Thèse n° 11 082

Wind, Hail, and Climate Extremes: Modelling and Attribution Studies for Environmental Data

Présentée le 29 septembre 2023

Faculté des sciences de base
Chaire de statistique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Ophélia Mireille Anna MIRALLES

Acceptée sur proposition du jury

Prof. F. Nobile, président du jury
Prof. A. C. Davison, Prof. V. Panaretos, directeurs de thèse
Prof. V. Chavez-Demoulin, rapporteuse
Prof. E. Fischer, rapporteur
Prof. M. Lehning, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

Are we going to look our [grand]children in the eye and tell them that we understood the issues, that we recognized the dangers and the opportunities, and still we failed to act?

— Nicholas Stern

To my beloved husband and children . . .

Acknowledgements

I am grateful to my Ph.D. supervisor for letting me define my objectives and path during those three years. Thank you for the very meaningful insights and feedback, for introducing me to key relationships in climate science, and for creating a climate of trust and bilateral exchanges. Thank you for being supportive during both of my pregnancies and for giving me the time to recover and get back to work.

Regarding the conciliation of maternity and a Ph.D., despite a shockingly high number of online testimonies describing the experience as a nightmare, I was lucky to be actively supported by my institution through reasonably lengthy maternity leave, special spots in a nearby nursery, financial help with childcare, on-demand reduced working rate and other perks. Thank you to EPFL for making it as easy as possible for (female) Ph.D. students to have a family life. Special thanks to Darlene Goldstein for being supportive and understanding.

I was part of interdisciplinary projects during my Ph.D. and would like to thank the people from the climate science community I was lucky to work with for their insightful discussions and interest in my work: Daniel Steinfeld, Olivia Martius, and Daniele Nerini. Thanks to the Federal Office of Meteorology and Climatology MeteoSwiss for being the main data provider for this Ph.D., and in particular thanks to the MeteoSwiss postprocessing and verification team for trusting me enough to offer me a postdoctoral position after my Ph.D. and for giving me the opportunity to make my research useful in a practical way for everyone in Switzerland.

As a young mother, I did not have much time to hang out on campus but still want to thank the people in my past and present group for the — unfortunately rare — cheerful moments we shared: thank you Sonia, Mario, Jonathan, Marcelo, and Stefano.

Figuring out the direction I wanted to take with this thesis, and my career in general, was paved with more or less successful experiences. I wish to say thank you to some specific persons met along the way. Thank you to Fadhel Ben Atig, my colleague from

Goldman Sachs, for making me feel that I was not the only one in my workplace caring about the climate crisis. Thank you to Johannes Muhle-Karbe, professor of mathematical finance at Imperial College, for giving me a chance to start a Ph.D. in financial mathematics, even though it was not pursued. Thank you to Anna Mauvernay, Thiago Vargos, and Dener Giovanini, my team at RENCTAS (renctas.org.br), for your warm welcome and initiation to the fight against wildlife trafficking. Thanks also to Tanya Wyatt and Francis Masse from Northumbria University for giving me the opportunity to work on and publish a study on online wildlife trafficking (Wyatt et al., 2022). Thank you to Michael Vincent from the NGO Greentervention (greentervention.org) for your implication in making our institutions more resilient to climate change and thank you for letting me contribute to your purpose by writing on the topic of carbon accounting (Miralles, 2021). Last but not least, thank you to Marie Schellens and Albert Martinez from the United Nations Environment Program for granting me the opportunity of a 2-month internship in a very interesting and challenging environment.

Merci à mes parents pour m’avoir écoutée, encouragée et soutenue financièrement tout au long de mon parcours chaotique à la recherche de ma voie.

Finally, a very special thanks to my husband, without whom I probably would not have finished this Ph.D. You provided feedback on my work, took care of the housework and the baby more often than not, and came with me to conferences so that I could breastfeed. You surrounded me with all of the support, patience, kindness, and love in the world. I am infinitely grateful for having you in my life. To Georgi, my son, thank you for existing and giving me a reason to wake up (too early) in the morning.

