 15 minutes.

We consider the period of the 2005 flood in Bern in more detail. Figure 1.8 shows the two discharge time series and precipitation at the closest meteorological station before and after the event. On the evening of August 21, 2005, the day preceding the first exceedance of the 100-year GEV return level (horizontal dashed line), the prediction of our EQRN already indicated a sudden increase in the probability of this exceedance; see bottom panel of Figure 1.2. Equivalently, the blue point in Figure 1.8 shows the increased value of the conditional 100-year quantile predicted by the model. The diamonds mark the observations of the previous 10 days that were used for this prediction.

In this case the high precipitation values on August 21 and the preceding days seem to have driven this prediction, possibly together with high values of the rivers. It is interesting to note that a similar situation on August 2 has not resulted in a "flood warning" since the forecasted exceedance probability and return level are not exceptionally high—as a matter of fact, there was no exceedance on the next day. This means that the predictions are driven by a complex combination of the risk factors $\tilde{\mathbf{X}}_t$ in space and time that are well-captured by the recurrent architecture of the EQRN.

At the time the FOEN used an adapted version of the hydrological model HBV (Lindström et al., 1997) for forecasting river discharges based on several inputs such as precipitation forecasts. However, the forecasts prior to this event underestimated the flood risk and resulted in too-late warnings. Physical models for discharges and precipitation do not use explicit extrapolation in

[5]Intel Core i5-8265U 1.6 GHz processor with four cores, and eight gigabytes of RAM memory.

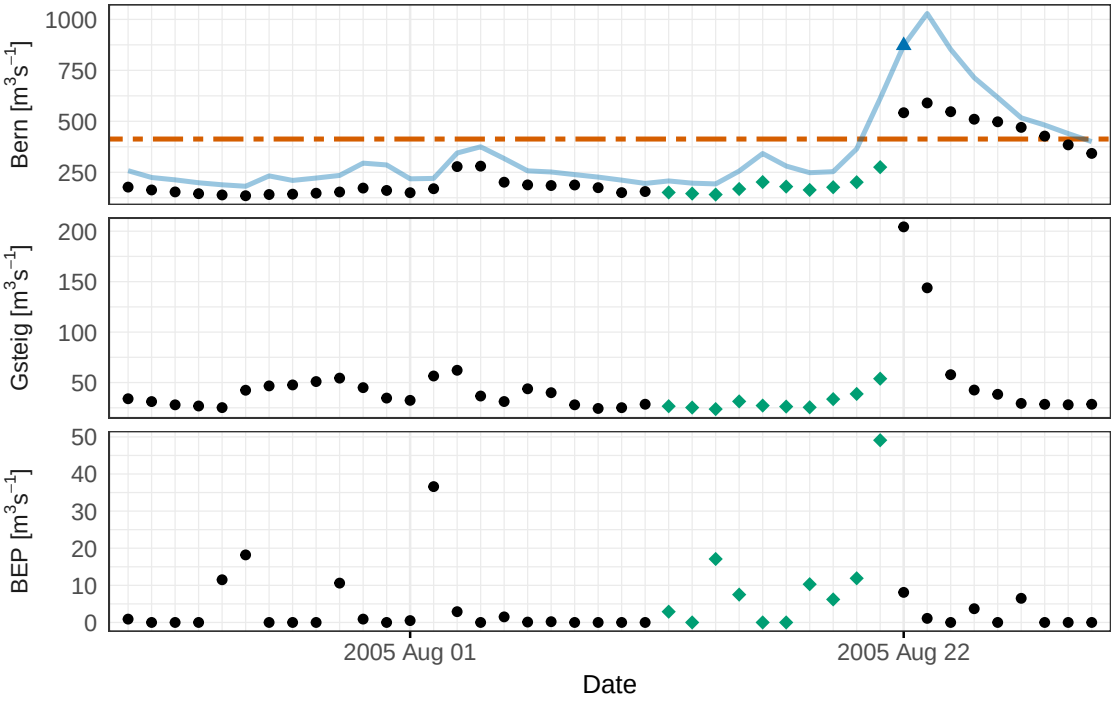

Figure 1.8: Discharge at Bern–Schönau (62) and Gsteig (42) and precipitation at BEP during the period of the 2005 flood, where diamonds indicate covariates used for prediction of the 100-year conditional quantile (triangle) on August 22, 2005; other precipitation stations are not shown. The top panel also shows the unconditional 100-year return level (dashed line) fitted on the training data and predictions on other days (solid line).

the extreme tails and often have poor performance in the largest predictions. In fact, during the 2005 flood, a main reason for the late warnings was that forecasters did not trust the predicted precipitation amounts during this extreme scenario. In the aftermath of this flood, the FOEN, therefore, published a detailed analysis of the internal forecasting procedures (Bezzola and Hegg, 2007). The exact forecasts from that time are not available, but the above discussion of the results shows that our statistical methodology is a competitive alternative to physical models for the forecasting of flood risk. We also note that our model uses much less information, as it relies only on observed discharge and precipitation and does not require forecasts of atmospheric variables.

In Section A.5 of the Supplementary Material, we compare and discuss the forecasts of our EQRN method with those of some of the competitors. Figures A.8–A.11 show that all methods seem to capture at least some of the temporal structure based on the past covariates. Except for the GBEX method, the forecasts of the competitors suffer, however, from a low sensitivity to changes in the conditional tail or from a too erratic behaviour as a function of time. Overall, the recurrent structure of our EQRN method, therefore, seems to be the best suited model for this kind of sequentially dependent data. Since in real-world applications the true quantiles are unknown, direct computation of the prediction error is difficult, and model assessment plots as in the right-hand panel of Figure 1.7 are crucial. This highlights the importance of simulation studies to evaluate and compare quantitatively the accuracy of different methods. In particular, in situations with temporal dependence, our EQRN method clearly outperforms the competitors (e.g., Figure 1.5). This is another indicator to trust the EQRN forecasts in applications.

1.6 Conclusion

Our EQRN model combines extrapolation results from extreme value theory with the prediction power of neural networks. It provides a flexible and versatile method for extreme quantile regression that is capable of prediction beyond the range of the data in the presence of a large number of covariates. Customised network architectures can be used in our open-source "`EQRN`" R package, allowing for tailor-made models capturing all types of potential dependencies between covariates and between observations.

The main focus in this paper was the case of sequential dependence to develop a tool for risk forecasting in time series that can be used for effective early warning systems in flood management. Our model already performs well in issuing sparse warnings for the days with increased risk of flooding, as illustrated in the case study of the Aare catchment in Switzerland. A further improvement could be attained by using additional covariates as input of the model. This could include observations of variables that are typically used in hydrological models, such as soil moisture, or forecasts of atmospheric variables such as precipitation and temperature.

Many other applications seem pertinent. Even in the case of independent data, our simulation study in Supplementary Material A.3 shows that neural networks outperform tree-based methods, such as ERF and GBEX, if the quantile function is more complex. Applications are financial risk assessment in insurance companies or banks. For spatial data, images, or graphs, convolutional or graph neural networks (LeCun et al., 2015; Scarselli et al., 2009; Wu et al., 2021) are known to perform extremely well in capturing neighbourhood structures. Our EQRN method can, therefore, be applied to quantify the risk of climate extremes where the predictor space contains spatiotemporal observations of meteorological variables (e.g., Boulaguiem et al., 2022). Transformer architectures (Vaswani et al., 2017) can also perform well for spatiotemporal or more complex dependencies.

The price for the high flexibility of machine learning methods, which focus on prediction accuracy, is limited statistical interpretability. However, feature-importance identification methods are becoming increasingly popular for interpreting neural network predictions (Lundberg and Lee, 2017). There is also active research on the construction of prediction intervals for black-box methods, for instance, through conformal inference (Lei et al., 2018; Romano et al., 2019). How such techniques can be adapted to assess uncertainty for extreme quantile regression is an interesting future research question.

Declarations

Acknowledgements

The authors would like to thank Daniel Viviroli for his valuable insights as well as the editorial team of the Annals of Applied Statistics and the anonymous reviewers for their helpful comments.

Funding

Both authors were supported by the Swiss National Science Foundation Eccellenza Grant 186858.

Published article

This document is the peer-reviewed "Author's Accepted Manuscript" of an article published in the Annals of Applied Statistics (Pasche and Engelke, 2024), with the DOI `https://doi.org/10.1214/24-AOAS1907`. When citing this work, please refer to the published version.

Supplementary material

Supplementary results

The Supplementary Material appended to this document contains additional information on Algorithms 1 and 2, the simulation study on independent data, additional results for the simulation study on dependent data, an analysis of the EQRN sensitivity to the intermediate probability level and competitor approaches to the application.

Reproducibility and R package

An open-source "`EQRN`" R package implementation of the proposed methodology is available on `https://github.com/opasche/EQRN`. The code and data with detailed instructions to reproduce the results presented in this paper, and more, are available on `https://github.com/opasche/EQRN_Results`.

2 Extreme conformal prediction: Reliable intervals for high-impact events

Olivier C. Pasche[1,2], Henry Lam[2], Sebastian Engelke[1]

[1]*Research Institute for Statistics and Information Science, University of Geneva, Switzerland*
[2]*Department of Industrial Engineering and Operations Research, Columbia University, New York, USA*

This chapter is a preprint of the homonymous invited article in press for publication in the special issue 'Bridging Heavy Tails and Artificial Intelligence' of *Extremes* (Pasche et al., 2025a).

Abstract

Conformal prediction is a popular method to construct prediction intervals with marginal coverage guarantees from black-box machine learning models. In applications with potentially high-impact events, such as flooding or financial crises, regulators often require very high confidence for such intervals. However, if the desired level of confidence is too large relative to the amount of data used for calibration, then classical conformal methods provide infinitely wide, thus, uninformative prediction intervals. In this paper, we propose a new method to overcome this limitation. We bridge extreme value statistics and conformal prediction to provide reliable and informative prediction intervals with high-confidence coverage, which can be constructed using any black-box extreme quantile regression method. A weighted version of our approach can account for nonstationary data. The advantages of our extreme conformal prediction method are illustrated in a simulation study and in an application to flood risk forecasting.

Keywords: conformal prediction, extreme value theory, prediction intervals, high confidence, generalized Pareto distribution, quantile regression.

2.1 Introduction

Conformal prediction is a simple approach to producing prediction sets from any regression or classification model. For a covariate vector $\mathbf{X}$ with values in $\mathcal{X} \subseteq \mathbb{R}^p$ and corresponding response variable Y, the goal of classical conformal prediction is to build a prediction set $\hat{C}(\mathbf{x})$ satisfying

marginal coverage

$$\mathbb{P}\{Y_{\text{test}} \in \hat{C}(\mathbf{X}_{\text{test}})\} \geq 1-\alpha, \tag{2.1}$$

for a desired confidence level $1-\alpha \in (0,1)$, for any new observations $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$. To obtain such a prediction set, we assume that a prediction model $\hat{f}$ was fitted on a training data set from the distribution $\mathcal{P}$ of $(\mathbf{X}, Y)$, and that a new calibration data set $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n_c}, Y_{n_c})$ of n_c observations from the same distribution is available. For a specific nonconformity score function $s : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ that may depend on $\hat{f}$ and acts on the predictors and responses, consider the calibration scores $S_i = s(\mathbf{X}_i, Y_i)$ for $i = 1, \ldots, n_c$. For some level $\alpha \in (0,1)$, denoting by $\hat{q}_\alpha$ the $\{\lceil (n_c+1)(1-\alpha) \rceil / n_c\}$-quantile of the scores $S_1, \ldots, S_{n_c}$, the prediction set

$$\hat{C}(\mathbf{x}) = \{y : s(\mathbf{x}, y) \leq \hat{q}_\alpha\} \tag{2.2}$$

has the desired $1-\alpha$ coverage. That is, it satisfies Eq. (2.1) for any new test observation $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$ from the same distribution $\mathcal{P}$, under a remarkably weak exchangeability assumption. This well-established framework is the so-called split conformal approach (Papadopoulos et al., 2002; Lei et al., 2018). More generally, originally started in Vovk et al. (1999, 2005), the conformalization idea that leverages quantile-based construction of prediction sets elicits a range of variants, with focus on optimal data usage and applying to different problems, including jackknife+ (Alaa and van der Schaar, 2020; Barber et al., 2021), cross-conformal prediction (Vovk, 2015) and ensemble-based approaches (Kim et al., 2020; Gupta et al., 2022). In combination with machine learning methods, it can efficiently capture predictive uncertainty (Shafer and Vovk, 2008; Zhou et al., 2025), and provide intervals with highly adaptive lengths (Romano et al., 2019). Extensions to nonexchangeable data are also well-studied (Oliveira et al., 2024). In particular, weighted conformal methods can account for distribution shifts and drifts (Tibshirani et al., 2019; Barber et al., 2023).

Conformal prediction intervals are widely used for confidence levels $1-\alpha$ of moderate value relative to the sample size n_c (e.g., 90% and $n_c = 1000$), where enough of the calibration scores are above this quantile. In many applications, however, test points $Y_{\text{test}} \notin \hat{C}(\mathbf{X}_{\text{test}})$ that fall outside of the prediction set correspond to a high-impact event with serious consequences for the environment, human lives or the economy. Examples for such risk-sensitive applications are the protection of cities and energy infrastructure from flooding (Keef et al., 2009; Asadi et al., 2015) or the financial reserves of banks and insurance companies (van Oordt and Zhou, 2019; Dupuis et al., 2023). In these cases, much larger values of confidence $1-\alpha$, close to one, will be required, sometimes even by law. Classical methods from conformal prediction fail for those requirements, since the quantile $\hat{q}_\alpha$ as defined above is not a useful estimator when the level $\alpha < 1/(n_c+1)$. Indeed, in this case, less than one observation then exceeds the $(1-\alpha)$-quantile on average, and $\hat{q}_\alpha$ is infinite (or ill-defined). Even for slightly larger values α close to that limit, the variance of $\hat{q}_\alpha$ can be huge.

Extreme value theory provides statistical tools for accurate estimation beyond the data range (de Haan and Ferreira, 2006). The tools have proven successful for improving extrapolation properties of machine learning methods in regression (de Carvalho et al., 2022a; Huet et al., 2024; Buriticá and Engelke, 2024), classification (Jalalzai et al., 2018), and generative methods (Boulaguiem et al., 2022).

In this paper, we propose a new methodology that bridges the wide applicability of conformal prediction with extrapolation tools from extreme value statistics to construct reliable prediction sets for high-impact events. In a first step, in order to obtain a good pretrained model $\hat{f}$ beyond the data range, we rely on flexible machine learning methods from extreme quantile regression (Velthoen et al., 2019; Gnecco et al., 2024; Pasche and Engelke, 2024; Richards and Huser, 2024). Second, we rely on the classical and theoretically justified peaks-over-threshold approach, which consists of using the generalised Pareto distribution (GPD) to extrapolate, for example, quantile estimates beyond the range of empirical observations (Balkema and de Haan, 1974; Pickands, 1975). For a confidence level $1-\alpha$ close to one, we leverage the GPD fitted to the calibration scores $S_1, \ldots, S_{n_c}$ to obtain a reliable estimate of q_α beyond the calibration data. The resulting extreme conformal prediction intervals have better properties compared to those from the classical empirical approach for large confidence requirements. In a simulation study, we show that our method improves existing approaches in terms of better coverage, in the sense of Eq. (2.1), and of informativeness of the prediction interval. We also consider a weighted version of our extreme conformal approach to deal with nonexchangeable data, and discuss its usage with several classical types of conformal procedures.

The advantages of our approach are illustrated in an application to flood risk prediction. Using several of the flexible machine learning methods as base predictions, it provides high-confidence one-day-ahead interval forecasts of the conditional range for water flow. We show that using conformal prediction intervals based on extreme value theory improves the coverage of the classical method, which either yields uninformative intervals or exhibits undercoverage and, therefore, seriously underestimates the risk of high-impact events.

2.2 Background on conformal prediction

2.2.1 Split conformal prediction

As in the introduction, let $\mathbf{X}$ be a covariate vector taking values in $\mathcal{X} \subseteq \mathbb{R}^p$, and Y the response random variable of interest. We will consider, here and in the sequel, the regression case where the response Y is real-valued in $\mathbb{R}$. We also suppose that we have access to a calibration set of n_c observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n_c}, Y_{n_c})$. Classical split conformal prediction builds prediction intervals (PIs) as in Eq. (2.2) that have desired coverage under very weak assumptions. Specifically, if the joint distribution of the calibration set and the test point $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$ is exchangeable, then the unconditional coverage guarantee Eq. (2.1) holds (Papadopoulos et al., 2002).

Importantly, the probability measure in Eq. (2.1) is with respect to the randomness in the calibration set jointly with the test point, that is, $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n_c}, Y_{n_c}), (\mathbf{X}_{\text{test}}, Y_{\text{test}})\}$. In fact, when conditioning on the calibration set, the distribution of the coverage is a beta distribution, that is,

$$\mathbb{P}\left[Y_{\text{test}} \in \hat{C}(\mathbf{X}_{\text{test}}) \mid \{(\mathbf{X}_i, Y_i)\}_{i=1}^{n_c}\right] \sim \text{Beta}(n_c + 1 - l, l), \quad l := \lfloor (n_c + 1)\alpha \rfloor. \tag{2.3}$$

Running the conformal prediction twice on different calibration sets, therefore yields PIs with different coverage probabilities. The guarantee in Eq. (2.1) says that when averaging out the

calibration set, the coverage is at least $1-\alpha$.

Furthermore, the marginal coverage property in Eq. (2.1) only guarantees "overall" marginal coverage of the prediction set $\hat{C}(\mathbf{x})$, but does not imply the conditional coverage property

$$\mathbb{P}\{Y_{\text{test}} \in \hat{C}(\mathbf{x}) \mid \mathbf{X}_{\text{test}} = \mathbf{x}\} \geq 1-\alpha, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{2.4}$$

The latter is generally impossible to guarantee in such a general setting. How close $\hat{C}(\mathbf{x})$ is to satisfying the conditional coverage property depends on the quality of the given pretrained model. For example, for the conformalized quantile regression approach described in Section 2.2.2, it depends on the accuracy of the initial quantile regression model $\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})$.

2.2.2 Conformalized quantile regression

Conformalized quantile regression, first proposed by Romano et al. (2019) and also described in Angelopoulos and Bates (2023), is one of the most popular conformal methods, in particular thanks to its ability to provide varying-length prediction intervals with competitive adaptivity. Suppose that we have access to a black-box quantile regression model trained to estimate the conditional quantiles $\hat{Q}_{\alpha/2}(\mathbf{x})$ and $\hat{Q}_{1-\alpha/2}(\mathbf{x})$ of Y, given $\mathbf{X} = \mathbf{x}$, at probability levels $\alpha/2$ and $1-\alpha/2$, respectively. Then, conformalized quantile regression uses the score function

$$s(\mathbf{x}, y) := \max\{\hat{Q}_{\alpha/2}(\mathbf{x}) - y,\ y - \hat{Q}_{1-\alpha/2}(\mathbf{x})\}. \tag{2.5}$$

Following the general procedure described in the introduction, this leads to the final prediction set in Eq. (2.2) being the interval

$$\hat{C}(\mathbf{x}) = \left[\hat{Q}_{\alpha/2}(\mathbf{x}) - \hat{q}_\alpha,\ \hat{Q}_{1-\alpha/2}(\mathbf{x}) + \hat{q}_\alpha\right], \tag{2.6}$$

where $\hat{q}_\alpha$ is the empirical $\{\lceil (n_c+1)(1-\alpha)\rceil / n_c\}$-quantile of the calibration scores $S_1, \ldots, S_{n_c}$, i.e. the order statistic $S_{(\lceil (n_c+1)(1-\alpha)\rceil)}$. As an equivalent definition, $\hat{q}_\alpha$ equals the empirical $(1-\alpha)$-quantile of $\{S_1, \ldots, S_{n_c}\} \cup \{+\infty\}$, the calibration score sample augmented with a point at infinity.

Intuitively, the procedure either widens (with a positive $\hat{q}_\alpha$) or narrows (with a negative $\hat{q}_\alpha$) the initial interval $\left[\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})\right]$ so that it covers $\lceil (n_c+1)(1-\alpha)\rceil$ of the n_c calibration observations. Note that the resulting prediction intervals satisfy the marginal coverage Eq. (2.1), but there is no guarantee that the conditional coverage in Eq. (2.4) is satisfied. In fact, the more accurate the initial quantile regression models $\hat{Q}_{\alpha/2}(\mathbf{x})$ and $\hat{Q}_{1-\alpha/2}(\mathbf{x})$ are, the better the conditional coverage will be.

2.2.3 Limitation for extreme confidence levels

For high-impact events, regulators often require predictions with very high coverage probabilities to ensure that protective infrastructures or measures are sufficient. In particular, in such risk-sensitive applications, the level α in Eq. (2.1) is typically close to 0 and may satisfy $\alpha < 1/(n_c+1)$.

This is generally referred to as an extreme confidence or probability level since, on average, there is less than one observation above the $(1-\alpha)$-quantile in a sample of size n_c. Note that the size n_c of the calibration set is typically fairly small, since these data cannot be used for model fitting, often resulting in extreme scenarios even for relatively moderate levels of α.

The classical construction of the conformal prediction intervals described in the introduction requires the computation of $\hat{q}_\alpha$, the empirical $\{\lceil (n_c+1)(1-\alpha)\rceil/n_c\}$-quantile of the calibration set scores. For extreme confidence levels $1-\alpha$, this quantile level

$$\lceil (n_c+1)(1-\alpha)\rceil/n_c > \lceil n_c\rceil/n_c = 1,$$

in which case $\hat{q}_\alpha$ is ill-defined and, by convention, set to infinity (Romano et al., 2019; Angelopoulos and Bates, 2023). This results in degenerate trivial prediction intervals $\hat{C}(\mathbf{x}) = (-\infty, \infty)$, for all $\mathbf{x} \in \mathcal{X}$. Although this interval satisfies the coverage Eq. (2.1), it is of no practical utility.

2.3 Extreme conformal prediction

We propose an approach based on extreme value statistics to construct nondegenerate conformal prediction intervals at extreme confidence levels $1-\alpha > n_c/(n_c+1)$. Similarly to classical conformalized quantile regression (Romano et al., 2019), our method requires two steps:

1. fitting a quantile regression model at level $1-\alpha$ on a training data set of size n;
2. calibrating based on the scores $S_1, \ldots, S_{n_c}$ on an independent data set.

For extreme confidence levels, both steps typically require extrapolation beyond the data range. Indeed, if α is close to 0, and in particular if $\alpha < 1/(n+1)$ is also extreme in the training data, then usual quantile regression will not be accurate. Instead, extreme quantile regression methods should be used. There is large literature on such methods based on linear models (Chernozhukov, 2005), additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019; de Carvalho et al., 2022b), or more flexible machine learning models such as gradient boosting (Velthoen et al., 2023; Koh, 2023), random forest (Gnecco et al., 2024) or neural networks (Pasche and Engelke, 2024; Richards and Huser, 2024; Allouche et al., 2024). Importantly, the model in step 1 can be a black-box, in the sense that we do not require theoretical guarantees. We discuss the extrapolation for step 2 in Section 2.3.2 and come back to examples of extreme quantile regression models in Sections 2.4 and 2.5. Furthermore, we discuss extensions of our approach to nonexchangeable data and alternative conformal procedures in Sections B.1 and 2.3.3.

2.3.1 Single-sided prediction intervals

Extreme conformal prediction intervals are most relevant in cases where very large values of the response variable Y lead to severe negative impacts. In such cases, reliable prediction intervals for $Y \mid \mathbf{X} = \mathbf{x}$, which contain the realisation of Y with very high marginal probability, are a crucial forecasting tool. They can be used to determine whether a dangerous level of the response could

potentially be reached. This also allows the reduction of false negatives (e.g., in the form of missing warnings) that can be critical to the system. In those risk assessment scenarios, it is often a single of the two tail directions which is of risk. This is the case for the risk of flooding discussed in Section 2.5 but also, for instance, for high temperatures in dry areas at risk of wildfires, and financial asset returns at risk of large losses. Without loss of generality, we, thus, suppose that one is interested in single-sided prediction intervals; two-sided intervals can be constructed analogously.

The classical procedure, yielding two-sided intervals, can be adapted to obtain single-sided prediction intervals by using the score function

$$s(\mathbf{x}, y) := y - \hat{Q}_{1-\alpha}(\mathbf{x}), \tag{2.7}$$

instead of Eq. (2.5), where $\hat{Q}_{1-\alpha}(\mathbf{x})$ is the pretrained quantile regression model at level $1-\alpha$. Let $y_{\min} \in \mathbb{R} \cup \{-\infty\}$ be the lower endpoint of the distribution of Y (or of the conditional distribution $Y \mid \mathbf{X} = \mathbf{x}$, if known). Then, following the usual procedure, the resulting interval is

$$\hat{C}(\mathbf{x}) = \left(y_{\min},\ \hat{Q}_{1-\alpha}(\mathbf{x}) + \hat{q}_\alpha\right], \tag{2.8}$$

where $\hat{q}_\alpha$ is, in classical conformal prediction, the $\{\lceil (n_c+1)(1-\alpha)\rceil / n_c\}$-quantile of the calibration scores $S_1, \ldots, S_{n_c}$, i.e., the order statistic $S_{(\lceil (n_c+1)(1-\alpha)\rceil)}$.

2.3.2 Calibrative extrapolation

As discussed in Section 2.2.3, when an extreme confidence level $1-\alpha > n_c/(n_c+1)$ is required, using order statistics to estimate $\hat{q}_\alpha$ would lead to degenerate intervals. Therefore, an alternative approach is needed to estimate a finite value $\hat{q}^e_\alpha$ from the calibration set such that

$$\mathbb{P}\left(S_{\text{test}} \leq \hat{q}^e_\alpha\right) \geq 1-\alpha. \tag{2.9}$$

Substituting $\hat{q}^e_\alpha$ for $\hat{q}_\alpha$ in Eq. (2.2) (or in Eqs. (2.6) and (2.8)) would, then, yield nondegenerate prediction intervals that satisfy the marginal coverage guarantee in Eq. (2.1). We propose to rely on the classical peaks-over-threshold methodology from extreme value theory to find such a quantile estimate. The tail of the distribution of the calibration score $S := s(\mathbf{X}, Y)$ can be approximated by the generalized Pareto distribution (GPD) above a high threshold u by

$$\mathbb{P}(S > y) = \mathbb{P}(S > u)\mathbb{P}(S > y \mid S > u) \approx \mathbb{P}(S > u)\left\{1 + \xi\frac{y-u}{\sigma(u)}\right\}_+^{-1/\xi}, \quad y \geq u, \tag{2.10}$$

where $\xi \in \mathbb{R}$ and $\sigma(u) > 0$ are the shape and scale parameters and u is an intermediate threshold. Under very mild assumptions on the distribution F_S of S, this approximation is theoretically justified as u tends to the upper endpoint of F_S (Balkema and de Haan, 1974; Pickands, 1975). In practice, u is typically chosen as the empirical τ_0-quantile $\hat{Q}_{\tau_0}$ of S, for $\tau_0 = (1 - k/n_c)$ and some $k < n_c$. The tuning parameter k is the number of exceedances used for estimation of the parameters σ and ξ, for instance, by maximum likelihood. Quantiles of S can then be estimated

at probability levels beyond the data range using the approximation

$$\hat{Q}^{\mathrm{GPD}}_{\tilde{\tau}} := \hat{Q}_{\tau_0} + \frac{\hat{\sigma}}{\hat{\xi}} \left\{ \left(\frac{1-\tau_0}{1-\tilde{\tau}} \right)^{\hat{\xi}} - 1 \right\}, \quad \tilde{\tau} > \tau_0. \tag{2.11}$$

Theoretical guarantees of these estimators typically require that $k \to \infty$ and $k/n_c \to 0$ as $n_c \to \infty$ to ensure the correct tradeoff between bias and variance; see de Haan and Ferreira (2006) for more details.

Asymptotically, as $n_c \to \infty$ and under additional second-order conditions, using $\hat{q}^e_\alpha := \hat{Q}^{\mathrm{GPD}}_{1-\alpha}$ would satisfy Eq. (2.9) with equality (in a suitable limiting sense as $\tau_0 \to 1$). But, since the calibration sample is finite and the level τ_0 fixed, $\hat{Q}^{\mathrm{GPD}}_{1-\alpha}$ can underestimate the true quantile due to estimation and approximation biases, respectively (Roodman, 2018; Bücher and Zhou, 2021; Zeder et al., 2023). We, therefore, follow here a more conservative approach. Alternatively to choosing $\hat{q}^e_\alpha$ as $\hat{Q}^{\mathrm{GPD}}_{1-\alpha}$, we may use the upper endpoint of a $(1-\alpha_2)$-confidence interval for $F_S^{-1}(1-\alpha_1)$, the $(1-\alpha_1)$-quantile of the calibration scores S, for two suitable levels $\alpha_1, \alpha_2 \in (0,1)$. The following proposition shows that, if this confidence interval has correct coverage, then the resulting extreme conformal prediction interval satisfies Eq. (2.1).

Proposition 2.3.1. *Let $\alpha_1, \alpha_2 \in (0,1)$, and $[L_Q, U_Q]$ be a $(1-\alpha_2)$-confidence interval for $F_S^{-1}(1-\alpha_1)$, the $(1-\alpha_1)$-quantile of the calibration scores S. If the confidence interval has correct coverage, i.e. $\mathbb{P}\{L_Q \le F_S^{-1}(1-\alpha_1) \le U_Q\} \ge 1-\alpha_2$, and if*

$$1 - \alpha \le (1-\alpha_1)(1-\alpha_2), \tag{2.12}$$

then $\mathbb{P}\left(S_{test} \le U_Q\right) \ge 1-\alpha$. That is,

$$\mathbb{P}\{Y_{test} \in \hat{C}^e(\mathbf{X}_{test})\} \ge 1-\alpha, \quad \textit{where} \quad \hat{C}^e(\mathbf{x}) := \{y : s(\mathbf{x}, y) \le U_Q\}. \tag{2.13}$$

The proposition applies more broadly to other types of base predictions and nonconformity score functions than to those of conformalized quantile regression. Its proof is presented in Section B.2.

A natural choice for α_1 and α_2 that satisfies Eq. (2.12) with equality is $\alpha_1 = \alpha_2 = 1 - (1-\alpha)^{1/2}$, which is analogous to the Šidák correction (Šidák, 1967). Another notable choice is $\alpha_1 = \alpha_2 = \alpha/2$, which is analogous to a Bonferroni correction (Bonferroni, 1936). Although the latter is, in this case, overconservative, the difference is negligible for small values of α.

There are several well-studied approaches for obtaining extreme quantile confidence intervals (CI) using the GPD approximation (Coles, 2001; Davison and Hinkley, 1997; Davison et al., 2003; de Haan and Zhou, 2022), including the profile likelihood, the bootstrap and the normal-approximation delta method. The profile-likelihood CI typically represents the inherently asymmetrical uncertainty best and yields the most conservative CI upper endpoint. For very small α and small sample sizes, it sometimes suffers from numerical difficulties for estimating the CI's upper endpoint. They arise as the derivative of the profile log-likelihood can get close to zero on its right-hand side, which sometimes makes finding the crosspoint between the profile curve and the chi-square confidence line numerically difficult. Slightly varying the value of

the GPD threshold can sometimes solve the issue. The bootstrap approach comes in different variations, both for the sampling step (nonparametric, parametric) and for the aggregation step (basic, percentile, normal, etc.). We here consider the nonparametric bootstrap with percentile aggregation, as it is the most commonly used. The bootstrap can give reliable confidence intervals, but, compared to the profile-likelihood approach, its upper endpoint estimates might not always be conservative enough. Finally, the delta-method CI is computationally less expensive than the other two alternatives, but it provides intervals that are symmetric around the quantile estimate, which is not realistic for large quantiles. Moreover, it can also suffer from numerical instability issues, due to its matrix inversion step, and fail to yield meaningful CIs. These alternative extreme CI methods are further discussed and compared in Section 2.4.

To summarise, when given a high confidence level $1-\alpha$, above or close to $n_c/(n_c+1)$, our proposed extreme conformal prediction interval takes the form

$$\hat{C}^e(\mathbf{x}) = \left(y_{\min},\, \hat{Q}^e_{1-\alpha}(\mathbf{x}) + \hat{q}^e_\alpha\right], \tag{2.14}$$

where $\hat{Q}^e_{1-\alpha}(\cdot)$ is a prefitted extreme quantile regression model and $\hat{q}^e_\alpha$ is the upper endpoint of an appropriate GPD profile-likelihood confidence interval for a high quantile of the calibration nonconformity scores, as discussed above. If a two-sided extreme PI is preferred, one can use the classical nonconformity scores from Eq. (2.5), instead of Eq. (2.7), to obtain

$$\hat{C}^e(\mathbf{x}) = \left[\hat{Q}^e_{\alpha/2}(\mathbf{x}) - \hat{q}^e_\alpha,\, \hat{Q}^e_{1-\alpha/2}(\mathbf{x}) + \hat{q}^e_\alpha\right], \tag{2.15}$$

where a second prefitted extreme quantile regression model $\hat{Q}^e_{\alpha/2}(\cdot)$ is required for extrapolating the lower-tail conditional quantiles.

2.3.3 Extensions to other conformal approaches and nonexchangeable data

We introduce our extreme conformal prediction approach as a variant of the well-established split conformalized quantile regression, as its use of base quantile predictions naturally results in varying-length intervals, which, in addition to valid marginal coverage, can also achieve good conditional coverage (Romano et al., 2019; Angelopoulos and Bates, 2023), and as the split procedure is significantly less computationally costly than full-conformal or k-fold alternatives. Nevertheless, Proposition 2.3.1 applies more broadly, and our approach is, in principle, adaptable to other conformal approaches, including different base regression models and nonconformity scores, the alternative full-conformal and k-fold procedures, and weighted methods for nonexchangeable data.

The extension to different base predictive models and scores is the most straightforward, as the extreme conformalization described in Section 2.3.2 is agnostic to their definitions. This includes the classical split-conformal approach, using a conditional-mean base regression model $\hat{\mu}(\mathbf{x})$, instead of the quantile predictions $\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})$ (Papadopoulos et al., 2002; Papadopoulos, 2008; Lei et al., 2018), and its heteroscedastic variants (Papadopoulos et al., 2008, 2011; Lei et al., 2018). Although these alternatives would not require extreme quantile regression, they either yield fixed-length PIs or tend to underestimate conditional variability (Romano et al., 2019).

The extension to full-conformal procedures (Vovk et al., 1999, 2005; Shafer and Vovk, 2008) is also straightforward, but requires repeating both the predictive model fit and our extreme conformalization from Section 2.3.2 for a dense grid of potential response values. Although it makes a more efficient use of the data than the split variant, it has an extremely high computational cost. The same applies, but less extremely, to k-fold approaches, such as Jackknife+/CV+ and cross-conformal prediction (Barber et al., 2021; Vovk, 2015). The extensions of our method to these alternative conformalisation approaches are described in more detail in Section B.1.

Weighted conformal approaches allow for relaxing the exchangeability assumption to account, for example, for distribution shifts or drifts. In those approaches, each calibration observation is assigned a weight $w_i \in [0,1]$, $i = 1,\dots,n_c$, which reflects its similarity to the test observation, or more generally, the similarity between its score S_i and the test score S_{test}. The classical empirical quantile of the calibration scores $\hat{q}_\alpha$ is then replaced by a weighted sample quantile. More precisely, it is redefined as the $(1-\alpha)$-quantile of the weighted empirical distribution $\bar{w}^n \sum_{i=1}^{n_c} w_i \delta_{s_i} + \bar{w}^n \delta_\infty$, where $\bar{w}^n = (\sum_{i=1}^{n_c} w_i + 1)^{-1}$ and δ_x is the Dirac measure at x (Barber et al., 2023; Tibshirani et al., 2019). As our proposed $\hat{q}^e_\alpha$ from Section 2.3.2 is, for extreme confidence levels, based on likelihood inference, the natural analogous extension to nonexchangeable data would be to rely on weighted likelihood inference instead. Using the sample weights $w_i \in [0,1]$, $i = 1,\dots,n_c$, the procedure would remain the same as in Section 2.3.2, only using the weighted

$$\ell_w^{\text{GPD}}(\sigma,\xi) = -\log(\sigma)\sum_{i=1}^{n_c} w_i - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{n_c} w_i \log\left(1+\xi\frac{s_i-u}{\sigma}\right)_+, \tag{2.16}$$

instead of the classical GPD log-likelihood, to infer the $(1-\alpha_2)$-CI endpoint for $F^{-1}_{S_{\text{test}}}(1-\alpha_1)$. This weighted alternative should achieve a similar effect as classical weighted conformal prediction methods: proportionally use the calibration scores most similar to the test point during conformalization, to correct for nonexchangeable distribution drifts and shifts. This weighted extension is further discussed and applied in Section 2.5.

2.4 Simulation study

2.4.1 Experimental setup

To assess the different conformalization methods, we perform a simulation study with several scenarios. The data is generated from

$$\begin{cases} \mathbf{X} \sim \text{Unif}\left([-1,1]^{10}\right), \\ Y \mid \mathbf{X} = \mathbf{x} \sim \sigma(\mathbf{x}) \cdot \varepsilon_Y, \end{cases} \tag{2.17}$$

with $\sigma(\mathbf{x}) := 1 + 6\phi(x_1, x_2)$, where ϕ is the bivariate Gaussian density with correlation 0.9. We consider two scenarios for the noise variable ε_Y: a heavy-tailed Student t distribution $t_{\alpha(\mathbf{x})}$, with covariate-dependent tail index $\alpha(\mathbf{x}) = 1/\xi(\mathbf{x}) := 7 \cdot \{1 + \exp(4x_1 + 1.2)\}^{-1} + 3$, and a light-tailed Gaussian $\text{N}(0,1)$ distribution. The former choice corresponds exactly to the generating process used in two extreme quantile regression benchmark studies (Velthoen et al., 2023; Pasche and

Engelke, 2024).

We consider several sizes for the calibration sets, with $n_c \in \{10^3, 10^{3.5}, 10^4\}$ observations. For each calibration size, we repeat the experiments 100 times. We consider extreme PI confidence levels $1-\alpha$, with $\alpha \in \{10^{-3}, 10^{-3.5}, 10^{-4}, 10^{-4.5}, 10^{-5}\}$. We consider three choices for the base quantile predictions $\hat{Q}^e_{1-\alpha}(\cdot)$: the conditional-quantile ground truth and two different pretrained extreme quantile regression models. The ground-truth choice aims at assessing the methods with ideal initial predictions. As all conformalization methods are translation invariant, adding first-order bias to the pretrained model would always lead to the same final PIs. For the first pretrained extreme quantile predictions, we use the extreme quantile regression neural networks (EQRN) model, as it performed best on this benchmark dataset (Pasche and Engelke, 2024), aiming at assessing our procedure with accurate but realistic initial predictions. The EQRN model is pretrained on 5,000 observations generated from Eq. (2.17). Its hyperparameters and architecture were selected based on validation GPD deviance with a grid search. Lastly, to investigate the performance of our extreme conformalization procedure for a poorly-performing model, which could happen in practice due to poor historical modelling choices, we also consider a linear GPD quantile model as base predictions. More precisely, we use linear quantile regression to obtain a conditional threshold $\hat{u}(\mathbf{x})$, and model the exceedances with a GPD having a linear parametrization $\sigma = \sigma_0 + \sigma_1 x_1 + \sigma_2 x_2$ for its scale parameter. Although having the GPD extrapolation potential, its linear parametrization is a bad fit for the highly nonlinear dependence of Y on $\mathbf{X}$. Experiments with Gaussian-noise data were only performed with the ground-truth base predictions.

For each calibration size, repetition, confidence level, and initial predictions, we perform the following conformalization procedures and assess their average population coverage on a separate test set of 10^6 observations. The methods compared for conformalization are:

- GPD profile: $\hat{C}^e(\mathbf{x})$ from Eq. (2.14), using the GPD profile-likelihood CI for $\hat{q}^e_\alpha$.
- GPD bootstrap: $\hat{C}^e(\mathbf{x})$, using the GPD nonparametric bootstrap percentile CI for $\hat{q}^e_\alpha$.
- GPD delta: $\hat{C}^e(\mathbf{x})$, using the GPD delta-method CI for $\hat{q}^e_\alpha$.
- GPD simple: $\hat{C}^e(\mathbf{x})$, but using the simple GPD $(1-\alpha)$-quantile estimate of the calibration scores instead of the endpoint of a CI for $\hat{q}^e_\alpha$.
- Classical: Single-sided version of the classical split conformalized quantile regression (still using the extreme quantile predictions $\hat{Q}^e_{1-\alpha}(\cdot)$, for a fair comparison).

Each of the GPD-based procedures use the empirical 0.95-quantile as the threshold u.

2.4.2 Coverage results

Figure 2.1 shows the distribution of the computed test coverage for each considered conformalization method, confidence level, and calibration size, for the Student t noise and ground truth original predictions. The chosen confidence levels are particularly large relative to the size of the calibration sets. Hence, for most scenarios, $\alpha < 1/(n_c + 1)$. In those cases, the classical conformalization method always yields trivial infinite intervals. They have a coverage of 1, but are uninformative and of no practical use. On the other hand, the other methods, relying on peaks

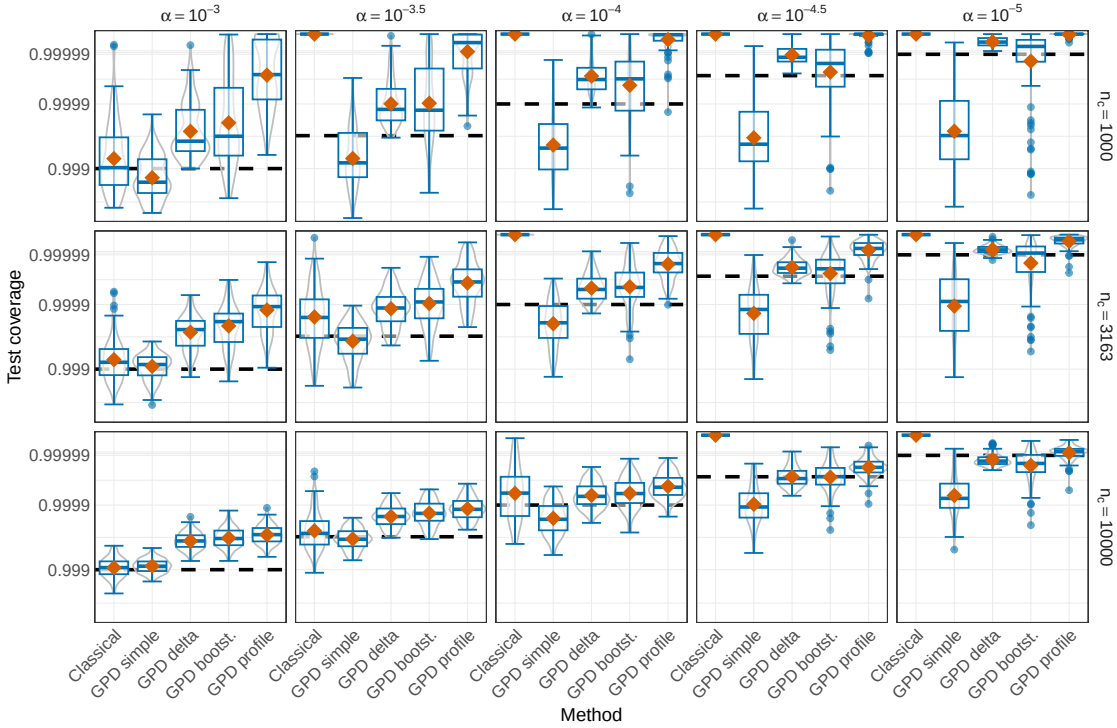

Figure 2.1: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over X) for different conformalization methods, conformal confidence values $1-\alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Student t distributed noise and quantile ground-truth predictions.

over threshold extrapolation instead of empirical quantiles, are able to yield finite PIs, even when $\alpha \ll 1/(n_c+1)$.

The simple GPD estimates of the calibration-score $(1-\alpha)$-quantile do not seem to provide sufficient coverage with small calibration samples and for the larger confidence levels, likely due to the GPD estimation error or approximation biases. The other three methods, relying on the confidence intervals for the score quantiles (and Proposition 2.3.1), achieve much better coverage and, in most cases, satisfy Eq. (2.1) as their coverage is larger than $1-\alpha$ on average. However, those GPD CI-based methods seem consistently overconservative for lower confidence levels.

In general, the profile likelihood method seems the most conservative, compared to the non-parametric bootstrap and delta method alternatives, as anticipated. It also satisfies the coverage guarantee in all scenarios. Its downside is the numerical difficulty, described in Section 2.3.2. With the implementation at hand, this issue arose, in the worst case, in 85% of the repetitions for $n_c = 1000$ and the lowest α value, but quickly decreased to, at most, 2% for $n_c = 3163$ and 0% for $n_c = 10000$. This instability is understandable in the former truly extreme case, as the confidence level is more than two orders of magnitude larger than the level for which PIs are obtainable with the classical conformal method and as the likelihood only relies on 50 observations to estimate a $(1-5\cdot10^{-6})$-confidence interval for a $(1-5\cdot10^{-6})$-quantile. However, this instability issue appears to often be resolved by slightly varying the choice of the GPD threshold, or of the α_1 and α_2 split.

The bootstrap and delta-method approaches seem less overconservative for the more moderate α values, but slightly undercover in the scenarios with the lowest α values. Nevertheless, they still significantly outperform the simple GPD approach and the infinite classical PIs. Contrary to the profile likelihood approach, the bootstrap method never fails to provide finite estimates. On the other hand, the delta-method approach also suffers from stability issues with small calibration sizes, regardless of the confidence level.

Figure B.1 in Section B.3 shows the same coverage distribution when the data-generating process has light-tailed Gaussian noise, instead of heavy-tailed Student t_4 noise. In comparison, all methods tend to result in significantly more conservative intervals in terms of the coverage. In particular, all three GPD CI-based approaches always result in more coverage than necessary.