Lausanne, September 11, 2023 O. M.

Abstract

This thesis presents work at the junction of statistics and climate science. We first provide methodology for use by climate scientists when performing fast event attribution using extreme value theory, and then describe two interdisciplinary projects in climate science that involve advanced statistical techniques.

The first chapter connects the climate literature on fast extreme event attribution studies with the statistical literature on selection effects. It provides simulations in the univariate and bivariate settings showing that not accounting for the stopping rule can lead to misestimation of return levels, but that bias can be reduced by more appropriate analysis. We discuss the spatial selection bias induced by the systematic analysis of data from the location of the extreme event and show that the estimated return period for the "trigger event" based on a dataset that contains this event can be both biased and very uncertain. We illustrate the impact of timing and spatial selection bias on return level estimation with analysis of environmental data inspired by real use cases. The Appendix describes a Python package for likelihood inference that was useful for the simulations and case studies in this chapter.

The rest of the thesis describes two applications of machine learning and statistics in climate science. The first topic studied is downscaling of historical wind fields in Switzerland. High-resolution wind maps are essential to climate scientists looking to study past climate events such as wildfires and avalanches. The deep learning model proposed in the second chapter provides realistic-looking high-resolution (1.1km) historical maps of gridded hourly wind fields over Switzerland from ERA5 input on a 25km grid. The downscaled wind fields demonstrate physically plausible orographic effects, such as ridge acceleration and sheltering, which are not resolved in the original ERA5 fields. The prediction of the aggregated wind speed distribution is very good and robust. Regionally averaged image-specific metrics measure generally better for locations over the flatter Swiss Plateau than for Alpine regions.

The third chapter proposes a random line process for hail impact modelling. Hail damage is crucial for insurance companies because big hailstones tend to produce large economic losses. Appropriate modelling and uncertainty quantification for hail impact could also be a good starting point for the study of the sensitivity of our economy to a changing climate. A two-step Bayesian hierarchical framework incorporating the random line process and extreme value theory is built to model the counts and value of hail impacts for individual buildings in the canton of Zürich and fitted using insurance data for buildings. The results are compared to the use of a benchmark deterministic hail impact function. The random line model with extreme marks proves better at capturing hail spatial patterns than the benchmark and allows for localized and extreme damage, which is observed in the insurance data.

Résumé

Cette thèse présente un travail à la jonction des statistiques et de la climatologie. Nous proposons dans un premier temps une méthodologie claire et concise utilisable par les climatologues pratiquant l'Attribution rapide d'événements extrêmes (AEE) lors de la présence de biais temporel ou spatial dans les données étudiées. Sont ensuite décrits deux projets interdisciplinaires importants pour la communauté scientifique et impliquant des connaissances avancées en statistiques.

Le premier chapitre connecte la littérature en climatologie sur les études d'attribution d'événements extrêmes (AEE) rapides avec la littérature statistique sur les biais de sélection. Le premier article présente des simulations dans les cas univarié et bivarié montrant que ne pas tenir compte du temps d'arrêt peut entraîner une estimation erronée des évènements extrêmes futurs, mais que le biais peut être réduit par une analyse plus appropriée. Nous discutons du biais de sélection spatiale induit par l'étude systématique du lieu où l'événement extrême s'est produit, et montrons que la durée de retour estimée pour l'"événement déclencheur" basée sur un jeu de données contenant cet événement peut être à la fois biaisée et très incertaine. Nous illustrons l'impact des biais de sélection temporels et spatiaux sur la prédiction de futurs évènements extrêmes en analysant des données environnementales utilisées dans des études récentes de climatologie. Un package Python facilitant l'inférence basée sur la vraisemblance est décrit dans l'Appendix. Ce package a été très utile pour réaliser les simulations et études de cas présentées dans ce chapitre.

Le reste de la thèse décrit deux applications de l'apprentissage automatique et des statistiques en climatologie. Le premier sujet étudié est la réduction d'échelle des champs éoliens historiques en Suisse. Les cartes de vents en haute résolution sont essentielles pour les climatologues qui cherchent à étudier les événements climatiques passés tels que les incendies de forêt et les avalanches. Le modèle de *deep learning* proposé dans le premier chapitre de cette thèse produit des cartes des champs de vent horaires historiques en Suisse en haute résolution (1,1 km) réalistes à partir de grilles de champs éoliens d'une résolution d'environ 25x25km^2. Les champs de vent prédits avec le modèle démontrent des effets orographiques physiquement plausibles qui ne sont pas résolus par les champs éoliens ERA5 d'origine. La prédiction de la distribution agrégée de la vitesse du vent est très bonne et robuste. Les métriques spatiales évaluées sur les cartes de vent prédites mesurent généralement mieux sur le Plateau Suisse que dans les régions Alpines.

Dans le troisième chapitre, un processus stochastique linéaire dans l'espace est proposé pour la modélisation de l'impact de la grêle dans le canton de Zürich en Suisse. Les dommages causés par la grêle sont cruciaux pour les compagnies d'assurance, car les gros grêlons ont tendance à produire des pertes économiques conséquentes. Une modélisation appropriée et une quantification de l'incertitude pour l'impact de la grêle pourraient également être un bon point de départ pour l'étude de la sensibilité de notre économie au changement climatique. Un modèle hiérarchique Bayésien en deux étapes incorporant le processus linéaire dans l'espace et la théorie des valeurs extrêmes est construit pour modéliser le nombre et la valeur des impacts individuels liés à la grêle sur des immeubles dans le canton de Zürich calibré avec des données d'assurance. Les résultats sont comparés à l'utilisation d'une fonction d'impact de grêle déterministe de référence. Le modèle de lignes aléatoires avec des marques extrêmes s'avère meilleur pour capturer les trajectoires des orages de grêle dans l'espace que le modèle de référence et permet la prédiction de dommages localisés et extrêmes, ce qui est également observé dans les données d'assurance.

Contents

Acknowledgements i

Abstract (English/Français) iii

Introduction 1

1 Timing and Spatial Selection Bias in Rapid Extreme Event Attribution 5

- 1.1 Introduction . . . 7
- 1.2 Accounting for timing bias . . . 10
  - 1.2.1 Preliminaries . . . 10
  - 1.2.2 Stopping and estimation . . . 11
  - 1.2.3 Discussion of the stopping rule . . . 14
- 1.3 Accounting for spatial selection . . . 15
- 1.4 Simulation studies . . . 17
  - 1.4.1 Setup . . . 17
  - 1.4.2 Timing bias with univariate extremes . . . 18
  - 1.4.3 Timing bias with correlated extremes . . . 20
  - 1.4.4 Bias in return period estimation . . . 23
- 1.5 Real data analyses . . . 25
  - 1.5.1 1999 flooding in Vargas state, Venezuela . . . 25
  - 1.5.2 2016 heatwave in Phalodi, India . . . 27
  - 1.5.3 2021 heatwave in the Pacific Northwest . . . 31
- 1.6 Discussion . . . 32
- 1.7 Conclusion . . . 34
- 1.8 Data statement . . . 34
- 1.9 Supplementary material . . . 35
  - 1.9.1 Data-dependent thresholds . . . 35
  - 1.9.2 Further simulation results . . . 36
  - 1.9.3 Supplementary material related to case studies . . . 37