Figure 2.2 shows the coverage distributions for the EQRN predictions and Student t_4 noise. We observe that, although being accurate predictions of the conditional quantile, in terms of integrated squared error (Pasche and Engelke, 2024), the EQRN predictions, here, undercover when considered as a PI endpoint. Note that even very accurate quantile predictions are still likely to lead to undercoverage, as a local quantile underestimation typically leads to a larger coverage loss than the coverage gain from a local overestimation of the same amplitude, due to the generally decreasing probability density in the tails. The conformalization results closely match those for the ground-truth quantile predictions, although the coverage seems smaller for the largest confidence levels, for all methods. The scenario with $\alpha = 10^{-5}$ and the largest sample size is the only one for which the GPD profile approach seems to slightly undercover. All the other finite alternatives also undercover for this largest confidence level. The CI-based extreme conformal prediction methods all outperform the original EQRN prediction in all scenarios.

Figure B.2 in Section B.3 shows the same coverage distributions when using the poorly-fitting linear GPD quantile predictions, described in Section 2.4.1, instead of the EQRN model. The base predictions severely undercover, when used as PIs. We observe that, even with poor quantile modelling choices, in all considered scenarios, our extreme conformalization approach results in a mean test coverage always greater than the desired levels, in this case for all three GPD CI-based variants. Interestingly, the intervals seem overall more marginally conservative than with the EQRN base predictions. Thus, even with this choice of a poorly-fit base model, undoubtingly worsening local adaptivity and coverage, our method still results in nontrivial PIs consistently achieving sufficient marginal coverage.

As a takeaway, our practical recommendation for conservative and informative high-confidence PIs is to use the profile-likelihood version of our method. In case it suffers from numerical issues and varying the GPD threshold or the α_1 and α_2 split slightly does not solve them (or if those are desired fixed), the bootstrap-based CI can be used instead. This combination results in the profile-likelihood conservativeness, in most scenarios, and avoids the cases of potential numerical difficulties by using the bootstrap estimate, which is still conservative enough in the majority of scenarios. We call this method the "GPD safeprofile" PI. Alternatively, considering the maximum of the bootstrap and delta-method PI endpoints as a replacement in unstable profile likelihood situations could be more conservative but might be less stable.

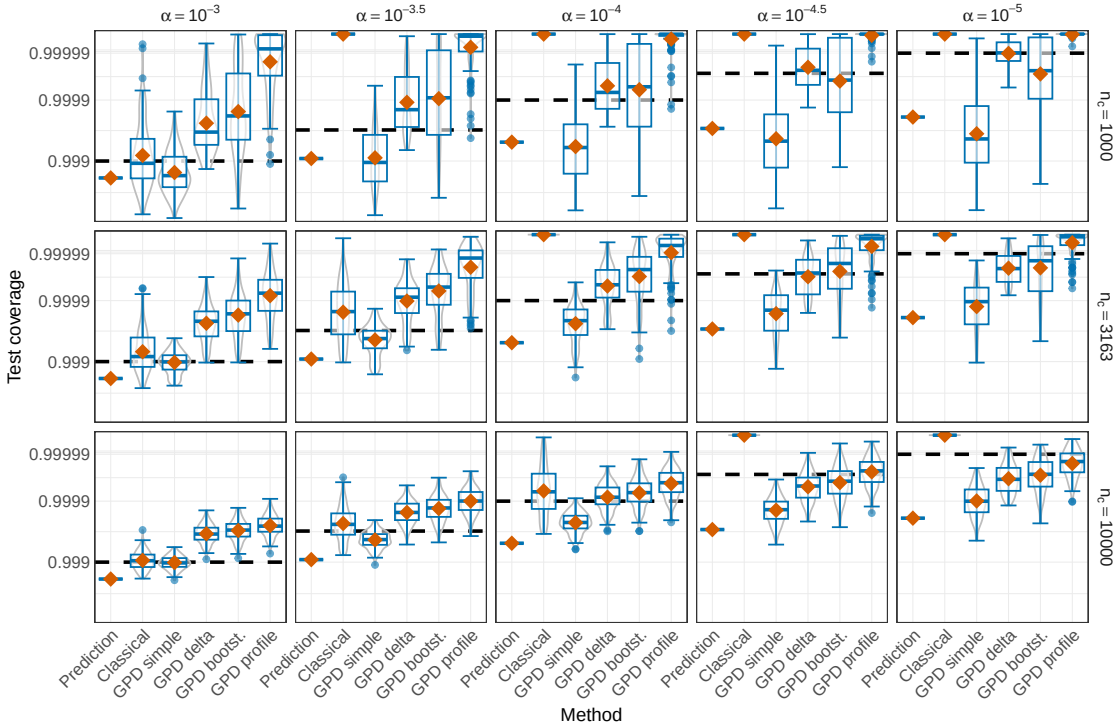

Figure 2.2: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over $\mathcal{X}$) for different conformalization methods, conformal confidence values $1-\alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Student t distributed noise and EQRN predictions.

2.5 Application to flood risk forecasting

2.5.1 Description and aim

Flooding is one of the most impactful natural hazards in terms of infrastructure and economic damage, and of the endangerment of human lives. Methods from extreme value theory have proven successful for assessing flood risk and providing reliable worst-case scenarios (e.g., Katz et al., 2002; Keef et al., 2009; Asadi et al., 2018; Engelke and Hitz, 2020; Engelke and Ivanovs, 2021).

In Switzerland, the Federal Office for the Environment (FOEN) monitors the river flow with numerous gauging stations throughout the river network. In its capital city, Bern, extreme water flow events of the Aare river led to several major floods, causing some of the most severe infrastructural and economic flooding damages recorded in the country. The main driver of strong water-flow events is the cumulative amount of upstream precipitation. In this study, we rely on the average daily discharge measures of the Aare river (in m^3s^{-1}) provided by the FOEN[1], and on recordings of daily precipitation (in mm) at various meteorological stations, obtained from MeteoSwiss[2]. This version of the dataset was preprocessed and analysed in previous studies (Pasche et al., 2023; Pasche and Engelke, 2024).

[1] `https://www.hydrodaten.admin.ch/`.
[2] `https://opendatadocs.meteoswiss.ch/`.

With our proposed extreme conformal approach, we aim to provide high-confidence one-day-ahead interval forecasts of the conditional range for water flow. We rely on several extreme quantile regression models pretrained to forecast the one-day-ahead extreme quantiles of the Aare water flow in Bern, given observations of upstream precipitation at six locations in Bern's water catchment and of the average daily water flow at an upstream gauging station, during the previous 10 days. Figure B.3 in Section B.3 shows the location of those meteorological and gauging stations.

2.5.2 Methodology

The river data exhibits a strong seasonal pattern with more water flow, and with more frequent and intense extreme events, during the late spring and summer months, due to snowmelt and heavy precipitation. This seasonality is likely to propagate to the residual nonconformity scores, violating the exchangeability assumption underlying classical conformal prediction. To account for it, we use the weighted variation of our extreme conformalization approach, discussed in Section 2.3.3. As a general procedure to account for distribution drift, Barber et al. (2023) suggest choosing large weights for recent calibration observations, close in time to the test period, and having weights decay for earlier observations, either exponentially or as a cutoff. Here, we make use of the inherent periodicity of the seasonal river discharge behaviour, by matching it with sinusoidal weights. Each year is partitioned into 24 roughly equal seasonal blocks of 15 to 16 days each. Let $B(i) \in \{1, \dots, 24\}$ denote the block in which an observation indexed by i falls into. For each block $b = 1, \dots, 24$, $\hat{q}^e_\alpha(b)$ is estimated as the upper-endpoint of the appropriate extreme score quantile CI, as described in Section 2.3.2, but using the weighted GPD likelihood in Eq. (2.16), with calibration sample weights

$$w_i = \cos\left[\frac{2\pi}{24}\{B(i) - b\}\right] + 1, \quad i = 1, \dots, n_c. \tag{2.18}$$

This choice of weights gives the highest importance to calibration observations in the same seasonal block b as the test point, and decreases the importance of observations in blocks further away in the year, with a yearly periodicity. For the estimation of $\hat{q}^e_\alpha(b)$, we use the GPD safeprofile CI procedure recommended in Section 2.4, although almost no numerical issues were encountered with the profile likelihood.

Our final one-day-ahead extreme PI for the average daily discharge, during seasonal block $b = 1, \dots, 24$ of the year, given past observations of upstream precipitation and water flow $\mathbf{X}_{\text{test}} = \mathbf{x}$, is then $\hat{C}^e(\mathbf{x}) = \left[0, \hat{Q}^e_{1-\alpha}(\mathbf{x}) + \hat{q}^e_\alpha(b)\right]$. We compare this PI to the unconformalized quantile predictions $\hat{Q}^e_{1-\alpha}(\mathbf{x})$, and to a single-sided (see Section 2.3.1) and weighted (Barber et al., 2023) variation of the classical split conformalized quantile regression PI $\hat{C}(\mathbf{x})$ (Romano et al., 2019), using the same sinusoidal weights from Eq. (2.18).

Apart from its seasonality, the data is also sequentially dependent. However, the residual dependence between the scores S_i is likely weak and short-term, which should not significantly affect the marginal coverage guarantee of the conformal PIs (Oliveira et al., 2024).

We consider several choices of extreme quantile regression models for $\hat{Q}^e_{1-\alpha}(\cdot)$: EQRN, GBEX (Velthoen et al., 2023), EGAM (Youngman, 2019), and EXQAR (Li and Wang, 2019). We also consider the constant unconditional GPD quantile estimates as a comparative baseline. We emphasise results using the recurrent version of EQRN, which is specifically designed for sequential dependence, and seems to fit the data best (Pasche and Engelke, 2024).

The quantile models were pretrained on data from 1939 to 1951. They were all fine-tuned with a grid search for hyperparameter selection. We use data from 1951 to 1999 for calibration and testing, i.e. 48 years of daily data. The observations after 1999 are not considered, due to a major distribution shift[3]. We choose the first 10 years as the default calibration set in the first part of the analysis, but vary its size from 3 to 15 years in the second part, and use the rest for estimating PI coverage empirically. We use multiples of complete years to keep a seasonal balance in the calibration and test sets.

2.5.3 Results

Figure 2.3 compares the number of observations exceeding the PIs from each method during the test period, using predictions of the different pretrained models, for a range of moderate to extreme confidence levels. The number of observations expected to exceed the PIs during the 38-year test period varies from 2,776, for $1-\alpha = 0.8$, to only 1.4, for the largest level $1-\alpha = 0.9999$. Using the original model predictions as PIs leads, in most cases, to undercoverage. Although the best-performing predictions seem to vary around the target coverage, they fail to provide satisfactory coverage consistently. The classical conformalization seems effective for the lowest two confidence levels but worsens the coverage for the following moderately extreme levels, compared to the initial predictions. At the two largest levels, the classical method yields uninformative infinite PIs. The extreme conformalization method yields finite PIs with significantly better coverage for confidence levels above 0.95, for which the approaches differ. The PI coverage consistently strictly satisfies the target confidence levels for each initial prediction model. The profile likelihood procedure was stable for almost all models, confidence levels and seasonal blocks. A numerical instability only occurred once, for the unconditional predictions at the largest confidence level, during a single one of the seasonal blocks.

Figure 2.4 shows the initial EQRN predictions, which seem to fit the data variation best, and their conformalized PI endpoints for two of the considered confidence levels, including the largest, during a test year. At $1-\alpha = 0.999$, even though it has more marginal coverage, the extreme PIs are not excessively larger than the classical PIs. The extreme intervals are even tighter during some of the seasonal blocks. At the largest level, the infinite classical-method PIs are uninformative. On the other hand, the extreme PIs, satisfying the desired test coverage, are again not overly large compared to the data variation, the original predictions, and the unconditional quantile estimates. Extreme conformal corrections for the EQRN predictions at the other unshown levels are smaller in magnitude. During the considered year, the original predictions at the largest confidence level are exceeded once, on 25th July 1973. This exceedance is covered by the conformalized extreme PI. Finally, we observed the localised adaptivity of weighted conformalization, as, for example,

[3]See the flood report of the FOEN at `https://www.hydrodaten.admin.ch/en/2135.html`.

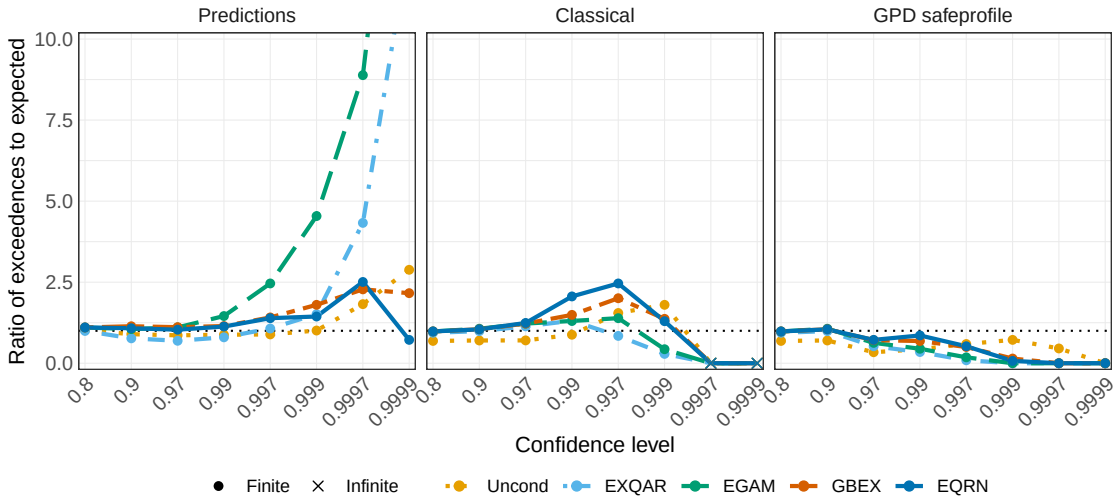

Figure 2.3: Number of observations exceeding the PIs during the test period as a ratio to the expected number of exceedances for different confidence levels and each pretrained prediction model, for the original predictions (left), the classical conformalization (center), and the GPD safeprofile method (right).

the conformal correction is larger during late summer than during winter for both methods and confidence levels.

Figure 2.5 shows the evolution of the test coverage with the calibration size, for all predictions, methods, and different confidence levels. We observe that the extreme PIs significantly outperform the classical conformalization in terms of empirical test coverage, for all relevant levels. It always provides informative finite PIs, contrary to the classical approach that yields infinite PIs for calibration sizes up to 5 years with $1-\alpha = 0.999$, and for all sizes at the two largest levels. The extreme PI has valid coverage in almost all combinations, including when the classical approach and/or original predictions significantly undercover. In the few undercovered situations, its coverage is closer to the target $1-\alpha$ than both the original predictions and the classical approach. For some of the lower levels, there seems to be a pattern of decreasing coverage with increasing calibration size for the conformalized PIs. The largest discharge events happening in years one, three and 15 of the chosen calibration period is a likely explanation for this decrease in between, although there might also be a small effect from the decreasing test size, as observed with the coverage of the unconformalized predictions.

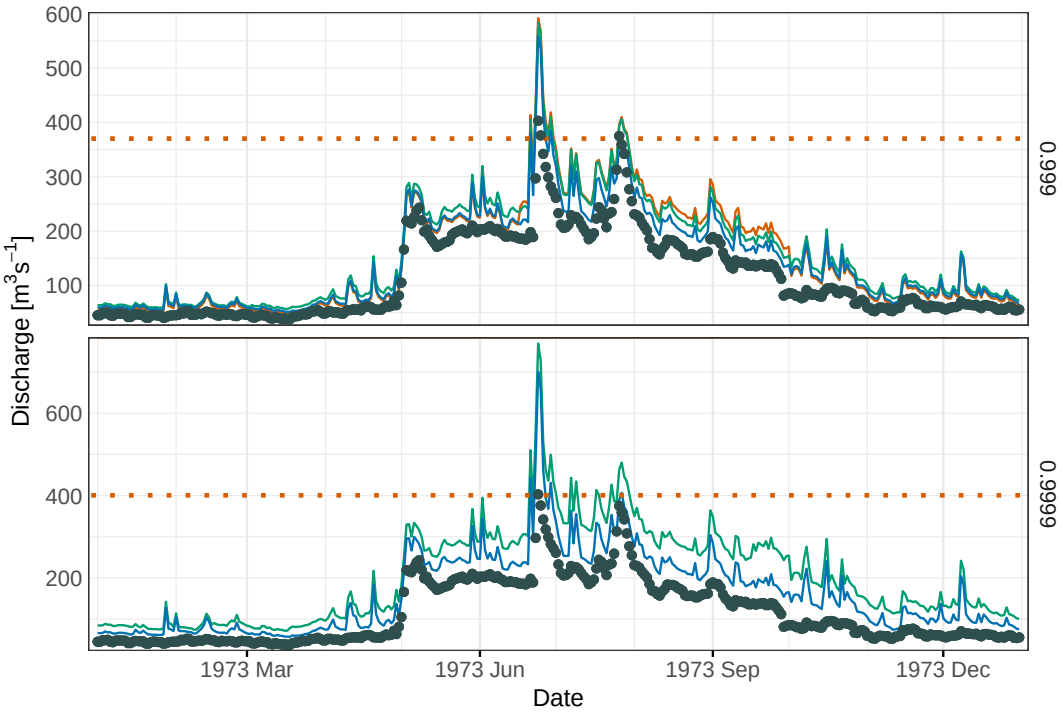

Figure 2.4: Original EQRN prediction (blue), classical conformal PI (red), and extreme conformal PI (green), at confidence levels 0.999 (top panel) and 0.9999 (bottom panel), during one of the test years. The classical conformal PI is infinite at level 0.9999. The observations (points) and the unconditional GPD $(1-\alpha)$-quantile estimates (dotted lines) are also shown.

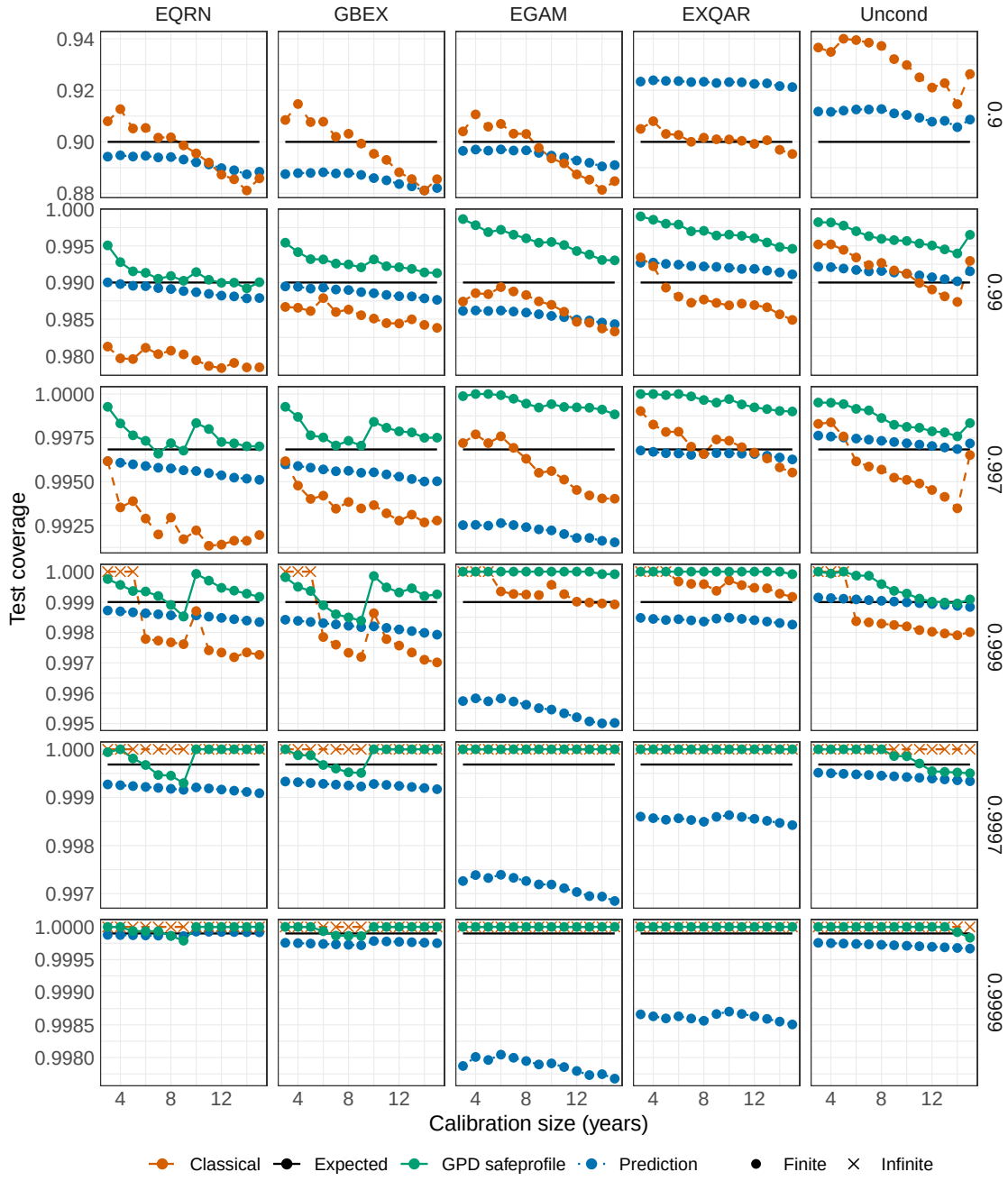

Figure 2.5: Empirical test coverage of the prediction intervals for a range of calibration-set sizes, for each conformalization method, pretrained model (columns), and confidence level $1-\alpha$ (rows). For $1-\alpha < 0.95$, corresponding to the quantile level of the GPD threshold u, the GPD approach coincides with the classical.

2.6 Conclusion

We propose a conformalization method which relies on extreme value statistics to provide conservative and nondegenerate prediction intervals for reliable risk assessment of high-impact events under extreme-confidence requirements. The novel method uses the well-studied peaks-over-threshold approach, which leverages the generalized Pareto distribution to extrapolate the necessary conformal correction from the calibration data to the required extreme confidence levels. It uses a conservative confidence interval solution for robustness against estimation and approximation biases.

In the simulation study and the application to forecasting river flow, at large confidence levels, our extreme conformal prediction method consistently provides PIs with significantly better coverage properties than both the original predictions and classical conformalization. For the largest levels, classical approaches result in infinite (or undefined) PIs, which are of no practical use. On the other hand, our recommended method yields informative finite intervals that consistently achieve the desired coverage for confidence levels up to several orders of magnitude larger than what is feasible with classical methods. The weighted version of our approach can account for nonstationary data drifts, such as the seasonal behaviour of the river flow. Importantly, our extreme conformal prediction method can be used in combination with any extreme quantile regression model, including black-box machine learning methods without known asymptotic guarantees.

One downside of our approach is its potential overconservativeness in certain scenarios, e.g., for moderately extreme confidence levels and some lighter-tailed data distributions. This is, at least in part, a consequence of our conservative CI-based solution, used to circumvent possible estimation and approximation biases in the GPD estimation. Moreover, peaks-over-threshold approaches rely on asymptotic properties. Thus, our approach lacks the finite-sample guarantees of classical conformal prediction. However, this is an unavoidable tradeoff for the ability to extrapolate beyond the moderate levels for which the classical empirical-quantile-based methods are feasible. In fact, obtaining finite-sample bounds for quantiles extrapolated beyond the data range is not theoretically possible without additional assumptions on the data distribution (Boucheron and Thomas, 2015; Thomas, 2015; Lhaut et al., 2022). A similar tradeoff exists in conformal prediction for conditional coverage, which is impossible to ensure on finite samples, resulting in the formulation of asymptotic conditional coverage guarantees as an alternative (Lei and Wasserman, 2014).

This work establishes a first method for extreme-confidence PIs, as an extension of the well-established conformal regression framework. As our experiments mainly focused on the split-conformal approach, due to its computational efficiency and practical popularity, further work is possible with alternative or more specialised conformal approaches. One direction is to enhance the efficiency of data usage, using, for example, the full conformal version of our approach, described in Section B.1, or potential variants of jackknife (Barber et al., 2021; Steinberger and Leeb, 2016), infinitesimal jackknife (Alaa and van der Schaar, 2020), or cross-validation (Vovk, 2015). However, such approaches typically have a much greater computational cost, as they require refitting the predictive model many times. Whether the gain in statistical efficiency for

our extreme conformal approach from using these alternative conformalization procedures is worth the extra computational cost remains an open question. Other directions include further extensions to nonexchangeable data. The weighted version of our extreme conformal approach seems suitable for nonstationary covariate drifts and shifts (analogously to Barber et al., 2023; Tibshirani et al., 2019). However, alternative solutions might be necessary to account for other problematic scenarios, such as long-term dependence or drifts in the conditional distribution of the response given the covariates. Finally, approaches other than conformalization could also warrant investigations, such as building PIs based on the so-called high-quality criterion and using deep learning (Pearce et al., 2018; Khosravi et al., 2011; Chen et al., 2021). How to incorporate extreme value statistics to extrapolate prediction intervals to cover high-impact events in these methodologies appears to be largely open.

Supplementary material

Implementation and reproducibility

To facilitate its practical use, the proposed extreme conformal procedure, including its weighted variant for nonstationary data, is implemented as an open-source R package, available at `https://github.com/opasche/ExtremeConformal`. Furthermore, a new versatile and efficient algorithm for estimating profile-likelihood confidence intervals for extreme quantiles (and return levels) is used as a dependency and is available as a separate R package at `https://github.com/opasche/ExtremeCI`. The code and data, with detailed instructions for reproducing the results presented in this paper, are available at `https://github.com/opasche/Reprod_ExtremeConformalPred`.

Declarations

Acknowledgements

This research project was conducted while the first author, O. C. Pasche, was a visiting scholar at the Department of Industrial Engineering and Operations Research, at Columbia University. He thanks the department and the university for their hospitality during this period. We also thank the reviewers and the associate editor for their valuable comments.

Funding

O. C. Pasche and S. Engelke were supported by the Swiss National Science Foundation Eccellenza Grant 186858. H. Lam was supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, Laboratory for AI-Powered Financial Technologies, and the Columbia Innovation Hub Award.

Availability of supporting data

In the Application to flood risk forecasting, we use river discharge and precipitation data recorded in Switzerland between 1930 and 1999, in the Rhine and Aare basins. The precipitation records can be freely obtained from MeteoSwiss, on `https://opendatadocs.meteoswiss.ch/`, and the discharge records from the Swiss Federal Office for the Environment (FOEN), on `https://www.bafu.admin.ch/bafu/en/home/topics/water/data-and-maps/water-monitoring-data/hydrological-data-service-for-watercourses-and-lakes.html`.

3 Causal modelling of heavy-tailed variables and confounders with application to river flow

Olivier C. Pasche[1,3], Valérie Chavez-Demoulin[2], Anthony C. Davison[3]

[1]*Research Institute for Statistics and Information Science, University of Geneva, Switzerland*
[2]*Faculty of Business and Economics, University of Lausanne, Switzerland*
[3]*Institute of Mathematics, EPFL, 1015 Lausanne, Switzerland*

This chapter is a postprint of the homonymous article published in *Extremes* (Pasche et al., 2023), with doi:10.1007/s10687-022-00456-4.

Abstract

Confounding variables are a recurrent challenge for causal discovery and inference. In many situations, complex causal mechanisms only manifest themselves in extreme events, or take simpler forms in the extremes. Stimulated by data on extreme river flows and precipitation, we introduce a new causal discovery methodology for heavy-tailed variables that allows the effect of a known potential confounder to be almost entirely removed when the variables have comparable tails, and also decreases it sufficiently to enable correct causal inference when the confounder has a heavier tail. We also introduce a new parametric estimator for the existing causal tail coefficient and a permutation test. Simulations show that the methods work well and the ideas are applied to the motivating dataset.

Keywords: causation, causal tail coefficient, confounder, extreme value statistics, generalized Pareto distribution.

3.1 Introduction

The field of causal inference has developed massively in recent decades (e.g., Pearl, 2009; Peters et al., 2017), with much recent work on the detection of causality from observational data (e.g., Maathuis and Nandy, 2016). Most of this literature concerns central quantities such as expectations, but certain causal mechanisms manifest themselves only in rare events and/or may

simplify in distribution tails. Standard methods of causal inference are ill-suited for such situations, and recent work has begun to link causality and extreme value theory. Examples are Gissibl and Klüppelberg (2018), who define recursive max-linear models on directed acyclic graphs, Klüppelberg and Krali (2021), who propose a scaling technique to determine the causal order of the variables in such graphs, Kiriliouk and Naveau (2020), who use multivariate generalized Pareto distributions to study probabilities of necessary and sufficient causation as defined in the counterfactual theory of Pearl, and Mhalla et al. (2020), who construct a causal inference method for tail quantities relying on Kolmogorov complexity of extreme conditional quantiles. See surveys by Naveau et al. (2020) on extreme event attribution and by Engelke and Ivanovs (2021) on the detection and modeling of sparse patterns in extremes.

Our work stems from that of Gnecco et al. (2021), who propose an estimator of the causal tail coefficient and an algorithm that, under mild conditions, consistently retrieves a causal order on an underlying graph even in the presence of hidden confounders. Such an order helps to exclude some causal structures, but does not provide evidence for the existence of a specific structure, as in general a given order is causal for several possible graphs; in particular, all orders are causal for the empty graph corresponding to absence of causality. Although it is asymptotically invariant to hidden confounders, this estimator can suffer from confounding in finite samples when inference on the direct relationship between two variables is needed, when these effects are too strong or when the confounders have heavier tails than the two variables.

This paper addresses a central challenge in causal inference: the presence of confounders. In theoretical development it is often assumed that all the relevant variables are observed and can be included in the model, but in practice one can rarely be sure of this. The available variables are often subject to external influences, observed or unobserved, that affect the variables of interest and can make it harder or even impossible to infer a correct causal relationship. Our goals are to mitigate the effect of a set of known confounders on an extremal causal analysis by treating them as covariates, and to present a permutation test for direct causality between the two observed variables. Our approach relaxes the assumption of Gnecco et al. (2021) that the confounders have the same tail index as the two main variables of interest, and thus encompasses a much broader range of situations, such as that in our application. Such a model enables causal discovery and inference for a greater variety of situations.

Our work was stimulated by average daily discharge data from 68 gauging stations along the Rhine and Aare catchments in Switzerland, see Figure 3.1. The data were collected by the Swiss Federal Office for the Environment (`https://hydrodaten.admin.ch/`), but were provided by the authors of Engelke and Ivanovs (2021), with some useful preliminary insights. We focus on the causal relationship between extreme discharges, for which precipitation is an obvious confounder, and use daily precipitation data from 105 meteorological stations, provided by the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss (`https://opendatadocs.meteoswiss.ch/`). Unlike in our simulation experiments, we know neither the true tail properties of the discharges and precipitation nor the effect of the confounder. We use precipitation as a covariate in our test, allowing inference on the direct causal relationships between discharges for the majority of the station pairs, with at least 95% estimated confidence, which was impossible without our proposed approach.

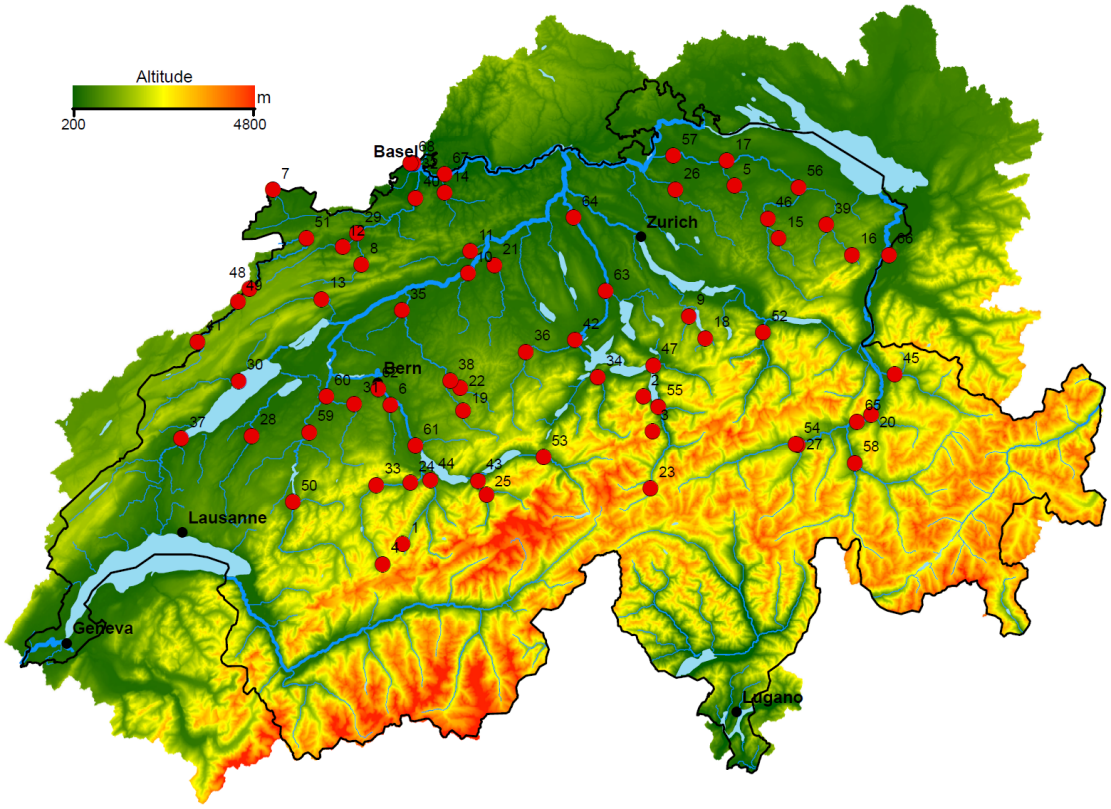

Figure 3.1: Topographic map of Switzerland showing the 68 gauging stations (red dots) along the Rhine, the Aare and their tributaries. Water flows towards station 68. Adapted from Engelke and Ivanovs (2021).

The paper is organised as follows. Section 3.2 discusses the causal tail coefficient, its interpretation and its properties. Section 3.3 introduces a new parametric estimator for it based on generalized Pareto modelling of threshold excesses, which allows a known confounder to be used as a covariate. A simulation study in Section 3.4 underlines the strengths and limitations of the two estimators. Section 3.5 presents a permutation test intended to detect direct causality between two heavy-tailed variables, which is also assessed via simulation. Section 3.6 applies the methodology to the river discharges, and Section 3.7 gives a brief discussion.

3.2 Causal tail coefficient and its estimation

3.2.1 Existing work

We first give some basic notions needed to describe the setting in which causal relationships between random variables can be recovered.

Definition 3.2.1. A *linear structural causal model (LSCM)* over a set of random variables $X_1, \dots, X_p$ satisfies

$$X_j = \sum_{k \in \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V,$$

where $V := \{1, \dots, p\}$ is a set of nodes representing the corresponding random variables, $\mathrm{pa}(j) \subseteq V$ is the set of parents of j, $\beta_{jk} \in \mathbb{R} \setminus \{0\}$ is called the *causal weight* of node k on node j, and $\varepsilon_1, \dots, \varepsilon_p$

are jointly independent noise variables. We suppose that the *associated graph* $G = (V, E)$, in which the directed edge $(i, j) \in V \times V$ belongs to E if and only if $i \in \mathrm{pa}(j)$, is a directed acyclic graph (DAG).

In a DAG $G = (V, E)$, we say that $i \in V$ *is an ancestor of* $j \in V$ in G, if there exists a directed path from i to j. The set of the ancestors of j in G is denoted by $\mathrm{An}(j, G)$, and we define $\mathrm{an}(j, G) := \mathrm{An}(j, G) \setminus \{j\}$. In a LSCM over random variables $X_1, \ldots, X_p$, with associated DAG $G = (V, E)$, we say that X_i *causes* X_j, if $i \in \mathrm{an}(j, G)$. We call X_i a *confounder* (or *common cause*) of X_j and X_k if there exist directed paths from i to j and from i to k in G that do not include k and j, respectively. We say that there is *no causal link* between X_i and X_j if $\mathrm{An}(i, G) \cap \mathrm{An}(j, G) = \emptyset$. For any $i, j \in V$ we let $\beta_{i \to j}$ denote the sum of the products of the causal weights along the distinct directed paths from vertex i to vertex j; we set $\beta_{j \to j} := 1$ and $\beta_{i \to j} := 0$ if $i \notin \mathrm{An}(j, G)$.

Let X_i and X_j be random variables from a LSCM with respective distributions F_i and F_j. The *causal (upper) tail coefficient* of a random variable X_i on another random variable X_j is defined as (Gnecco et al., 2021)

$$\Gamma_{ij} := \lim_{u \to 1^-} \mathbb{E}\left\{ F_j(X_j) \mid F_i(X_i) > u \right\}, \tag{3.1}$$

if the limit exists. This coefficient lies between zero and one and captures the causal influence of X_i on X_j in their upper tails: if X_i has a linear causal effect on X_j, $\Gamma_{1,2}$ will be close to unity. The coefficient is asymmetric, as extremes of X_j need not lead to extremes of X_i, and in that case, Γ_{ji} will be appreciably smaller than Γ_{ij}. As Γ_{ij} only depends on the rescaled margins of the variables, it is invariant to monotone increasing marginal transformations.

If both tails are of interest, the causal tail coefficient can be generalized to capture the causal effects in both directions, by considering the *symmetric causal tail coefficient* of X_i on X_j, i.e.,

$$\Psi_{ij} := \lim_{u \to 1^-} \mathbb{E}\left[\rho\left\{ F_j(X_j) \right\} \mid \rho\left\{ F_i(X_i) \right\} > u \right]$$

if the limit exists, where $\rho : x \mapsto |2x - 1|$. As $F_i(X_i) \sim \mathrm{Unif}(0, 1)$,

$$\Psi_{ij} = \underbrace{\lim_{u \to 1^-} \frac{1}{2} \mathbb{E}\left[\rho\left\{ F_j(X_j) \right\} \mid F_i(X_i) > u \right]}_{=: \Psi_{ij}^{+}} + \underbrace{\lim_{u \to 0^+} \frac{1}{2} \mathbb{E}\left[\rho\left\{ F_j(X_j) \right\} \mid F_i(X_i) < u \right]}_{=: \Psi_{ij}^{-}}.$$

The interpretation and properties of Ψ_{ij} are similar to those of Γ_{ij}. The symmetric version captures the causal influence of X_i on X_j in both of their tails.

For simplicity we focus on Γ_{ij} in this paper, though all of our results and methods can be generalized to both tails by considering Ψ_{ij} instead, if the assumptions for the upper tails are also satisfied in the lower tails of the variables considered.

Before stating the theorem that describes how the underlying causal relationships in a set of random variables can be recovered, we define the concept of regular variation.

Definition 3.2.2. A positive measurable function f is said to be *regularly varying* with index $\alpha \in \mathbb{R}$, written $f \in \mathrm{RV}_\alpha$, if for all $c > 0$, $\lim_{x \to \infty} f(cx)/f(x) = c^\alpha$. If $f \in \mathrm{RV}_0$, then f is said to

be *slowly varying*.

Definition 3.2.3. The random variable X_j is said to be *regularly varying* with index $\alpha > 0$, if, for some $\ell \in \mathrm{RV}_0$, $\mathbb{P}(X_j > x) \sim \ell(x)x^{-\alpha}$ as $x \to \infty$.

Independent regularly varying random variables $X_1, \ldots, X_p$ are said to have *comparable upper tails* if there exist $c_1, \ldots, c_p > 0$, $\alpha > 0$ and $\ell \in \mathrm{RV}_0$ such that, for each $j \in \{1, \ldots, p\}$, $\mathbb{P}(X_j > x) \sim c_j \ell(x) x^{-\alpha}$ as $x \to \infty$.

The following theorem describes how the causal relationships underlying a set of random variables can be recovered from their causal tail coefficients.

Theorem 3.2.4 (Gnecco et al., 2021)**.** *Let $X_1, \ldots, X_p$ be random variables from a LSCM, with associated directed acyclic graph $G = (V, E)$ and suppose that*

(a) the coefficients β_{jk} of the linear structural causal relationship $X_j = \sum_{k \in \mathrm{pa}(j,G)} \beta_{jk} X_k + \varepsilon_j$ are strictly positive for all $j \in V$ and $k \in \mathrm{pa}(j, G)$, and

(b) the real-valued noise variables $\varepsilon_1, \ldots, \varepsilon_p$ are independent and regularly varying with comparable upper tails.

Then the values of Γ_{ij} and Γ_{ji} allow one to distinguish between the different possible causal relationships between X_i and X_j summarized in Table 3.1.

Table 3.1: Equivalence of the possible values of Γ_{ij} and Γ_{ji} with the underlying causal relationship between X_i and X_j.

	$\Gamma_{ji} = 1$	$\Gamma_{ji} \in (1/2, 1)$	$\Gamma_{ji} = 1/2$
$\Gamma_{ij} = 1$		X_i causes X_j	
$\Gamma_{ij} \in (1/2, 1)$	X_j causes X_i	common cause only	
$\Gamma_{ij} = 1/2$			no causal link

Under the theorem's assumptions, the blank entries in Table 3.1 cannot occur. Theorem 3.2.4 is generalizable to the Ψ_{ij} variant of the coefficient and possibly negative β_{ij} values if the assumptions are also satisfied in the lower tails of the variables.

Gnecco et al. (2021) show that under the setup and assumptions of Theorem 3.2.4, the causal tail coefficient (3.1) for any distinct $i, j \in V$, and with $A_{ij} := \mathrm{An}(i, G) \cap \mathrm{An}(j, G)$, is

$$\Gamma_{ij} = \frac{1}{2} + \frac{1}{2} \frac{\sum_{h \in A_{ij}} \beta_{h \to i}^{\alpha}}{\sum_{h \in \mathrm{An}(i,G)} \beta_{h \to i}^{\alpha}}. \tag{3.2}$$

Without loss of generality we set $i = 1$ and $j = 2$ in what follows, and thus consider the causal effect of X_1 on X_2.

If $\{(X_{i,1}, X_{i,2})\}_{i=1}^{n}$ are independent replicates of (X_1, X_2), with the random variables X_i and X_j from the LSCM, then the *non-parametric estimator* of $\Gamma_{1,2}$ is defined to be

$$\hat{\Gamma}_{1,2} = \frac{1}{k} \sum_{i=1}^{n} \hat{F}_2(X_{i,2}) \mathbb{1}(X_{i,1} > X_{(n-k),1}) \tag{3.3}$$

for some $k \in \{1, \ldots, n-1\}$, where $\mathbb{1}(\cdot)$ denotes the indicator function, $X_{(h),1}$ denotes the h^{th} order statistic and $\hat{F}_j$ is the empirical cumulative distribution function of X_j, i.e.,

$$\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{i,j} \leq x), \quad j = 1, 2.$$

This estimator is the empirical counterpart to (3.1), as $X_{(h),1} = \hat{F}_1^{\leftarrow}(h/n)$ is a quantile of the corresponding empirical distribution. The value of k controls the number of data pairs in the upper tail of X_1 that contribute to the estimator. Under the assumptions of Theorem 3.2.4 and a "very mild assumption that is satisfied by most univariate regularly varying distributions of interest", estimator (3.3) is consistent as $n \to \infty$, for a choice of k such that $k \to \infty$ and $k/n \to 0$ (Gnecco et al., 2021).

3.2.2 Practical limitations

A strength of the causal tail coefficient approach is its asymptotic robustness to hidden confounders. Studies of causation frequently presuppose that all the relevant variables have been observed, which is usually moot, but Theorem 3.2.4 holds even when some variables in the underlying LSCM are unobserved. This capacity to deal with confounders both when studying the causal relationship between two variables and when retrieving a causal order is not generally shared by other approaches in causal inference, as argued by Gnecco et al. (2021, Section 4.2), but the unobserved variables must satisfy a regular variation assumption that is hard to check and may be unrealistic. In practice, moreover, the tail behaviour of the confounders may differ from that of X_1 and X_2, violating assumption (b) of Theorem 3.2.4. In our motivating setting, for example, the tail of the confounder, precipitation, may not behave like the tails of the river discharges. This problem worsens when the confounder has a heavier tail than the variable of interest. Furthermore, distinguishing between different causal situations using empirical estimates may be difficult; an increase in the strength of the causal effect of a common confounder of X_1 and X_2 will increase $\Gamma_{1,2}$, making it harder to tell whether a high value of $\hat{\Gamma}_{1,2}$ indicates that $\Gamma_{1,2} = 1$ or that $\Gamma_{1,2} \lesssim 1$, as we shall see in Section 3.4.

The discussion above suggests that conditioning on the values of known confounders might be valuable. In the presence of a vector $\mathbf{H}$ of potential confounders we therefore define

$$\Gamma_{1,2|\mathbf{H}} := \lim_{u \to 1^-} \mathbb{E}_{(X_1, X_2, \mathbf{H})} \left\{ F_2(X_2 \mid \mathbf{H}) \mid F_1(X_1 \mid \mathbf{H}) > u \right\}. \tag{3.4}$$

If there is no direct dependence of X_2 on X_1, then X_2 is independent of X_1 conditional on $\mathbf{H}$, so $\Gamma_{1,2|\mathbf{H}} = 1/2$, whereas $\Gamma_{1,2}$ lies in $[1/2, 1)$ but might be close to unity. Thus $\Gamma_{1,2|\mathbf{H}} < \Gamma_{1,2}$ unless there are no confounders. If X_1 causes X_2, on the other hand, then $\Gamma_{1,2|\mathbf{H}} = \Gamma_{1,2} = 1$. In the presence of potential confounders, therefore, (3.4) seems preferable to $\Gamma_{1,2}$. The difficulty is that the estimation of (3.4) requires the modelling of the dependence of both X_1 and X_2 on $\mathbf{H}$. The first is more straightforward, because for large u only the upper tail of X_1 need be considered, whereas the second ostensibly requires a model for the entire distribution of X_2, and this may be complex. We compromise by fitting similar models to both variables, letting the upper tails alone vary with $\mathbf{H}$. As we shall see below, this can greatly improve estimation of the causal

dependence structure relative to the original approach. Moreover fitting such a model should highlight simpler, potentially linear, structures in the tails, rather than more complex ones in the body of the data. This leads us to propose a peaks-over-threshold approach to estimating the conditional dependence of X_1 and X_2 on $\mathbf{H}$ (Section 3.3). Another useful tool, a reliable statistical test for direct causality, is discussed in Section 3.5.