2 Downscaling of Historical Wind Fields over Switzerland using Generative Adversarial Networks 41
2.1 Statement of interest 42
2.2 Introduction 43
2.3 Data 46
2.3.1 Geographical setting and typical wind systems in Switzerland . 46
2.3.2 Low-resolution input fields: ERA5 reanalysis 46
2.3.3 Topographic descriptors 47
2.3.4 High-resolution target fields: COSMO-1 48
2.4 Generative Adversarial Network (GAN) 48
2.4.1 General architecture 49
2.4.2 Modelling the wind time series 50
2.4.3 Generator network 51
2.4.4 Discriminator network 53
2.4.5 Input data 55
2.5 Training 55
2.5.1 Adversarial training 55
2.5.2 Transfer learning 56
2.5.3 Technical challenges 58
2.6 Metrics 58
2.6.1 Root mean squared error (RMSE) 59
2.6.2 Angular cosine distance (ACD) 60
2.6.3 Spatially convolved Kolmogorov–Smirnov statistic (SKSS) 60
2.6.4 Log spectral distance (LSD) 61
2.7 Results 61
2.7.1 Training phase 1: downscaling COSMO-1 blurred wind fields . . 62
2.7.2 Training phase 2: downscaling ERA5 wind fields 66
2.8 Conclusion 69
2.9 Acknowledgements 72
2.10 Data statement 72
2.11 Appendix 73
2.11.1 GAN structure 73
2.11.2 Examples of wind mean daily patterns from COSMO-1 blurred test sample 74
2.11.3 Examples of GAN prediction from COSMO-1 blurred test sample 75
2.11.4 Examples of GAN prediction from ERA5 test sample 76

3 Bayesian Modelling of Insurance Claims for Hail Damage 77
3.1 Introduction 79
3.2 Data and initial analysis 80
  3.2.1 Data 80
  3.2.2 Exploratory analysis 83
  3.2.3 Spatiotemporal correlation 85
3.3 Key model elements 85
  3.3.1 Random line process 86
  3.3.2 Marginal model for extreme claim values 87
3.4 Modeling extreme hailstorms 87
  3.4.1 Hail damage count 88
  3.4.2 Hail damage values 88
3.5 Model fitting and validation 91
  3.5.1 Technical challenges 91
  3.5.2 Metrics 92
3.6 Results 93
  3.6.1 Claim counts 93
  3.6.2 Combination of claim counts and values 95
  3.6.3 Claim values 95
3.7 Conclusion 96
3.8 Authorship contribution statement 97
3.9 Acknowledgments 97
3.10 Funding 97
3.11 Data statement 97
3.12 Supplementary material 101

4 Perspectives 107

A Appendix: `pykelihood` 111

Bibliography 117

Curriculum Vitae 131

Introduction

This doctoral thesis is a contribution from the statistical side to links between statistics and climate science. Both fields would benefit from more interactions, yet applied statistics papers are rarely mentioned in IPCC reports. At the heart of the division between statistics and climate science lies the question of the *ground truth*. While statisticians usually think of observed data as being “true” realizations of a random process, physicists and climate scientists tend to prefer to use reanalysis data, i.e., observed events post-processed to make more physical sense. One issue with reanalysis is that extreme events can be considered outliers and removed during the post-processing. This makes it hard to apply the statistics of extremes directly to reanalysis data, despite its relevance to climate science as a source of tools for accounting for and analyzing rare events. This thesis discusses applications of the statistics of extremes in climate science, modelling environmental phenomena with complex spatiotemporal features requiring the use of advanced statistical methods. It consists of three main chapters, two of which have been published; the third is submitted for publication.

The first part of this thesis contributes to the dialogue between statisticians from the field of extreme value theory (EVT) and climate scientists practicing extreme event attribution (EEA). Indeed, both fields have known very recent improvements, but almost independently of each other. Chapter 1 makes the connection between the literature on timing bias in fast EEA studies on the statistical side (Barlow et al., 2020; Naveau et al., 2020) and climate science (Philip et al., 2020; van Oldenborgh et al., 2021). Recent work has shown that the upward bias in return levels estimated from an analysis performed immediately after an extreme event may stem from the timing of the analysis, whereas excluding this “trigger event” will lead to a downward bias in such levels (Philip et al., 2020; Barlow et al., 2020). Introducing a stopping rule that appropriately reflects the timing of the analysis can account for such biases without requiring the analyst to decide whether to exclude or include the trigger event. This chapter sketches notions of inference using stopping rules, uses simulation to compare different approaches to data analysis, and reanalyses examples from the recent literature on climate event attribution. It also discusses the estimation of return periods for specific events and the effects of spatial selection, when the trigger event might have taken place in any of several related time series. The study of timing and spatial selection biases in rapid EEA in this chapter appears to have no equivalent in the recent literature on climate event attribution.

Environmental statistics has greatly evolved in recent years and can now deal with spatiotemporal interactions for large datasets (Cressie and Wikle, 2015; Berrocal, 2017), assimilation of several data sources (Berrocal, 2017), attribution of extreme events (Hammerling et al., 2019), spatially or temporally correlated extremes (Kropp and Schellnhuber, 2011), computational efficiency for multi-layer Bayesian models (Rue et al., 2017), and the most recent machine learning and deep learning techniques (Efron and Hastie, 2016). Nevertheless, many climate models do not incorporate these recent techniques and might misestimate the future frequency or amplitude of extreme environmental hazards. Furthermore, the coarse grids on which most climate models are run do not allow for an accurate analysis of any specific local climate event. Chapters 2 and 3 focus on these two issues, describing two interdisciplinary projects in which both the topic of interest and problem to explore were decided by a climate science practitioner and advanced statistical knowledge was required to solve it.

Chapter 2 concerns the downscaling of historical wind fields over Switzerland. High-resolution wind maps are essential to climate scientists looking to study past climate events such as wildfires and avalanches. However, climate models are run on very coarse grids: wind field predictions for a specific location are unreliable, especially on complex terrain, as airflow is strongly modified by underlying topography. The resolution of such models is usually of the order of 0.5 to 1°, which corresponds to grids of resolution between 25 and 80 square kilometers and is too coarse for detailed analysis of specific climate events. Higher-resolution climate models are very computationally expensive to calibrate and run, but downscaling of coarse climate model predictions allows faster and easier access to high-resolution wind field maps. Grid-to-point probabilistic (Winstral et al., 2017; Nerini, 2020) and grid-to-grid deep learning-based downscaling (Höhlein et al., 2020; Leinonen et al., 2021; Ramon et al., 2021) have been studied in the recent literature. Nevertheless, the model presented in

Chapter 2 is the first deep learning model that can efficiently perform such an extreme (from 25 square km to 1.1 square km resolution) downscaling of wind fields from two different data sources.

In Chapter 3, a random line process is proposed for hail impact modelling. Hail damage is crucial for insurance companies, because big hailstones tend to produce large economic losses. Furthermore, appropriate modelling and uncertainty quantification for hail impact could be helpful in studying the sensitivity of our economy to a changing climate. The literature on the statistical modelling of hailstorm impact on buildings is very limited. Although stochastic models for hailstorm risk (Deepen, 2006; Otto, 2009; Punge et al., 2014) or hailstone size (Liu et al., 2021; Perera et al., 2018) exist, most open-source studies on hailstorm impact use deterministic impact functions to link the intensity of a hail hazard to its local monetary damage. In Chapter 3 the probability and intensity of a hail event are input to the model, and stochasticity comes from the various ways a hail storm can impact different locations. The model developed seems to be the first to combine a random line process and an extreme value model to accurately represent hail damage tracks.

The thesis ends with a brief chapter with perspectives for future work and an appendix describing a Python package for likelihood inference.

1 Timing and Spatial Selection Bias in Rapid Extreme Event Attribution

Ophélia Miralles[1], Anthony C. Davison[1]

1 – Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

My Ph.D. supervisor provided feedback and helpful insights during this project and wrote Sections 1.4.4 and 1.3 of the following paper. I provided the rest of the work, i.e., the simulations, the expansion of the framework proposed in Barlow et al. (2020) for bivariate data, real data analyses, and writing of the remaining sections.