3.3 Parametric tail causality and confounder dependence

3.3.1 Generalized Pareto causal tail coefficient

As mentioned above, we use the generalized Pareto distribution (GPD) to model the tails of our variables (Coles, 2001, Chapter 4). For $j = 1,2$, and under mild conditions on X_j, for a large enough threshold u_j large enough, we have

$$\mathbb{P}(X_j - u_j \leq x \mid X_j > u_j) \approx G(x;\sigma_j,\xi_j) = 1 - \left(1 + \xi_j x/\sigma_j\right)_+^{-1/\xi_j}, \quad x > 0, \tag{3.5}$$

with a scale parameter $\sigma_j > 0$ and a shape parameter $\xi_j \in \mathbb{R}$:

- $\xi_j = 0$ corresponds to light-tailed distributions, and then X_j lies in the maximum domain of attraction of the Gumbel distribution;
- $\xi_j > 0$ corresponds to heavy-tailed distributions, and then X_j lies in the maximum domain of attraction of the Fréchet distribution; and
- $\xi_j < 0$ corresponds to distributions with bounded upper tails, and then X_j lies in the maximum domain of attraction of the (reverse) Weibull distribution.

Any random variable satisfying the assumptions of Theorem 3.2.4 satisfies (3.5), as a regularly varying random variable with index $\alpha > 0$ lies in the Fréchet maximum domain of attraction. If the threshold u_j is chosen to be the q quantile of X_j for some $q \in (0,1)$, then we can write

$$\mathbb{P}(X_j \leq x) \approx \left\{G(x-u_j;\sigma_j,\xi_j)(1-q)+q\right\}\mathbb{1}(x > u_j) + \mathbb{P}(X_j \leq x)\mathbb{1}(x \leq u_j),$$

and using the empirical distribution $\hat{F}(x)$ to estimate $\mathbb{P}(X_j \leq x)$ and maximum likelihood estimation using the excesses of u_j to obtain $\hat{\sigma}_j$ and $\hat{\xi}_j$ yields a hybrid estimator of the distribution function $F_j(x)$ of X_j, i.e.,

$$\hat{F}_j(x;\hat{\sigma}_j,\hat{\xi}_j) = \hat{F}(x)\mathbb{1}(x \leq u_j) + \left\{G(x-u_j;\hat{\sigma}_j,\hat{\xi}_j)(1-q)+q\right\}\mathbb{1}(x > u_j).$$

The choice of q involves a bias–variance trade-off: q should be chosen large enough for the tail to be well approximated by a GPD, thus reducing the bias, but small enough to have enough exceedances, thus reducing the variance of the estimator. Using hybrid estimators for F_1 and F_2 for an integer $k \in \{1,\ldots,n-1\}$ yields the parametric *GPD causal tail coefficient* estimator for $\Gamma_{1,2}$,

$$\hat{\Gamma}_{1,2}^{\mathrm{GPD}} = \frac{1}{k_g}\sum_{i=1}^{n} \hat{F}_2(X_{i,2};\hat{\sigma}_2,\hat{\xi}_2)\mathbb{1}\left\{\hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1) > 1 - k/n\right\}, \tag{3.6}$$

where $k_g := |\{i \in \{1,\dots,n\} : \hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1) > 1-k/n\}|$. Unlike with the non-parametric estimator (3.3), the number of data pairs k_g used in (3.6) may not equal k, as it depends on the fit of $\hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1)$.

The GPD model can be extended to allow dependence on covariates of interest by expressing its parameters in the form $\theta(i) = h\{\boldsymbol{\gamma}^\top \mathbf{Z}(i)\}$, where θ denotes one or both of σ and ξ, h is an inverse link function, $\boldsymbol{\gamma}$ is a vector of parameters and $\mathbf{Z}(i)$ is the vector of explanatory variables on which the model might depend (Davison and Smith, 1990).

We wish to reparametrise the model to reduce or remove the effect on $\Gamma_{1,2}$ of a vector of potential confounders $\mathbf{H}$ of X_1 and X_2. If $\mathbf{H}$ is part of the LSCM then under the setup in Section 3.2 it is straightforward to show that $\mathbf{H}$ affects the scale parameters of the GPD model that applies to X_1 and X_2 above high thresholds, but not their shapes, so we write

$$\sigma_j(i) := \sigma_j^0 + \boldsymbol{\sigma}_j^{1\top}\mathbf{H}_i, \quad i = 1,\dots,n,\ j = 1,2, \tag{3.7}$$

where $\mathbf{H}_i$ is the replicate of $\mathbf{H}$ corresponding to the observations $(X_{i,1}, X_{i,2})$ of (X_1, X_2).

This yields, for $k \in \{1,\dots,n-1\}$, the parametric $\mathbf{H}$-*conditional linear generalized Pareto distribution (LGPD) causal tail coefficient estimator*,

$$\hat{\Gamma}^{\text{GPD}}_{1,2|\mathbf{H}} = \frac{1}{k_l}\sum_{i=1}^{n} \hat{F}_2\{X_{i,2};\hat{\sigma}_2(i),\hat{\xi}_2\}\mathbb{1}\left[\hat{F}_1\{X_{i,1};\hat{\sigma}_1(i),\hat{\xi}_1\} > 1-k/n\right]. \tag{3.8}$$

where $k_l := |\{i \in \{1,\dots,n\} : \hat{F}_1\{X_{i,1};\hat{\sigma}_1(i),\hat{\xi}_1\} > 1-k/n\}|$. Estimation of σ_j^0, σ_j^1 and ξ_j is performed by maximum likelihood. In applications it is preferable to center and rescale each confounder in $\mathbf{H}$ componentwise to unit variance and zero mean, to avoid numerical issues. Although the confounder is here assumed to be part of the LSCM, this does not seem to be necessary in practice, as non-linear effects can be approximated linearly, especially in the tail region. We investigate the effect of varying the tail index in Section 3.4.2.

3.3.2 The positive linear scale issue

Linear modelling of the GPD scale parameter may not yield positive scale estimates $\hat{\sigma}_j(i) > 0$ for each $i = 1,\dots,n$ and $j = 1,2$. The use of a nonlinear link function to ensure that the scale estimates were positive would not agree with the assumption of extremal linearity of the causal relationships, as the effect of $\mathbf{H}$ on the scale is also necessarily linear. We now describe two different solutions to this problem, which we compare by simulation in Section 3.4.

The first solution, *post-fit correction*, replaces $\hat{\sigma}_j(i)$ in (3.8) by $\max\{\hat{\sigma}_j(i), \epsilon\}$ for some arbitrary but small positive ϵ. The second solution, the *constrained approach*, applies the following linear constraints to the estimates when maximizing the likelihood

$$\sigma_j^0 + \boldsymbol{\sigma}_j^{1\top}\min_{i=1,\dots,n}\mathbf{H}_i > 0, \quad \sigma_j^0 + \boldsymbol{\sigma}_j^{1\top}\max_{i=1,\dots,n}\mathbf{H}_i > 0, \quad j = 1,2, \tag{3.9}$$

where $\min_{i=1,\dots,n}\mathbf{H}_i$ and $\max_{i=1,\dots,n}\mathbf{H}_i$ represent the vectors of componentwise minima and

maxima. When the data have a known distribution, box constraints can be used instead of (3.9). For example, in the case of a single confounder H and if X_1, X_2 and $X_h = H$ have t_ν distributions, then $\sigma_j^0 = u_j/\nu$ and $\sigma_j^1 = -\beta_{h\to i}/\nu$. Thus, if $\sigma_j(i) = \sigma_j^0 + \sigma_j^1 H_i > 0$ $(j = 1,2; i = 1,\dots,n)$, then

$$-\frac{u_j}{\nu \max_{i=1,\dots,n} H_i} < \sigma_j^1 < -\frac{u_j}{\nu \min_{i=1,\dots,n} H_i}, \tag{3.10}$$

where the lower and upper bounds are needed for positive and negative H_i, respectively.

3.4 Simulation study

Here we perform a simulation study using the Student t, Pareto and log-normal noise distributions. The first two lie in the Fréchet maximum domain of attraction and are regularly varying with index $\alpha = 1/\xi > 0$. We write Pareto(a,α) for the Pareto model with scale parameter a and tail index α; recal that lower values of α indicate heavier tails. This distribution satisfies Definition 3.2.3 exactly, so one might expect Pareto data to show better behaviour than Student data. The log-normal distribution, LogN(μ,σ^2) lies in the maximum domain of attraction of the Gumbel distribution and is not regularly varying, but finite samples from it can appear to be heavy-tailed.

We focus on the behaviour of the causal tail coefficient estimators (3.3) and (3.8) between two variables X_1 and X_2 in their causal configurations, as shown in Figure 3.2. As we study the estimators of causal effects of both X_1 on X_2 and of X_2 on X_1, we generated simulations only for the four causal cases, A, B, C and D. The LSCM causal weights $\beta_{2,1}$, β_{1h} and β_{2h} were chosen to equal 1.0, by default, for each existing edge in all four cases. Hence, in D, X_2 is caused by X_1 and the single confounder H with equal strength, even though H has another effect on X_2 through X_1.

Unless stated otherwise, each estimate is based on a random sample of $n = 10^6$ triples (X_1, X_2, H), of which $k = 2\lfloor n^{0.4} \rfloor = 502$ were chosen — Gnecco et al. (2021) found that the optimal fractional exponent of n for choosing k seems to lie between 0.3 and 0.4. The factor 2 doubles the number of data pairs used in the estimator, thus decreasing its variability, but does not introduce much bias for such large values of n. The GPD-based estimators are based on the top $(1-q)n$ observations, where we take $q = 0.9$, though only around k of the largest observations are used to estimate the coefficients Γ_{ij}. Setting $q = 0.95$ yields similar results. One thousand independent replicates were generated for each of the four causal configurations and three distributions.

We present only the highlights of the study; the code and all the results are available from `https://github.com/opasche/ExtremalCausalModelling`.

3.4.1 Variables with comparable tails

Detailed results for variables with comparable tails may be found in Section C.1 of the Supplementary Material. In this case it is essentially always possible to infer the existence and direction of any causality between X_1 and X_2, based on the non-parametric or **H**-conditional LGPD estimators, (3.3) or (3.8), of $\Gamma_{1,2}$ and $\Gamma_{2,1}$ alone. When the causal effects of H on X_1 and X_2, i.e., β_{1h} and β_{2h}, are increased relative to the noise variance and any causal effect $\beta_{2,1}$ of

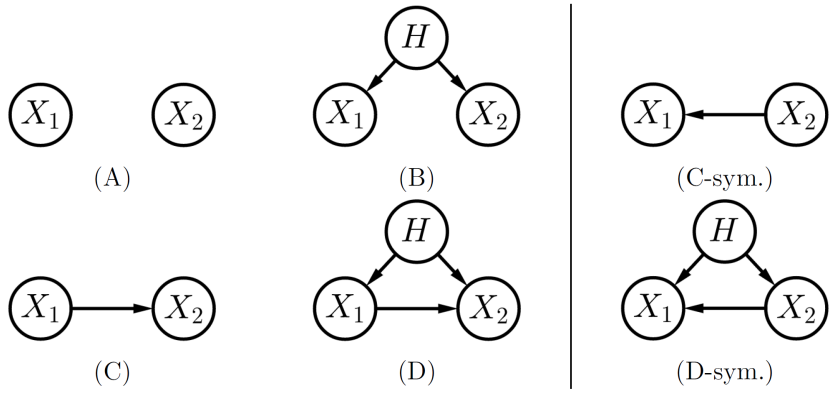

Figure 3.2: The six possible causal configurations between X_1 and X_2 with a possible confounder H, separated into the four cases studied in the simulations, and the two omitted by symmetry.

X_1 on X_2, both $\Gamma_{1,2}$ and $\Gamma_{2,1}$ increase in configuration B, and $\Gamma_{2,1}$ increases in configurations C and D. This increase is larger with the non-parametric estimators of $\Gamma_{1,2}$ and $\Gamma_{2,1}$, which are biased upwards in these configurations. When the confounder has a high causal impact, inference based on the non-parametric estimator (3.3) for direct causal link between X_1 and X_2 can fail, as $\hat{\Gamma}_{1,2}, \hat{\Gamma}_{2,1} \approx 1$ and hence $|\hat{\Gamma}_{1,2} - \hat{\Gamma}_{2,1}| \approx 0$ in configurations B and D.

Use of the **H**-conditional LGPD estimator (3.8) greatly reduces the effect of H on the coefficient estimates in configurations B and D. For Pareto and log-normal data, the results are indistinguishable from those without the confounder, both in terms of location and variability, as if the effect of H had been entirely removed. The estimates based on Student data are also shifted to around the same values as in the corresponding confounder-free configurations, though their upper tails are marginally heavier. These few greater values remain appreciably lower than without H as a covariate. For configurations A and C, unlike for B and D, the estimator is almost unaffected by the addition of H as a covariate when it is not a confounder. This is also a useful property, as it could allow tests of whether a specific covariate is a confounder of two variables, based on changes to the estimated coefficients.

3.4.2 Confounder with a different tail

One generalisation allows the tail of the distribution of H to be heavier or lighter than those of X_1 and X_2. A lighter tail does not negatively affect whether the non-parametric and **H**-conditional LGPD estimators can infer a direct causal relationship between X_1 and X_2, as the tails of X_1 and X_2 then dominate. Figure 3.3 shows the sampling distributions of $\hat{\Gamma}_{1,2}$ and $\hat{\Gamma}_{2,1}$ for all four causal structures when the tail of H is heavier than those of X_1 and X_2. The true coefficient values are unknown, as assumption (b) of Theorem 3.2.4 is not satisfied, though the coefficient for comparable tails, (3.2), is shown for comparison.

When H has a heavier tail than X_1 and X_2, the non-parametric estimators $\hat{\Gamma}_{1,2}$ and $\hat{\Gamma}_{2,1}$ in configuration B and $\hat{\Gamma}_{2,1}$ in configuration D are shifted well towards unity. With an even heavier-tailed, Student t_2, distribution for H (not shown here), the Student results resemble those for the Pareto and log-normal distributions. In all these cases it becomes impossible to infer a direct causal relationship between X_1 and X_2, owing to the effect of the heavier confounder tail

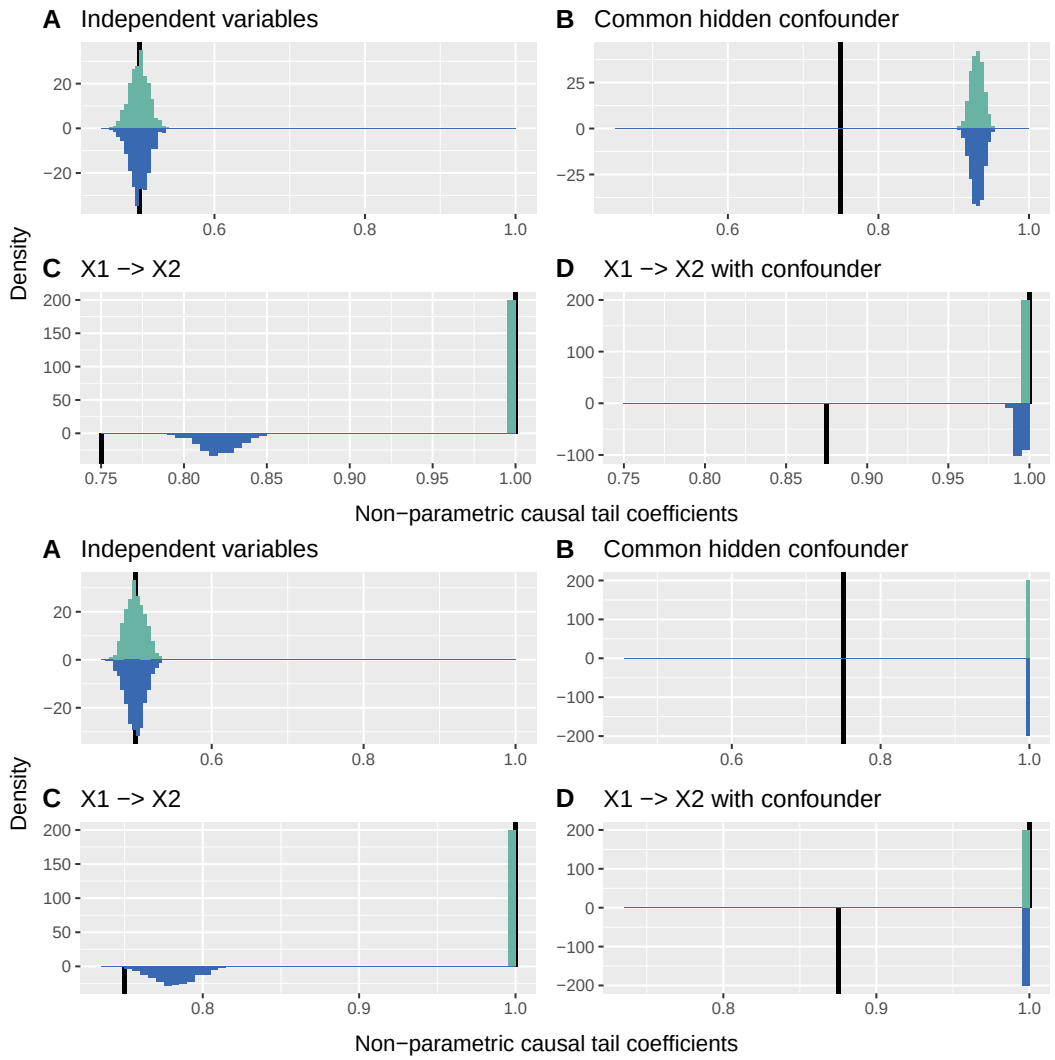

Figure 3.3: Histograms of $\hat{\Gamma}_{1,2}$ (turquoise) and $\hat{\Gamma}_{2,1}$ (blue) for t_4-distributed ε_1 and ε_2, and t_3-distributed H (top four panels) and for LogN$(0,1)$-distributed ε_1 and ε_2, and LogN$(0,1.5)$-distributed H (bottom four panels). Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails. The panels for Pareto$(1,3)$ distributed ε_1 and ε_2, and Pareto$(1,1.5)$ distributed H are very similar to the lower four panels.

on the non-parametric estimators.

Figure 3.3 shows that in configurations B and D the non-parametric estimator is badly affected by the heavier tail of H. Figure 3.4, which displays the sample distributions of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ with post-fit correction when the tail of H is heavier than those of X_1 and X_2, shows that the use of H as a covariate solves this problem: the estimates shift towards the coefficient values in the corresponding confounder-free cases, and consistently yield positive values of the difference of estimates $\hat{\Gamma}^{\text{GPD}}_{1,2|H} - \hat{\Gamma}^{\text{GPD}}_{2,1|H}$ for configuration D and differences centred at zero for configuration B; see also Section C.1 of the Supplementary Material. The estimates in configurations A and C, without the confounder causal effect, are barely changed by using H as a covariate.

Simulation results for $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ with the constrained fit are very similar to those for post-fit correction for the Pareto and log-normal distributions, but not for the Student distribution. Figure 3.5 shows the sample distribution of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ with the constrained fit, for a heavier

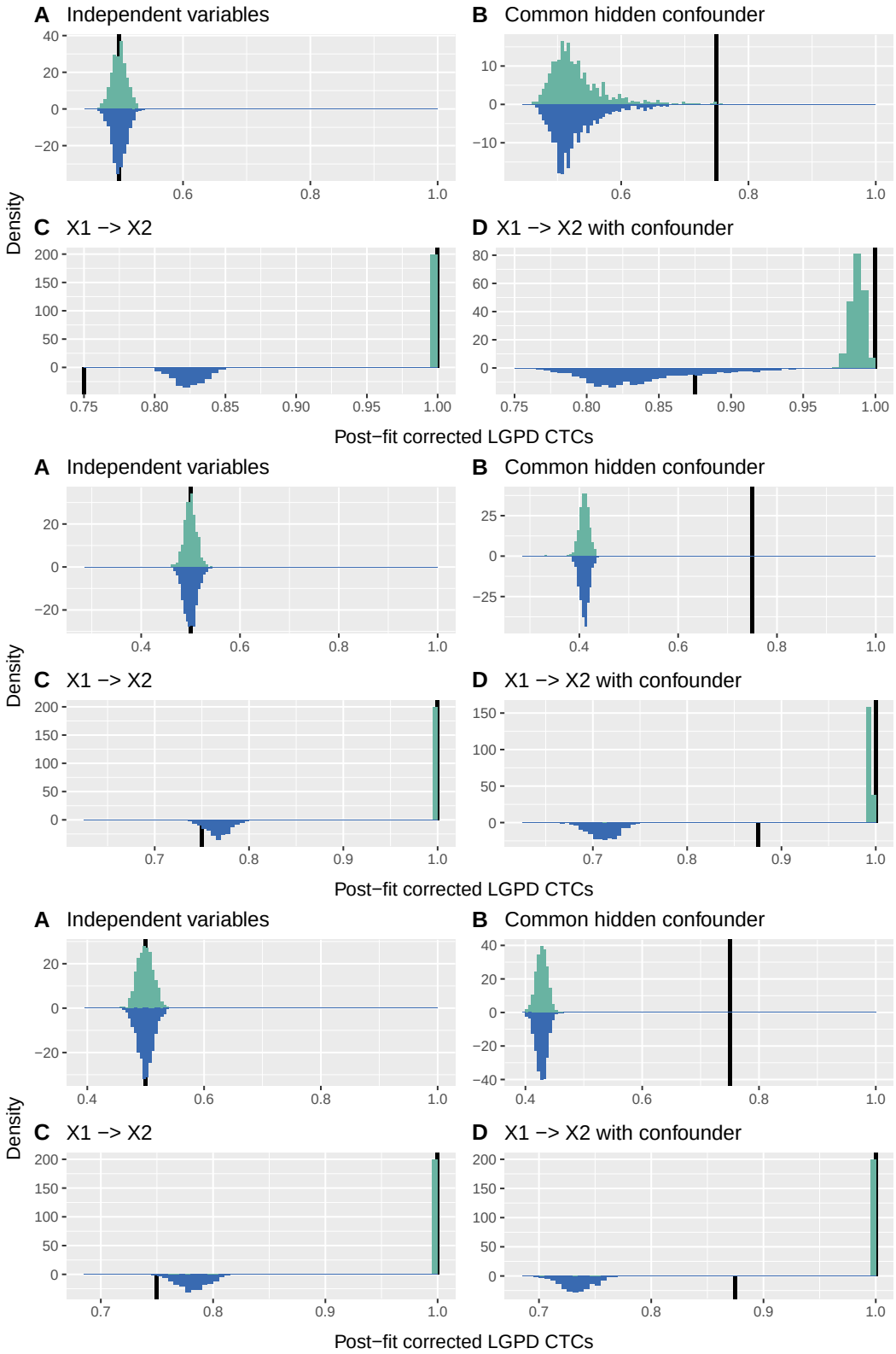

Figure 3.4: Histograms of $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ (blue) with post-fit correction for t_4 distributed ε_1 and ε_2, and t_3 distributed H (top four panels), for Pareto$(1,3)$ distributed ε_1 and ε_2, and Pareto$(1,1.5)$ distributed H (middle four panels), and LogN$(0,1)$ distributed ε_1 and ε_2, and LogN$(0,1.5)$ distributed H (lower four panels). Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails.

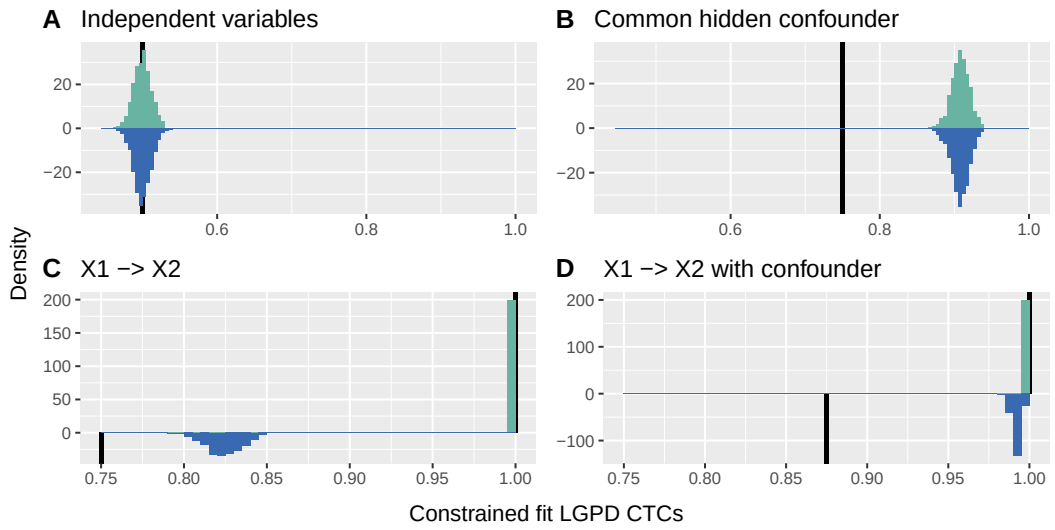

Figure 3.5: Histograms of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ (blue) with constrained fit for t_4 distributed ε_1 and ε_2, and t_3 distributed H. Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails.

confounder tail. For the Student distribution, the confounder affects the estimator appreciably more for the constrained fit than for post-fit correction, compared to the non-parametric results. As the Student distribution is heavy in both tails, the lower constraint in (3.9) forces $\hat{\sigma}_j(i)$ $(j = 1,2)$ to have an appreciably smaller slope, explaining this reduced effect. In configurations with a confounder, the absolute values of the constrained $\hat{\sigma}^1_j$ may be up ten times smaller than those for post-fit correction. With both approaches $\hat{\sigma}^1_j$ rarely differs greatly from zero for configurations without a confounder.

Both types of constraint yield very similar estimates for the Student distribution; see `https://github.com/opasche/ExtremalCausalModelling`.

To summarize, the simulations show that both the non-parametric estimator (3.3) and the **H**-conditional LGPD estimator (3.8) perform well when the theoretical assumptions are met and the influence of a hidden confounder is limited. When this influence grows, it becomes increasingly difficult to confidently infer the causal relationship between the variables using the non-parametric estimator, but the **H**-conditional LGPD estimator allows us to detect this relationship by reducing the effect of the confounding.

3.5 Testing for direct causality

3.5.1 Permutation test

In situations such as the causal analysis presented in Section 3.6, the distributions of the $\Gamma_{1,2}$ and $\Gamma_{2,1}$ estimators must be estimated to be used for inference. One way to obtain such distributions would be bootstrap resampling, but the extremal nature of the causal tail coefficient would require an unrealistically large sample size for its bootstrap distributions to be trustworthy, as these distributions tend to be too discrete in the extremes.

We therefore propose a permutation test (Davison and Hinkley, 1997, Chapter 4) for direct causality between two observed variables, measuring the asymmetry in their direct causal relationship. Suppose we have a sample $\{(X_{i,1}, X_{i,2})\}_{i=1}^{n}$ from a LSCM and wish to test the null hypothesis of no direct causal relationship between X_1 and X_2, $H_0 : \beta_{2,1} = 0$, versus the alternative that X_1 causes X_2, $H_A : \beta_{2,1} > 0$. Our proposed procedure is as follows:

1. Rescale values $\tilde{X}_{i,j} = \tilde{F}_j(X_{i,j})$ $(i = 1, \ldots, n,\ j = 1, 2)$, where known confounders can be used in the distribution estimator $\tilde{F}_j$, as for $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$.
2. For $r = 1, \ldots, R$, obtain $\tilde{X}^{(r)}_{i,1}$ and $\tilde{X}^{(r)}_{i,2}$ by randomly permuting the indices $j = 1, 2$ for each pair $(\tilde{X}_{i,1}, \tilde{X}_{i,2})$ $(i = 1, \ldots, n)$.
3. Compute $\tilde{\Delta}_{1,2} = \tilde{\Gamma}_{1,2} - \tilde{\Gamma}_{2,1}$ on the transformed original data $\{(\tilde{X}_{i,1}, \tilde{X}_{i,2})\}_{i=1}^{n}$ and $\tilde{\Delta}^{*r}_{1,2} = \tilde{\Gamma}^{*r}_{1,2} - \tilde{\Gamma}^{*r}_{2,1}$ on their bootstrapped values $\{(\tilde{X}^{(r)}_{i,1}, \tilde{X}^{(r)}_{i,2})\}_{i=1}^{n}$ $(r = 1, \ldots, R)$.
4. Obtain the Monte Carlo p-value, by comparing the value of the test statistic on the original rescaled data with the permutation distribution,

$$p_{\mathrm{mc}} = \frac{1 + \#_r\{\tilde{\Delta}^{*r}_{1,2} \geq \tilde{\Delta}_{1,2}\}}{R + 1}.$$

If there are no asymmetric confounding effects on the two variables, i.e. $\beta_{1h} = \beta_{2h}$ in the case of a single confounder, then $\Delta_{1,2} := \Gamma_{1,2} - \Gamma_{2,1} = 0$ under H_0, whereas $\Delta_{1,2} > 0$ under H_A; see equation (3.2) and Theorem 3.2.4. This does not hold generally with asymmetric confounding. The direct causal relationship is symmetric under H_0, i.e., X_2 is as likely to take extreme values when X_1 is extreme as is X_1 when X_2 is extreme. If so, then permutations such as those performed in step 2. are equally likely, so $\tilde{\Delta}_{1,2}, \tilde{\Delta}^{*1}_{1,2}, \ldots, \tilde{\Delta}^{*R}_{1,2}$ have a common distribution centered around zero, and p_{mc} will be uniformly distributed. Under the alternative, the direct causal relationship is "asymmetric", as X_2 is more likely to be extreme when X_1 is extreme than conversely; then $\tilde{\Delta}_{1,2}$ is more likely to lie in the upper tail of $\tilde{\Delta}^{*1}_{1,2}, \ldots, \tilde{\Delta}^{*R}_{1,2}$. Thus the distribution of p_{mc} will become increasingly skewed towards zero as the causal strength of X_1 on X_2 increases.

If all asymmetric confounding effects are captured in $\tilde{F}_j$ by estimating the distribution conditionally, X_1 and X_2 have comparable tails and causal effects behave linearly in the extremes, then the proposed procedure should provide a reliable p-value for testing direct causality of X_1 on X_2.

3.5.2 Simulations

We used simulation from different data distributions and for different causal configurations involving X_1, X_2 and a potential confounder H to assess our proposed test. We used values of $0, 0.01, 0.05, 0.1, 0.2$ for the causal strength $\beta_{2,1}$ of X_1 on X_2, with confounding effects both present and absent. Symmetric ($\beta_{1H} = \beta_{2H} = 1$) and asymmetric ($\beta_{1H} = 0.8$ and $\beta_{2H} = 1$, or $\beta_{1H} = 1$ and $\beta_{2H} = 0.8$) confounding effects were considered, and the noise variable were Pareto, Student t and log-normal. We generated $m = 10^3$ replicate samples of $n = 10^4$ independent triples $(X_{i,1}, X_{i,2}, H_i)$ for each causal configuration and noise distribution. The sample size n was chosen closer to practical orders of magnitude, compared to our large-sample study in Section 3.4. Three versions of the permutation test were performed for each sample, corresponding to the

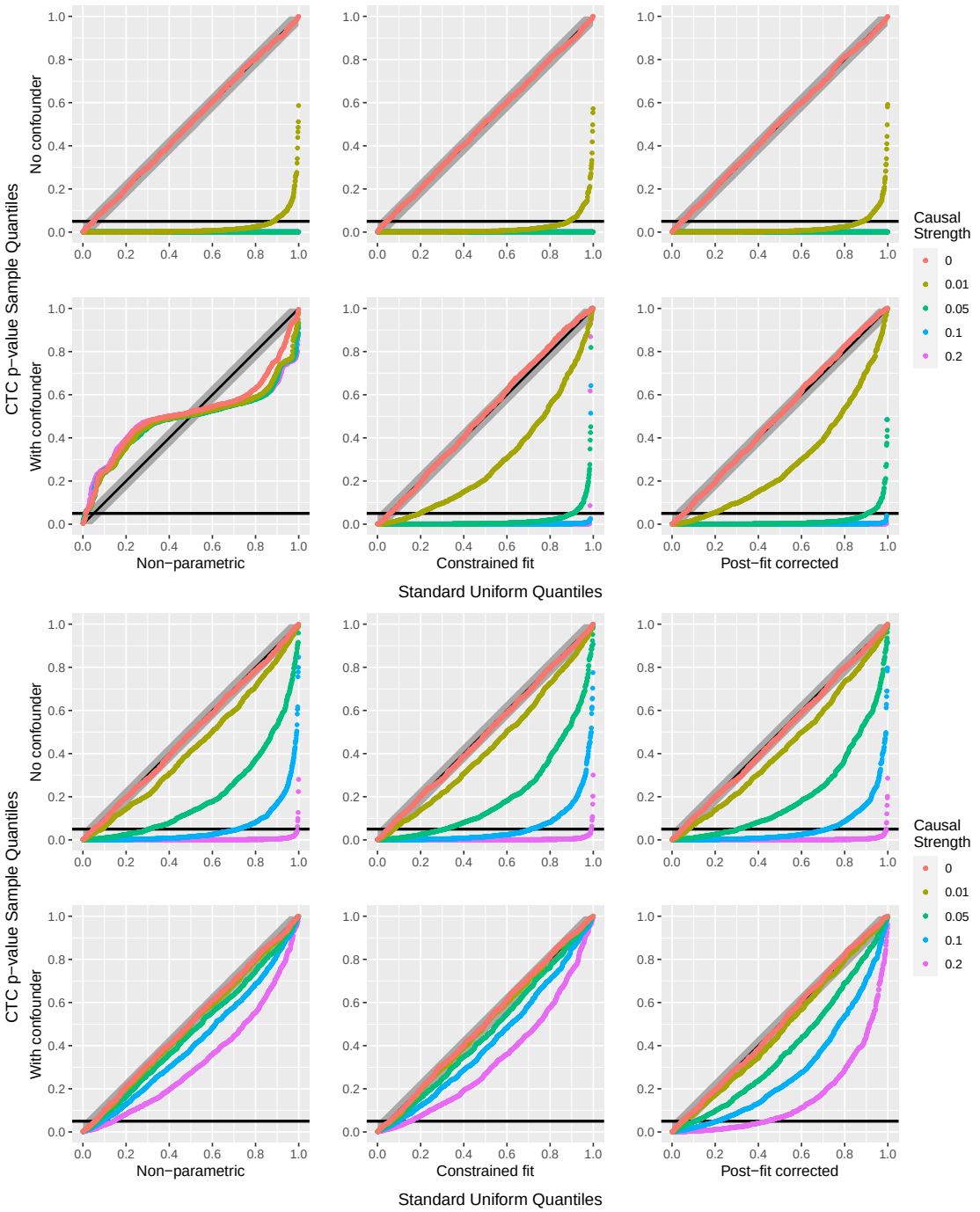

Figure 3.6: Uniform QQ-plots of Monte Carlo p-values p_{mc}, with Kolmogorov–Smirnov confidence bands for different causal strengths $\beta_{2,1}$ (colors), the three estimators (columns) and optional symmetric confounding effects, $\beta_{1H} = \beta_{2H} = 1$ (rows). Top six panels: Pareto$(1,2)$ distributed ε_1 and ε_2, and Pareto$(1,1)$ distributed H. Bottom six panels: t_4 distributed ε_1 and ε_2, and t_3 distributed H.

causal tail coefficient estimators discussed in Sections 3.2 and 3.3: the non-parametric (3.3), and **H**-conditional LGPD (3.8) with either post-fit correction or constrained fit. Each used $R = 10^3$ permutations and the estimator hyper-parameters were set to $k = 2\lfloor n^{0.4} \rfloor = 78$ and $q = 0.9$.

Figure 3.6 shows uniform QQ-plots of p_{mc} for the Pareto and Student distributions, in the case of

heavier confounder tail, with symmetric effects. In the absence of confounding the test behaves as expected in both cases, and adding dependence on the independent H variable in the modelling through the parametric estimators has no visible effect on the distribution of p_{mc} compared to the non-parametric approach. For the Pareto distribution, the test has a power of almost 0.9 for a direct causal strength of 0.01, and it behaves perfectly for higher causal strengths. For the Student distribution, the test reaches a power of 0.3 for a direct causal strength of 0.05, of 0.7 for causal strength of 0.1 and of near 1.0 for a causal strength of 0.2.

When the confounding effects are added, the test based on the non-parametric estimator fails for the Pareto distribution, as most of the p_{mc} then lie outside the 95% confidence bands, indicating that the distribution of p_{mc} is highly non-uniform. This is corrected when the value of the confounder is taken into account using the parametric approaches, with power 0.9 for a direct causal strength of only one twentieth of the confounder's marginal effects. In the Student case, p_{mc} seems to be close to uniformity in the absence of direct causality (the difference in tail shape is much greater in the Pareto case), but post-fit correction increases the power from below 0.2 to above 0.4 for a direct causal strength of one fifth of the confounder's marginal effects. Similar conclusions to those of Section 3.4.2 about the constrained fit for distributions with both tails heavy apply, as the constrained fit estimator is not significantly better than the non-parametric estimator compared to post-fit correction.

Unlike in the corresponding symmetric case, the test here fails when using the non-parametric estimator owing to the asymmetry induced by the confounder, but both parametric approaches remove this unwanted effect by enough that p_{mc} nearly has a uniform distribution, with almost perfect power, for a causal strength of one sixteenth and one twentieth of the marginal confounding effects.

Figure 3.7 shows the uniform QQ-plots with asymmetric confounding effects for the Pareto distribution with comparable tails.

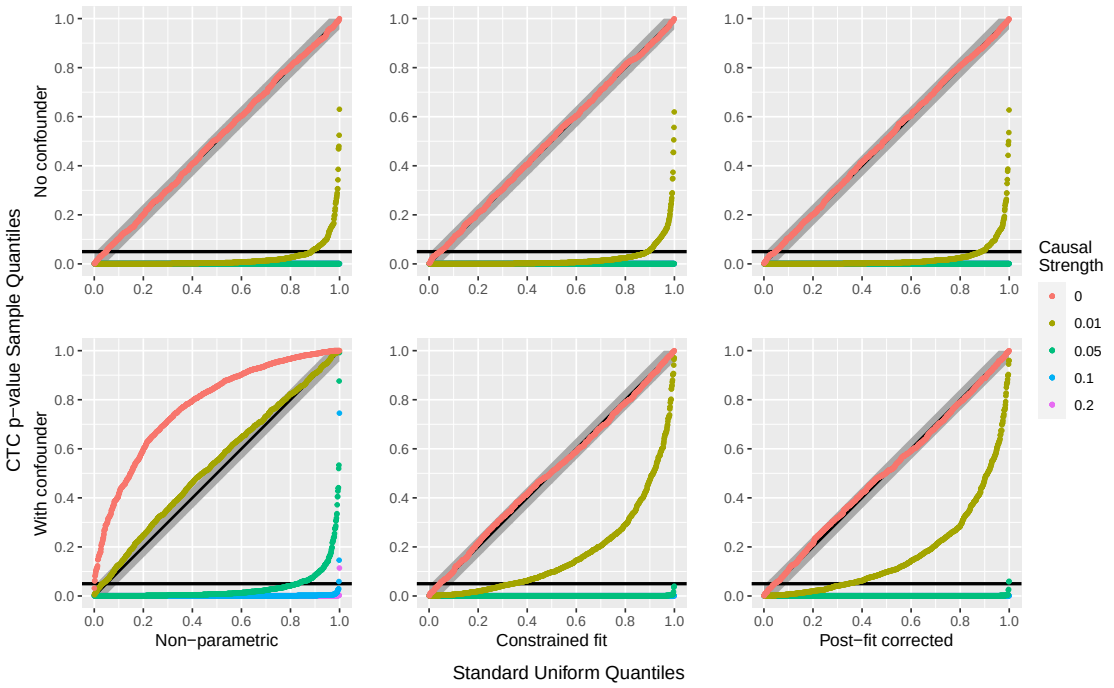

Figure 3.7: QQ-plot of the p_{mc} estimates against the standard uniform distribution, with Kolmogorov–Smirnov confidence bands, for Pareto$(1,2)$ distributed ε_1, ε_2 and H, for different causal strengths $\beta_{2,1}$ (colors), the three estimators (columns) and optional asymmetric confounding effects, $\beta_{1H} = 0.8$, $\beta_{2H} = 1$ (rows).

3.6 Application to Swiss rivers

We now illustrate how our method can discover direct causal relationships between the discharge extremes of pairs of river stations. This illustrates our method on a real example for which we know the 'ground truth' of extremal causality, but unlike in the simulations of Section 3.4, we cannot control and do not know the true tail behaviour of the station discharges and their potential confounders.

3.6.1 Data sources and additional collection

We use the average daily discharges between January 1913 and December 2014 at the 68 Swiss gauging stations shown in Figure 3.1, and add daily precipitation data from 105 meteorological stations during the same period. Some additional information, such as the station elevation, catchment surface area and mean elevation, glaciation percent and coordinates, was collected from the Federal Office for the Environment's website. To reduce any seasonal effects due to unobserved confounders, we only consider data during June, July and August, as the more extreme observations happen during this period when mountain rivers are less likely to be frozen. Temporal clustering is likely to appear for average daily discharge data but can be captured by considering the average catchment precipitation as a covariate in the model for the GPD scale parameter (3.7).

Figure 3.8 shows relationships between the estimates, station altitudes and average discharges.

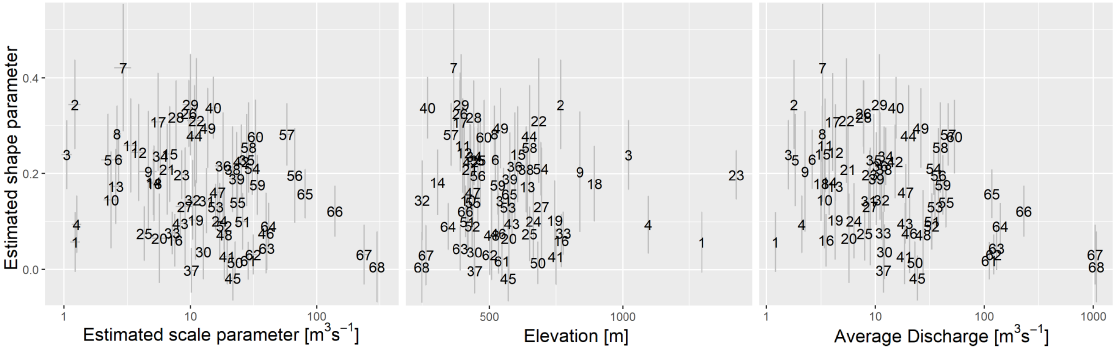

Figure 3.8: Relation between shape parameter estimates, scale parameter estimates (log scale), station elevation and average discharge (log scale), with standard errors (±SE) shown as error bars.

Altitude does not greatly affect the estimates, but the shape parameter estimates broadly decrease with increased average river discharge volume.

3.6.2 Choice of stations and comonotonicity

For the causal analysis, we consider pairs of stations with known direct causal relationships, and pairs with no direct causal relationship. Causal pairs are ordered by the flow of water, with one downstream of the other. The river volumes for the pairs should be as similar as possible, as our exploratory analysis indicated different tail behaviours for rivers with very different average discharges. There should also be enough confluences between the two stations, otherwise one would observe *comonotonicity*, i.e., almost perfect dependence, between their discharges. If there is comonotonicity between X_1 and X_2, then $F_1(X_{i,1}) \approx F_2(X_{i,2})$, for all $i = 1, \ldots, n$, and it is impossible to know which variable causes which based on the data alone regardless of the approach, even if one is certain both of direct causality and of its direction. Confluences between the two stations reduce comonotonicity and make it possible to detect the direction of causality.

As we shall use precipitation as the confounding covariate, the stations must share likely meteorological effects and must lie in regions where precipitation data is available. Based on these criteria, we chose seven causal station pairs: $(43, 62)$, $(42, 63)$, $(36, 63)$, $(24, 61)$, $(44, 61)$, $(22, 38)$, $(22, 35)$, where the first station of each pair lies upstream from the second.

The non-causal station pairs were selected to have similar average volume and similar shape parameter estimates. Pairs with stations separated by long distances and pairs relatively close to each other were both considered. The 13 pairs selected are $(30, 45)$, $(36, 39)$, $(42, 34)$, $(32, 33)$, $(62, 63)$, $(57, 60)$, $(13, 14)$, $(17, 22)$, $(12, 21)$, $(26, 28)$, $(27, 31)$, $(23, 39)$, $(23, 35)$.

The choice of covariate for the causal pairs was the mean daily precipitation among the meteorological stations in the area and the catchment of the two stations. The choice of covariate was less meaningful for the non-causal pairs with large separating distances, which have different meteorological conditions, so the average daily precipitation over the whole country was used. For the pair $(42, 34)$, which has the closest stations and local precipitation data available, the daily average in the local catchments was also considered. In the latter case, the pair will be highlighted with an asterisk to avoid confusion.

3.6.3 Causal analysis results

For each station pair, the permutation test for direct causality was performed using the non-parametric (3.3) and **H**-conditional LGPD (3.8) estimators with post-fit correction or constraints, with $R = 10^4$ permutations and estimator hyper-parameters $k = 1.5\lfloor n^{0.4} \rfloor$ and $q = 0.9$. Table 3.2 shows the values of p_{mc}, the covariate shape estimate and its estimated extremal linear effects for the two stations, the latter estimated without constraints. The number of common observations for the pairs varies from 2,024 to 8,464, and k lies between 31 and 55. With precipitation covariates added, the number of common observations ranges from 1,483 to 7,820, and k lies between 27 and 54.

With the non-parametric approach for the causal stations, the absence of direct causality was rejected for four of the seven station pairs at significance level 5%, and for two of these four at level 2.5%. Adding daily precipitation as a covariate by either parametric approach decreases the p-values but two pairs remain non-significant; both lie in the same region and contain station 22.

With the non-parametric approach, the absence of direct causality was not rejected for ten of the 13 non-causal station pairs. Adding precipitation as a covariate with the two parametric approaches 'corrected' the p-value for another station. For the pair $(42, 34)$ using local instead of global precipitation as a covariate gave a higher p-value.

We also considered using an exponential rather than a linear inverse-link function, i.e., taking $\log \sigma_j(i) = \sigma_j^0 + \sigma_j^1 H_i$ $(i = 1, \ldots, n; j = 1, 2)$, to avoid any need for correction or constraints. The resulting p_{mc} values, also shown in Table 3.2, lead to the same conclusions as with the linear approaches.

Using the usual normal approximation, every $\hat{\sigma}_1^1$ is significantly positive for the causal pairs and 10 of the 14 estimates are positive for the non-causal pairs, with the highest confidence for the pair using local precipitation. Standard errors for $\hat{\sigma}_2^1$ are systematically larger than those for $\hat{\sigma}_1^1$ for the causal pairs, perhaps owing to the double causal effect of the covariate on the downstream station, both direct and indirect through the upstream station, as we do not observe this systematically for non-causal pairs. Consequently, the $\hat{\sigma}_1^1$ estimates are significantly positive for only four of the seven causal pairs, to be contrasted with 12 of the 14 estimates for the non-causal pairs. In particular, only the local precipitation effect is significant for the pair $(42, 34)$.

We compare our results to two classical causal inference approaches appropriate to our problem. These are a non-Gaussian method for estimating causal linear structures based on results from independent component analysis, ICA-LiNGAM (Shimizu et al., 2006), and the PC algorithm, which retrieves the completed partially directed acyclic graph by performing conditional independence tests on the variables. For the latter, we consider both the classic PC algorithm (Spirtes et al., 2000), which uses Gaussian conditional independence tests, and the Rank PC algorithm (Harris and Drton, 2013), which uses rank-based Spearman correlation to perform the independence tests and thus is more robust to non-normal variables. The results for the ICA-LiNGAM method are presented in Table C.1 in the Supplementary Material, which shows the linear causal coefficients for the discharge station pairs estimated with the ICA-LiNGAM algorithm using either the station pair only (two variables) or the station pair and precipitation

Table 3.2: Permutation p-values p_{mc} for station pairs using the non-parametric approach (NP), the **H**-conditional post-fit corrected (PFC) and constrained fit (CF) LGPD approaches, and an **H**-conditional exponential inverse-link GPD approach (Exp). The shape estimate $\hat{\xi}_H$ for the precipitation covariate and the unconstrained scale slope estimates are also shown (with standard errors of at most 0.03 for the former and in parentheses for the latter).

Stations	Pair type	NP	PFC	CF	Exp	$\hat{\xi}_H$	$\hat{\sigma}_1^1$	$\hat{\sigma}_2^1$
43-62	causal	0.01	0.01	0.01	0.01	0.06	0.88(0.3)	1.91(1.3)
42-63	causal	0.03	0.02	0.02	0.04	0.06	6.49(1.1)	8.60(2.2)
36-63	causal	0.03	0.02	0.02	0.03	0.06	5.03(1.1)	7.25(2.8)
24-61	causal	0.06	0.01	0.01	0.00	−0.01	3.42(1.2)	−2.34(2.4)
44-61	causal	0.01	0.00	0.00	0.01	0.01	1.89(0.7)	−1.21(2.0)
22-38	causal	0.58	0.40	0.40	0.33	0.07	3.43(0.8)	8.00(2.0)
22-35	causal	0.22	0.17	0.17	0.10	0.03	3.43(0.9)	11.67(3.0)
30-45	non-caus.	0.56	0.47	0.47	0.46	0.01	1.01(0.4)	0.89(0.9)
36-39	non-caus.	0.80	0.70	0.70	0.69	0.01	4.61(1.1)	4.17(1.6)
42-34	non-caus.	0.23	0.04	0.04	0.10	0.01	5.97(1.2)	0.43(0.3)
42-34*	non-caus.	0.23	0.13	0.13	0.11	0.05	6.29(1.1)	0.66(0.3)
32-33	non-caus.	0.01	0.01	0.01	0.00	0.01	0.63(0.4)	1.00(0.3)
62-63	non-caus.	0.10	0.49	0.48	0.30	0.01	1.08(1.4)	7.67(2.1)
57-60	non-caus.	0.99	1.00	1.00	1.00	0.01	6.31(3.7)	5.23(1.8)
13-14	non-caus.	0.32	0.56	0.56	0.53	0.01	0.59(0.2)	1.19(0.3)
17-22	non-caus.	0.01	0.05	0.06	0.05	0.01	0.78(0.5)	2.18(0.7)
12-21	non-caus.	0.51	0.50	0.50	0.72	0.01	0.71(0.3)	1.33(0.4)
26-28	non-caus.	0.63	0.90	0.89	0.92	0.01	1.90(0.5)	1.63(0.4)
27-31	non-caus.	0.40	0.63	0.62	0.75	0.01	1.71(0.7)	2.91(1.1)
23-39	non-caus.	0.80	0.91	0.92	0.93	0.01	2.50(0.6)	4.27(1.5)
23-35	non-caus.	0.65	0.88	0.89	0.86	0.01	2.50(0.6)	6.66(1.7)

(three variables). Non-null values indicate significant causal effects. The upper-script arrows indicate the estimated direct causal direction between the station pair. Although in both cases of the two or three variables, ICA-LiNGAM retrieves all the correct causal pairs, with correct direction, all the non-causal pairs are indicated by non-null values as significantly causal. Both versions of the PC algorithm, once applied to our 21 pairs, provide existing direct causal links (without weights nor direction) between all the pairs of stations. Apparently both ICA-LiNGAM and PC methods are too eager to detect causality, unlike the tail coefficients. One explanation could be a set of unobserved confounders related to common global weather conditions triggering causal effects even between stations that are far apart. Extreme discharges depend more on local weather conditions, and particularly on heavy precipitation. Another explanation could be that causal effects are only linear in the tails, perhaps due to ground saturation by precipitation.

3.7 Discussion and conclusion

This paper addresses the reduction or removal of the unwanted effect of known confounders from the extremal causal analysis between two variables and the discovery of extremal causal relationships using a parametric estimator of the causal tail coefficient, based on generalized Pareto modelling, and a permutation test for direct causality. Both allow the use of known

confounders as covariates.

In our simulation study, the new estimator removed the confounder's unwanted effect almost entirely for variables with comparable tails, and reduced its effect enough to allow correct causal inference on the direct causal relationship in the case of a confounder with a heavier tail. The permutation test was shown to provide reliable p-values when all asymmetric confounding effects are captured in the model.

When applied to Swiss river discharge data, our methodology allowed correct inference on the direct causal relationships between discharges for the majority of the chosen station pairs, and the parametric approach captured the confounding effect of precipitation.

In many real-life situations, statistically significant covariates need not correspond to causal effects. Peters et al. (2016) have proposed a methodology for causal discovery, for when data from different settings or regimes are observed. Their method constructs invariant causal regression or classification models that should still make accurate predictions under interventions on the covariates or a change of environment. Adapting this approach to our setting would lead to a better understanding of causality of extremes.

Declarations

Acknowledgements

We thank the referees and associate editor for their helpful remarks.

Funding

This work was supported by the Swiss National Science Foundation. Open access funding was provided by University of Geneva.

Data availability

The data that support the findings of this study may be obtained from the Swiss Federal Office for the Environment (`https://hydrodaten.admin.ch/`) and the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss (`https://opendatadocs.meteoswiss.ch/`), but restrictions applied, as the data were used under licence for the study. The combined and preprocessed version of the data used in this study are, however, available from the authors upon reasonable request.

4 Validating deep-learning weather forecast models on recent high-impact extreme events

Olivier C. Pasche[1], Jonathan Wider[2,3], Zhongwei Zhang[1],
Jakob Zscheischler[2,3,4], Sebastian Engelke[1]

[1]*Research Institute for Statistics and Information Science, University of Geneva, Switzerland*
[2]*Department of Compound Environmental Risks, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany*
[3]*Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany*
[4]*Department of Hydro Sciences, TUD Dresden University of Technology, Dresden, Germany*

This chapter is a postprint of the homonymous article published in *Artificial Intelligence for the Earth Systems* (Pasche et al., 2025b), with doi:10.1175/AIES-D-24-0033.1.

Abstract

The forecast accuracy of machine learning (ML) weather prediction models is improving rapidly, leading many to speak of a "second revolution in weather forecasting". With numerous methods being developed and limited physical guarantees offered by ML models, there is a critical need for a comprehensive evaluation of these emerging techniques. While this need has been partly fulfilled by benchmark datasets, they provide little information on rare and impactful extreme events or on compound impact metrics, for which model accuracy might degrade due to misrepresented dependencies between variables. To address these issues, we compare ML weather prediction models (GraphCast, PanguWeather, and FourCastNet) and ECMWF's high-resolution forecast system (HRES) in three case studies: the 2021 Pacific Northwest heatwave, the 2023 South Asian humid heatwave, and the North American winter storm in 2021. We find that ML weather prediction models locally achieve similar accuracy to HRES on the record-shattering Pacific Northwest heatwave but underperform when aggregated over space and time. However, they forecast the compound winter storm substantially better. We also highlight structural differences in how the errors of HRES and the ML models build up to that event. The ML forecasts lack important variables for a detailed assessment of the health risks of the 2023 humid heatwave. Using a possible substitute variable, prediction errors show spatial patterns with the highest danger levels over Bangladesh being underestimated by the ML models. Generally, case-study-driven,

impact-centric evaluation can complement existing research, increase public trust, and aid in developing reliable ML weather prediction models.

Significance statement

With the performance of machine-learning-based weather forecasting models improving rapidly, thorough analyses are needed to ensure that their forecasts are accurate and reliable before deploying them in operational settings. Existing evaluations often reduce forecast performance to a few metrics, potentially obscuring rare but systematic errors. This is especially problematic for high-impact extreme events, which, by definition, are rare in the data but often substantially affect society. In a detailed analysis of three extreme events, we observe that, although machine learning (ML) models generally outperform the best physics-based numerical weather prediction (NWP) model on benchmark datasets, they do not consistently do so for the studied extreme events or compound impact metrics and lack some impact-relevant variables.

Keywords: extreme events, heat wave, humidity, forecast verification/skill, numerical weather prediction/forecasting, deep learning.

4.1 Introduction

In recent years, the performance of machine learning (ML) weather forecast models has improved drastically (Rasp et al., 2024), leading some authors to speak of a "rise" of ML methods in weather forecasting (Ben Bouallègue et al., 2024), or even a second revolution of the field. While some studies focused on short-term predictions ("nowcasting", Espeholt et al., 2022; Leinonen et al., 2023; Andrychowicz et al., 2023), and long-term subseasonal-to-seasonal forecasting (two weeks to two months ahead, Weyn et al., 2021; Lopez-Gomez et al., 2023), much work concentrated on the medium range (Rasp and Thuerey, 2021; Pathak et al., 2022; Nguyen et al., 2023a; Chen et al., 2023a; Bi et al., 2023; Kochkov et al., 2024; Lam et al., 2023; Chen et al., 2023b; Nguyen et al., 2023b; Price et al., 2023), i.e., forecasting days to two weeks into the future.

The established technique for weather forecasting in the medium range is numerical weather prediction (NWP), which is based on evolving an estimate of the current weather state constructed from observations through time under differential equations. Therefore, the point of comparison for ML approaches is ECMWF's Integrated Forecasting System (Owens and Hewson, 2018), including its high-resolution forecast system (HRES), which is generally considered to be the most reliable NWP model for global deterministic weather forecasts. The latest ML weather models match or even outperform HRES in terms of overall summary scores across many variables, pressure levels, and prediction lead times (Rasp et al., 2024). In addition to performance, other reasons to consider ML-based weather forecasting include energy efficiency during operations and improved inference speed. ML models that supplement or replace parts of the weather forecasting pipeline are increasingly seen as a realistic possibility (Bauer, 2024; Ben Bouallègue et al., 2024). ECMWF is already publishing forecast data produced with its own AI model, AIFS (Lang et al., 2024), as part of its experimental suite.

Given these recent advances and the immense importance of accurate and robust weather forecasting to many aspects of human life, thorough analyses are necessary before operationalizing ML weather prediction models. As extreme weather events often have severe impacts (Zscheischler et al., 2020; Seneviratne et al., 2021), such as crop loss, wildfires, and floods, effective mitigation measures require accurate predictions in the tails of the distribution.

While ML-based weather forecasts can achieve high overall accuracy, their performance for extreme events is not well understood. ML models generally face fundamental difficulties during extrapolation and generalization to unseen domains, and good test accuracy estimates do not guarantee good performance outside the range of previous observations, or in regions of the input space where observations were scarce (Hastie et al., 2009; Watson, 2022).

Summary scores, like the root mean squared error (RMSE), play a central role in the evaluation of ML weather prediction models. Typically, one score is computed for each lead time, predicted variable, and (pressure) level to quantify the model's performance over the entire test set (see, e.g., scorecards in Rasp et al., 2024). Several other aspects of ML forecasts have also been studied in the literature. For instance, Bonavita (2024), Lam et al. (2023) and Rasp et al. (2024) examined the smoothness of the predictions and found that most ML models tend to blur predictions for long lead times as a consequence of the way these models are conceptualized and trained. Bonavita (2024) studied PanguWeather, one of the best-performing ML models, and found it to be worse at maintaining physical balances than ECMWF's HRES.

Olivetti and Messori (2024a) summarized some of the extreme-event evaluations performed for the latest generation of ML weather forecast models. In previous work, extreme temperatures (both hot and cold) were studied by comparing threshold exceedances of predictions and ground truth data (Ben Bouallègue et al., 2024; Lam et al., 2023; Olivetti and Messori, 2024b). Other types of investigated extreme events include tropical cyclones, atmospheric rivers, and storm systems (Magnusson, 2023; Ben Bouallègue et al., 2024; Lam et al., 2023; Charlton-Perez et al., 2024). While some studies have looked into individual events, many types of extremes are still under-explored, especially on a case-study level.

In addition, little attention has been paid to impact metrics that combine multiple predicted variables, or to events where accurate assessment of their spatial or temporal extent is important. The compounding effect of multiple variables in space and time can lead to particularly large impacts (Zscheischler et al., 2020). Examining prediction performance for these events in case studies is necessary to increase public trust in ML models, and also has the potential to uncover rare systematic errors in the ML model predictions that might be hidden by summary scores.

This study evaluates the ability of three popular ML weather prediction models to accurately forecast relevant impact metrics of extreme weather events through three case studies. The ML models GraphCast (Lam et al., 2023), PanguWeather (Bi et al., 2023), and FourCastNet (Pathak et al., 2022) are compared to IFS HRES (Owens and Hewson, 2018) for the 2021 Pacific Northwest heatwave, the 2023 South Asian humid heatwave, and the 2021 North American winter storm.

4.2 Data and models

4.2.1 Data

In this paper, we use two kinds of data: ERA5 reanalysis data (Hersbach et al., 2020) and ECMWF HRES analysis data. All ML models considered in this study were trained on ERA5, which is produced using data assimilation, i.e., by combining observations with short-range forecasts to obtain a "best guess" of the actual weather state. ERA5 has a horizontal resolution of $0.25° \times 0.25°$, an hourly temporal resolution, and provides estimates of many atmospheric, land, and oceanic climate variables over the globe from 1940 to present. The ML models GraphCast and FourCastNet have an internal time step of 6 h and were trained on a subset of ERA5 at 00:00, 06:00, 12:00, and 18:00 UTC. Therefore, we also restrict our analyses to these times of day.

We use HRES forecasts of versions 47r1, 47r2, and 47r3 from ECMWF's Integrated Forecasting System. They have a horizontal resolution of $0.1° \times 0.1°$, and we downsample them to the $0.25° \times 0.25°$ grid using the default Meteorological Interpolation and Regridding (MIR) library in the ECMWF Meteorological Archival and Retrieval System (MARS). ERA5 and HRES data can be retrieved from online archives. HRES forecasts initialized at 00:00 UTC or 12:00 UTC are archived for lead times up to 10 d, while forecasts initialized at 06:00 UTC and 18:00 UTC are only available for lead times up to 3.75 d. To ensure a fair comparison, we use "HRES forecast at step 0" (HRES-fc0) as the ground truth for HRES forecasts. If ERA5 was used instead, HRES would have a non-zero error at lead time 0 h. We use HRES forecast data with lead times ranging from 0 h to the maximum available length in steps of 6 h, so that the lead times match those of the ML forecasts.

A difference between our comparison study and Ben Bouallègue et al. (2024) is the data used for initializing and evaluating ML models. In both cases, they used HRES data to ensure fairness in an operational context, since ERA5 reanalysis data are simply not available at the time of an operational prediction. From a ML perspective, this will unfortunately bring disadvantages to ML models because they are trained on ERA5 data. Here, we choose the conventional approach in ML studies (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023b) and use ERA5 data to initialize and evaluate ML models.

4.2.2 Machine learning models for weather forecasting

We focus on three recent ML models in this work: FourCastNet, PanguWeather, and GraphCast. Table 4.1 summarizes the main characteristics of these models. More details on the variables predicted by these models are presented in Section E.1 of the Supplementary Materials.

Bi et al. (2023) trained four models with different lead times (1, 3, 6, and 24 h). These models are combined during inference to achieve the minimum number of model executions for a given forecast lead time ("hierarchical temporal aggregation strategy"), thereby minimizing error accumulation (Bi et al., 2023). For instance, to forecast the weather state in 36 hours, using the 24h-model once plus two iterations of the 6h-model gives a more accurate forecast than simply using the 1h-model iteratively for 36 times. Because we only consider lead times that

Table 4.1: Summary of key features of recent ML-based weather forecasting models.

	FourCastNet	PanguWeather	GraphCast
Resolution	$0.25° \times 0.25°$	$0.25° \times 0.25°$	$0.25° \times 0.25°$
Architecture	Fourier Neural Operator	Transformer	Graph Neural Network
# surface variables	6	4	5
# atmospheric variables	5 at 4 pressure levels (pl)	5 at 13 pl	6 at 37 pl
Training and validation	1979–2017	1979–2017, 2019	1979–2017
# parameters (millions)	74.7	256	36.7
Fundamental timestep (h)	6	1, 3, 6, 24	6

are multiples of 6 h in our study, we use sequences of 6 h- and 24 h-model calls to achieve the smallest possible forecast error for PanguWeather.

We also highlight that GraphCast requires the weather states at two consecutive time points as inputs to each forecast, while the other two models only need one. Furthermore, as shown in Table 4.1, for each time step the dimensionality of the input data (especially the number of pressure levels) of GraphCast is substantially higher than that of PanguWeather and FourCastNet. These two distinctions might advantage GraphCast in comparison to the other two ML models, since using additional covariate information in principle tends to increase the potential predictive accuracy.

Although forecast data from various ML models are available on WeatherBench 2 (Rasp et al., 2024), they do not cover all periods we investigate (e.g., GraphCast forecasts are only available for 2018 and 2020). We thus produced additional forecast data for more recent events by running the ML models ourselves. More precisely, we implemented the inference of the three ML models by directly leveraging their pre-trained models released on GitHub. Alternatively, the forecast data can be generated using the ECMWF library "ai-models", but note that the GraphCast model in the library is GraphCast operational (a smaller version than the one described in Lam et al. (2023)), which was pre-trained on ERA5 data from 1979 to 2017 and fine-tuned on HRES data from 2016 to 2021, and only includes atmospheric variables at 13 pressure levels as input and output.

4.2.3 Initialization times

ERA5 and HRES-fc0 differ in their assimilation windows (Lam et al., 2023). While observations up to three hours into the future are included in the assimilation for HRES-fc0, the lookahead for ERA5 varies between initialization times: three hours for forecasts initialized at 06:00/18:00 UTC and nine hours for forecasts initialized at 00:00/12:00 UTC. To ensure an equal lookahead, Lam et al. (2023) compared ML-based and HRES forecasts initialized at 06:00/18:00 UTC for lead times up to the availability of HRES (3.75 d). Beyond this time limit, ML forecasts initialized at 06:00/18:00 UTC are compared with HRES forecasts initialized at the preceding 00:00/12:00 UTC. We follow this mixed initialization methodology in one of our analyses in Section 4.3.1.

As shown in Section S.5.2 in the Supplementary materials of Lam et al. (2023), the effect of unequal lookahead is small, particularly for long lead times. Therefore, for all analyses except the RMSE comparison in the first case study, we include all forecasts (00:00, 06:00, 12:00, and

18:00 UTC initialization times) in our analysis. Additionally, we extend the short HRES forecasts initialized at 06:00/18:00 UTC beyond lead times of 4 d; these forecasts are augmented with data from the forecasts initialized 6 h prior to the 06:00/18:00 UTC initialization time, while increasing the lead time by 6 h, so that the validity time of the forecast remains the same. This filling might disadvantage HRES, but it enables the analysis of a denser set of initialization and lead times.

4.3 Case studies

4.3.1 2021 Pacific Northwest heatwave

In this first case study, we investigate a record-shattering extreme temperature event. In late June 2021, a heatwave of unprecedented magnitude hit the Pacific Northwest with temperatures reaching up to 49.6 °C, beating the all-time record for Canada by 4.6 K (Fig. 4.1A). Even in hindsight, quantifying the return period of the event is challenging (Bartusek et al., 2022; Philip et al., 2022; Zeder et al., 2023). While the impacts caused by such extreme events can be substantially large, prediction of such events is also challenging for ML models, due to the scarcity of similar events in training data. On the other hand, even though NWP models are more directly bound to follow physical laws, their forecast accuracy is not guaranteed either.

The heatwave impacted ecosystems, infrastructure, and human health considerably (with more than 1400 deaths), and attracted massive public attention and scientific interest (Neal et al., 2022; Schumacher et al., 2022; White et al., 2023; Röthlisberger and Papritz, 2023). In the analyzed region, temperatures peaked between June 27 and June 29. We analyze the heatwave in terms of the temperature at 2 m above the surface (T_{2m}), which is a standard variable for studying temperature extremes.

In the grid cells closest to three major population centers affected by the heatwave (Vancouver, Seattle, Portland), the prediction error of HRES and all tested ML models reaches at least twice the size of a typical HRES 10-day prediction error and exceeds the typical HRES 10-day error in Portland by a factor of four. This is consistent with the results of Lin et al. (2022), who examined the predictions of subseasonal-to-seasonal NWP models for the Pacific Northwest heatwave and found that all models failed to predict the magnitude of the heatwave for forecasts initialized on June 17, i.e., ten days before temperatures began to peak. We visualize our prediction errors in predictability barrier plots in Fig. 4.2, using HRES-fc0 as ground truth for the HRES forecasts and ERA5 as ground truth for the ML-based predictions. We aggregate to daily scale by computing RMSEs as described in Section D.1.1. A version of the plot showing $T_{2m,prediction} - T_{2m,groundtruth}$ without this aggregation is presented in the supplementary material. Maps of the temperature anomaly patterns predicted for the peak of the heatwave for forecasts with different initialization times are shown in Fig. D.2.

FourCastNet has the largest error among all models, while the errors of PanguWeather, GraphCast, and HRES are of a similar magnitude visually (Fig. 4.2). For all models, the prediction errors are largest during the peak of the heatwave. The predictability barrier plots exhibit prominent vertical structures (i.e., forecasts for the same validity day), suggesting that the dominant factor is

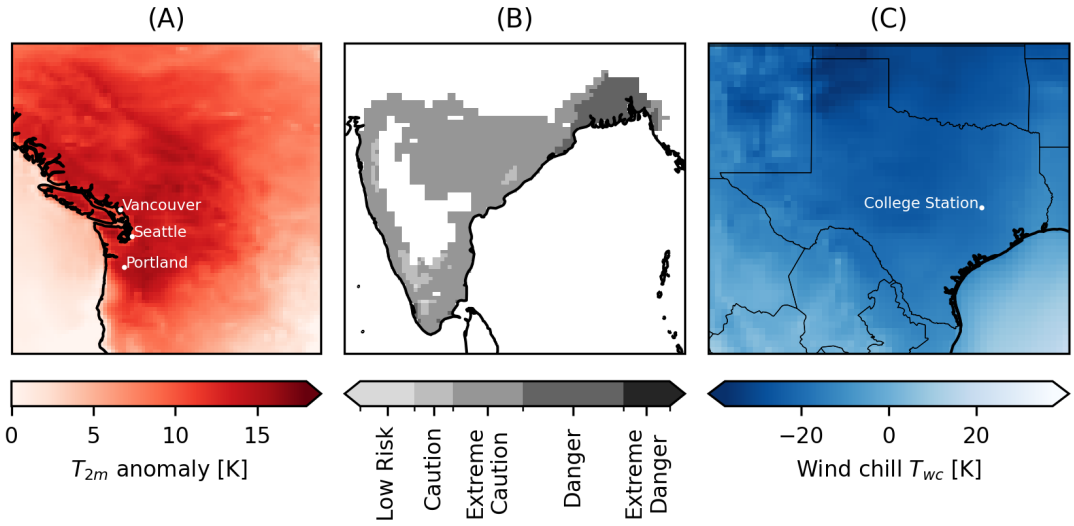

Figure 4.1: Magnitudes of the three events analyzed in this paper. (A) 2021 Pacific Northwest heatwave. Shown is the 2 m temperature anomaly averaged over 27–29 June 2021, the peak of the heatwave. (B) 2023 South Asian humid heatwave. Shown is the category of maximum daily Heat Index HI, as defined in Section D.2.1, averaged over 17–20 April 2023 in India and Bangladesh. (C) 2021 North American winter storm. Shown is the wind chill index T_{wc}, as defined in Section 4.3.3, on 12:00 UTC, 15 February 2021.

the predictability of the weather situation rather than the forecast initialization. However, HRES seems to also exhibit hints of diagonal error structures. This structural difference in error patterns is further discussed in Section 4.3.3, where it is more significant.

For HRES, the prediction errors in the first days of July 2021, when temperatures started to fall again, are larger than for PanguWeather and GraphCast, especially for the grid cells closest to Seattle and Portland. For long lead times, however, the HRES errors reach their largest values during the heatwave peak around June 27-29 in all three grid cells. The predictability barrier plots for PanguWeather appear very patchy, likely due to the "hierarchical temporal aggregation strategy" of PanguWeather (described in Section 4.2.2).

The best and worst performing models across various lead and validity times are visualized in Fig. D.3. Conclusions match those from Fig. 4.2; FourCastNet has the largest errors during the heatwave, and HRES has comparatively high errors after the peak of the heatwave. During many of the time steps, especially for short lead times, GraphCast and HRES yield the smallest error. However, there is no clear best-performing model overall.

To assess the models' performance in predicting the extreme event, we compute the forecast RMSEs of all models during the peak of the heatwave and compare them to the RMSEs during the summer of 2022, a baseline year without extreme heatwaves in the region. We vary the lead time and study the event in the region defined by Philip et al. (2022). The RMSE aggregation here follows Lam et al. (2023): it includes latitude-based weights, and only forecasts initialized at 06:00/18:00 UTC and lead times in multiples of 12 h are considered to ensure equal assimilation windows between ERA5 and HRES-fc0. The results, shown in Fig. 4.3, again highlight the

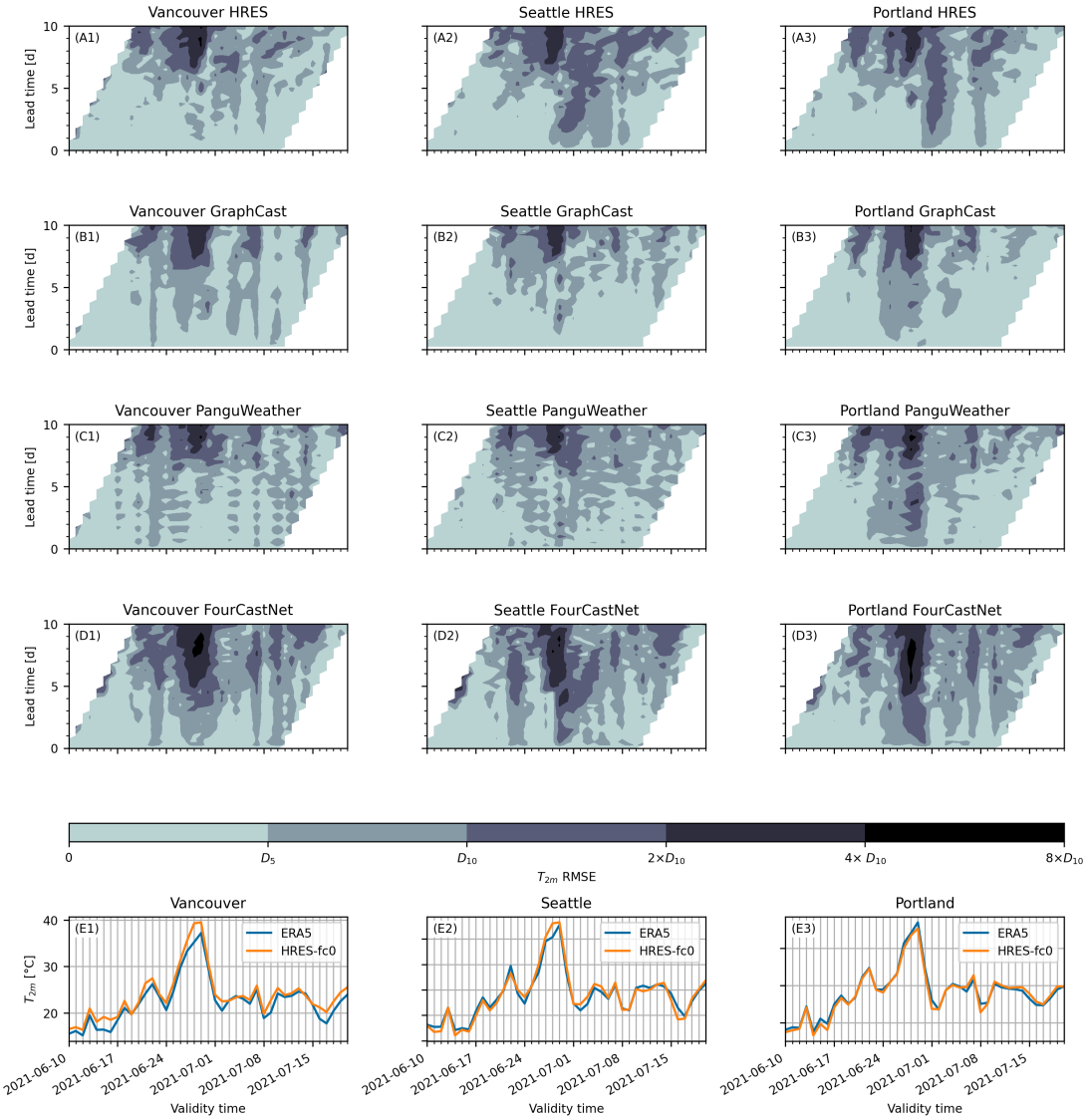

Figure 4.2: Panels (A1) to (D3): Predictability barrier plots for the grid cells closest to major cities affected by the 2021 heatwave. For HRES, HRES-fc0 is used as ground truth, for the ML models, we use ERA5 instead. In the color bar, D_5 and D_{10} indicate long-term multi-year average HRES 5-day and 10-day prediction errors. For the computation of the RMSE, D_5, and D_{10} see Section D.1.1. Numerical values for D_5 and D_{10} are given in Table D.1. Panels (E1) to (E3): time series of daily maximum T_{2m} for the data sets used as ground truth.

difficulty of predicting the extreme temperatures during the event: for lead times beyond one week, all models perform substantially worse than for the summer 2022 baseline. More precisely, all ML models perform up to at least three times worse and HRES up to two times worse. Given the small sample size, these numerical values should be interpreted with care, however. We also observe that for lead times up to 6.5 d, the forecast errors of HRES are smaller than ML models, contrary to their performance in the baseline year. This might be a consequence of extrapolation, as discussed in the introduction. As the evaluated baseline period is rather short, mainly for computational reasons, the baseline might not precisely represent the typical performance. However, the baseline results are in line with other studies (Lam et al., 2023) and additional baseline data from the year

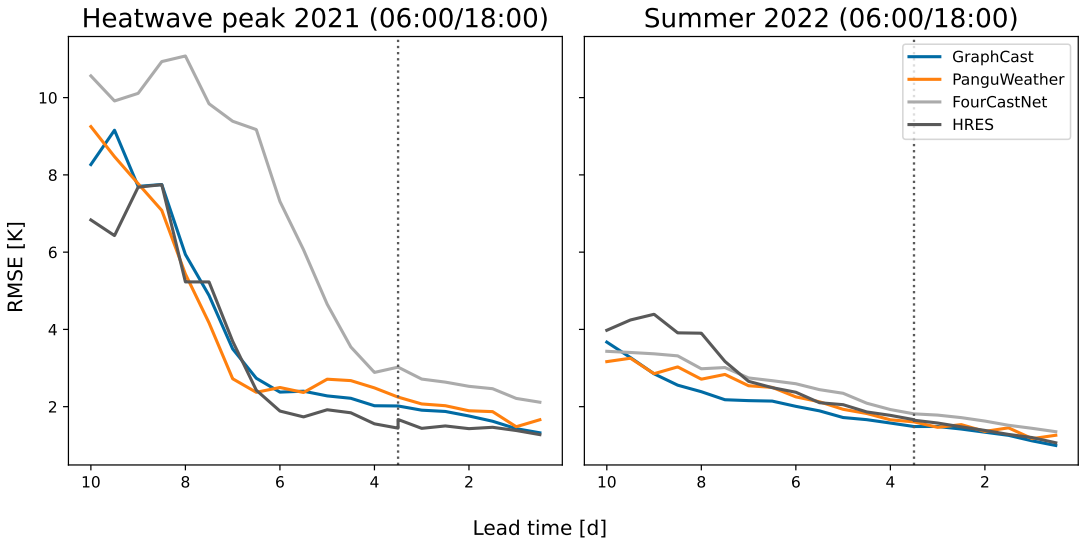

Figure 4.3: Evolution of the T_{2m} prediction RMSE with lead-time for the three ML models and HRES in the event region during the peak of the heatwave (June 27–29 2021, left) compared to summer 2022 as a baseline (June 20–July 10, right). Observations in the considered box region, 45°–52°N, 119°–123°W, are weighted to correct for differences in grid-cell area. ML models use 06:00/18:00 UTC initial conditions and evaluation times only, and the HRES forecasts use the mixed initialization described in Section 4.2.3 after 3.75 d (dotted line).

2020 (Fig. D.1 in Section D.1).

A further analysis focusing on the spatial aspect of the event is presented in Section E.2 of the Supplementary Materials. A main conclusion is that FourCastNet under-predicts the area in which temperature anomalies exceed a given threshold, while in some PanguWeather forecasts, the area predicted to exceed the thresholds is too large.

4.3.2 2023 South Asian humid heatwave

In April 2023, high temperature and humidity levels were reached simultaneously in South Asia (Fig. 4.1B). The human tolerance to high temperatures decreases with increasing humidity, mainly due to the inability of the body to self-regulate its temperature through transpiration. Heat stress associated with this type of event can therefore be particularly harmful to human health (Buzan and Huber, 2020; Lo et al., 2023).

The heat index (HI) is an impact metric quantifying this hazard to human health. It estimates the apparent temperature (i.e., how hot the temperature feels) for given values of temperature (T_{2m}) and relative humidity (RH). While many metrics have been proposed to combine the influence of these two variables (Lo et al., 2023), we follow Zachariah et al. (2023), who employ the modified version of the heat index (Rothfusz and Headquarters, 1990) used by the NOAA Weather Prediction Center in an attribution study on the 2023 South Asian humid heatwave. The detailed computations, including information on how we convert predicted specific humidity to relative humidity, are given in Section D.2.3.

Following Zachariah et al. (2023), we focus on two study regions in South Asia: Laos-Thailand (for

which results are presented in Section E.3 of the Supplementary Materials), and India-Bangladesh. For the latter, a sub-region with dry and semi-arid climate is excluded from the analysis (see Section D.2.1 for details). We select a temporal range of April 17–20, 2023 (UTC time, inclusive range) for the India-Bangladesh region, corresponding to the period in which the heat stress peaked.

With existing ML weather prediction models, HI at the surface cannot be calculated correctly because humidity is only modeled at higher pressure levels, and none of the models predict a variable that could enable the calculation of relative humidity at the surface level. This presents a strong limitation on the utility of ML models in forecasting humid heat waves. While GraphCast and PanguWeather forecast variables at the 1000 hPa level, FourCastNet predictions only include variables at pressure levels starting from 850 hPa. In the following, we exclude FourCastNet from the analysis and use relative humidity at the 1000 hPa level as an approximation for humidity at the surface.

The HI prediction error during the peak of the heatwave in the India-Bangladesh region is shown in Fig. 4.4. For each day, we select the time when the ground truth HI is maximal, and then average the errors over April 18–20. In all cases, the predicted HI is computed using T_{2m} and $RH_{1000\,\mathrm{hPa}}$. This setup is the simplest possible substitute that forecasters could use with the ML models available at the time of writing. The forecasts show deviations from the ground truth data sets, especially over Bangladesh. The under-prediction for ML models is stronger than for HRES. Looking at the prediction errors of $RH_{1000\,\mathrm{hPa}}$, we find a matching pattern: predictions for relative humidity over Bangladesh are too low, especially for PanguWeather (Fig. D.4). For T_{2m}, the values at the time of the HI peak are mostly smaller than the corresponding ground truth for the ML methods, while HRES T_{2m} predictions are larger than the HRES-fc0 ground truth (Fig. D.4).

For HRES, HRES-fc0, and ERA5 it is possible to compute relative humidity at the surface level (RH_{sfc}) from 2m temperature and 2m dewpoint temperature (see Section D.2.3). We found rather large differences between the ground truth data sets ERA5 and HRES-fc0 for the studied event however, therefore we used $RH_{1000\,\mathrm{hPa}}$ in the computations for Fig. 4.4.

Large fractions of the India-Bangladesh region experienced a mean daily maximum HI during April 17–20 2021 that falls in the "extreme caution" or "danger" category (Fig. 4.5, see Table D.2 for the definition of the categories). In Fig. 4.5, the HI distribution computed from ERA5 data using the input variables T_{2m} and RH_{sfc} differs strongly from the other ground truth data sets, mainly due to ERA5 RH_{sfc} values being higher during daily maximum HI. This may be a consequence of differences in the assimilation procedure used to produce the ground truth data. The ML-based HI forecasts (computed using $RH_{1000\,\mathrm{hPa}}$) underestimate the ERA5 HI values both when $RH_{1000\,\mathrm{hPa}}$ or RH_{sfc} is used. This is especially the case for high values of HI. Results for the Laos-Thailand region are also in line with these findings (see Section E.3 in supplementary information).

Figure D.6 visualizes the heat index forecast for the peak of the heatwave by forecasts with different initialization times (in terms of threat categories, see Table D.2). Results match the other analyses in this section.

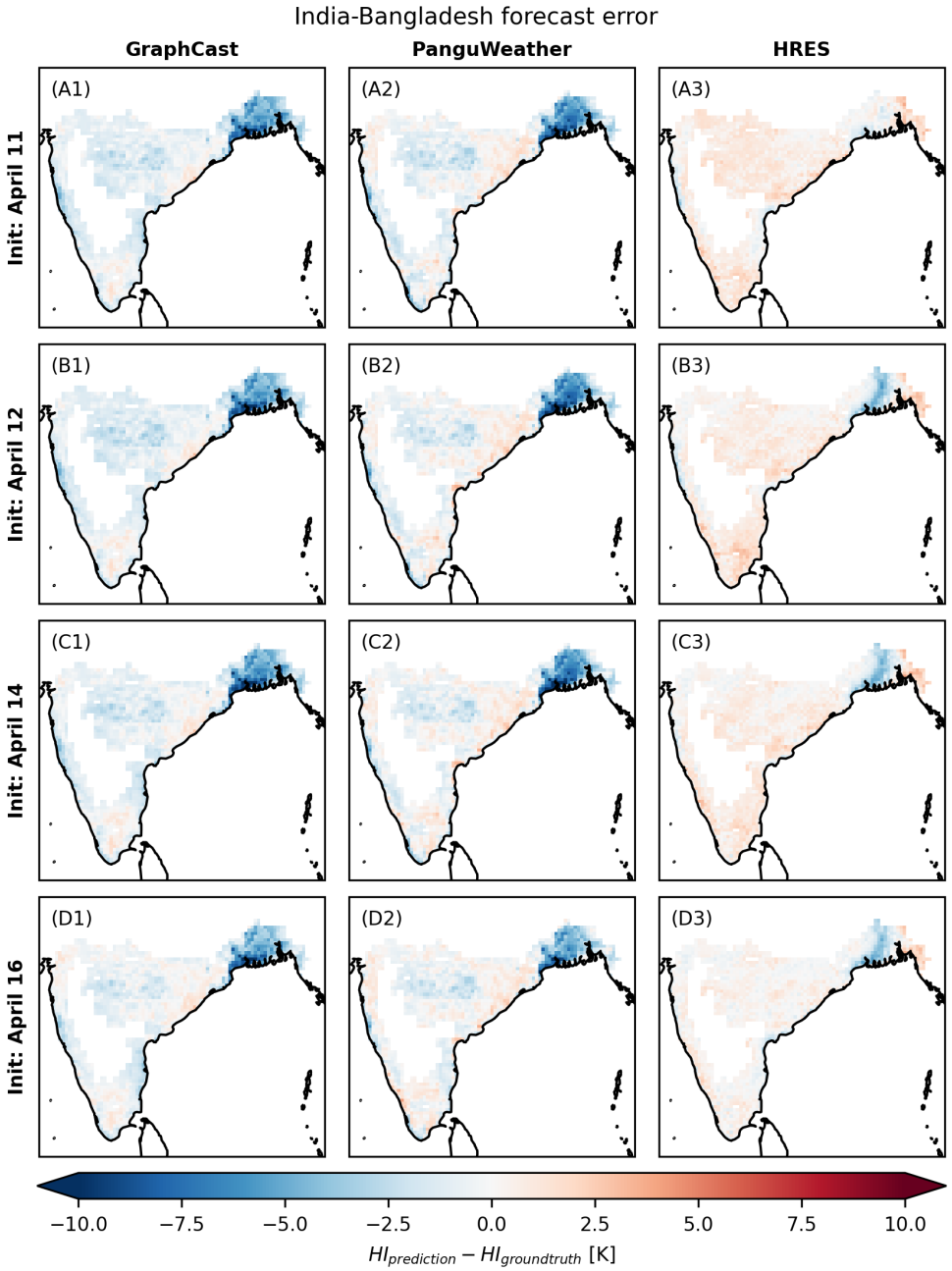

Figure 4.4: Error of the HI prediction, for the time step of each day during which HI peaked in the ground truth data set, averaged over April 17–20, 2023. For all forecasting methods and ground truth data sets, HI is computed using $RH_{1000\,\mathrm{hPa}}$ rather than the value at the surface.

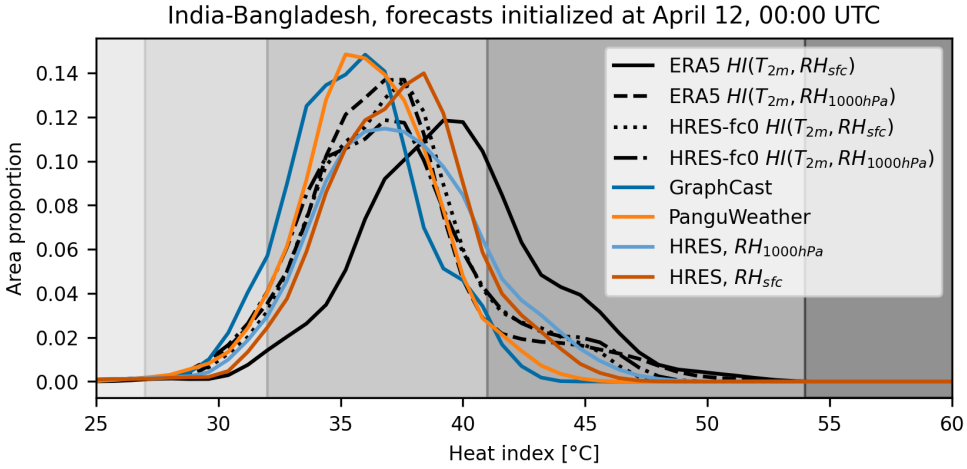

Figure 4.5: Proportion of area in study region with given mean daily maximum heat index during April 17–20, 2023, computed using area-weighted kernel density estimation. Shaded areas in the background indicate threat levels (see Section D.2.3). Light gray to dark gray: low risk, caution, extreme caution, danger, extreme danger. Compared are distributions resulting from forecasts initialized 6 days prior to the start of the event and different ground truths: ERA5 and HRES-fc0, each in two versions of computing the heat index either using RH_{sfc} or using the substitute $RH_{1000\,\mathrm{hPa}}$. For HRES forecasts, we show versions computed with $RH_{1000\,\mathrm{hPa}}$ and RH_{sfc} as well.

4.3.3 2021 North American winter storm

While heatwaves often receive a lot of media attention, especially in light of anthropogenic climate change, cold spells are also hazardous. Under the current climate, cold extremes lead to more human deaths overall than hot extremes (Gasparrini et al., 2015). In mid-February 2021, a winter storm hit large parts of the United States, Northern Mexico, and Canada (Fig. 4.1C). Rapidly falling temperatures were accompanied by snow, sleet, freezing rain, and strong winds, causing damages to human livelihood and infrastructure (National Weather Service, 2021). In Texas, which was strongly affected by the event, pipes burst, interrupting the water distribution, and energy infrastructure failed, resulting in power outages and ordered rolling blackouts. Impacts were amplified by inadequate wintering of energy infrastructure (Gruber et al., 2022).