This chapter was published as a paper by the journal Weather and Climate Extremes in June 2023. It can be accessed via the link doi.org/10.1016/j.wace.2023.100584 and cited using the following bibliographical reference.

Miralles, O. and Davison, A. C. (2023). Timing and spatial selection bias in rapid extreme event attribution. *Weather and Climate Extremes*, 41:100584

Abstract

Selection bias may arise when data have been chosen in a way that subsequent analysis does not account for. Such bias can arise in climate event attribution studies that are performed rapidly after a devastating “trigger event”, whose occurrence corresponds to a stopping rule. Intuition suggests that naïvely including the trigger event in a standard fit in which it is the final observation will bias its importance downwards, and that excluding it will have the opposite effect. In either case the stopping rule leads to bias recently discussed in the statistical literature (Barlow et al., 2020) and whose implications for climate event attribution we investigate. Simulations in a univariate setting show substantially lower relative bias and root mean squared error for estimation of the 200-year return level when the timing bias is accounted for. Simulations in a bivariate setting show that not accounting for the stopping rule can lead to both over- and under-estimation of return levels, but that bias can be reduced by more appropriate analysis. We also discuss biases arising when an extreme event occurs in one of several related time series but this is not accounted for in data analysis, and show that the estimated return period for the “trigger event” based on a dataset that contains this event can be both biased and very uncertain. The ideas are illustrated by analysis of rainfall data from Venezuela and temperature data from India and Canada.

1.1 Introduction

An important objective of extreme event attribution (EEA) studies is to quantify the change in the probability of an extreme event due to external forcing, such as anthropogenic climate change (Allen, 2003; Stott et al., 2016; Naveau et al., 2020). Most such studies focus on the extent to which increased greenhouse gas (GHG) levels in the atmosphere affect the risk ratio for a specified extreme event (Stott et al., 2004; Fischer and Knutti, 2015, 2016; Jones et al., 2020). The risk ratio is commonly defined as the ratio of the probability p_1 of exceeding an extreme threshold u in the factual world, and the corresponding probability p_0 of doing so in a counterfactual world, often taken to be the pre-industrial era (Naveau et al., 2020).

The National Academies of Sciences, Engineering, and Medicine (2016) report divides EEA studies into two types: observation-based and climate-model-based. The first type uses series of historical and recent data to capture temporal changes in the probabilities and magnitudes of extreme events and thus to infer the effects of anthropogenic climate change. The second type uses a data-generating process to simulate two different worlds that are intended to be identical except for a "treatment" variable, usually GHG levels (Stott et al., 2004; Fischer and Knutti, 2015), or fine particulate matter (Larsen et al., 2020), and thereby assesses how the "treatment" affects phenomena such as temperature, precipitation or wildfires. In this framework causal inference techniques are required to efficiently capture the causality while reducing the signal-to-noise ratio in a complex and noisy climate system (Reich et al., 2021). The literature on causal statistical analysis has greatly evolved in recent decades and now has many applications (Hernan and Robins, 2023).

Observation-based EEA studies can be further divided into two groups: return-level-based studies intended to assess temporal changes in the data distribution, and meteorology-oriented studies that explore how long-term trends in large-scale circulation patterns affect local extreme events sharing common meteorological characteristics (National Academies of Sciences, Engineering, and Medicine, 2016).

The purpose of this paper is to bridge recent improvements in inference using extreme value theory (EVT) and EEA studies that employ EVT. Thus it focuses on return-level-based EEA studies such as those using the approach developed within the World Weather Attribution (WWA) initiative (see worldweatherattribution.org/about), which performs rapid return-level-based EEA.