While it would be interesting to look at a metric that directly relates to vulnerabilities of the Texas power grid, defining such a metric is not straightforward. Therefore, we restrict our analyses to T_{2m} and the wind chill index T_{wc} as defined in Osczevski and Bluestein (2005). The wind chill index is a metric that describes the apparent temperature in the presence of wind. It is defined as

$$T_{wc} = 13.12 + 0.6215 T_{2m} - 11.37 v^{0.16} + 0.3965 T_{2m} v^{0.16}, \tag{4.1}$$

where T_{wc} and T_{2m} are given in °C, and the wind-speed at 10 m height v in kilometers per hour (computed from the horizontal wind components). The formula was obtained by modeling heat transfer from the human body to the atmosphere (Osczevski and Bluestein, 2005). Note that T_{wc} is only defined for temperatures below 10 °C and wind speeds above 4.8 $\mathrm{km\,h^{-1}}$. In a particularly affected Texas city, College Station, HRES-fc0 and ERA5 closely follow observational records for both T_{wc} and T_{2m} (see Fig. D.7 in Section D.3), demonstrating the suitability of those datasets for this index.

Looking at predictions of T_{wc} for the grid cell closest to College Station in Fig. 4.6, one can see that all models struggle to predict the minimum wind chill index, with forecast errors being largest for FourCastNet (sometimes exceeding 40 K at minimum T_{wc} for large lead times). In general, errors are larger for the winter storm than for the Pacific Northwest heatwave, which might be due to a potential seasonality in prediction errors (as suggested by Figure 2 in Ben Bouallègue et al., 2024) or, simply, to the different nature of the events. Errors for PanguWeather and GraphCast are substantially lower than for HRES, especially between February 9 and February 17, after the peak of the winter storm.

The vertical structures in the plot (which are particularly prominent for GraphCast and PanguWeather on February 15–17) hint at the difficulty of predicting weather situations, while the diagonal structures (strong for FourCastNet and HRES) suggest variation between individual forecasts caused by their initial conditions. In general, HRES seems to produce stronger diagonal error structures, while the ML models tend to exhibit vertical error patterns. While the data filling we use to extend HRES forecasts might affect this finding, it could also be a result of fundamental differences between ML-based and NWP forecasts. Assuming a very extreme event that is easily predictable following physical laws, the predictions of HRES would steadily improve when approaching the event. ML methods, however, might not be able to extrapolate to such extreme conditions, even at very short lead times, resulting in vertical error patterns. On the other hand, if an extreme event that is difficult to predict with (first-order) physical laws is somewhat hidden in the atmospheric state of the initial-conditions day, HRES would have trouble forecasting both the event and its buildup, leading to a diagonal pattern. Then, when the event becomes more apparent from the conditions, the prediction will improve. The ML methods might be able to model such 'hidden' (second-order) conditions better because of their flexibility, leading to less strong diagonal patterns.

When the predicted temperatures are too high, or the predicted wind speeds too low, the thresholds in the definition of T_{wc} are not exceeded, and T_{wc} is thus not defined. This is the case during the wind chill minimum between February 15 and February 17, even though these were the most hazardous days. In Figs. 4.6 and 4.7 we ignore the thresholds in the definition and still compute the T_{wc} expression for visual clarity.

Patterns in the prediction of T_{2m} look similar to those of T_{wc} (Fig. D.8), with FourCastNet errors being largest, and HRES errors during the event larger than PanguWeather and GraphCast errors. For the surface windspeed, which also enters Equation (4.1), the patterns are not as clear (Fig. D.9). Therefore, the T_{wc} errors seem to be dominated by T_{2m}.

Figure 4.7 depicts the forecast errors in a bounding box around Texas (107 °W to 93 °W, 25 °N to 37 °N) for 12:00 UTC on February 15, 2021 (6:00 am, Houston time), when the average wind chill index in the box reached a minimum in the ground truth data. For long lead times, forecast errors of GraphCast and PanguWeather seem smaller than those of FourCastNet and HRES, although all forecasts appear to be too warm. One can also notice a slightly "patchy" structure in the FourCastNet predictions, with notable discontinuities between 8x8 patches. 8x8 is the patch size used internally by FourCastNet.

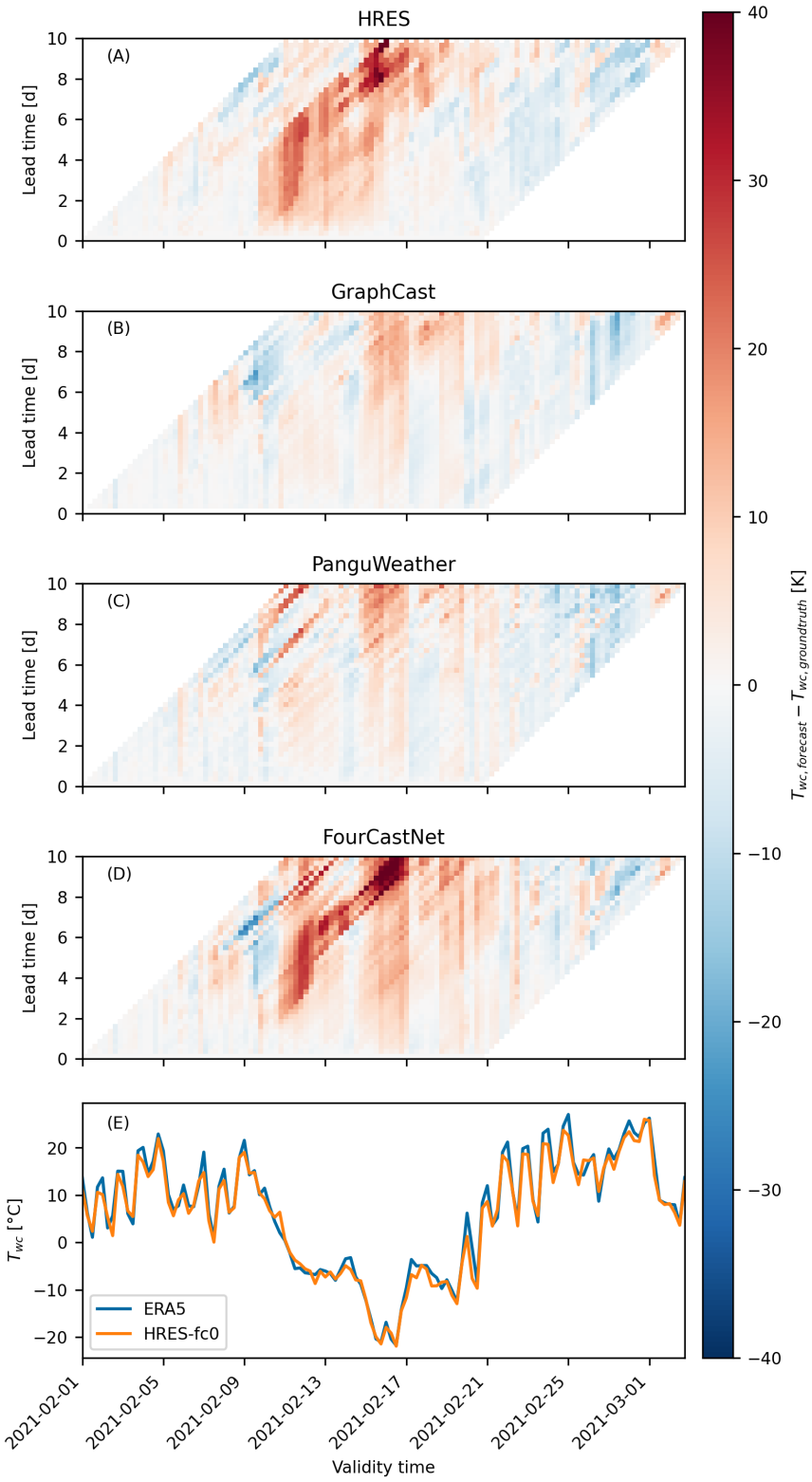

Figure 4.6: Panels (A) to (D): T_{wc} prediction errors for different validity and lead times. Data from grid cell closest to College Station, Texas. Times and dates are given in UTC. Panel (E): time series of the ground truth data sets used in the computation: ERA5 is used for the ML forecasts, HRES-fc0 is used for HRES.

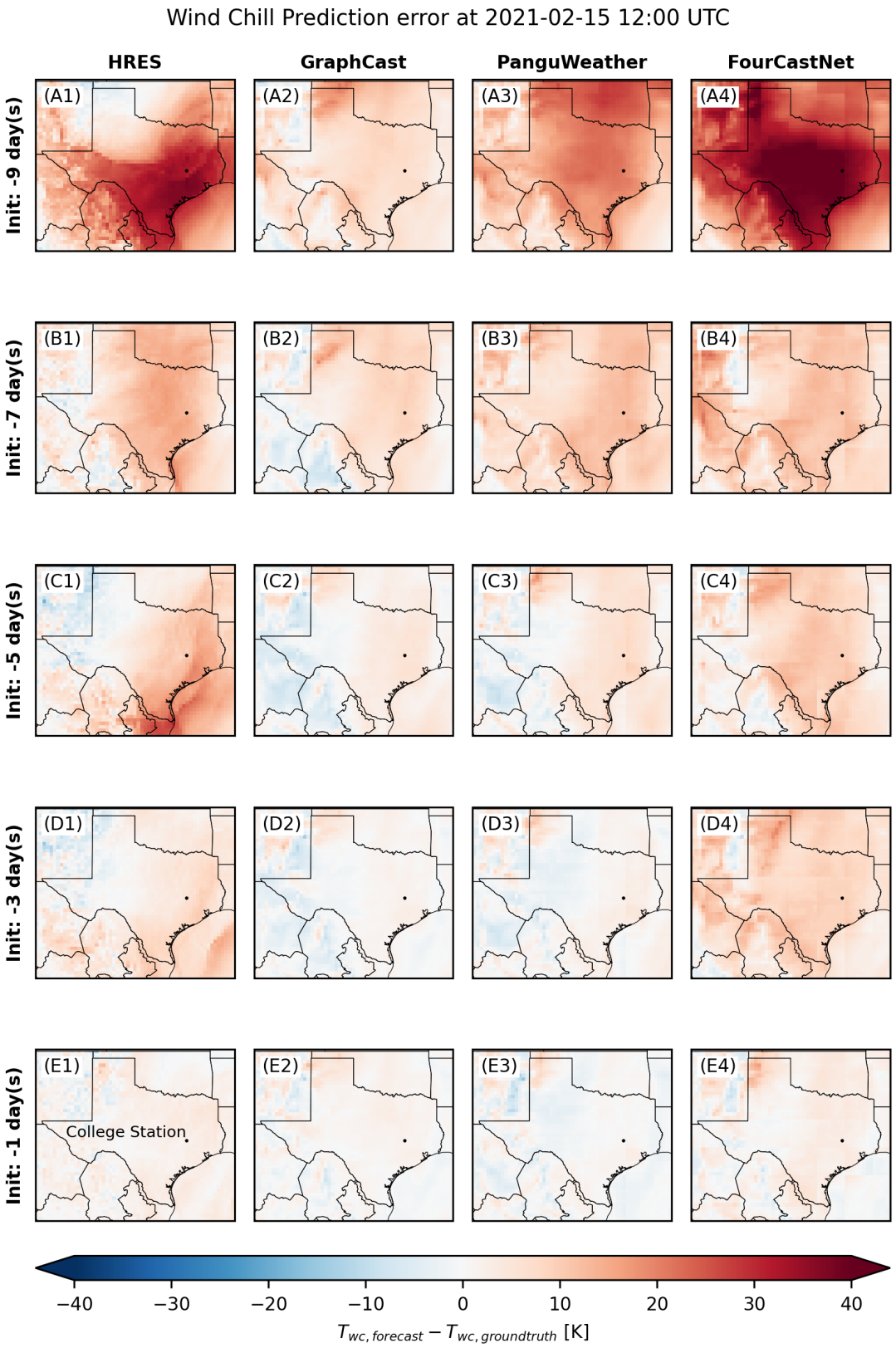

Figure 4.7: Errors of T_{wc} forecasts for 12:00 UTC, February 15 2021 (6:00 am Houston time). The ground truth used to compute results for the ML forecasts is ERA5, while HRES-fc0 is used for HRES.

4.4 Discussion and conclusions

The three case studies conducted highlight different aspects of comparison between the ML models GraphCast, PanguWeather, FourCastNet, and the NWP system HRES. For the 2021 Pacific Northwest heatwave, predictions of PanguWeather and GraphCast maintained comparable quality to HRES in terms of the evaluated metrics. However, for short lead times, HRES showed smaller forecast errors in both the predictability barrier plot and RMSE plot than ML models, contrary to their performances in the baseline summer of 2020 or 2022, indicating that ML models might be more severely impacted by the extrapolation to those extreme conditions. We also observe that HRES has more difficulties predicting the sharp drop in the temperature after the peak of the heatwave than ML models. When analyzing the South Asia humid heatwave substituting RH_{sfc} with $RH_{1000\,\mathrm{hPa}}$, prediction errors show spatial patterns with the highest danger levels over Bangladesh being underestimated by the ML models. For many lead times and initial conditions, the North American Winter Storm is forecast more accurately by PanguWeather and GraphCast than by HRES. From those predictions, we observe structurally different error patterns: HRES and FourCastNet are potentially more impacted by subtle signals in the initial conditions than GraphCast and PanguWeather, leading to errors that build up before the event. We emphasize that our findings are limited to the three case studies, and more systematic analyses need to be conducted to reach definitive conclusions about general extreme weather event forecasts.

None of the ML models predict a variable that enables the computation of surface-level humidity. This would have allowed us to better study the effects of the 2023 humid heatwave in South Asia, as surface humidity alters the effect of temperature on the human body. Looking at substitute variables, ML models seem to perform worse for this event overall, potentially due to extrapolation. Whether this effect persists for ML models that do predict surface humidity remains to be answered in future research. The rather large differences in relative humidity at the surface level between the ERA5 data set and HRES-fc0 complicate the estimation of "true" heat index forecasting errors. One way to resolve this would be to directly compare against station observations.

Comparing forecast systems on only a subset of extreme events incurs the danger of favoring alarmist forecasts, a phenomenon termed the "forecaster's dilemma" (Lerch et al., 2017). This is a general problem of case studies and even applies to a broader class of analyses. Such results should not be used alone to judge the overall quality of weather forecasting systems. The existing literature on evaluating ML-based weather forecasts described in Section 4.1 can be combined with our findings to obtain a more complete picture.

Our study only uses single forecasts and disregards probabilistic forecasting. In NWP, forecast uncertainty is accounted for by running ensemble forecasts. While including NWP ensemble forecasts is possible, this would have caused further complications in the analyses, e.g. due to differing model resolutions. Furthermore, producing ensemble forecasts with the given ML models is non-trivial. Attempts have been made, e.g., by perturbing initial conditions or model parameters (Weyn et al., 2021; Bi et al., 2023; Bülte et al., 2024), but problems capturing the right scaling of uncertainties have been found in the literature. Selz and Craig (2023) investigated PanguWeather and found that the error growth for small perturbations is too small ("no butterfly effect"). Recently, Price et al. (2023) explored generative modeling to obtain better ensembles.

Because of the generative training objective, these models can better capture the spectrum of the weather at long lead times, avoiding the over-smoothing that occurs for autoregressive models like GraphCast, PanguWeather, and FourCastNet. For the evaluation of generative ML-based weather forecast models, proper scoring rules (Gneiting and Katzfuss, 2014), like the class described by Allen et al. (2023) for probabilistic forecasts, will be an important analysis tool. The comparison of ML models with HRES is also limited by differences between the ground truth data sets ERA5 and HRES-fc0 (differing assimilation times, short forecasts for 06:00/18:00 UTC initializations). However, this does not affect the comparison between the three ML models. ML weather prediction models are typically trained using ERA5 data, which does not correspond to an "operational setting" and complicates the comparison with HRES. While ML models could be trained or fine-tuned with HRES-fc0 data directly, the IFS version used to produce the forecasts varies over time, therefore characteristics and biases of the "ground truth" HRES-fc0 would also vary.

Most ML models employ a large autoregressive time step (6h for GraphCast and FourCastNet). This coarse temporal resolution might affect the forecast of impacts for which the daily maximum or minimum is relevant, such as short-term heat stress peaks or severe wind gusts. The most extreme values might be missed due to an unfortunate combination of forecast time step and daily cycle or event time. Some important variables for impact assessments are forecast by few or no ML models. These include humidity at the surface, solar radiation reaching the surface (potentially relevant for solar energy production forecasts), and precipitation. While some ML models (including GraphCast and FourCastNet) do predict precipitation, authors have advised caution in the interpretation of this variable, citing issues with the ERA5 precipitation ground truth (Lavers et al., 2022).

While case studies can only provide anecdotal evidence, testing ML models under individual extreme events can reveal unexpected deficiencies (or advantages) of these models in comparison to well-established techniques. The rather small number of meteorological variables predicted by ML models, as well as the available forecast lead time, limits the types of impactful extreme events that can be studied for these models. While longer forecasts would be interesting and would allow the study of more complex types of extreme events (Zscheischler et al., 2020), one would likely need to include new processes and variables into the models, such as feedback from soil moisture and the influence of sea-surface temperatures.

Non-linear combinations of predicted output variables (e.g., wind chill, see Eq. (4.1)) have the potential to reveal weaknesses of ML models; Price et al. (2023) investigated horizontal surface wind speed (a non-linear function of the horizontal wind components) and found that GraphCast tends to perform worse in terms of this combined metric than for the individual components. They hypothesized that this might be due to the tendency of a certain type of ML architecture to predict close to the mean under forecast uncertainty and the non-commutativity of non-linear function applications and averaging. However, in our case studies (T_{wc} during the 2021 North American winter storm and HI during the 2023 South Asia humid heatwave), the large differences in prediction errors for individual input variables (T_{2m} during winter storm, relative humidity during humid heatwave) and the necessity of having to substitute relative humidity at the surface level seem to outweigh this effect. Nevertheless, the described problem is an interesting target for

future work, as impacts often are not simply determined by linear combinations of the variables predicted by the ML models. Price et al. (2023) suggest using generative modeling to overcome this systematic problem.

For theoretically-justified extrapolation to extremes and when interested in risk assessment, a natural approach is the use of extreme value statistics (Coles, 2001). Recently, various approaches have combined machine learning and extreme value statistics to improve predictive extrapolation of extreme risk for the predicted variable (Pasche and Engelke, 2024; Richards and Huser, 2024; Velthoen et al., 2023; Cisneros et al., 2024; Allouche et al., 2024; Gnecco et al., 2024). Methods for extrapolation in the predictor space also exist but require stronger dependence assumptions (Chen and Meinshausen, 2023; Pfister and Bühlmann, 2024). Including physical domain knowledge in ML-based models, for example through architectural restrictions of explicit equations, could be another approach to improving generalization (Kochkov et al., 2024).

Releasing raw predictions instead of aggregates or summaries, or even pre-trained models, is valuable (Burnell et al., 2023). As considering all metrics that stakeholders deem important during model development and testing is challenging or impossible, releasing the full predicted data or trained models allows assessing domain-specific model skill even after the development of the model. WeatherBench 2 (Rasp et al., 2024) already partially addressed this point. Building a continuously-updated database of extreme event case study set-ups (similar to ECMWF's Severe Event Catalogue, ECMWF, 2024), including domains and impact metrics, might be a valuable contribution to existing literature. A focus on re-usability for new models would be important, potentially through the integration of a framework like WeatherBench 2. One could hypothesize that the selection of extreme events based on their real-world impacts leads to a "selection bias", as the larger impacts might have been caused by poor forecasts by operational models, potentially resulting in a biased estimate of the relative performance of HRES.

The evaluation of ML models typically focuses on meteorological variables. Putting a stronger focus directly on impacts has the potential to improve the practical value of ML models. To find suitable impact metrics, researchers could, for instance, look at warnings issued by weather services and analyse how warnings based on NWP compare to forecasts using ML models. Coupling ML weather forecasts to impact models, such as models for floods (Nearing et al., 2024), crop loss, or fires, might also be valuable, although the analysis would then also depend on the impact model's fidelity. While ML models have shown impressive skill in forecasting key meteorological variables, it is worthwhile investigating whether their predictions can lead to similarly impressive results when assessing impacts.

Declarations

Acknowledgements

We thank Gloria Buriticá and Guohao Li for the discussions in early stages of the project, and Lily-Belle Sweet for her feedback on this manuscript. We are grateful to the developers of FourCastNet, PanguWeather, and GraphCast for sharing their code, and thank ECMWF for making their data sets publicly available, which has allowed us to conduct this study.

Funding

SE, OP, and ZZ acknowledge funding from the Swiss National Science Foundation Eccellenza grant "Graph structures, sparsity and high-dimensional inference for extremes" (grant no. 186858). JW acknowledges financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the program Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI. JW and JZ acknowledge the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537).

Author roles

Conceptualization: SE, OP, JW, ZZ, JZ; Data curation: OP, JW, ZZ; Formal analysis: OP, JW, ZZ; Funding acquisition: SE, JZ; Investigation: OP, JW, ZZ; Methodology: SE, OP, JW, ZZ, JZ; Project administration: SE, JZ; Resources: SE, JZ; Software: OP, JW, ZZ; Supervision: SE, JZ; Validation: OP, JW, ZZ; Visualization: OP, JW, ZZ; Writing - original draft: OP, JW, ZZ; Writing - review and editing: SE, OP, JW, ZZ, JZ. OP led the analysis and implementation of PanguWeather and FourCastNet, ZZ led the analysis and implementation of GraphCast, and JW led the case study data analysis.

Published article

This document is the peer-reviewed "Author's Accepted Manuscript" of an article published in Artificial Intelligence for the Earth Systems (Pasche et al., 2025b), with the DOI `https://doi.org/10.1175/AIES-D-24-0033.1`. When citing this work, please refer to the published version.

Supplements

Supplementary materials

The Supplementary Materials related to this paper are appended to this document.

Data and code availability statement

We use data from the ECMWF products ERA5, HRES, and TIGGE, which are published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. ERA5 is available on the Copernicus Climate Data Store. HRES forecasts initialized at 00:00/12:00 UTC can be accessed through ECMWF's TIGGE Data Retrieval portal. HRES forecasts initialized at 06:00/18:00 UTC are accessible through ECMWF's MARS, which requires access to be granted. Recently, Rasp et al. (2024) published cloud-optimized versions of the ERA5 and HRES data. We used these data sets for the case studies in 2021, and accessed their versions of ERA5, ERA5 climatology, and HRES forecasts initialized at 00:00/12:00 UTC.

Code to produce forecasts with GraphCast (`https://github.com/google-deepmind/graphcast`), PanguWeather (`https://github.com/198808xc/Pangu-Weather`), and FourCastNet (`https://github.com/NVlabs/FourCastNet`) is available.

We published the preprocessed ground-truth data and model forecasts for the periods and regions studied (Pasche et al., 2024), under CC BY 4.0 licence. The code to reproduce the analyses and figures discussed in this work is available on `https://github.com/jonathanwider/DLWP-eval-extremes` (release v1.0).

Conclusion and perspectives

One of the main scientific contributions of this thesis is the development of novel methodologies for accurately forecasting the conditional risk of extreme events by combining machine learning with extreme value statistics. This objective contrasts with the traditional use of extreme value theory, which primarily focuses on inference and static risk estimation. While Chapter 1 emphasizes extreme quantiles and high-threshold exceedance probabilities as the predicted conditional risk metrics, the proposed method models the entire conditional tail of the response distribution. This enables the prediction of any tail-density-based risk metrics, such as the conditional expected shortfall. The extreme conformalization procedure introduced in Chapter 2 extends this framework by providing high-confidence prediction intervals as an additional forecast metric. Another contribution of the thesis are the proposed confounder-mitigating causal tail estimator and the permutation test for discovering extreme causal effects, discussed in Chapter 3. Those tools could enable a better understanding of the causal mechanisms behind extreme events, and building more robust extreme-event prediction models through causal covariate selection. Finally, the study presented in Chapter 4 contributes to the assessment of state-of-the-art deep-learning global weather prediction models in forecasting extreme weather events, highlighting potential limitations in extrapolation and in the lack of crucial impact-relevant metrics. Together, these contributions also raise new questions and open several directions for future research.

The applications studied in this thesis highlight the intrinsic difficulty of evaluating predictive accuracy for extreme statistics, such as high quantiles, in observational settings. As the prediction targets are unobserved and often lie beyond the data range, standard comparative metrics such as the mean squared error or the quantile loss are inapplicable. This difficulty extends to hyperparameter tuning and model selection, as classical validation strategies typically rely on comparing the respective accuracies of the considered alternatives. In this thesis, the assessment of our methods and comparison to alternative approaches could be rigorously made through simulation studies, where true quantiles are known. In those studies, we observe that, on held-out data, higher generalized Pareto distribution (GPD) likelihood values typically correspond to more accurate extreme quantile predictions, motivating the use of the validation-set GPD likelihood for model selection and tuning in practice. However, as in the unconditional setting, the likelihood-based metric cannot detect tail misspecification when the asymptotic approximation is inaccurate. Existing alternative scoring rules, to our knowledge, are either inapplicable due to similar data-scarcity issues as the quantile loss or essentially reduce to calibration metrics, thus, unable to satisfactorily assess model fit and accuracy. Whether principled alternatives for the validation and comparison of extreme quantile prediction exist, therefore, remains an important question. Similar issues of validation extend more broadly to most statistical models for which

the target or estimand is unobserved.

Regarding hyperparameter tuning, the intermediate quantile level τ_0 is central to peaks-over-threshold methods. While well-established procedures exist in the univariate case, tuning τ_0 in conditional settings is more complex, as the threshold is localized and impacts the effective evaluation dataset for the other hyperparameters. However, we find that flexible machine-learning-based GPD models, such as EQRN, allow for substantially lower τ_0 values than unconditional or simpler GPD approaches, and that finely tuning τ_0 becomes unnecessary. It seems their flexibility compensates for the potential bias otherwise introduced by low thresholds, which is a significant practical advantage.

Chapter 2 addresses a key limitation of classical conformal prediction methods: their inability to provide informative prediction intervals when the desired confidence level is too high. Our proposed solution to this limitation relies on peaks-over-threshold asymptotic theory. While our empirical results demonstrate effective performance and valid finite-sample marginal coverage across the various practical settings considered, the method, thus, does not share the finite-sample theoretical guarantees of classical conformal prediction. This limitation reflects a fundamental tradeoff; one can show that extrapolation beyond the observed data necessarily requires sacrificing finite-sample guarantees, unless strong additional assumptions are imposed. A similar tradeoff in conformal prediction exists for conditional coverage, which cannot be guaranteed on finite samples, but asymptotic guarantees are attainable (Lei and Wasserman, 2014). In particular, theoretically guaranteeing finite-sample coverage for our framework would essentially require explicit bounds on extrapolated quantiles, which are not theoretically possible without restrictive assumptions on tail distribution (Boucheron and Thomas, 2015; Thomas, 2015; Lhaut et al., 2022). Such assumptions might, for example, include second-order assumptions on the data distribution, a bounded GPD approximation bias, or specific tail decay properties. However, these assumptions would all be difficult to verify in practice.

In the flood risk forecasting applications of Chapters 1 and 2, the proposed EQRN, and its conformalized variant, seem to deliver accurate one-day-ahead forecasts of multiple risk metrics, including extreme quantiles, extreme-event probabilities, and high-confidence prediction intervals, with well-calibrated uncertainty quantification, even for unprecedented events requiring extrapolation. The resulting early warnings successfully foresee all major flood events in the test period without excessive conservatism. Notably, our approaches capture the 2021 flood event, which was, at that time, missed by the operational hydrological model, despite the latter using more predictive variables and physical information than our proposed approach. To make our method's predictions operational, the EQRN's prediction accuracy could likely be further improved by including the additional covariates used in operational hydrological models, including soil moisture, snowpack information, and crucial weather forecasts such as the expected precipitation. It could also be used in combination with the operational hydrological model to potentially benefit from its additional physical information, for example by using the hydrological model's predictions as additional covariates. This combinative strategy has already been shown to be effective at improving mean river-flow predictions (Martel et al., 2025). Online fine-tuning or periodic retraining as new data become available could also be considered.

The data distribution of crucial high-impact variables may change over time, for example, due to climate change, leading to distributional shifts between the historical data, used for training predictive models, and future test data. In such contexts, the ability of prediction models to adapt, or to be adapted, to these changes is crucial for maintaining accuracy and reliability. In our flood forecasting application, the EQRN model seems able to accurately adapt to certain structural changes, including the breaking point in extreme river flow behaviour occurring during the testing period. However, the model's ability to adapt likely depends on the nature of the distribution shift. If all relevant, ideally causal, covariates are observed and the conditional distribution of the response given covariates remains unchanged, the predictions might stay reasonably accurate for moderate changes in covariate behaviour. Extreme causal discovery methods, such as those proposed in Chapter 3, may help identify such covariates and build causal predictive models. If the conditional relationship itself changes or covariates move outside the training support, continuously retraining the models with the new data might, for example, be necessary.

Chapter 4 highlights some challenges of extreme-event prediction and extrapolation for deep-learning weather prediction models. While only deterministic models were considered in our study, probabilistic or ensemble approaches alone seem unlikely to fully resolve these challenges. Capturing low-probability events would require very large ensembles of forecasters. Otherwise, limitations similar to those described in Chapter 2 for classical prediction intervals would arise. Regarding generative models, the distributional accuracy and realism far in the generated tails still seem unclear and are likely not yet well understood. Specific model tuning might be possible, but would certainly require trading central accuracy for tail and rare-event performance. A solution could be to use the mean-prediction weather models in combination with specifically designed extreme-value or extreme prediction-interval methods, such as the ones proposed in Chapters 1 and 2.

More broadly, the methods introduced in Chapters 1 and 2 can complement any mean-prediction model by providing high-confidence conditional uncertainty quantification around the point predictions, making them a useful tool in all classical regression settings. For example, LSTM-based mean-prediction models often tend to underpredict peak events, such as extreme river flow, when trained on observational data only, due to the scarcity of such events (Martel et al., 2025). On the other hand, our methods incorporating extreme-value extrapolation with LSTMs yield accurate extreme-quantile and high-confidence prediction-interval forecasts, even for extreme events of unprecedented magnitude. As an addition to mean predictions, they capture the increased uncertainty during high-variability periods, when extreme events are more likely, providing dynamic risk assessment and early warning systems in settings where mean predictions alone are insufficient. Being widely applicable tools, our methods can, in that regard, be more broadly beneficial than for flood and extreme weather forecasting. Other possible applications include, for instance, predicting financial risk, or forecasting the risk of collapse of essential infrastructures and services such as electrical grids, hospitals, insurances, urban buildings and structures, and internet services, due to exceptional demand or overload.

Bibliography

Alaa, A. and van der Schaar, M. (2020). Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *Proc. 37th Int. Conf. Mach. Learn.*, volume 119, pages 165–174.

Allen, S., Ginsbourger, D., and Ziegel, J. (2023). Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):906–940. doi:10.1137/22M1532184.

Allouche, M., Girard, S., and Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Stat. and Comput.*, 34(12). doi:10.1007/s11222-023-10331-2.

Andres, N., Steeb, N., Badoux, A., and Hegg, C., editors (2021). *Extremhochwasser an der Aare. Hauptbericht Projekt EXAR. Methodik und Resultate. [Extreme flooding of the Aare. Main Report on the EXAR Project. Methodology and results.]*, volume 104 of *WSL Berichte*.

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., Agrawal, S., and Kalchbrenner, N. (2023). Deep learning for day forecasts from sparse observations. *ArXiv preprint*. doi:10.48550/arXiv.2306.06079.

Angelopoulos, A. N. and Bates, S. (2023). Conformal Prediction: A Gentle Introduction. *Found. Trends in Mach. Learn.*, 16(4):494–591. doi:10.1561/2200000101.

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *Ann. Appl. Stat.*, 9(4):2023–2050. doi:10.1214/15-aoas863.

Asadi, P., Engelke, S., and Davison, A. C. (2018). Optimal regionalization of extreme value distributions for flood estimation. *J. Hydrol.*, 556:182–193. doi:10.1016/j.jhydrol.2017.10.051.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Ann. Stat.*, 47(2):1148–1178. doi:10.1214/18-AOS1709.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *Ann. Probab.*, 2(5):792–804. doi:10.1214/aop/1176996548.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *Ann. Stat.*, 49(1):486–507. doi:10.1214/20-AOS1965.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *Ann. Stat.*, 51(2):816–845. doi:10.1214/23-AOS2276.

Bartusek, S., Kornhuber, K., and Ting, M. (2022). 2021 North American heatwave amplified by climate change-driven nonlinear interactions. *Nature Climate Change*, 12:1143–1150. doi:10.1038/s41558-022-01520-4.

Bauer, P. (2024). What if? Numerical weather prediction at the crossroads. *ArXiv preprint*. doi:10.48550/arXiv.2407.03787.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F. (2024). The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883. doi:10.1175/BAMS-D-23-0162.1.

Bezzola, G. R. and Hegg, C. (2007). Ereignisanalyse Hochwasser 2005, Teil 1 — Prozesse, Schäden und erste Einordnung [Event analysis of the 2005 flood, Part 1 — Processes, damage and initial classification]. Technical report, Federal Office for the Environment FOEN, Swiss Federal Institute for Forest, Snow and Landscape Research WSL. Umwelt-Wissen Nr. 0707. 215 S.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538. doi:10.1038/s41586-023-06185-3.

Blazejczyk, K., Epstein, Y., Jendritzky, G., Staiger, H., and Tinz, B. (2012). Comparison of UTCI to selected thermal indices. *International journal of biometeorology*, 56:515–535. doi:10.1007/s00484-011-0453-2.

Bodik, J. and Pasche, O. C. (2024). Granger causality in extremes. *ArXiv preprint*. doi:10.48550/arXiv.2407.09632.

Bonavita, M. (2024). On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters*, 51(12):e2023GL107377. doi:10.1029/2023GL107377.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62.

Boucheron, S. and Thomas, M. (2015). Tail index estimation, concentration and adaptivity. *Electron. J. Stat.*, 9(2):2751–2792. doi:10.1214/15-EJS1088.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S. (2022). Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environ. Data Sci.*, 1:e5. doi:10.1017/eds.2022.4.

Breiman, L. (1996). Stacked Regressions. *Mach. Learn.*, 24:49–64. doi:10.1023/A:1018046112532.

Bücher, A. and Zhou, C. (2021). A Horse Race between the Block Maxima Method and the Peak-over-Threshold Approach. *Stat. Sci.*, 36(3):360–378. doi:10.1214/20-STS795.

Buriticá, G. and Engelke, S. (2024). Progression: an extrapolation principle for regression. *ArXiv preprint*. doi:10.48550/arXiv.2410.23246.

Buriticá, G., Hentschel, M., Pasche, O. C., Röttger, F., and Zhang, Z. (2025). Modeling extreme events: Univariate and multivariate data-driven approaches. *Extremes*, 28(1):75–99. doi:10.1007/s10687-024-00499-9.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., and Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138. doi:10.1126/science.adf6369.

Buzan, J. R. and Huber, M. (2020). Moist Heat Stress on a Hotter Earth. *Annual Review of Earth and Planetary Sciences*, 48(1):623–655. doi:10.1146/annurev-earth-053018-060100.

Bülte, C., Horat, N., Quinting, J., and Lerch, S. (2024). Uncertainty quantification for data-driven weather models. *ArXiv preprint*. doi:10.48550/arXiv.2403.13458.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, 37(9):1277–1284. doi:10.1016/j.cageo.2010.07.005.

Cannon, A. J. (2012). Neural networks for probabilistic environmental prediction: Conditional Density Estimation Network Creation and Evaluation (CaDENCE) in R. *Comput. Geosci.*, 41:126–135. doi:10.1016/j.cageo.2011.08.023.

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., Hunt, K. M. R., Lee, R. W., Swaminathan, R., Vandaele, R., and Volonté, A. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Clim. Atmos. Sci.*, 7(1):93. doi:10.1038/s41612-024-00638-w.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *J. R. Stat. Soc. C*, 54(1):207–222. doi:10.1111/j.1467-9876.2005.00479.x.

Chen, H., Huang, Z., Lam, H., Qian, H., and Zhang, H. (2021). Learning prediction intervals for regression: Generalization and calibration. In *Int. Conf. Artificial Intelligence and Stat.*, volume 130, pages 820–828.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W. (2023a). FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *ArXiv preprint*. doi:10.48550/arXiv.2304.02948.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H. (2023b). FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.*, 6(1):190. doi:10.1038/s41612-023-00512-1.

Chen, X. and Meinshausen, N. (2023). Engression: Extrapolation for nonlinear regression? *ArXiv preprint*.

doi:10.48550/arXiv.2307.00835.

Chernozhukov, V. (2005). Extremal quantile regression. *Ann. Stat.*, 33(2):806 – 839. doi:10.1214/009053604000001165.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. Conf. Empir. Methods Nat. Lang. Process.*, pages 1724–1734. doi:10.3115/v1/D14-1179.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme Australian wildfires. *Spat. Stat.*, 59. doi:10.1016/j.spasta.2024.100811.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J. Off. Stat.*, 6(1):3–73.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer. doi:10.1007/978-1-4471-3675-0.

Cox, D. R. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference (with discussion). *J. R. Stat. Soc. B*, 49:1–39.

Daouia, A., Gardes, L., Girard, S., and Lekina, A. (2011). Kernel estimators of extreme level curves. *TEST*, 20(2):311–333. doi:10.1007/s11749-010-0196-0.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. doi:10.1017/CBO9780511802843.

Davison, A. C., Hinkley, D. V., and Young, G. A. (2003). Recent Developments in Bootstrap Methodology. *Stat. Sci.*, 18(2):141–157. doi:10.1214/ss/1063994969.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. R. Stat. Soc. B*, 52:393–442. doi:10.1111/j.2517-6161.1990.tb01796.x.

de Carvalho, M., Kumukova, A., and Dos Reis, G. (2022a). Regression-type analysis for multivariate extreme values. *Extremes*, 25(4):595–622. doi:10.1007/s10687-022-00446-6.

de Carvalho, M., Pereira, S., Pereira, P., and de Zea Bermudez, P. (2022b). An extreme value bayesian lasso for the conditional left and right tails. *J. Agric. Biol. Environ. Stat.*, 27(2):222–239. doi:10.1007/s13253-021-00469-9.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer. doi:10.1007/0-387-34471-3.

de Haan, L. and Zhou, C. (2022). Bootstrapping extreme value estimators. *J. Am. Stat. Assoc.*, 119(545):382–393. doi:10.1080/01621459.2022.2120400.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Dupuis, D. J., Engelke, S., and Trapin, L. (2023). Modeling panels of extremes. *Ann. Appl. Stat.*, 17(1):498–517. doi:10.1214/22-AOAS1639.

ECMWF (2024). Severe Event Catalogue - Forecast User - ECMWF Confluence Wiki. https://confluence.ecmwf.int/display/FCST/Severe+Event+Catalogue. Accessed 08 April 2024.

Elman, J. L. (1990). Finding Structure in Time. *Cogn. Sci.*, 14(2):179–211. doi:10.1207/s15516709cog1402_1.

Engelke, S. and Hitz, A. (2020). Graphical models for extremes (with discussion). *J. R. Stat. Soc. B*, 82(4):871–932. doi:10.1111/rssb.12355.

Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annu. Rev. Stat. Appl.*, 8:241–270. doi:10.1146/annurev-statistics-040620-041554.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., Bell, A., and Kalchbrenner, N. (2022). Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13(1):5145. doi:10.1038/s41467-022-32483-x.

Fischer, E., Sippel, S., and Knutti, R. (2021). Increasing probability of record-shattering climate extremes. *Nat. Clim. Chang.*, 11:689–695. doi:10.1038/s41558-021-01092-9.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Math. Proc. Camb. Philos. Soc.*, volume 24, pages 180–190. Cambridge University Press.

Gardes, L. and Stupfler, G. (2019). An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT*, 17(1):109–144.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., Leone, M., De Sario, M., Bell, M. L., Guo, Y.-L. L., Wu, C.-f., Kan, H., Yi, S.-M., de Sousa Zanotti Stagliorio Coelho, M., Saldiva, P. H. N., Honda, Y., Kim, H., and Armstrong, B. (2015). Mortality risk attributable to high and low ambient temperature: A multicountry observational study. *The Lancet*, 386(9991):369–375. doi:10.1016/S0140-6736(14)62114-0.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Comput.*, 12(10):2451–2471. doi:10.1162/089976600300015015.

Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. (2003). Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.*, 3:115–143.

Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24:2693–2720. doi:10.3150/17-BEJ941.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021). Causal discovery in heavy-tailed models. *Ann. Stat.*, 49(3):1755–1778. doi:10.1214/20-AOS2021.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal random forests. *J. Am. Stat. Assoc.*, 119(548):3059–3072. doi:10.1080/01621459.2023.2300522.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151. doi:10.1146/annurev-statistics-062713-085831.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gruber, K., Gauster, T., Laaha, G., Regner, P., and Schmidt, J. (2022). Profitability and investment risk of Texan power system winterization. *Nature Energy*, 7(5):409–416. doi:10.1038/s41560-022-00994-y.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognit.*, 127:108496. doi:10.1016/j.patcog.2021.108496.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12):701–702.

Harris, N. and Drton, M. (2013). PC Algorithm for Nonparanormal Graphical Models. *Journal of Machine Learning Research*, 14:3365–3383.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition. doi:10.1007/978-0-387-84858-7.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049. doi:10.1002/qj.3803.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 4(2):251–257. doi:10.1016/0893-6080(91)90009-T.

Huet, N., Clémençon, S., and Sabourin, A. (2024). On regression in extreme regions. *ArXiv preprint*. doi:10.48550/arXiv.2303.03084.

Jalalzai, H., Clémençon, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *Adv. Neural Inf. Process. Syst.*, volume 31, pages 3092–3100.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer, 2nd edition. doi:10.1007/978-1-0716-1418-1.

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures. In *Proc. 32nd Int. Conf. Mach. Learn.*, volume 37, pages 2342–2350.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour.*, 25:1287–1304. doi:10.1016/S0309-1708(02)00056-8.

Keef, C., Tawn, J., and Svensson, C. (2009). Spatial risk assessment for extreme river flows. *J. R. Stat. Soc. C*, 58(5):601–618. doi:10.1111/j.1467-9876.2009.00672.x.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Netw.*, 22(9):1341–1356. doi:10.1109/TNN.2011.2162110.

Kim, B., Xu, C., and Barber, R. (2020). Predictive inference is free with the jackknife+-after-bootstrap. In *Adv. Neural Inf. Process. Syst.*, volume 33, pages 4138–4149.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *3rd Int. Conf. Learn. Repres.*

Kinsvater, P., Fried, R., and Lilienthal, J. (2016). Regional extreme value index estimation and a test of tail homogeneity. *Environmetrics*, 27(2):103–115.

Kiriliouk, A. and Naveau, P. (2020). Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. *Ann. Appl. Stat.*, 14(3):1342–1358. doi:10.1214/20-AOAS1355.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-Normalizing Neural Networks. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, pages 972–981.

Klüppelberg, C. and Krali, M. (2021). Estimating an extreme bayesian network via scalings. *Journal of Multivariate Analysis*, 181:104672. doi:10.1016/j.jmva.2020.104672.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S. (2024). Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066. doi:10.1038/s41586-024-07744-y.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.

Koh, J. (2023). Gradient boosting with extreme-value theory for wildfire prediction. *Extremes*, 26(2):273–299. doi:10.1007/s10687-022-00454-6.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421. doi:10.1126/science.adi2336.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F. (2024). AIFS - ECMWF's data-driven forecasting system. *ArXiv preprint*. doi:10.48550/arXiv.2406.01465.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165. doi:10.1002/qj.4351.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444. doi:10.1038/nature14539.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523):1094–1111. doi:10.1080/01621459.2017.1307116.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. B*, 76(1):71–96. doi:10.1111/rssb.12021.

Leinonen, J., Hamann, U., Nerini, D., Germann, U., and Franch, G. (2023). Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *ArXiv preprint*. doi:10.48550/arXiv.2304.12891.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's Dilemma: Extreme Events and Forecast Evaluation. *Statistical Science*, 32(1):106 – 127. doi:10.1214/16-STS588.

Lhaut, S., Sabourin, A., and Segers, J. (2022). Uniform concentration bounds for frequencies of rare events. *Statist. Probab. Lett.*, 189:109610. doi:10.1016/j.spl.2022.109610.

Li, D. and Wang, H. J. (2019). Extreme Quantile Estimation for Autoregressive Models. *J. Bus. Econ. Stat.*, 37(4):661–670. doi:10.1080/07350015.2017.1408469.

Lin, H., Mo, R., and Vitart, F. (2022). The 2021 Western North American Heatwave and Its Subseasonal Predictions. *Geophysical Research Letters*, 49(6):e2021GL097036. doi:10.1029/2021GL097036.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.*, 201:272–288. doi:10.1016/S0022-1694(97)00041-3.

Lo, Y. T. E., Mitchell, D. M., Buzan, J. R., Zscheischler, J., Schneider, R., Mistry, M. N., Kyselý, J., Lavigne, É., da Silva, S. P., Royé, D., Urban, A., Armstrong, B., Multi-Country Multi-City (MCC) Collaborative Research Network, Gasparrini, A., and Vicedo-Cabrera, A. M. (2023). Optimal heat stress metric for modelling heat-related mortality varies from country to country. *Int. J. Climatol.*, 43(12):5553–5568. doi:10.1002/joc.8160.

Lopez-Gomez, I., McGovern, A., Agrawal, S., and Hickey, J. (2023). Global Extreme Heat Forecasting Using Neural Weather Models. *Artificial Intelligence for the Earth Systems*, 2(1):e220035. doi:10.1175/AIES-D-22-0035.1.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Adv. Neural Inf. Process. Syst.*, volume 30.

Maathuis, M. H. and Nandy, P. (2016). *A review of some recent advances in causal inference. In P. Bhlmann, P. Drineas, M. Kane and M.J. van der Laan (Eds.), Handbook of Big Data*. Chapman and Hall.

Magnusson, L. (2023). Exploring machine-learning forecasts of extreme weather. https://www.ecmwf.int/en/newsletter/176/news/exploring-machine-learning-forecasts-extreme-weather.

Martel, J.-L., Arsenault, R., Turcotte, R., Castañeda Gonzalez, M., Brissette, F., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Lachance-Cloutier, S., Rondeau-Genesse, G., and Caron, L.-P. (2025). Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis. *Hydrol. Earth Syst. Sci.*, 29(19):4951–4968. doi:10.5194/hess-29-4951-2025.

Meinshausen, N. (2006). Quantile Regression Forests. *J. Mach. Learn. Res.*, 7(35):983–999.

Mhalla, L., Chavez-Demoulin, V., and Dupuis, D. (2020). Causal mechanism of extreme river discharges in the upper Danube basin network. *Applied Statistics*, 69:741–764. doi:10.1111/rssc.12415.

National Weather Service (2021). Valentine's Week Winter Outbreak 2021: Snow, Ice, & Record Cold. https://www.weather.gov/hgx/2021ValentineStorm. Accessed 26 January 2024.

Naveau, P., Hannart, A., and Ribes, A. (2020). Statistical methods for extreme event attribution in climate science. *Annu. Rev. Stat. Appl.*, 7:89–110. doi:10.1146/annurev-statistics-031219-041314.

Neal, E., Huang, C. S. Y., and Nakamura, N. (2022). The 2021 Pacific Northwest Heat Wave and Associated Blocking: Meteorology and the Role of an Upstream Cyclone as a Diabatic Source of Wave Activity. *Geophysical Research Letters*, 49(8):e2021GL097699. doi:10.1029/2021GL097699.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563. doi:10.1038/s41586-024-07145-1.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. (2023a). Climax: A foundation model for weather and climate. *ArXiv preprint*. doi:10.48550/arXiv.2301.10343.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Madireddy, S., Maulik, R., Kotamarthi, V., Foster, I., and Grover, A. (2023b). Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *ArXiv preprint*. doi:10.48550/arXiv.2312.03876.

Oliveira, R. I., Orenstein, P., Ramos, T., and Romano, J. V. (2024). Split conformal prediction and non-exchangeable data. *J. Mach. Learn. Res.*, 25(225):1–38.

Olivetti, L. and Messori, G. (2024a). Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geoscientific Model Development*, 17(6):2347–2358. doi:10.5194/gmd-17-2347-2024.

Olivetti, L. and Messori, G. (2024b). Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast. *Geoscientific Model Development*, 17(21):7915–7962. doi:10.5194/gmd-17-7915-2024.

Osczevski, R. and Bluestein, M. (2005). The new wind chill equivalent temperature chart. *Bulletin of the American Meteorological Society*, 86(10):1453–1458. doi:10.1175/BAMS-86-10-1453.

Owens, R. and Hewson, T. (2018). ECMWF Forecast User Guide. *Meteorological Bulletin*. doi:10.21957/M1CS7H.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*, chapter 18. IntechOpen. doi:10.5772/6078.

Papadopoulos, H., Gammerman, A., and Vovk, V. (2008). Normalized nonconformity measures for regression

conformal prediction. In *Int. Conf. Artificial Intelligence and Appl.*, pages 64–69.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *Mach. Learn.: ECML 2002*, volume 2430, pages 345–356. Springer. doi:10.1007/3-540-36755-1_29.

Papadopoulos, H., Vovk, V., and Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. *J. Artificial Intelligence Res.*, 40:815–840. doi:10.1613/jair.3198.

Pasche, O. C., Chavez-Demoulin, V., and Davison, A. C. (2023). Causal modelling of heavy-tailed variables and confounders with application to river flow. *Extremes*, 26(3):573–594. doi:10.1007/s10687-022-00456-4.

Pasche, O. C. and Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *Ann. Appl. Stat.*, 18(4):2818–2839. doi:10.1214/24-AOAS1907.

Pasche, O. C., Lam, H., and Engelke, S. (2025a). Extreme conformal prediction: Reliable intervals for high-impact events. *ArXiv preprint*. doi:10.48550/arXiv.2505.08578.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S. (2024). Data release: Validating deep-learning weather forecast models on recent high-impact extreme events. *Zenodo*. doi:10.5281/zenodo.14358212. (version 1.0).

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S. (2025b). Validating deep learning weather forecast models on recent high-impact extreme events. *Artif. Intell. Earth Syst.*, 4(1):e240033. doi:10.1175/AIES-D-24-0033.1.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *ArXiv preprint*. doi:10.48550/arXiv.2202.11214.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *Proc. 35th Int. Conf. Mach. Learn.*, volume 80, pages 4075–4084.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals (with Discussion). *J. R. Stat. Soc. B*, 78(5):947–1012. doi:10.1111/rssb.12167.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Pfister, N. and Bühlmann, P. (2024). Extrapolation-aware nonparametric statistical inference. *ArXiv preprint*. doi:10.48550/arXiv.2402.09758.

Philip, S. Y., Kew, S., van Oldenborgh, G. J., Anslow, F. S., Seneviratne, S. I., Vautard, R., Coumou, D., Ebi, K. L., Arrighi, J., Singh, R., van Aalst, M., Pereira Marghidan, C., Wehner, M., Yang, W., Li, S., Schumacher, D. L., Hauser, M., Bonnet, R., Luu, L. N., Lehner, F., Gillett, N., Tradowsky, J. S., Vecchi, G. A., Rodell, C., Stull, R. B., Howard, R., and Otto, F. E. L. (2022). Rapid attribution analysis of the extraordinary heat wave on the Pacific coast of the US and Canada in June 2021. *Earth System Dynamics*, 13(4):1689–1713. doi:10.5194/esd-13-1689-2022.

Pickands, III, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat.*, 3(1):119–131. doi:10.1214/aos/1176343003.

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M. (2023). GenCast: Diffusion-based ensemble forecasting for medium-range weather. *ArXiv preprint*. doi:10.48550/arXiv.2312.15796.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. (2020). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11). doi:10.1029/2020MS002203.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. (2024). WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019. doi:10.1029/2023MS004019.

Rasp, S. and Thuerey, N. (2021). Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on

Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, 13(2). doi:10.1029/2020MS002405.

Richards, J. and Huser, R. (2024). Regression modelling of spatiotemporal extreme U.S. wildfires via partially-interpretable neural networks. *ArXiv preprint*. doi:10.48550/arXiv.2208.07581.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. In *Adv. Neural Inf. Process. Syst.*, volume 32.

Roodman, D. (2018). Bias and size corrections in extreme value modeling. *Comm. Statist. Theory Methods*, 47(14):3377–3391. doi:10.1080/03610926.2017.1353630.

Rothfusz, L. P. and Headquarters, N. S. R. (1990). The heat index equation (or, more than you ever wanted to know about heat index). *Fort Worth, Texas: National Oceanic and Atmospheric Administration, National Weather Service, Office of Meteorology*, 9023:640.

Röthlisberger, M. and Papritz, L. (2023). Quantifying the physical processes leading to atmospheric hot extremes at a global scale. *Nature Geoscience*, 16(3):210–216. doi:10.1038/s41561-023-01126-1.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Trans. Neural Netw.*, 20(1):61–80. doi:10.1109/TNN.2008.2005605.

Schmocker-Fackel, P. and Naef, F. (2010). Changes in flood frequencies in Switzerland since 1500. *Hydrol. Earth Syst. Sci.*, 14(8):1581–1594.

Schumacher, D. L., Hauser, M., and Seneviratne, S. I. (2022). Drivers and Mechanisms of the 2021 Pacific Northwest Heatwave. *Earth's Future*, 10(12):e2022EF002967. doi:10.1029/2022EF002967.

Selz, T. and Craig, G. C. (2023). Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters*, 50(20):e2023GL105747. doi:10.1029/2023GL105747.

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., and Zhou, B. (2021). *Weather and Climate Extreme Events in a Changing Climate (Chapter 11)*, pages 1513–1766. Cambridge University Press. doi:10.1017/9781009157896.013.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9(3).

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Smith, A., Lott, N., and Vose, R. (2011). The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6):704–708. doi:10.1175/2011BAMS3015.1.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15:1929–1958.

Steinberger, L. and Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. *ArXiv preprint*. doi:10.48550/arXiv.1602.05801.

Thomas, M. (2015). *Concentration results on extreme value theory*. PhD thesis, Univeristé Paris Diderot Paris 7. `https://theses.hal.science/tel-01177197`.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Adv. Neural Inf. Process. Syst.*, volume 32.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5 – RMSProp. Technical report, COURSERA: Neural Networks for Machine Learning.

van Oordt, M. and Zhou, C. (2019). Systemic risk and bank business models. *J. Appl. Econometrics*, 34(3):365–384. doi:10.1002/jae.2666.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Adv. Neural Inf. Process. Syst.*, volume 30.

Velthoen, J., Cai, J.-J., Jongbloed, G., and Schmeits, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622. doi:10.1007/s10687-019-00355-1.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667. doi:10.1007/s10687-023-00473-x.

Vovk, V. (2015). Cross-conformal predictors. *Ann. Math. Artif. Intell.*, 74:9–28. doi:10.1007/s10472-013-9368-4.

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *Proc. 16th Int. Conf. Mach. Learn.*, pages 444–453.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, 1st edition. doi:10.1007/b106715.

Wang, H. J., Li, D., and He, X. (2012). Estimation of High Conditional Quantiles for Heavy-Tailed Distributions. *J. Am. Stat. Assoc.*, 107(500):1453–1464. doi:10.1080/01621459.2012.716382.

Watson, P. A. G. (2022). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17(11):111004. doi:10.1088/1748-9326/ac9d4e.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw.*, 1(4):339–356. doi:10.1016/0893-6080(88)90007-X.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N. (2021). Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, 13(7). doi:10.1029/2021MS002502.

White, R. H., Anderson, S., Booth, J. F., Braich, G., Draeger, C., Fei, C., Harley, C. D. G., Henderson, S. B., Jakob, M., Lau, C.-A., Mareshet Admasu, L., Narinesingh, V., Rodell, C., Roocroft, E., Weinberger, K. R., and West, G. (2023). The unprecedented Pacific Northwest heatwave of June 2021. *Nature Communications*, 14(1):727. doi:10.1038/s41467-023-36289-3.

Wolpert, D. H. (1992). Stacked generalization. *Neural Netw.*, 5(2):241–259. doi:10.1016/S0893-6080(05)80023-1.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):4–24. doi:10.1109/TNNLS.2020.2978386.

Youngman, B. D. (2019). Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts. *J. Am. Stat. Assoc.*, 114(528):1865–1879. doi:10.1080/01621459.2018.1529596.

Zachariah, M., Vautard, R., Chandrasekaran, R., Chaithra, ST., Kimutai, J., Arulalan, T., AchutaRao, K., Barnes, C., Singh, R., Vahlberg, M., Arrgihi, J., Raju, E., Sharma, U., Ogra, A., Vaddhanaphuti, C., Bahinipati, CS., Tschakert, P., Pereira Marghidan, C., Mondal, A., Schwingshackl, C., Philip, S., and Otto, F. (2023). Extreme humid heat in South Asia in April 2023, largely driven by climate change, detrimental to vulnerable and disadvantaged communities. Technical report, Imperial College London.

Zeder, J., Sippel, S., Pasche, O. C., Engelke, S., and Fischer, E. M. (2023). The effect of a short observational record on the statistics of temperature extremes. *Geophys. Res. Lett.*, 50(16):e2023GL104090. doi:10.1029/2023GL104090.

Zhang, W., Quan, H., and Srinivasan, D. (2019). An Improved Quantile Regression Neural Network for Probabilistic Load Forecasting. *IEEE Trans. Smart Grid*, 10(4):4425–4434. doi:10.1109/TSG.2018.2859749.

Zhou, X., Chen, B., Gui, Y., and Cheng, L. (2025). Conformal prediction: A data perspective. *ACM Comput. Surv.*, 58(2):49:1–49:37. doi:10.1145/3736575.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Maraun, D., Ramos, A. M., Ridder, N. N., Thiery, W., and Vignotto, E. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1(7):333–347. doi:10.1038/s43017-020-0060-z.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.*, 62(318):626–633. doi:10.1080/01621459.1967.10482935.

Appendices and Supplements

A Supplement to Neural networks for extreme quantile regression with an application to forecasting of flood risk

A.1 Additional LSTM illustration

Figure A.1 shows a schematic representation of a multilayer LSTM network.

A.2 Details on Algorithms 1 and 2

Algorithms 1 and 2 contain some abbreviated function calls. We give some details here:

RandomValidationSplit($\mathcal{I}$): For independent data, splits the index set $\mathcal{I}$ randomly into training set $\mathcal{T}$ and validation set $\mathcal{V}$ with prespecified proportions.

SequentialValidationSplit($\mathcal{I}$): For sequential data, splits the index set $\mathcal{I}$ sequentially into training set $\mathcal{T}$ and validation set $\mathcal{V}$ with prespecified proportions, such that all observations in $\mathcal{T}$ are before $\mathcal{V}$ in time.

InitializeNetworkWeights(Θ) / InitializeRecurrentNetWeights(Θ): Initializes the weights of the GPD (recurrent) neural network randomly. The number of weights is determined by Θ.

GetMiniBatches($\mathcal{T}$): Splits the training set $\mathcal{T}$ into mini-batches for stochastic gradient descent.

BackPropUpdate(ℓ, $\hat{\mathcal{W}}$, $\mathbf{x}_{\mathcal{B}}$, $\hat{Q}_{\mathbf{x}_{\mathcal{B}}}(\tau_0)$, Θ): Updates the parameter vector by a gradient step computed by backpropagation; may involve regularization methods such as L_2-penalty or dropout, specified in the hyperparameters Θ.

LossNotImproving($\hat{\mathcal{W}}$, $\mathbf{x}_{\mathcal{V}}$, $\hat{Q}_{\mathbf{x}_{\mathcal{V}}}(\tau_0)$, $z_{\mathcal{V}}$): If validation loss is tracked, then also a stopping criterion (e.g., for early stopping) is specified, and this function returns `TRUE` if this criterion is attained, indicating that the validation loss is not improving any more.

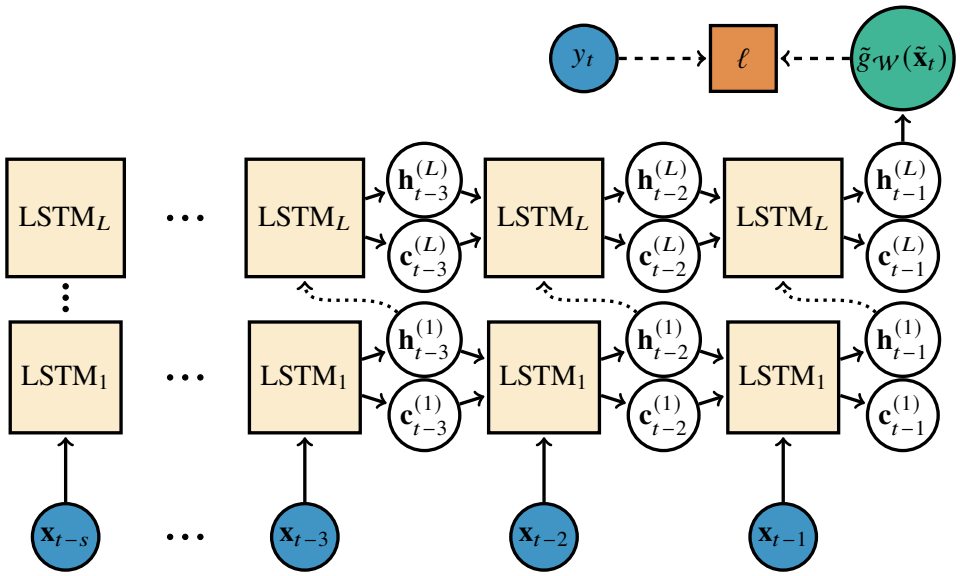

Figure A.1: Multilayer LSTM network flowchart from input $\tilde{\mathbf{x}}_t := (\mathbf{x}_{t-s}, \ldots, \mathbf{x}_{t-1})$ to output $\tilde{g}_{\mathcal{W}}(\tilde{\mathbf{x}}_t)$, with loss evaluation. The LSTM cells represent the transformation in (1.9).

A.3 Simulation study for independent observations

Three data-generating models with independent observations are considered. The training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn from

$$\begin{cases} \mathbf{X} \sim \mathcal{U}([-1,1]^p), \\ Y \mid \mathbf{X} = \mathbf{x} \sim \sigma(\mathbf{x}) \cdot t_{\alpha(\mathbf{x})}, \end{cases} \tag{A.1}$$

with $p = 10$, $\alpha(\mathbf{x}) = 1/\xi(\mathbf{x}) := 7 \cdot \{1 + \exp(4x_1 + 1.2)\}^{-1} + 3$, and three different models for $\sigma(\mathbf{x})$:

Model 1: $\sigma(\mathbf{x}) := 1 + 6\phi(x_1, x_2)$, where ϕ is the bivariate Gaussian density with correlation 0.9,

Model 2: $\sigma(\mathbf{x}) := 4 + 3\cos(7\,\|(x_1, x_2)^\top\|_2 + 3)$,

Model 3: $\sigma(\mathbf{x}) := 4 + 3\cos(6\,\|\mathbf{x}\|_2 + 3.5)$.

To avoid bias in the selection of the data models, $\alpha(\mathbf{x})$ and $\sigma(\mathbf{x})$ in Model 1 are the same as in the simulation study in Velthoen et al. (2023). The choices for Models 2 and 3 are designed to study more complex covariate dependencies. The constants are chosen to have positive scale values and enough variation for the inference task to be interesting.

As the focus of the experiments is on the extremal part of the model and since all extreme value competitors use the same intermediate quantile estimates, we use the true $Q_{\mathbf{x}}(\tau_0)$ as intermediate quantiles. Two main types of network architectures are considered for EQRN. The first type is an MLP with tanh activation functions, narrow architectures with between one and four hidden layers and optional L_2 weight penalty as training regularization. The second type is self-normalizing networks (Klambauer et al., 2017) using SELU activation functions, deeper architectures with between three and eight hidden layers and optional alpha dropout as regularization. This type of network is designed to maintain unit variance and zero mean across layers in deeper networks

trained for regression tasks, in order to avoid vanishing and exploding gradient issues. Results suggest that the additional flexibility of the second type was not necessary for the three tasks at hand, as they never yielded better validation scores than the best tanh models. Table A.1 summarizes the hyperparameters of the chosen EQRN networks.

To evaluate the accuracy of the best models over the full feature space $\mathcal{X} = [-1, 1]^p$, we use the integrated squared error (ISE) between the prediction $\hat{Q}_{\mathbf{x}}(\tau)$ and the true quantile $Q_{\mathbf{x}}(\tau)$,

$$\int_{\mathcal{X}} \left(\hat{Q}_{\mathbf{x}}(\tau) - Q_{\mathbf{x}}(\tau)\right)^2 d\mathbf{x}. \tag{A.2}$$

We generate test features using a Halton sequence (Halton, 1964) and compute the MSE between the corresponding predicted and true response quantiles, to estimate the p-dimensional integral.

Figure A.2 shows the accuracy of EQRN and the competitor methods for an increasingly large τ, for the three data models. The competitors' performances shown here were also significantly improved by using the intermediate $\hat{Q}_{\mathbf{x}}(\tau_0)$ as an additional covariate. For every model, EQRN outperforms all competitors, with a difference in accuracy increasing with τ. EGAM seems to suffer from the large dimension of the feature space at large τ, both when the actual quantile depends on only two or all of the 10 features. The difference in accuracy of EQRN and GBEX compared to the unconditional and semiconditional models is particularly significant for Models 1 and 3, and EQRN generally outperforms GBEX, especially at high probability levels.

We also define the quantile R squared of $\hat{Q}_{\mathbf{x}}(\tau)$ over the sample $\mathcal{D}$ as

$$R^2_\tau := 1 - \frac{\sum_{i=1}^n \left(Q_{\mathbf{x}_i}(\tau) - \hat{Q}_{\mathbf{x}_i}(\tau)\right)^2}{\sum_{i=1}^n \left(Q_{\mathbf{x}_i}(\tau) - \overline{Q_{\mathcal{D}}(\tau)}\right)^2}, \quad \text{with} \quad \overline{Q_{\mathcal{D}}(\tau)} := \frac{1}{n}\sum_{i=1}^n Q_{\mathbf{x}_i}(\tau). \tag{A.3}$$

The definition is similar to the classical R squared coefficient of determination in regression, but the true conditional quantile values are used as targets instead of the response observations. The R^2_τ is essentially the reversed MSE normalized by the variance of the true conditional quantile. A value close to unity indicates a very low MSE compared to the quantile variance, and negative values indicate a MSE larger than the quantile variance.

Figure A.3 shows the quantile R squared, the biases and the residual standard deviations of the same respective quantile predictions compared to the truth. The R squared lead to the same conclusions as the RISE. Regarding the bias-variance decomposition of the RISE, it seems that the variance term is dominating the square bias. Although EQRN is here not the least biased model for large τ values, it is its dominating performance in terms of residual variance that leads to it having the lowest RISE values for every data model.

Figure A.4 shows the predicted $\hat{Q}_{\mathbf{x}}(0.9995)$ for EQRN and the competitors, as a function of the two significant covariates (x_1, x_2) for Model 1. At that extreme level, EQRN still seems to capture the true conditional quantile function quite well, although the predictions show some residual noise. GBEX shows an elliptical stepwise approximation behaviour. and seems to underestimate the smaller quantiles and overestimate the largest quantiles. EGAM and GRF fail drastically at

Table A.1: Hyperparameters of the EQRN networks with the best validation loss for the three independent data models.

	Hidden activation function	Hidden layer dimensions	L_2 penalty
Model 1	tanh	$(128, 128, 128)$	10^{-5}
Model 2	tanh	$(20, 10, 10)$	10^{-5}
Model 3	tanh	$(10, 10, 10)$	0

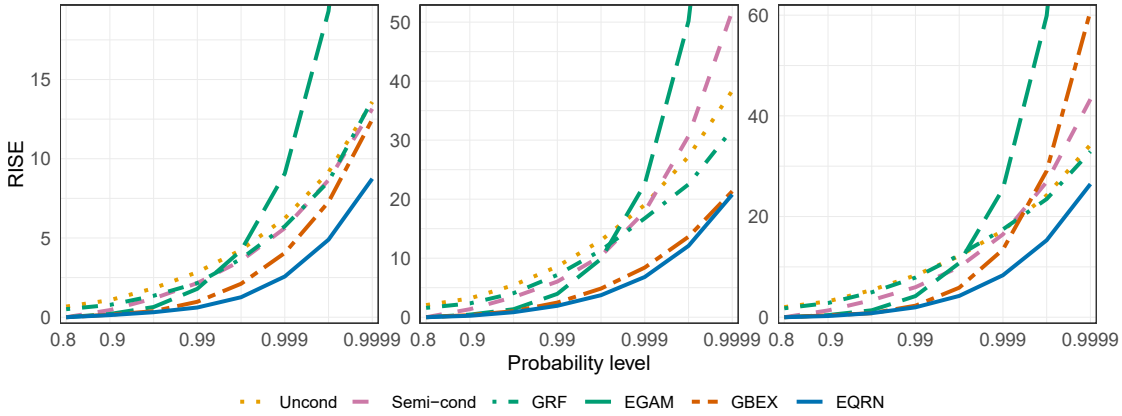

Figure A.2: Root integrated squared error between predicted and true conditional quantiles at different probability levels τ (log-scale) for the selected EQRN model and the improved competitors, for data Models 1–3 (left to right). The cropped-out RMISE for EGAM at level 0.9999 are around 43, 115 and 150, respectively.

recovering the conditional quantile function. The semiconditional estimates are a translation of the intermediate quantiles, which in particular fail to capture the varying shape along X_1.

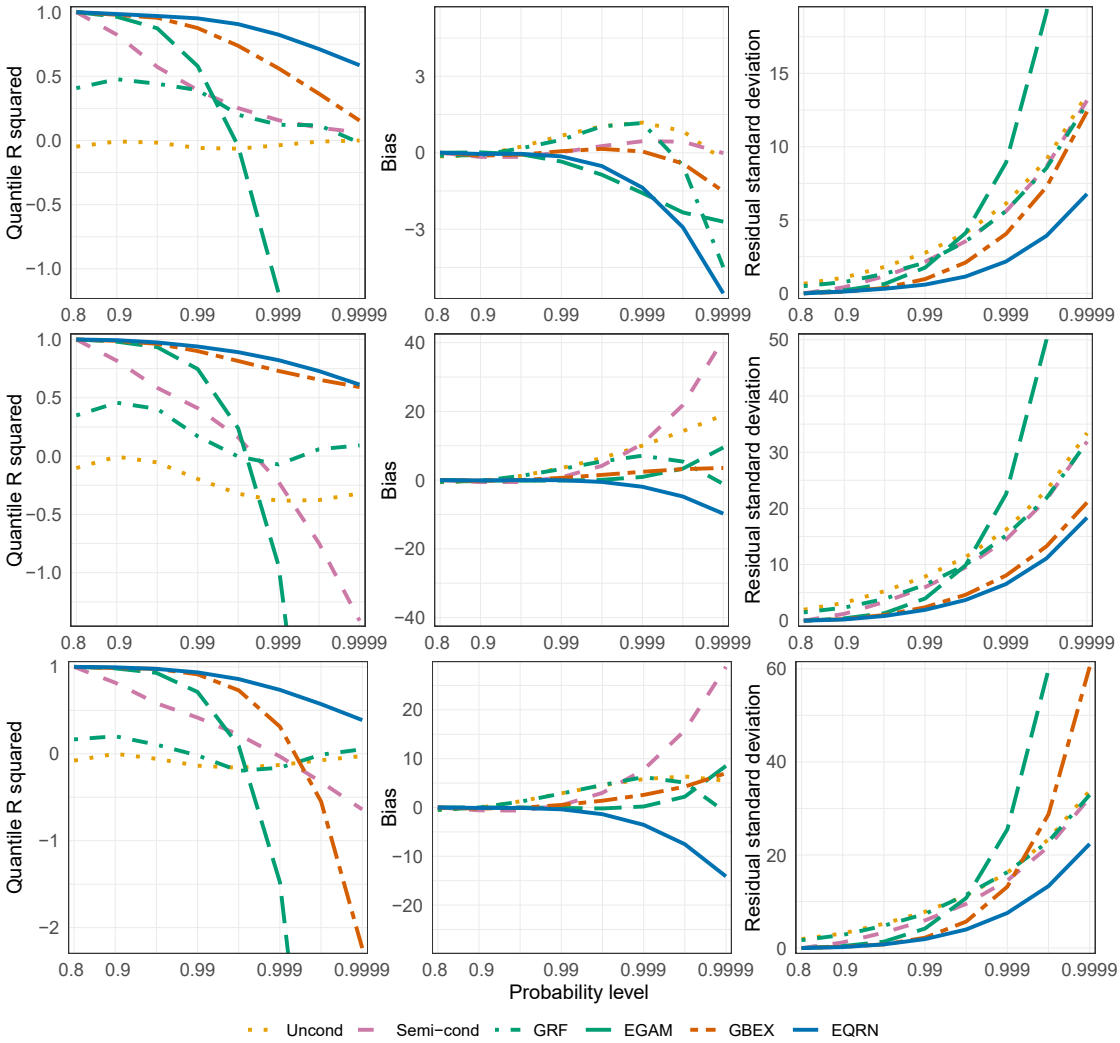

Figure A.3: Quantile R squared, bias and residual standard deviation of the predicted quantiles compared to the truth at different probability levels (log-scale) for the selected EQRN model and improved competitors, for data Models 1–3 (top to bottom).

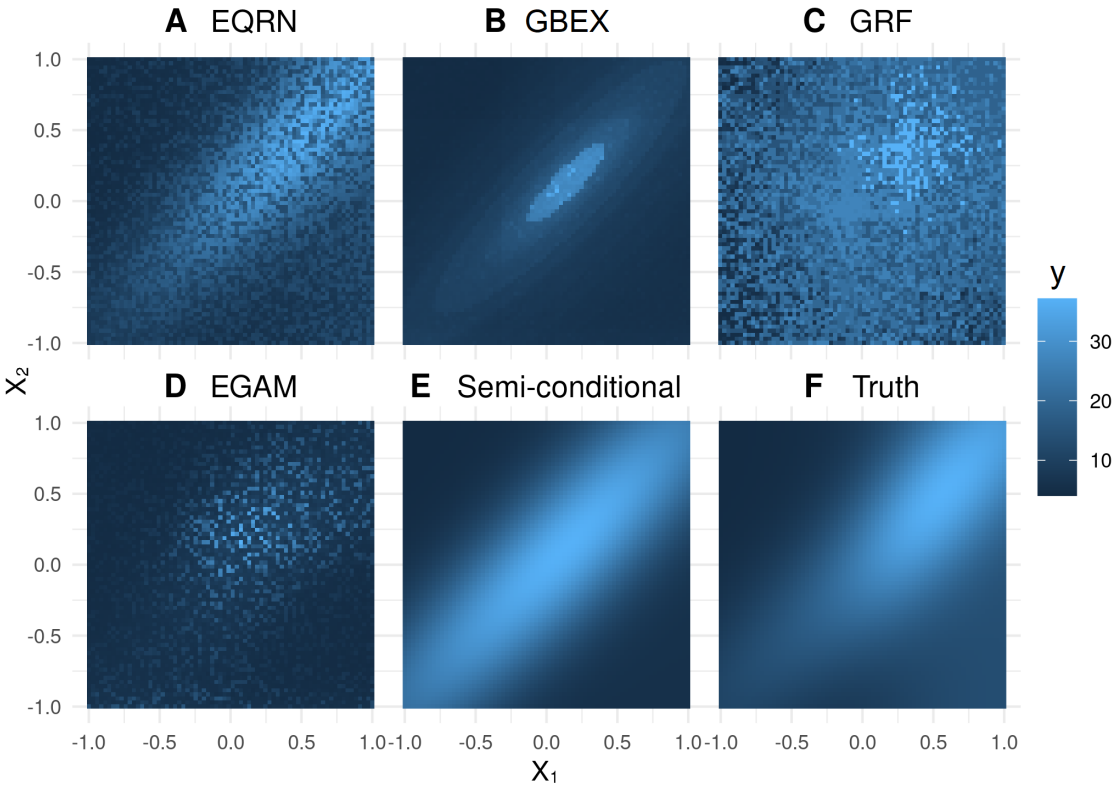

Figure A.4: Conditional quantile predictions of EQRN and the improved competitor models at probability level $\tau = 0.9995$, shown as a function of X_1 and X_2, for Model 1.

A.4 Simulation study for sequentially dependent data

The main results of the simulation study on sequentially dependent data are presented in the main paper. This section discusses additional results. Figure A.5 shows part of the sequential data simulated from the generating process described in the main paper.

Figure A.6 shows the quantile R squared (A.3), the biases and the residual standard deviations of the quantile predictions compared to the truth, for the two selected EQRN models and competitors. The R squared evolution again shows EQRN is the model that best captures the covariate sequential dependence in the tail, as it outperforms all competitors with a difference in accuracy increasing with τ. In terms of bias, the penalized EQRN here scores similar values as EXQAR, and both EQRN versions outperform all other methods. The unpenalized EQRN has the lowest residual variance, closely followed by both GBEX and the penalized EQRN, although GBEX has a bad accuracy overall, due to its large bias.

Figure A.7 shows the impact of the intermediate level τ_0 on the accuracy of the best EQRN model. The value $\tau_0 = 0.8$ used in the rest of the analysis leads to the best RMSE. This relatively low value shows that τ_0 can be chosen much lower than for the classical unconditional GPD model. An intuitive explanation for this fact is the following. In a situation without covariates, the choice of τ_0 is a trade-off between approximation bias (which favours larger thresholds) and variance (which favours lower thresholds); see Figure 3 in the main document. For covariate-dependent data, the distribution of the exceedances varies and more data is needed to accurately capture this

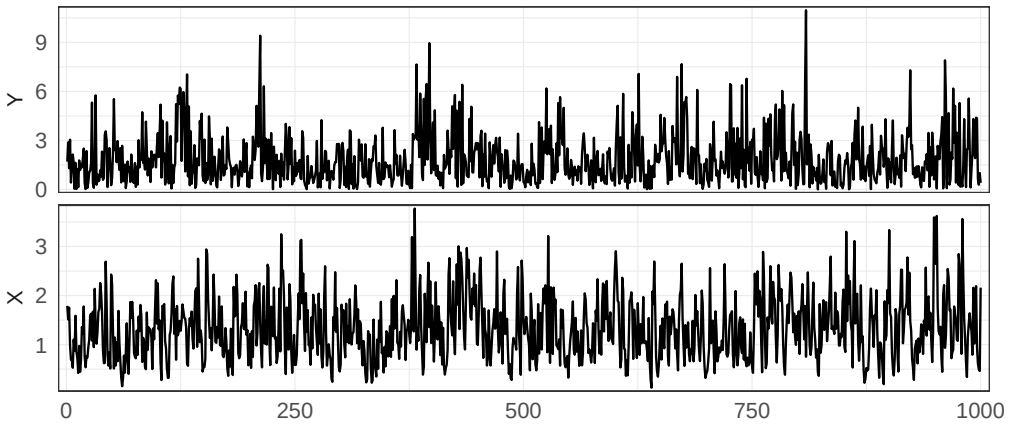

Figure A.5: First 1,000 observations of the sequential data simulated from (1.13).

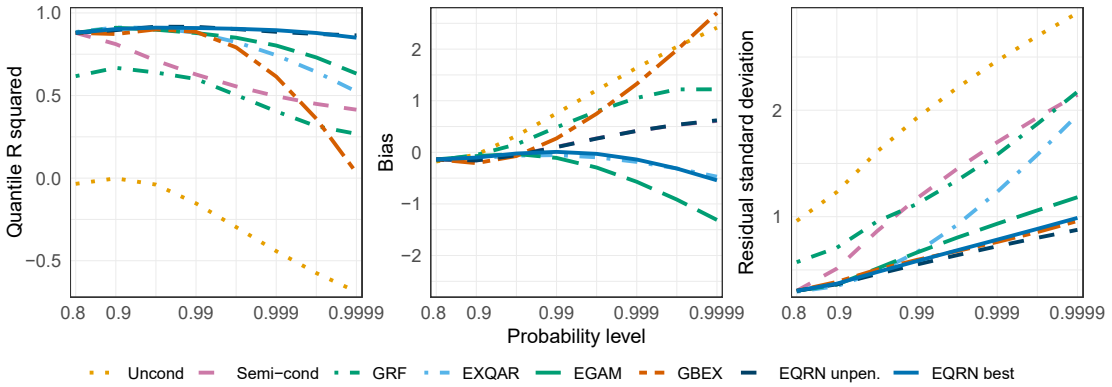

Figure A.6: Quantile R squared, bias and residual standard deviation of the predicted quantiles compared to the truth at different probability levels (log-scale) for the selected EQRN model and the improved competitors, for the sequential data model.

function of the covariates. The variance becomes more important than the approximation bias, therefore, lower thresholds are preferable. Moreover, the flexibility of the GPD regression neural network model seems to be able to absorb some of the approximation bias, also allowing for a low value of τ_0.

The final accuracy seems in fact to not be too sensitive to the choice for τ_0 compared to the network's grid-searched hyperparameters, discussed in the main analysis, as the differences in RMSE for τ_0 values close to the optimum are relatively small. As mentioned in Section 1.3 of the main paper, τ_0 cannot be treated as a classical tuning parameter, as different values for τ_0 generally yield different subsets of exceedances. Thus, the likelihood (1.11), which is used as a goodness of fit metric for hyperparameter tuning, would not be comparable between models. The RMSE cannot be computed in practice either, since $Q_{\mathbf{x}}(\tau)$ is generally unknown.

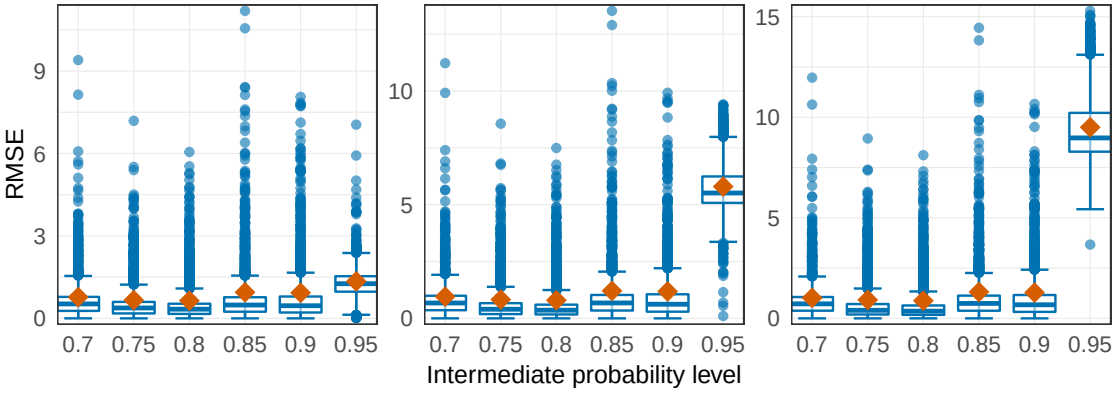

Figure A.7: Boxplot of the absolute residuals of the quantile predictions from the selected EQRN model (blue) and their RMSE (red diamond) at probability levels 0.995 (left), 0.999 (middle) and 0.9995 (right) for different choices of intermediate probability level τ_0, for the sequential data model.

A.5 Application: competitor results

The main results from our application to forecasting flood risk in Switzerland using our proposed EQRN methodology are presented in the main paper. This section discusses and compares additional results using the competitor methods, adapted to provide the same type of forecast as the EQRN approach, with a focus on the 2005 flood event (see Figure 1.2 in the main paper).

We first observe that the predictions have roughly the same behaviour, which shows that all methods capture at least some of the temporal structure based on the past covariates. The semiconditional method (Figure A.8) is clearly not flexible enough, since the predictions only show a weak sensitivity to the changes in covariates. It also fails to trigger any early warning during the main event, due to low probability ratio forecasts never exceeding the selected threshold value of 100. The reason is that, while the intermediate quantile is covariate-dependent, the GPD parameters are constant over time. We conclude that a covariate-dependent model for the tail is required in this application. Figure A.9 shows predictions from the EXQAR model (Li and Wang, 2019). This model is more sensitive to changes in the covariates, but the regression function looks fairly erratic, with sudden spikes at some time points. Those spikes are here due to unusually small shape estimates in combination with a large-scale estimate. This might be caused by an instability in the estimation of the local moments used in the model. EGAM (Figure A.10) fails to trigger an early warning for the first day of the flooding event as its quantile and probability ratio forecasts are very low, but then seems to severely overestimate the river flow during the rest of the event. The best competing model seems to be the GBEX (Figure A.11), as it yields a smooth prediction curve with a pronounced spike at the main event. Comparing this with our EQRN method (Figure 2 in the main paper), it reacts slightly later and its risk forecast (probability ratio) on the day before the first exceedance is significantly lower than for EQRN.

This qualitative way of checking the model is important since in real-world applications, the true extreme quantiles are unknown. Therefore, only some quantitative model checks can be performed, like the right-hand panel of Figure 1.7 in the main paper showing calibration in the number of quantile-exceeding test observations. Figure A.12 compares this number of quantile-exceedances for the competitor predictions. Although they can give evidence against the suitability of a model,

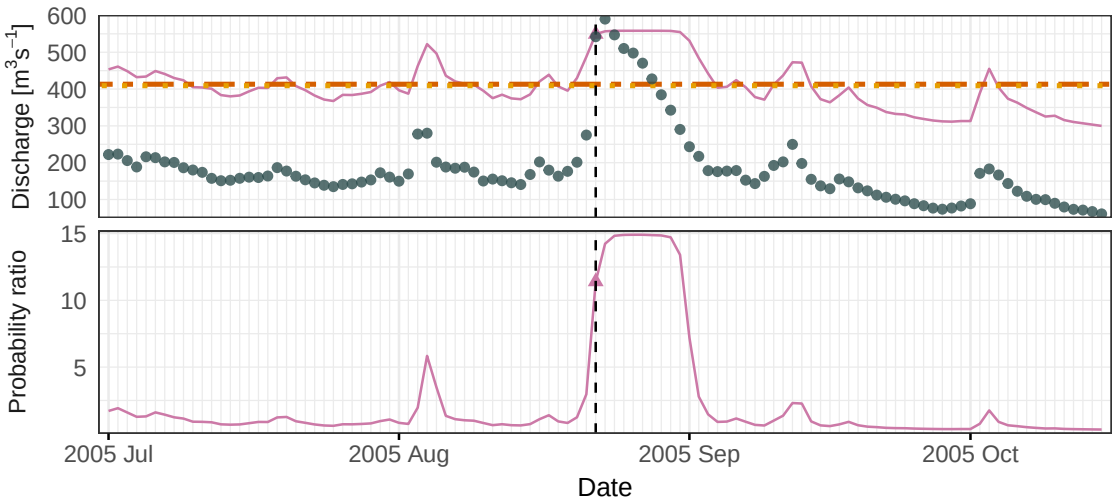

Figure A.8: Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead semiconditional forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the semiconditional parameter forecast. The vertical line indicates August 22, the day of the first exceedance.

it highlights that such metrics only assess calibration but not goodness-of-fit nor accuracy, as unconditional methods obtain similar values to flexible accurate methods. This underlines the importance of the simulation studies, which allow us to evaluate in several settings which methods are more accurate. In particular, in situations with temporal dependence, our EQRN method outperforms the competitors (e.g., Figure 1.5 in the main paper).

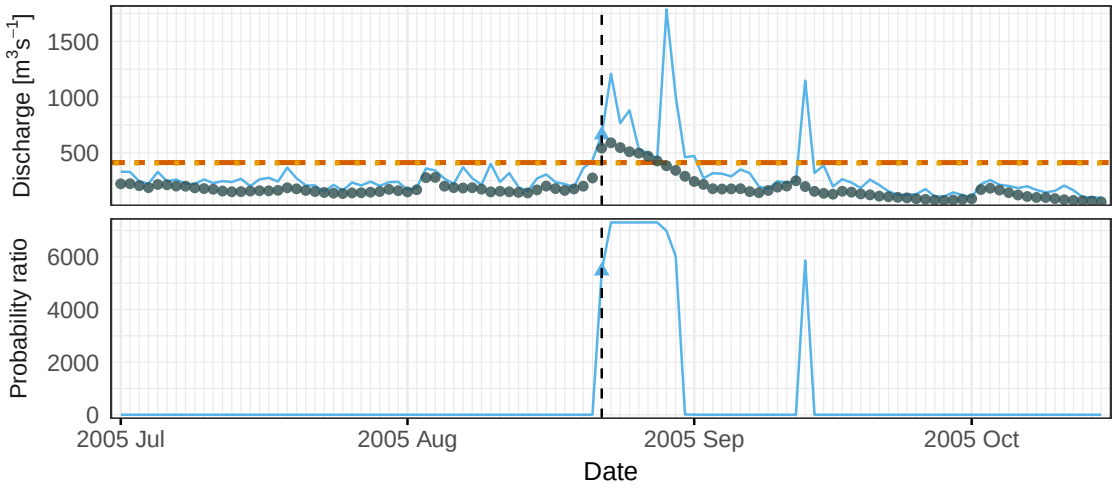

Figure A.9: Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead EXQAR forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the EXQAR forecast. The vertical line indicates August 22, the day of the first exceedance.

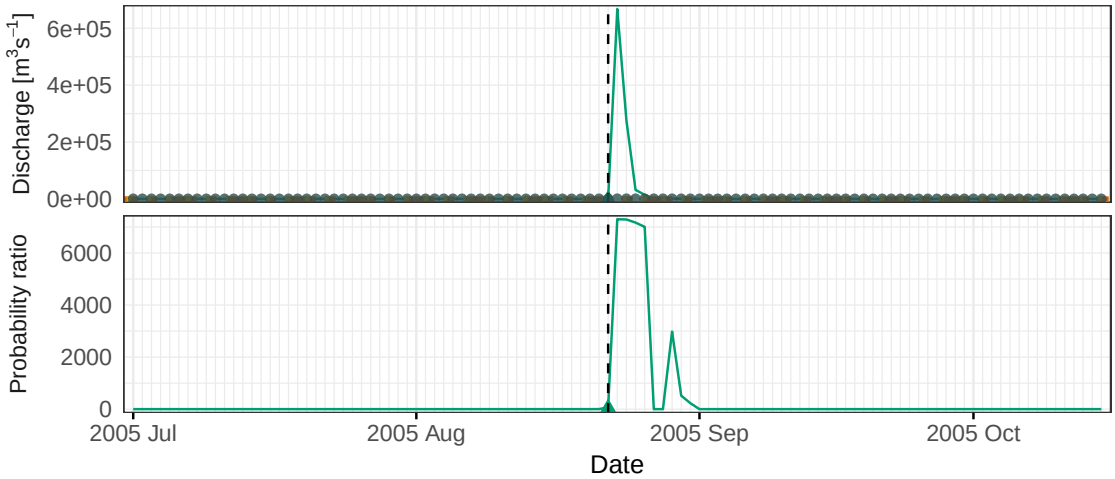

Figure A.10: Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead EGAM forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the EGAM parameter forecast. The vertical line indicates August 22, the day of the first exceedance.

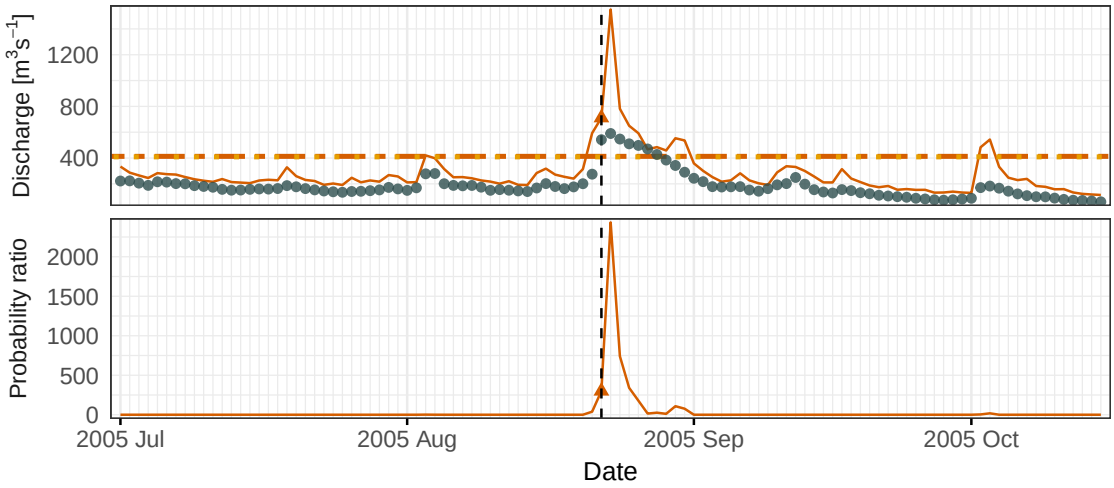

Figure A.11: Top: Daily average discharge (points) at the Bern–Schönau station (62) and one-day-ahead GBEX forecasts of conditional 100-year quantiles (solid line) during the 2005 flood. Horizontal lines show unconditional Q^{100} based on GEV (dashed) and GPD (dotted). Bottom: One-day-ahead forecast of the conditional probability of exceeding the GEV estimated Q^{100} as a ratio to the unconditional probability, using the GBEX parameter forecast. The vertical line indicates August 22, the day of the first exceedance.

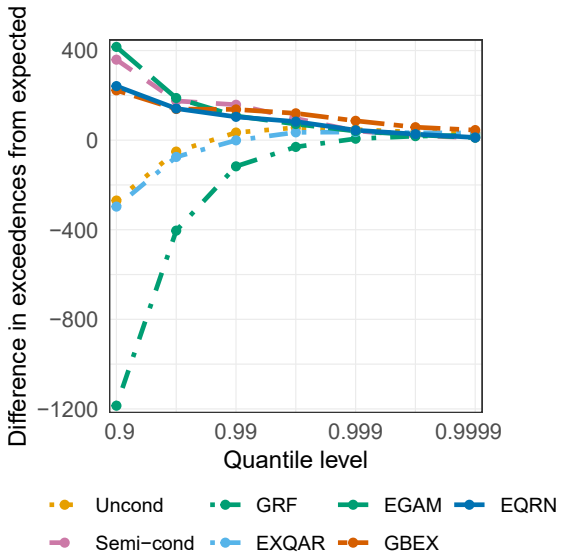

Figure A.12: Difference between the number of observations exceeding the EQRN and competitor quantile predictions on the test set and the expected number of exceedances, for different probability levels (log-scale).

B Appendix to Extreme conformal prediction: Reliable intervals for high-impact events

B.1 Extensions to other conformal approaches

This section describes in more detail the extensions of our proposed extreme conformal method to alternative conformal procedures discussed in Section 2.3.3.

As explained in the main text, the extension to different base predictive models and scores is the most straightforward, as our extreme conformalization described in Section 2.3.2 is agnostic to their definition. This includes the classical split-conformal approach that, instead of the quantile predictions $\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})$, relies on a conditional-mean base regression model $\hat{\mu}(\mathbf{x})$, and the residuals $s(\mathbf{x}, y) := |y - \hat{\mu}(\mathbf{x})|$ as nonconformity score (Papadopoulos et al., 2002; Papadopoulos, 2008; Lei et al., 2018). The procedure from Section 2.3.2 would then be performed with these alternative scores, resulting in the fixed-length extreme conformal PIs $\hat{C}^e(\mathbf{x}) = \left[\hat{\mu}(\mathbf{x}) - \hat{q}^e_\alpha, \hat{\mu}(\mathbf{x}) + \hat{q}^e_\alpha\right]$. If a residual-dispersion estimate $\hat{\sigma}(\mathbf{x})$ is also available, for instance, from a heteroscedastic regression model or a Bayesian approach, using the scaled residuals $s(\mathbf{x}, y) := |y - \hat{\mu}(\mathbf{x})| / \hat{\sigma}(\mathbf{x})$ as nonconformity scores (Papadopoulos et al., 2008, 2011; Lei et al., 2018) would result in the varying-length extreme PIs $\hat{C}^e(\mathbf{x}) = \left[\hat{\mu}(\mathbf{x}) - \hat{q}^e_\alpha \hat{\sigma}(\mathbf{x}), \hat{\mu}(\mathbf{x}) + \hat{q}^e_\alpha \hat{\sigma}(\mathbf{x})\right]$. Although the latter doesn't require extreme quantile regression to yield varying-length PIs, it tends to underestimate conditional variability and yield less adaptive predictions (Romano et al., 2019).

In full-conformal procedures, the data is not split into training and calibration sets. Instead, given a training set $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and the test covariates $\mathbf{X}_{\text{test}}$, the PI $\hat{C}(\mathbf{X}_{\text{test}})$ for Y_{test} is constructed by refitting the base model $\hat{f}$ (e.g., $\hat{Q}_{1-\alpha}$ or $\hat{\mu}$) on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n \cup \{(\mathbf{X}_{\text{test}}, y)\}$, for a dense grid of values y in the value space of Y. Each fit is denoted by $\hat{f}^y$. The desired nonconformity scores are then defined, for each y, as $S_i(y) := s(\mathbf{X}_i, Y_i; \hat{f}^y)$, $i = 1, \ldots, n$ (Vovk et al., 1999, 2005; Shafer and Vovk, 2008). Consequently, for a full-conformal variant, our proposed procedure from Section 2.3.2 should be performed, for each y, on the scores $\{S_i(y)\}_{i=1}^n$, to obtain a $\hat{q}^e_\alpha(y)$. The resulting PI is then $\hat{C}^e(\mathbf{X}_{\text{test}}) = \{y : s(\mathbf{X}_{\text{test}}, y; \hat{f}^y) \leq \hat{q}^e_\alpha(y)\}$. Although the latter makes a more efficient use of the data than the split variant, it is extremely computationally costly. The analogous extension to the middle-ground k-fold approaches, such as Jackknife+/CV+ (Barber et al., 2021) and cross-conformal prediction (Vovk, 2015), would also suffer, less extremely, from similar refitting expensiveness.

B.2 Proof of Proposition 2.3.1

Proof. Let $[L_Q, U_Q]$ be the $(1-\alpha_2)$-confidence interval for $q := F_S^{-1}(1-\alpha_1)$. Then, by assumption,

$$\mathbb{P}(q \leq U_Q) \geq \mathbb{P}(L_Q \leq q \leq U_Q) \geq 1-\alpha_2, \quad \text{and} \quad \mathbb{P}(S_{\text{test}} \leq q) \geq 1-\alpha_1.$$

As S_{test} and U_Q are independent,

$$\begin{aligned}\mathbb{P}\left(S_{\text{test}} \leq U_Q\right) &= \mathbb{P}\left(S_{\text{test}} - q \leq U_Q - q\right) \geq \mathbb{P}\left(\{S_{\text{test}} \leq q\} \cap \{U_Q \geq q\}\right) \\ &= \mathbb{P}(S_{\text{test}} \leq q)\,\mathbb{P}\left(U_Q \geq q\right) \geq (1-\alpha_1)(1-\alpha_2) \geq 1-\alpha.\end{aligned}$$

By the definitions of $\hat{C}^e(\mathbf{x}) = \{y : s(\mathbf{x}, y) \leq U_Q\}$ and of $S_{\text{test}} = s(\mathbf{X}_{\text{test}}, Y_{\text{test}})$, the events $\{Y_{\text{test}} \in \hat{C}^e(\mathbf{X}_{\text{test}})\}$ and $\{S_{\text{test}} \leq U_Q\}$ are equivalent. Therefore,

$$\mathbb{P}\{Y_{\text{test}} \in \hat{C}^e(\mathbf{X}_{\text{test}})\} = \mathbb{P}\left(S_{\text{test}} \leq U_Q\right) \geq 1-\alpha.$$

□

B.3 Additional figures

B.3.1 Simulation study

Figure B.1 shows the distribution of the computed test coverage for each considered conformalization method, confidence level, and calibration size, for the light-tailed Gaussian-noise data and the ground truth as base predictions. Figure B.2 shows the same test-coverage distribution, for each considered conformalization method, confidence level, and calibration size, for the Student t distributed noise and the linear GPD quantiles as the base predictions.

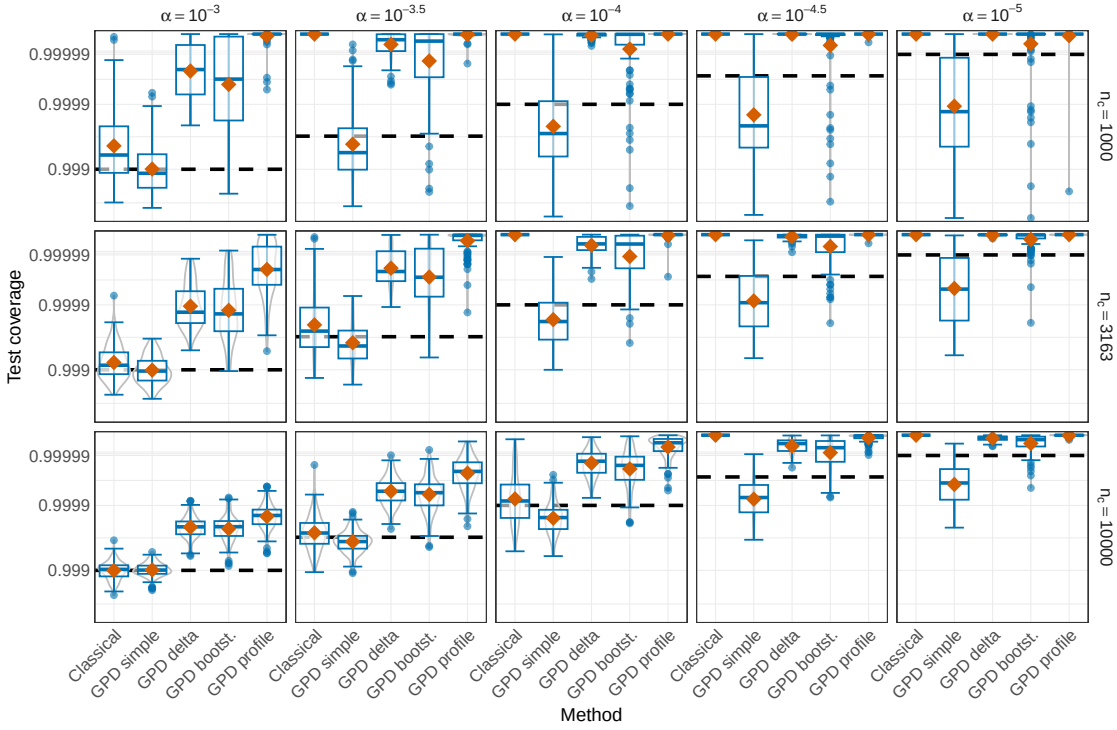

Figure B.1: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over X) for different conformalization methods, conformal confidence values $1-\alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Gaussian distributed noise and quantile ground-truth predictions.

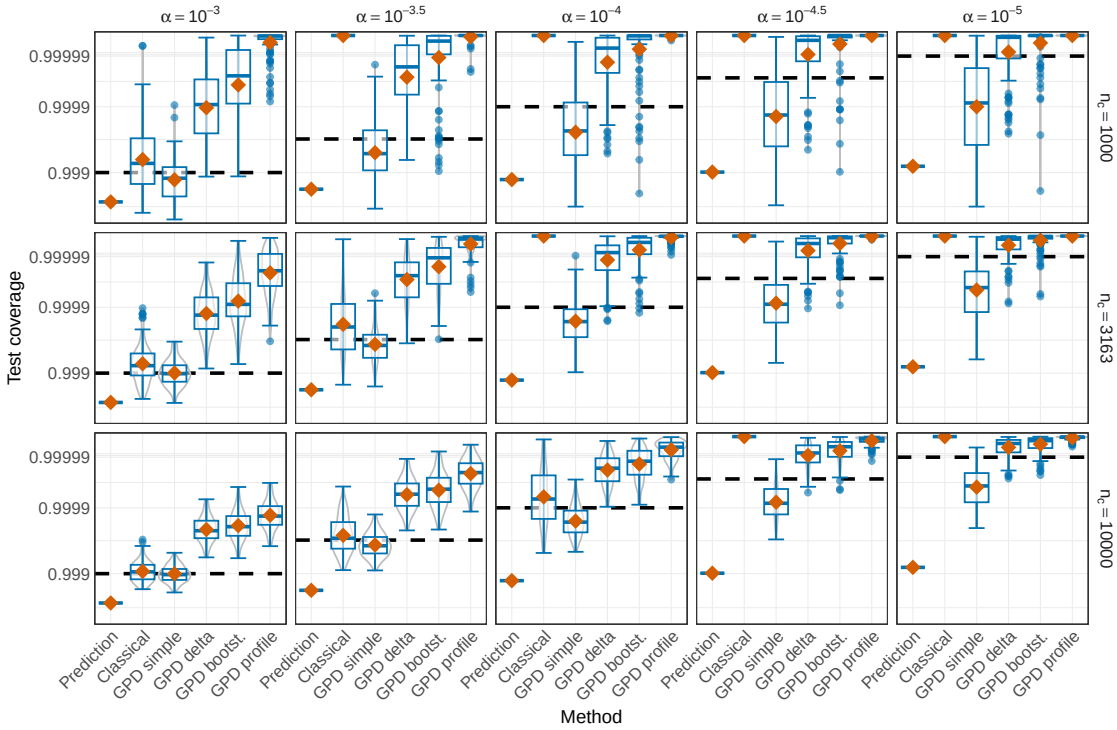

Figure B.2: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over X) for different conformalization methods, conformal confidence values $1-\alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Student t distributed noise and linear GPD quantile predictions.

B.3.2 Application to river-flow forecasts

Figure B.3 shows the locations of the meteorological and gauging stations corresponding to the variables used in the model forecasts.

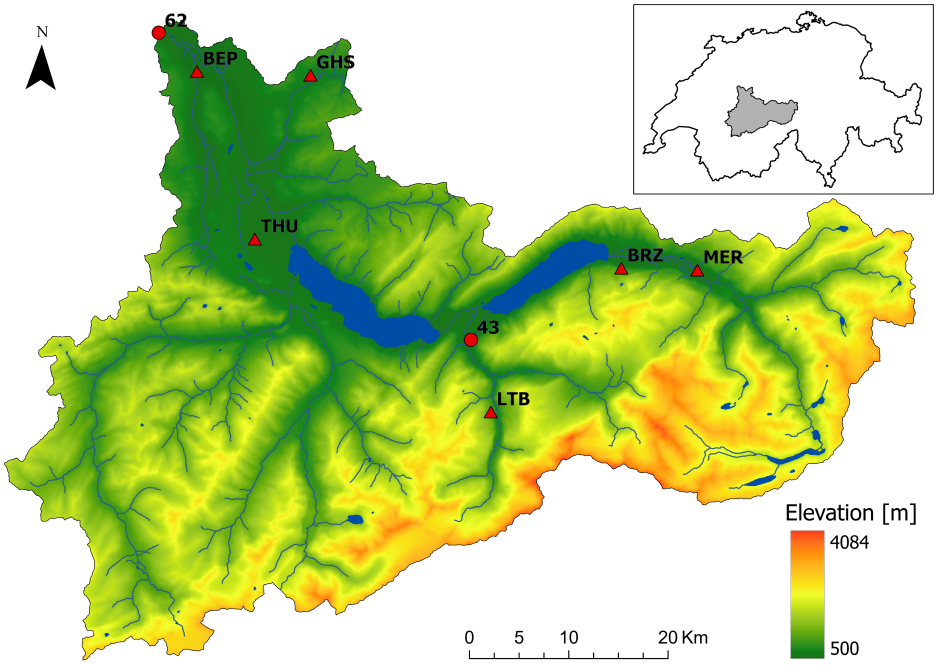

Figure B.3: Topographic map of water catchment of the gauging station in Bern–Schönau (62) on the Aare in Switzerland. Another upstream gauging station in Gsteig (43), on the Lütschine river, and six meteorological stations with precipitation measurements (triangles) are also shown (source: Pasche and Engelke, 2024).

C Supplement to Causal modelling of heavy-tailed variables and confounders with application to river flow

C.1 Variables with comparable tails

C.1.1 Non-parametric causal tail coefficient estimator

Figure C.1 shows the sample distributions of the non-parametric estimators $\hat{\Gamma}_{1,2}$ and $\hat{\Gamma}_{2,1}$ for all four causal structures, for the t_4, Pareto$(1,2)$ and LogN$(0,1)$ noise distributions, respectively. The true coefficient values $\Gamma_{1,2}$ and $\Gamma_{2,1}$ are obtained using (2). Figure C.2 shows the sample distribution of the coefficient difference estimator $\hat{\Delta}_{1,2} := \hat{\Gamma}_{1,2} - \hat{\Gamma}_{2,1}$ for the t_4 noise distribution.

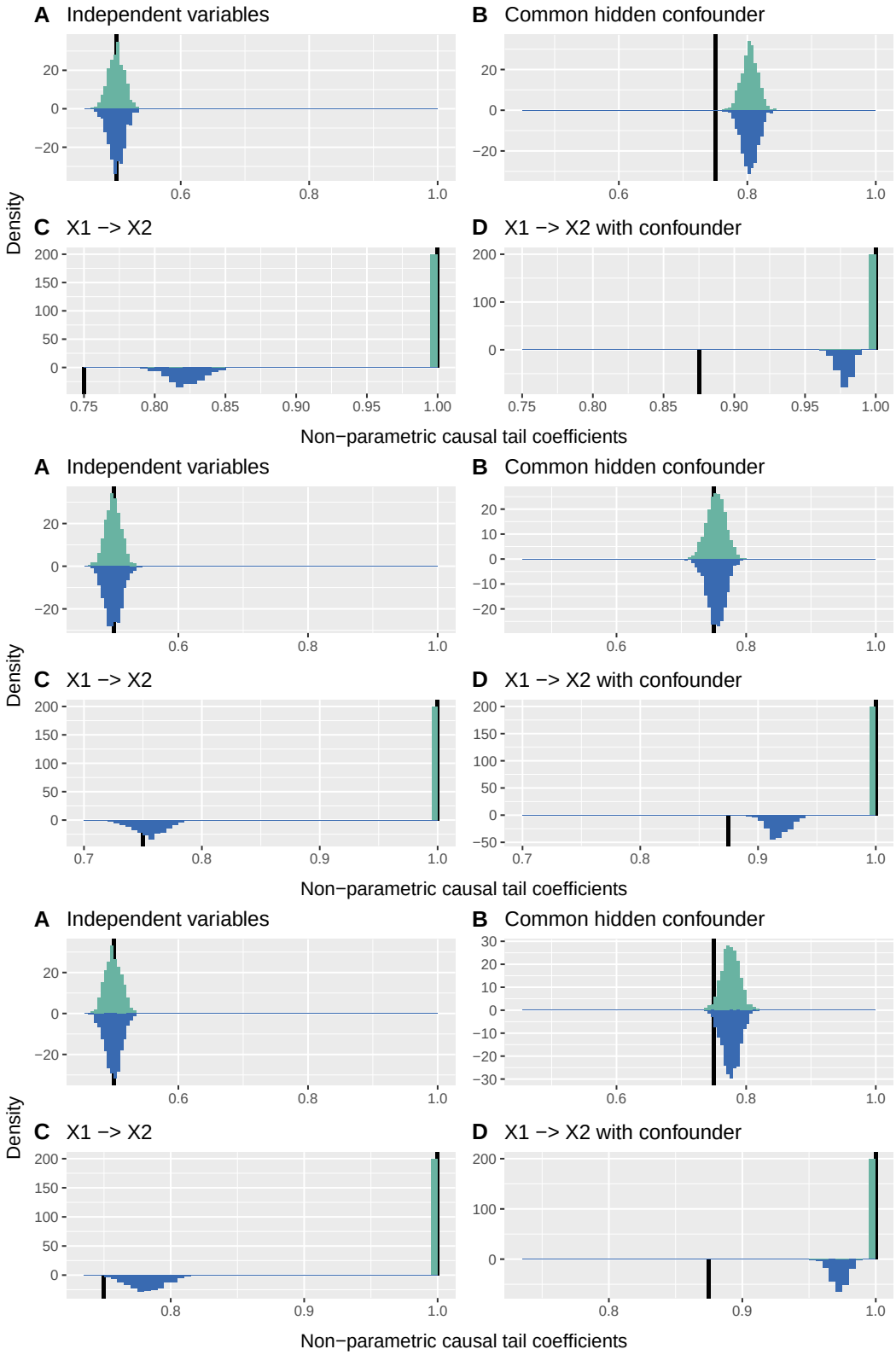

Figure C.1: Histograms of $\hat{\Gamma}_{1,2}$ (turquoise) and $\hat{\Gamma}_{2,1}$ (blue) for t_4 (top four panels), Pareto$(1,2)$ (middle four panels) and LogN$(0,1)$ (bottom four panels) distributed noise variables, for the four causal configurations. Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$.

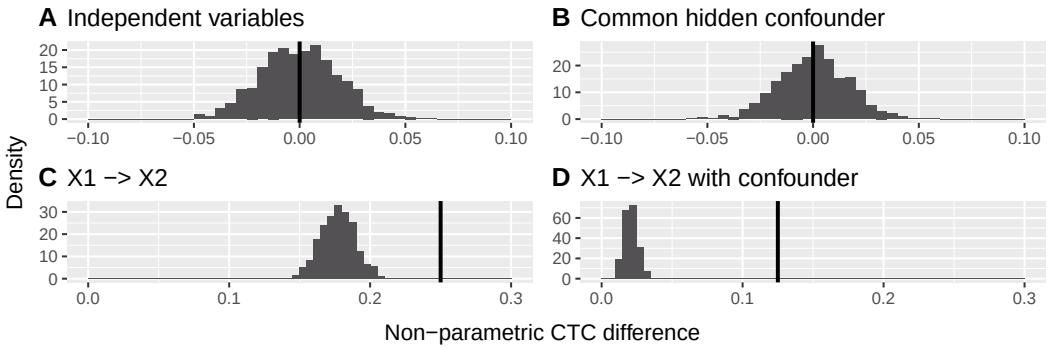

Figure C.2: Histogram of $\hat{\Delta}_{1,2}$ for t_4 distributed noise variables, for the four causal configurations. Lines indicate $\Delta_{1,2} = \Gamma_{1,2} - \Gamma_{2,1}$.

C.1.2 LGPD causal tail coefficient with post-fit and constrained fit corrections

Figure C.3 shows the sample distribution of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ with the constrained fit, for a comparable confounder tail. Figure C.4 shows the sample distribution of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ with post-fit correction for all four causal configurations, for the t_4, Pareto$(1,2)$ and LogN$(0,1)$ noise distributions.

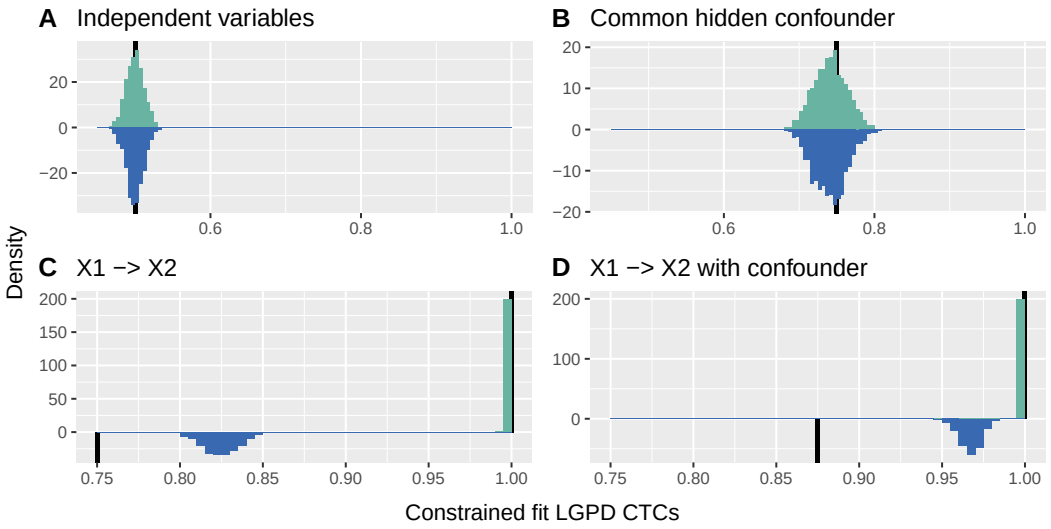

Figure C.3: Histograms of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ (blue) with constrained fit for t_4 distributed noise variables, for the four causal configurations. Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$.

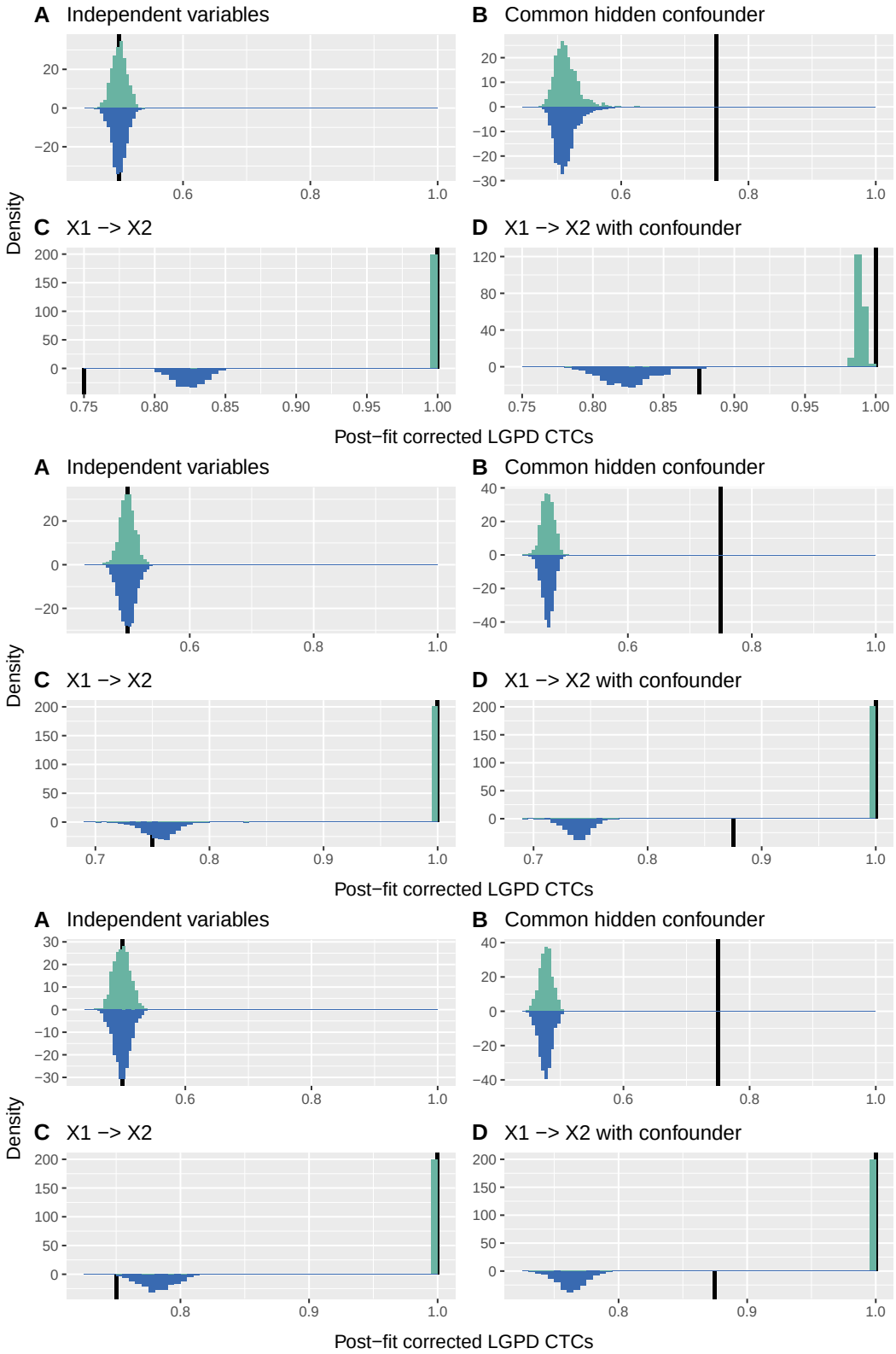

Figure C.4: Histograms of $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ (blue) with post-fit correction for t_4 (top four panels), Pareto$(1,2)$ (middle four panels) and LogN$(0,1)$ (bottom four panels) distributed noise variables, for the four causal configurations. Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$.

C.2 Application results for competitors

Table C.1 shows the causal coefficients between the discharge station pairs estimated using ICA-LiNGAM, with and without considering the average catchment precipitation variable.

Table C.1: Linear causal coefficients for the discharge station pairs estimated with the ICA-LiNGAM algorithm using either the station pair only (LiNGAM, two variables) or the station pair and precipitation (LiNGAM-H, three variables). Non-null values indicate significant causal effects. The arrows indicate the estimated direct causal directions between the stations.

Stations	Pair type	LiNGAM	LiNGAM-H
43-62	causal	$1.92^{\rightarrow}$	$2.02^{\rightarrow}$
42-63	causal	$2.08^{\rightarrow}$	$2.21^{\rightarrow}$
36-63	causal	$3.29^{\rightarrow}$	$3.61^{\rightarrow}$
24-61	causal	$2.96^{\rightarrow}$	$3.03^{\rightarrow}$
44-61	causal	$2.66^{\rightarrow}$	$2.83^{\rightarrow}$
22-38	causal	$2.35^{\rightarrow}$	$2.35^{\rightarrow}$
22-35	causal	$2.55^{\rightarrow}$	$2.55^{\rightarrow}$
30-45	non-caus.	$0.84^{\rightarrow}$	$0.87^{\rightarrow}$
36-39	non-caus.	$0.66^{\leftarrow}$	$0.66^{\leftarrow}$
42-34	non-caus.	$1.39^{\leftarrow}$	$1.29^{\leftarrow}$
42-34*	non-caus.	$1.39^{\leftarrow}$	$1.39^{\leftarrow}$
32-33	non-caus.	$0.59^{\rightarrow}$	$0.54^{\rightarrow}$
62-63	non-caus.	$1.02^{\rightarrow}$	$1.05^{\rightarrow}$
57-60	non-caus.	$0.68^{\rightarrow}$	$0.67^{\rightarrow}$
13-14	non-caus.	$0.50^{\leftarrow}$	$1.10^{\rightarrow}$
17-22	non-caus.	$1.80^{\rightarrow}$	$1.69^{\rightarrow}$
12-21	non-caus.	$1.04^{\rightarrow}$	$1.08^{\rightarrow}$
26-28	non-caus.	$0.75^{\leftarrow}$	$0.72^{\leftarrow}$
27-31	non-caus.	$0.54^{\rightarrow}$	$0.66^{\rightarrow}$
23-39	non-caus.	$0.25^{\rightarrow}$	$0.18^{\rightarrow}$
23-35	non-caus.	$0.42^{\rightarrow}$	$0.36^{\rightarrow}$

D Appendix to Validating deep-learning weather forecast models on recent high-impact extreme events

D.1 Further details and analysis of the 2021 Pacific Northwest heatwave

D.1.1 Computation of the root mean squared error

When computing the root mean squared error (RMSE), we follow Rasp et al. (2024), who choose to include the average over initialization time steps inside the square root of the RMSE formula, in opposition to earlier works in the field (Lam et al., 2023; Rasp et al., 2020). For a given variable x (say 2m temperature T_{2m}), its prediction $\hat{x}$, and a fixed prediction lead time τ, we compute its RMSE(τ) as

$$\mathrm{RMSE}_{\mathcal{T},\mathcal{P}}(\tau) = \sqrt{\frac{1}{|\mathcal{T}|}\frac{1}{|\mathcal{P}|}\sum_{t_0\in\mathcal{T}}\sum_{i\in\mathcal{P}} a_i \left(\hat{x}_i^{t_0+\tau} - x_i^{t_0+\tau}\right)^2}, \tag{D.1}$$

where $\mathcal{T}$ and $\mathcal{P}$ are the sets of initialization times and grid points of interest, $|\mathcal{T}|$ and $|\mathcal{P}|$ are their cardinalities, $\hat{x}_i^{t_0+\tau}$ is the forecast of the variable x at grid cell i and time $t_0+\tau$, with forecast initialized at time t_0, and $x_i^{t_0+\tau}$ is the ground truth for the same grid cell and time step $t_0+\tau$. The weight $a_i = \cos(\mathrm{lat}_i * \pi/180)$ is proportional to the area of the latitude-longitude grid cell i and varies with the latitude of i, lat_i, and $\{a_i, i \in \mathcal{P}\}$ are normalized to have unit mean across the included grid cells. To compute the average RMSE for a given day and lead time τ, we include only initialization times $\tilde{\mathcal{T}} \subseteq \mathcal{T}$ such that $t_0+\tau$ falls on the day of interest for $t_0 \in \tilde{\mathcal{T}}$.

In Fig. 4.2, two contours are determined by the long-term average HRES performance for 120 h (D_5) and 240 h (D_{10}) forecasts. To estimate these values, we use all HRES forecasts provided by WeatherBench 2 (Rasp et al., 2024), which at the time of the writing contain only initializations at 00/12 UTC between January 1, 2016, and January 10, 2023. We use HRES-fc0 as the ground truth and only consider predictions for days within a window size of 45 days around the day-of-year of June 28, 2021. Numerical values for the grid boxes closest to the three investigated cities are provided in Table D.1.

	Vancouver	Seattle	Portland
D_5 (T_{2m}RMSE at day 5)	1.53 K	1.61 K	1.93 K
D_{10} (T_{2m}RMSE at day 10)	2.80 K	2.82 K	3.73 K

Table D.1: Long-term average RMSE values of HRES predictions for lead times of 120 hours (D_5) and 240 hours (D_{10}), computed as described in Section D.1.1.

D.1.2 Additional figures

In this subsection, we show the additional Figs. D.1 to D.3 for analysis of the 2021 Pacific Northwest heatwave.

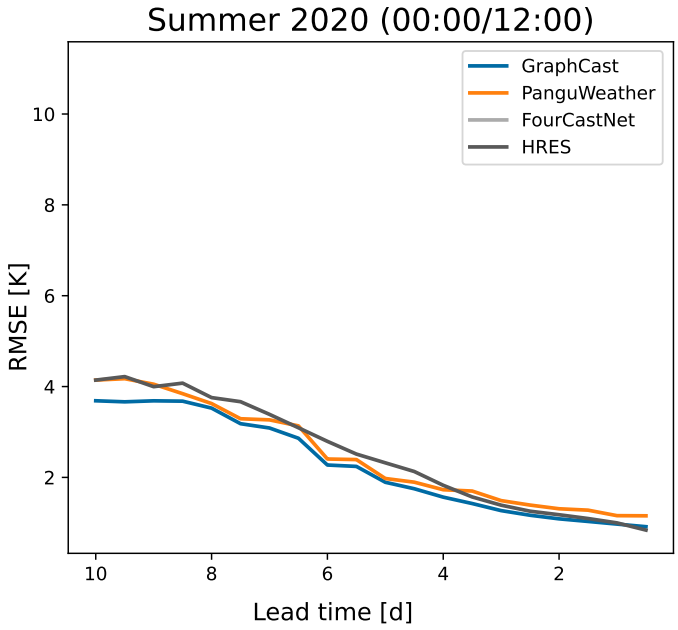

Figure D.1: Evolution of the T_{2m} prediction RMSE with lead-time for the three ML models and HRES in the 2021 Pacific Northwest heatwave region during the summer 2020 (June 1–July 31) as a baseline year. Observations in the considered box region, 45°–52°N, 119°–123°W, are weighted to correct for differences in grid-cell area. Both ML models and HRES use 00:00/12:00 UTC initial conditions and evaluation times only, as opposed to Fig. 4.3, since the forecast were downloaded from WeatherBench 2 for computational reasons, where only these initializations are available.

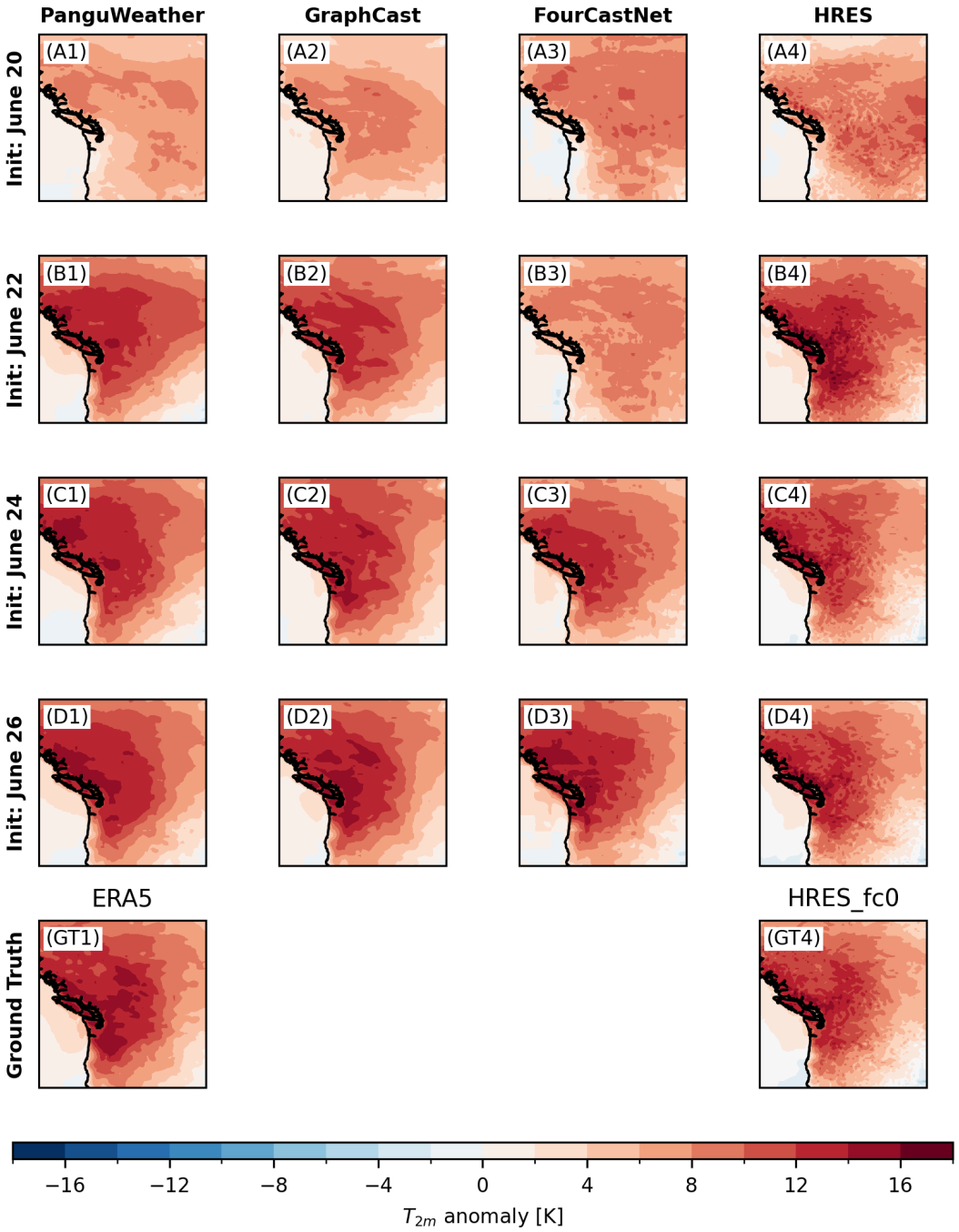

Figure D.2: Average temperature anomaly predicted for June 27–June 29 2021 (inclusive). All anomalies including those for HRES are calculated with respect to the ERA5 climatology given in Rasp et al. (2024). The fact that HRES anomalies are computed against ERA5 data explains the patchy small-scale structure visible in the HRES panels. Forecasts are initialized at 00 UTC on the day specified in the row title.

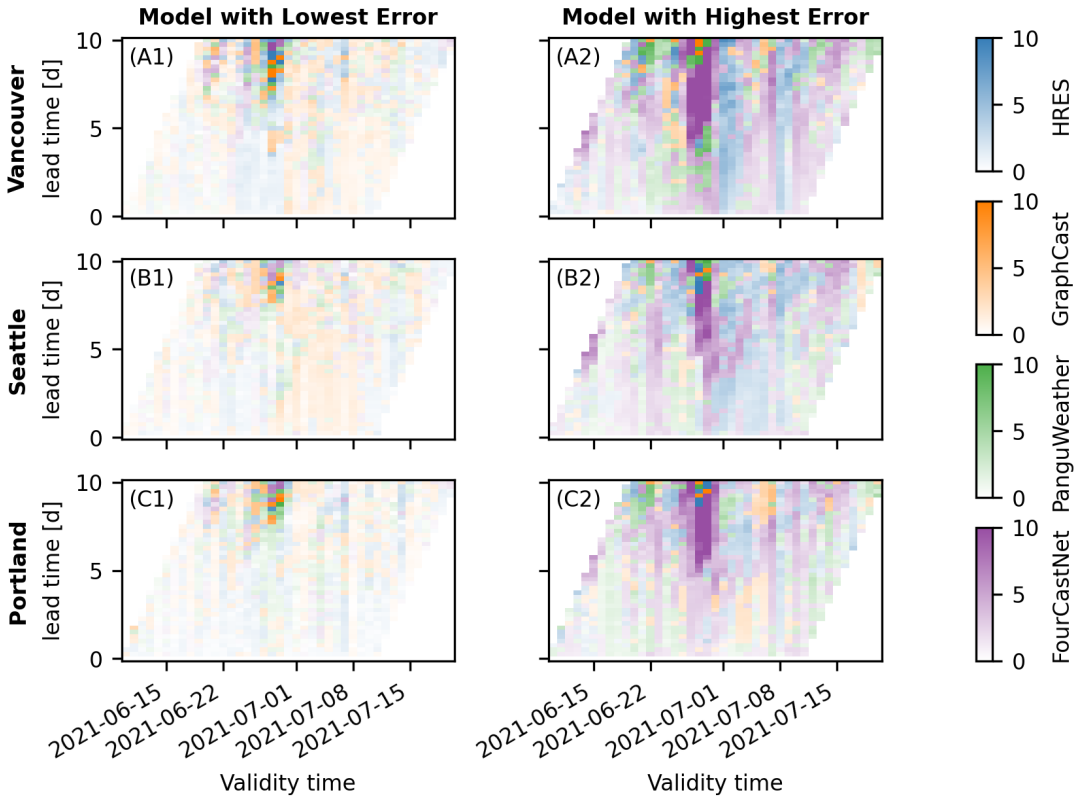

Figure D.3: Predictability barrier plots after taking the argmin (left) and argmax (right) over the RMSE (in K) of the different models. This means that the choice of colorbar in each pixel of the left panel indicates which model had the lowest RMSE for the given lead time and validity date, in the right panel it shows which model had the largest RMSE.

D.2 Further details and analysis of the 2023 South Asian humid heatwave

D.2.1 Shape files

In Section 4.3.2, we use shape files to subset the study regions of the South Asian humid heatwave. We use country boundaries from the "World Administrative Boundaries - Countries and Territories" (Open Government License 3.0, `https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/information/`) data set by the World Food Programme and for the India-Bangladesh region we additionally use the (1976-2000) map of "World Maps of the Köppen-Geiger Climate Classification" (Creative Commons Attribution 4.0, `https://datacatalog.worldbank.org/search/dataset/0042325`). We only include grid cells with Köppen-Geiger Class A ("tropical") in the India-Bangladesh region.

D.2.2 Relative humidity

Relative humidity is required to compute the heat index, but PanguWeather (Bi et al., 2023) and GraphCast (Lam et al., 2023) only produce specific humidity, and only on atmospheric pressure levels, not at the surface. Therefore we first need to compute relative humidity from specific humidity. We exclude FourCastNet (Pathak et al., 2022) from the case study in Section 4.3.2 because here humidity is only available at pressure levels higher than 850 hPa.

To convert specific humidity to relative humidity from the machine learning models, we first compute the saturation vapor pressure at the respective pressure level using the August–Roche–Magnus formula:

$$e_s(T) = 6.1094\,\text{hPa} \exp\left(\frac{17.625T}{T+243.04}\right) \tag{D.2}$$

where T is the temperature at the respective pressure level in °C. We compute the mixing ratio at saturation r_s (dimensionless) from pressure and saturation vapor pressure:

$$r_s = 0.622 \frac{e_s}{p - e_s} \tag{D.3}$$

where p is the pressure defining the pressure level. We then compute relative humidity RH from the mixing ratio at saturation and the specific humidity q:

$$RH = \frac{q}{(1-q)r_s}, \tag{D.4}$$

where specific humidity is given in $\text{g}\,\text{kg}^{-1}$, and relative humidity is given in percent throughout the rest of the study. We emphasize again that RH is not the relative humidity at the surface, which is not available for the ML models, but rather the humidity at pressure levels.

For HRES and ERA5, it is additionally possible to compute relative humidity from $2m$ temperature

T_{2m} and $2m$ dewpoint temperature T_d:

$$RH = \frac{e_s(T_d)}{e_s(T_{2m})}, \tag{D.5}$$

where e_s was computed using Eq. (D.2) with T_{2m} and T_d given in degrees Celsius.

D.2.3 Heat index

As described in Section 4.3.2, we follow Zachariah et al. (2023) during the computation of the heat index in using a modified version of the heat index (Rothfusz and Headquarters, 1990) used by NOAA Weather Prediction Center (WPC). The NOAA WPC formulation is accessible at `https://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml` (last accessed April 5 2024, page last modified May 12 2022 19:37:55 UTC).

In the following, the heat index formula is given for temperature in °F. We transform the temperature input to °F and the heat index output to °C in the study.

To obtain the heat index HI, a simplified version HI_s is computed first:

$$HI_s = 0.5(T_{2m} + 61 + 1.2(T_{2m} - 68) + 0.094RH) \tag{D.6}$$

where T_{2m} is the $2m$ temperature in °F, and RH is the relative humidity given as a percent value between 0 and 100. If $(HI_s + T_{2m})/2$ is smaller than or equal to 80 °F, the computation is complete, and the heat index is used as computed in Eq. (D.6). Otherwise, the heat index is recomputed using a more elaborate formula:

$$\begin{aligned} HI = &-42.379 + 2.04901523T_{2m} + 10.14333127RH - 0.22475541T_{2m}RH \\ &- 0.00683783T_{2m}^2 - 0.05481717RH^2 + 0.00122874T_{2m}^2RH \\ &+ 0.00085282T_{2m}RH^2 - 0.00000199T_{2m}^2RH^2 \end{aligned} \tag{D.7}$$

If $RH < 13\,\%$ and $80\,°\text{F} < T_{2m} < 112\,°\text{F}$, the heat index computed according to Eq. (D.7) is corrected:

$$HI = HI - \frac{13 - RH}{4}\sqrt{(17 - |T_{2m} - 95|)/17} \tag{D.8}$$

A different correction is applied if $RH > 85\,\%$ and $80\,°\text{F} < T_{2m} < 87\,°\text{F}$:

$$HI = HI + \frac{(RH - 85)}{10}\frac{87 - T_{2m}}{5} \tag{D.9}$$

We always use the 2 m temperature T_{2m} in the computation of the heat index. If the relative humidity is not available at the surface level, we specify in the text, how we substitute the value at the surface.

HI values are classified as in Zachariah et al. (2023), the classes are listed with potential health consequences in Table D.2.

Table D.2: Heat index categories, adapted from (Blazejczyk et al., 2012)

Category	Definition	Possible heat disorders for people in high-risk groups
Low risk	$HI < 27\,^{\circ}\mathrm{C}$	Fatigue possible with prolonged exposure and/or physical activity
Caution	$27\,^{\circ}\mathrm{C} \leq HI < 32\,^{\circ}\mathrm{C}$	Sunstroke, muscle cramps, and/or heat exhaustion possible with prolonged exposure and/or physical activity
Extreme caution	$32\,^{\circ}\mathrm{C} \leq HI < 41\,^{\circ}\mathrm{C}$	Sunstroke, muscle cramps, and/or heat exhaustion possible with prolonged exposure and/or physical activity
Danger	$41\,^{\circ}\mathrm{C} \leq HI < 54\,^{\circ}\mathrm{C}$	Sunstroke, muscle cramps, and/or heat exhaustion likely. Heatstroke possible with prolonged exposure and/or physical activity
Extreme danger	$54\,^{\circ}\mathrm{C} \leq HI$	Heat stroke or sunstroke likely

D.2.4 Additional figures

In this subsection, we show the additional Figs. D.4 to D.6 for our analysis of the 2023 South Asian humid heatwave.

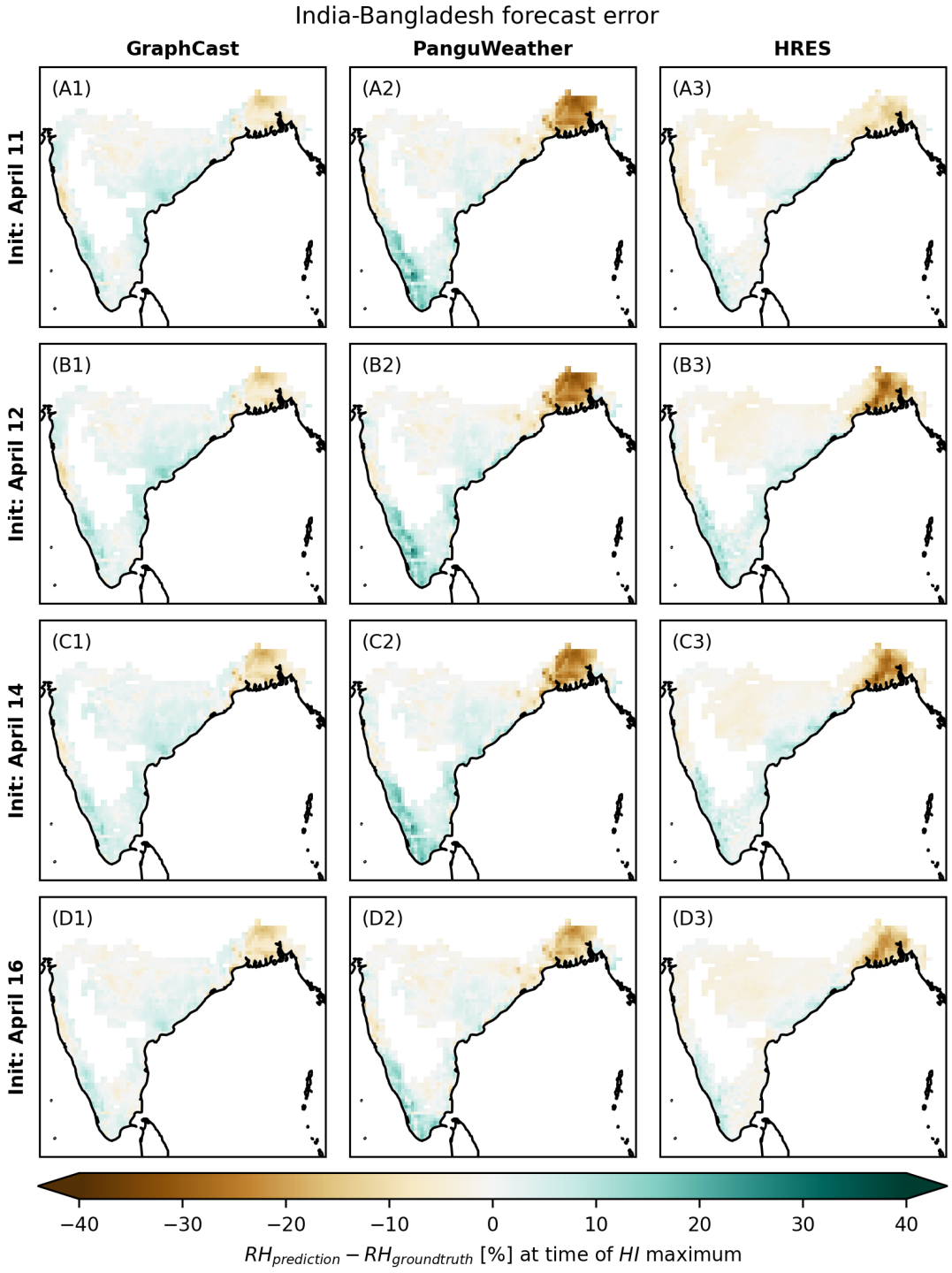

Figure D.4: Forecast error for $RH_{1000\,\mathrm{hPa}}$ at the time step of each day when observed HI peaked in the corresponding ground truth data set, averaged over April 17–April 20.

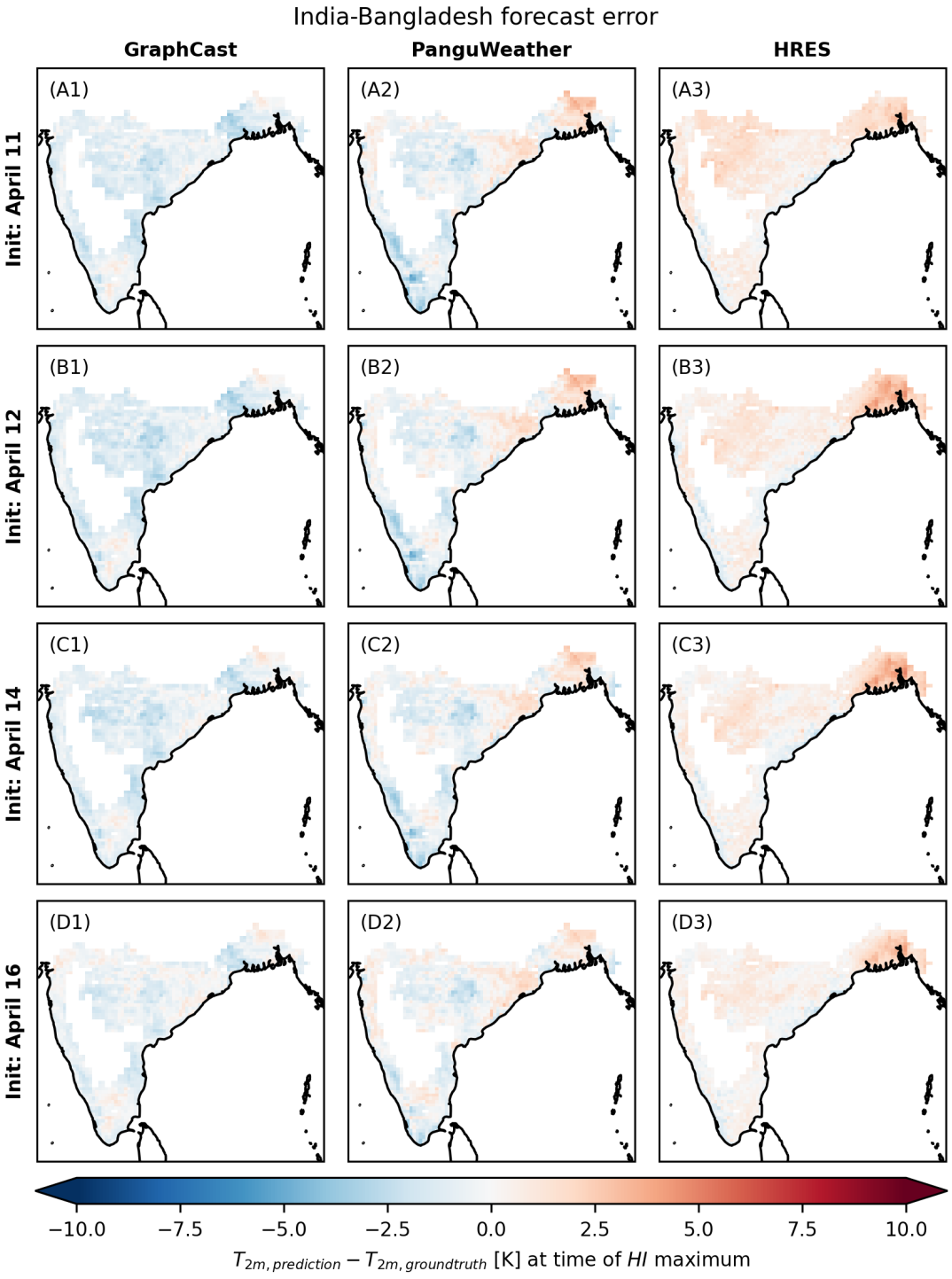

Figure D.5: Forecast error for T_{2m} at the time step of each day when observed HI peaked in the corresponding ground truth data set, averaged over April 17–April 20.

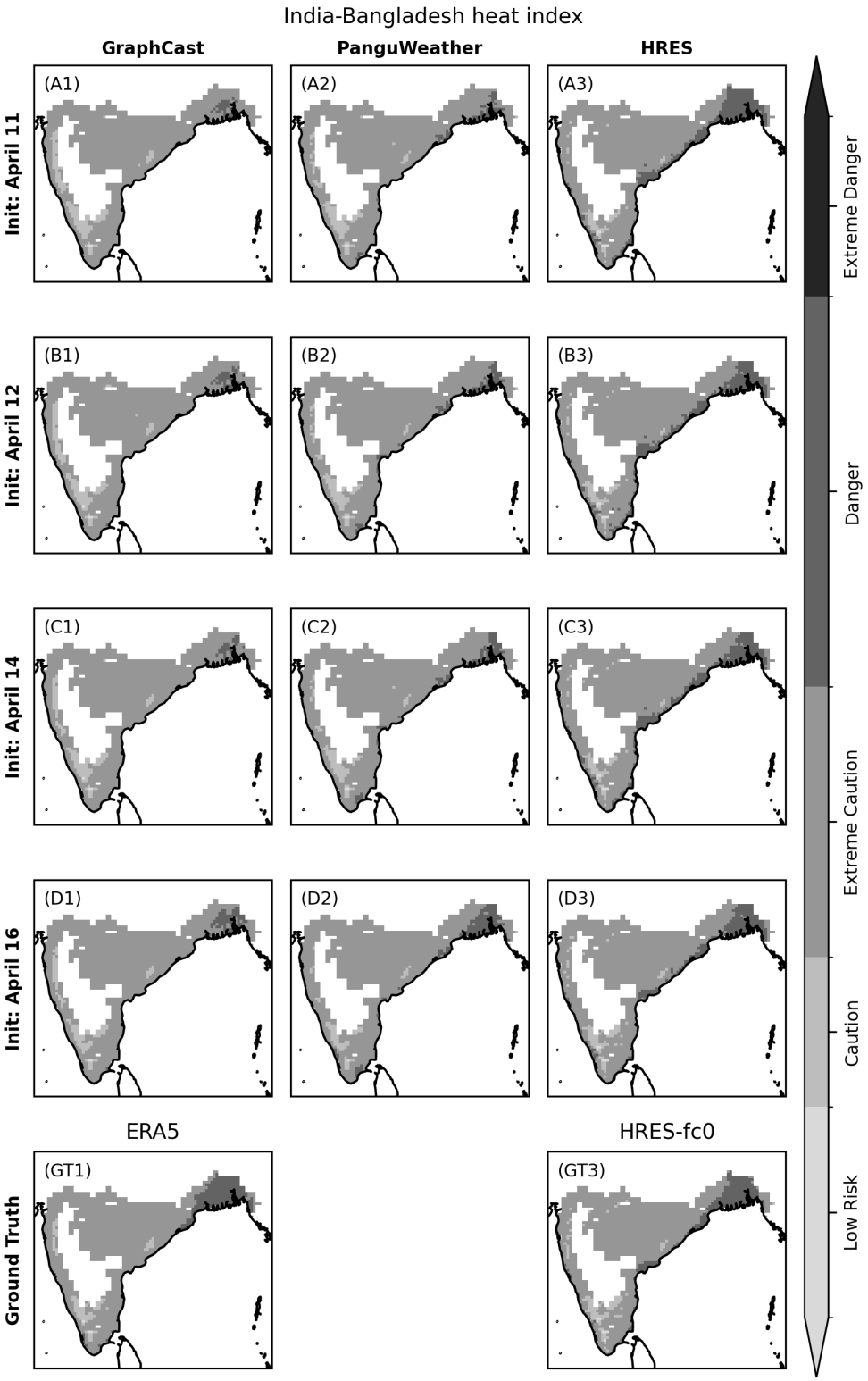

Figure D.6: Category of mean daily maximum HI (see Table D.2) predicted for April 17–April 20 with different models and for varying initialization times computed using $RH_{1000\,\mathrm{hPa}}$. Forecasts are started at 00:00 UTC on the specified initial date. Panels (GT1), (GT3): categories in ground truth data sets.For ML models, HI is computed using $RH_{1000\,\mathrm{hPa}}$, while for all other panels we use RH_{sfc}.

D.3 Further analysis of the 2021 North American winter storm

D.3.1 Additional figures

In this subsection, we show the additional Figs. D.7 to D.9 for our analysis of the 2021 North American winter storm.

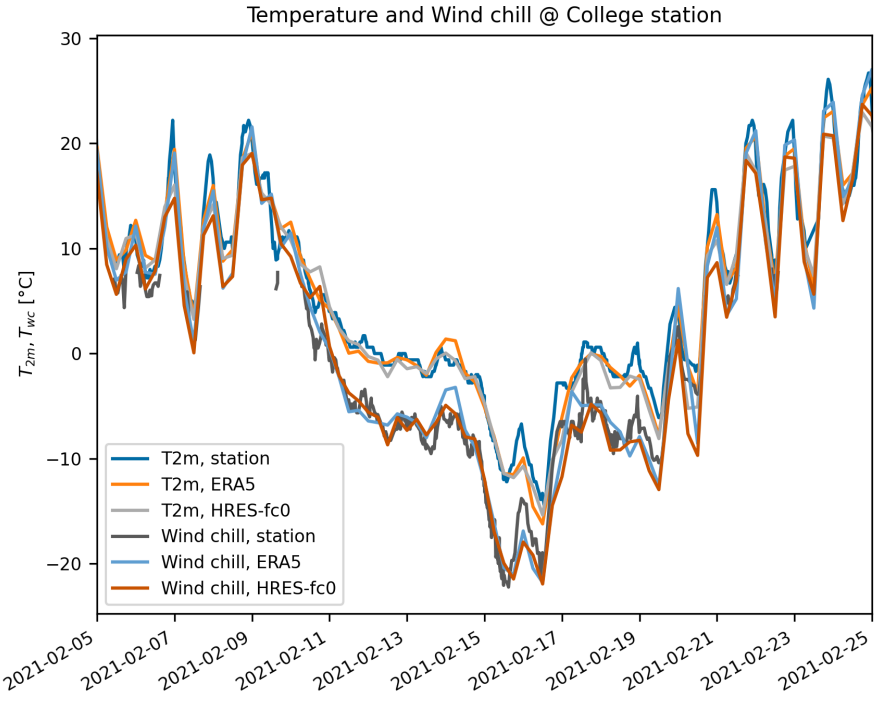

Figure D.7: ERA5 and HRES-fc0 T_{2m} and wind chill index T_{wc} time series in the grid box closest to College Station, Texas. Weather station data: Easterwood Field, College Station, Texas. Data retrieved from Integrated Surface Dataset (ISD) (Smith et al., 2011). Figure ignores thresholds in the definition of T_{wc} for better readability.

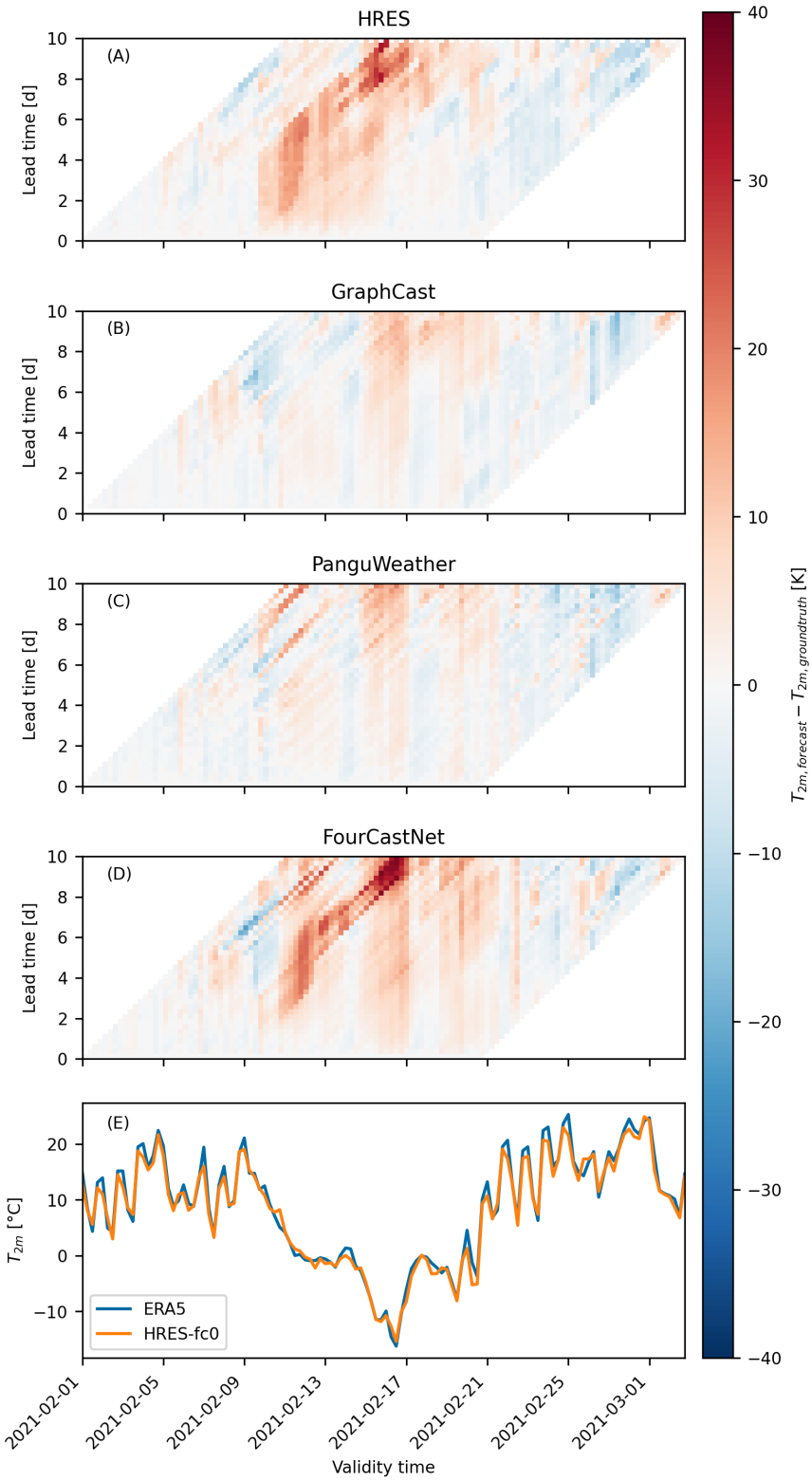

Figure D.8: Panels (A) to (D): T_{2m} prediction errors for different validity and lead times in the grid cell closest to College Station, Texas (UTC time used). Panel (E): ground truth T_{2m} time series in the same grid cell.

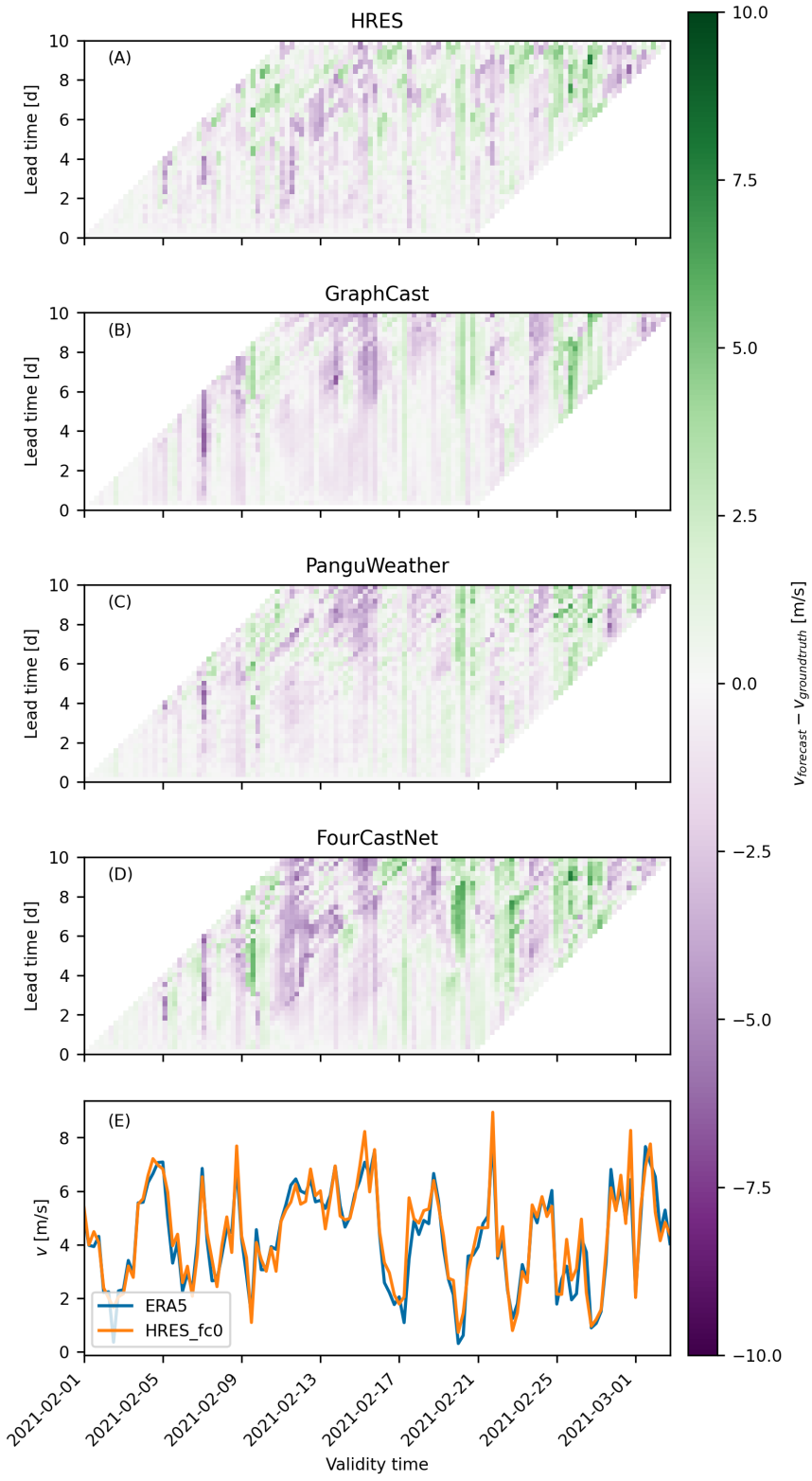

Figure D.9: Panels (A) to (D): Prediction errors for surface wind speed for different validity and lead times in the grid cell closest to College Station, Texas (UTC time used). Panel (E): ground truth wind speed time series in the same grid cell

E Supplement to Validating deep-learning weather forecast models on recent high-impact extreme events

E.1 Further details on ML models

In Table E.1 we list the surface and pressure-level variables predicted by the ML models.

Table E.1: Detailed list of variables predicted by ML models. T_{2m}: 2 m temperature, U_{10}, V_{10}: horizontal components of 10 m wind, MSL: mean sea-level pressure, TP: total precipitation, SP: surface pressure, Z: geopotential, T: Temperature, RH: relative humidity, q: specific humidity, U, V: horizontal components of wind, W: vertical component of wind

	FourCastNet	PanguWeather	GraphCast
Surface variables	T_{2m} U_{10} V_{10} MSL (TP) SP	T_{2m} U_{10} V_{10} MSL	T_{2m} U_{10} V_{10} MSL TP
Atmospheric variables	Z T RH U V	Z T q U V	Z T q U V W
Levels [hPa]	50, 500, 850, 1000	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000	1, 2, 3, 5, 7, 10, 20, 30, 50, 70, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000

E.2 Further analysis of the 2021 Pacific Northwest heatwave

In Fig. E.1, we show an alternative version of Figure 2 in the main text. Here, we don't aggregate to daily scale and instead show the residuals $T_{2m,prediction} - T_{2m,groundtruth}$. One can see patterns of the daily cycle in the predictions as well as negative forecast errors during the peak of the heatwave and positive forecast errors after the peak of the heatwave in some cases.

To further analyze the spatial aspect of the predictions, we first compute T_{2m} anomalies. We use the climatology provided by WeatherBench 2 (Rasp et al., 2024), which is computed from ERA5 data between 1990 and 2019 (inclusive) and uses a sliding window of 61 d around the day of interest. For simplicity, we also use the ERA5 climatology in the computation of the HRES and HRES-fc0 anomalies, acknowledging that there are differences in the climatologies of ERA5 and HRES-fc0. In panels (A1) to (A4) of Fig. E.2, we show in black the contour of the area in which the calculated true temperature anomaly, averaged between June 27 to 29 (inclusive), exceeds 12 K.

We then look at forecasts initialized from 7 d to 6 h (with a time step of 6 h) before June 27, 00 UTC. For each grid cell, we determine the percentage of the forecasts that predict the average temperature anomaly to exceed 12 K and plot this in panels (A1) to (A4) of Fig. E.2. An ideal prediction system would always predict an anomaly of ≥ 12 K everywhere within the black contour, and an anomaly < 12 K everywhere outside the contour.

Overall, the shapes of the predicted anomalies follow the shape of the true event well, i.e. heat was predicted in areas where it later actually occurred, and no large anomalies were predicted in regions where no large anomalies occurred. For FourCastNet, the anomaly is < 12 K everywhere for a large fraction of the forecasts - the yellow colors within the contour are faint. For GraphCast and HRES, the predicted area matches the contour well, meaning that a large fraction of the forecasts captured the right spatial distribution. For PanguWeather, more forecasts reach large positive anomalies than for GraphCast and HRES, this is reflected by brighter yellow colors. At the same time, PanguWeather is the only model whose forecasts "spill out" of the true 12 K contour notably, i.e. for PanguWeather some of the forecasts predicted anomalies ≥ 12 K in regions where they didn't occur in the ground truth data sets.

Next, we compute the prediction error of the June 27 to 29 temperature anomaly for each grid cell and then take the (area-weighted) average over all pixels within the area in which the true anomaly exceeded 12 K over these days. The results for the different models are shown in panel (B) of Fig. E.2. FourCastNet strongly under-predicts the temperature anomaly within the contour for initialization times before June 24, while the other models achieve better predictions of the magnitude of the heatwave sooner. PanguWeather's forecasts initialized on June 20 and June 21 predict notably warmer temperatures than HRES and GraphCast. PanguWeather and GraphCast slightly under-predict the size of the anomaly until the start of the event, while for HRES the under-prediction is smaller. One needs to keep in mind that the ERA5 ground truth and the HRES-fc0 ground truth don't coincide exactly, and thus it is not clear whether one can attribute this slight under-predictions of GraphCast and PanguWeather to model deficiencies.

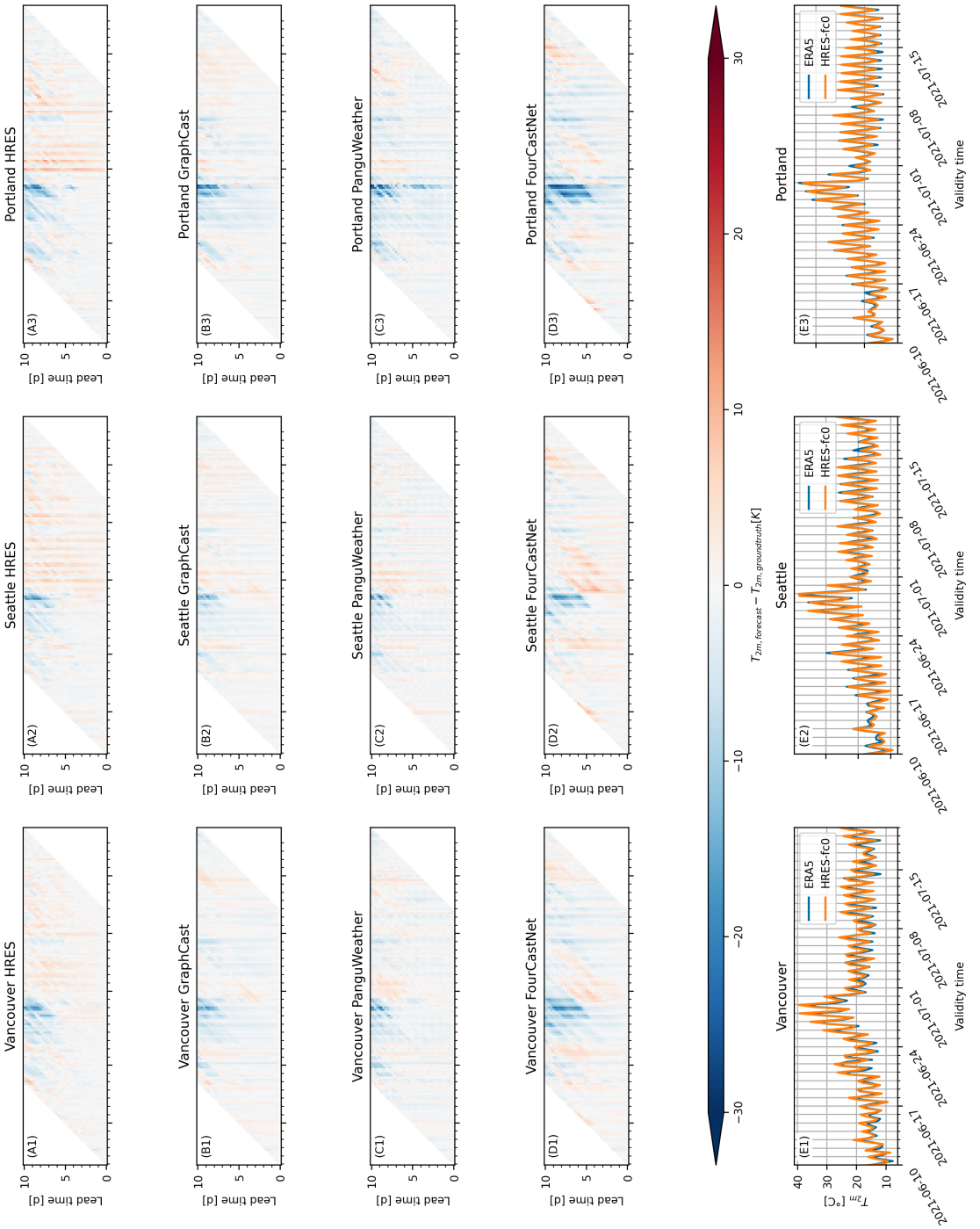

Figure E.1: Panels (A1) to (D3): Barrier plots of forecast residuals for the grid cells closest to major cities affected by the 2021 heatwave. For HRES, HRES-fc0 is used as ground truth, for the ML models, we use ERA5 instead. Panels (E1) to (E3): time series of daily maximum T_{2m} for the data sets used as ground truth.

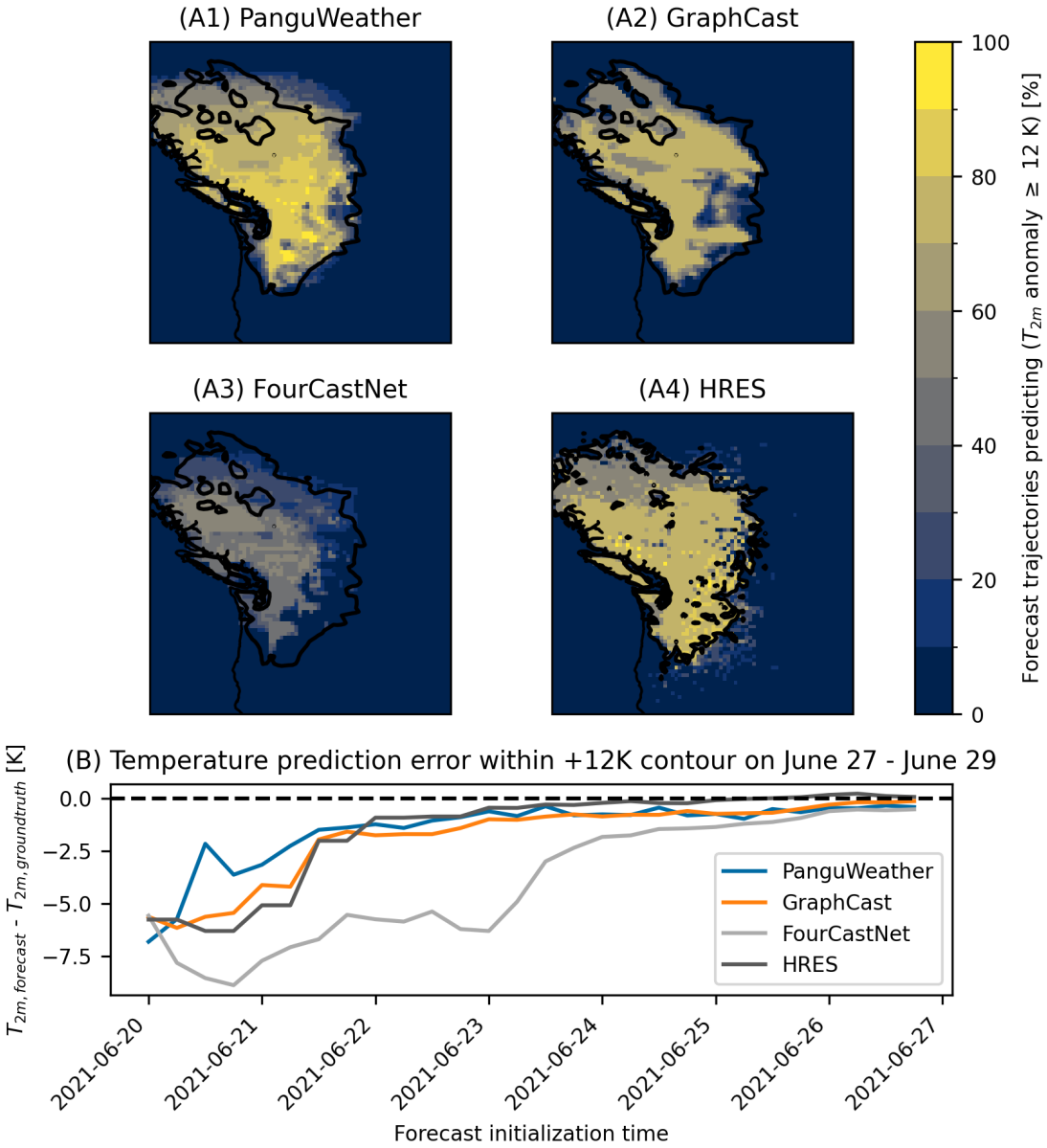

Figure E.2: Panels (A1) to (A4): Black contour delimits the region in which the ground-truth average T_{2m} anomaly between June 27 and June 29 exceeded 12 K. The colormap indicates what fraction of forecasts started before the event predicted an anomaly exceeding 12 K in the given grid box. Panel (B): error in the prediction of the June 27 to June 29 T_{2m} anomaly, area-weighted average over the area within the 12 K contours of panels (A1) to (A4).

E.3 Further analysis of the 2023 South Asian humid heatwave

E.3.1 Humid heatwave in the Laos-Thailand region

In section 3.2 of the main paper, we analyzed how well the 2023 South Asia humid heatwave is predicted in the India-Bangladesh study region. Conclusions for the Laos-Thailand study region are similar. Here the analyzed heatwave peak is April 18–April 21 2023.

In Fig. E.3, again the ground truth ERA5 and HRES-fc0 data sets don't yield the same HI values. HRES follows its ground truth well, while the ML prediction models under-predict the high HI values when compared to their ERA5 ground truth computed using $RH_{1000\,\mathrm{hPa}}$ and the version computed with RH_{sfc}.

An under-prediction of HI can also be seen in Fig. E.4, where the HI forecast errors for the different models are compared. All computations for this plot use $RH_{1000\,\mathrm{hPa}}$ rather than RH_{sfc}. There seems to be an under-prediction of HI by the ML methods, while HRES values tend to be slightly larger than the HRES-fc0 ground truth.

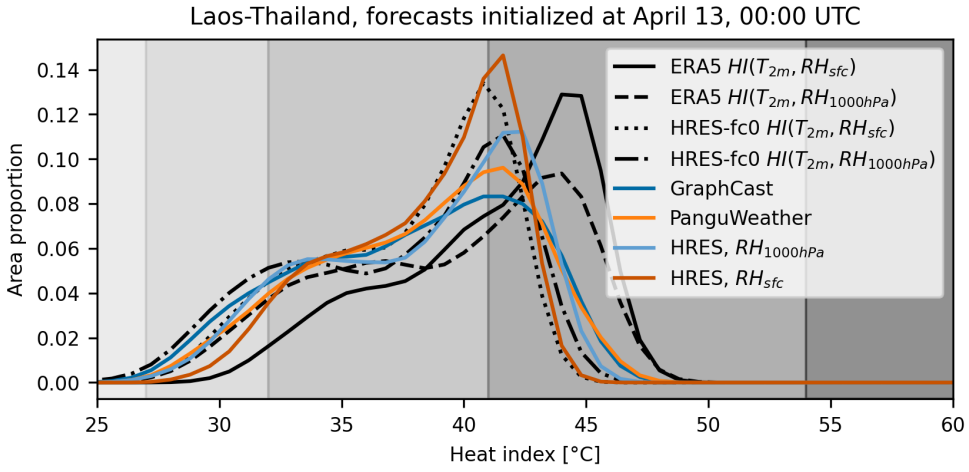

Figure E.3: Proportion of area in study region with given mean daily maximum heat index during April 18—21, computed using area-weighted kernel density estimation. Shaded areas in the background indicate threat levels (Blazejczyk et al., 2012). From light gray to dark gray: low risk, caution, extreme caution, danger, extreme danger. Compared are distributions resulting from forecasts initialized 6 days prior to the start of the event and different ground truths: ERA5 and HRES-fc0, each in two versions of computing the heat index, either using RH_{sfc} or using the substitute $RH_{1000\,\mathrm{hPa}}$. For HRES forecasts, we show versions computed with $RH_{1000\,\mathrm{hPa}}$ and RH_{sfc} as well.

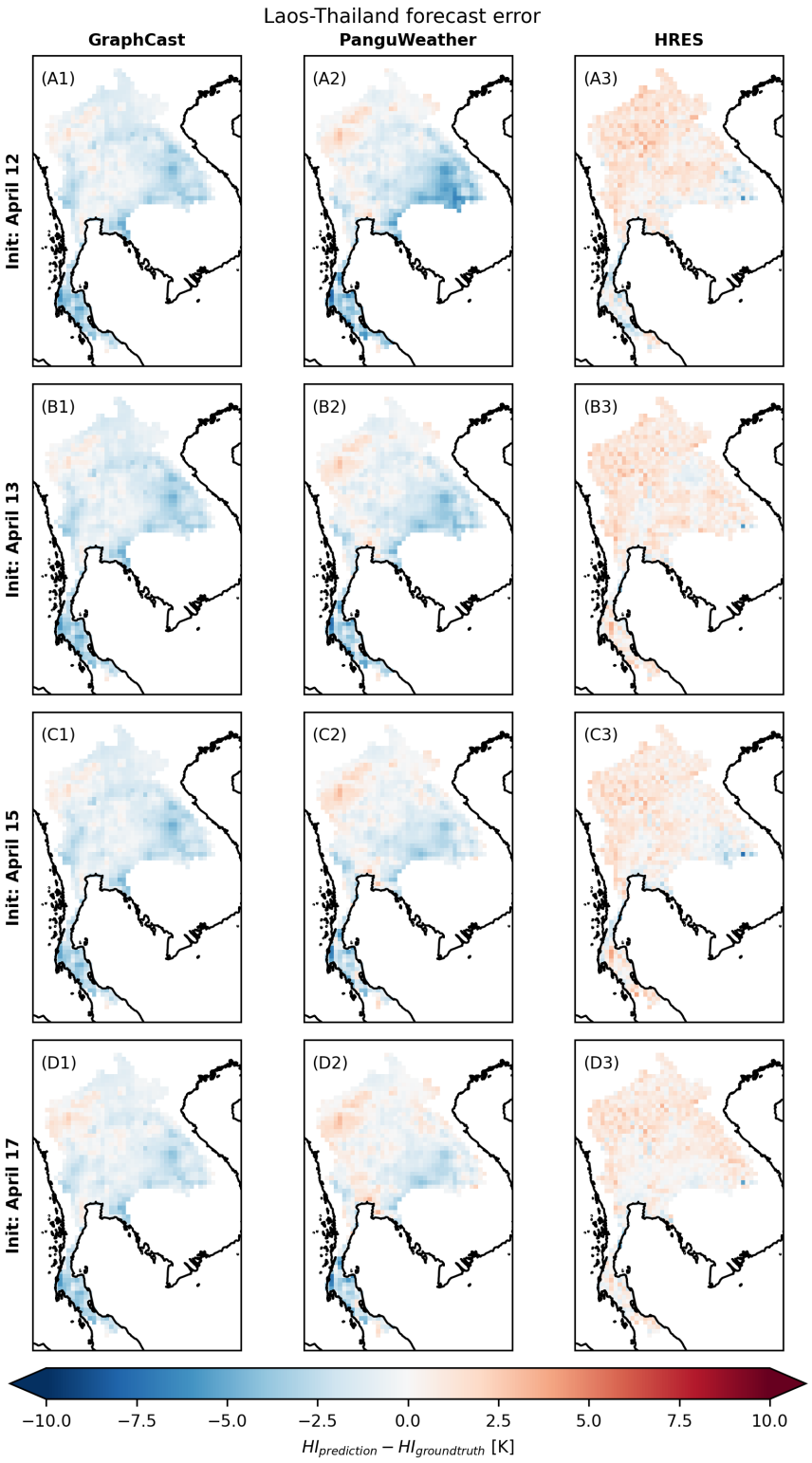

Figure E.4: Error of the HI prediction, for the time step of each day during which HI peaked in the ground truth data set, averaged over April 18–21. For all forecasting methods and ground truth data sets, HI is computed using $RH_{1000\,\mathrm{hPa}}$ rather than the value at the surface.

Links

Personal website: http://www.opasche.com
Google Scholar: https://scholar.google.com/citations?user=GRk09VQAAAAJ
GitHub: https://github.com/opasche
ORCID: http://orcid.org/0000-0002-1202-9199
ResearchGate: https://www.researchgate.net/profile/Olivier-Pasche-2
ArXiv: https://arxiv.org/a/pasche_o_1.html
LinkedIn: https://www.linkedin.com/in/olivier-pasche

UNIVERSITÉ
DE GENÈVE
GENEVA SCHOOL OF ECONOMICS
AND MANAGEMENT

LE DOYEN

A QUI DE DROIT

I M P R I M A T U R

Je, soussigné, Professeur Salvatore DI FALCO, Doyen de la Faculté d'Economie et de Management, confirme que Monsieur Olivier Colin PASCHE obtient l'imprimatur pour sa thèse N°163, suite à sa soutenance publique du 20 février 2026 pour le grade de docteur en Statistique.

Prof. Salvatore DI FALCO
Doyen

Genève, le 20 février 2026
SDF/GK/mc