Rapid event attribution usually takes place immediately after an extreme event, especially one with high economic or societal impact (Lerch et al., 2017). In Risser and Wehner (2017), changes in the likelihood of extreme rainfall near Houston, Texas, were analysed in September 2017, a month after Hurricane Harvey struck. Flooding in the United Kingdom (van Oldenborgh et al., 2015) and in Louisiana (van der Wiel et al., 2017) triggered immediate forecast evaluation studies for both events. The climate attribution study for the 2017 heatwave in India (van Oldenborgh et al., 2018) was conducted within a year.

Protocols for the attribution of extreme climate events have recently been improved (Philip et al., 2020; van Oldenborgh et al., 2021), but although the motivations for the choices of a relevant area, timescale, trend and distribution are well-documented, the question of whether or not to include the "trigger event" that led to the attribution study is rarely explored. Most recent studies include this event without further comment (van der Wiel et al., 2017; Risser and Wehner, 2017; Philip et al., 2018), but some exclude it because of a putative "positive bias" (van Oldenborgh et al., 2018). Sometimes it is excluded because the analysis takes place so soon after its occurrence that data are unavailable. An example of this is the study of the 2015 flooding in Northern England and Southern Scotland (van Oldenborgh et al., 2015) from which the extreme itself was initially excluded, but which was undertaken again after the data became available (Otto et al., 2018).

Guidelines for avoiding pitfalls in climate event attribution studies are provided by van Oldenborgh et al. (2021), who state:

> "There has been discussion on whether to include the event under study in the fit or not. We used not to do this to be conservative, but now realize that the event can be included if the *event definition does not depend on the extreme event itself.*"

Quote 1: Extract from van Oldenborgh et al. (2021) (our emphasis).

There can be confusion in the climate literature between events and realizations (Quote 1). The "event definition" section of most rapid EEA papers (van der Wiel et al., 2017; van Oldenborgh et al., 2018) relates to the random variable of concern, whereas the "extreme event" refers to the specific realization under study. For instance, in van der Wiel et al. (2017), "event definition" refers to the annual maximum 3-day precipitation average (a random variable in statistical terms) and the extreme event under study is the 3-day precipitation average of 216.1 mm.day^{-1} observed in Livingston, Louisiana, in August 2016 (a realization of the random variable). This raises three issues: potential for linguistic and hence conceptual confusion, in particular between random variables and events; the possibility that the random variables themselves are defined in light of an observed event; and the inclusion or not of the particular realisation in the data analysis. In this particular case, the event definition does not depend on the Louisiana level, but the latter is included in the analysis and thus influences the fitted generalized extreme value distribution. In this paper, we differentiate between random variables, their realisations and events using standard notation: realisations of a random variable X are designated by x and we refer to events using the letter $\mathscr{E}$.

We now focus on the third of the issues just mentioned, namely the inclusion or not of the trigger event in analysis. Certain discussions of protocols for extreme event attribution suggest that even if the extreme observation that stimulated the analysis is excluded from the dataset, the corresponding information can be incorporated by constraining the tail of the fitted distribution to be heavy enough to ensure that the return period for that observation is finite:

> “We do use the information that [the extreme event] occurred by demanding that the distribution has a non-zero probability of the observed event occurring [...]. This primarily affects the uncertainty estimates [...], which usually have upper bounds.”

Quote 2: Extract from Philip et al. (2020).

We explain below how appropriate statistical methods can account directly for the trigger event, thus removing any need for constraints of this sort.

Recent work has shown that the upward bias observed when an analysis is performed immediately after an extreme event may stem from the timing of the analysis, whereas excluding the trigger event will lead to a downward bias in estimated return levels (Philip et al., 2020; Barlow et al., 2020). Introducing a stopping rule that appropriately reflects the timing of the analysis can account for such biases without requiring a decision to exclude or include the trigger event, though we shall see below that it is best to exclude it if its return period is to be estimated.

Below we sketch notions of inference using stopping rules, use simulation to compare different approaches to data analysis and reanalyse examples from the recent literature on climate event attribution. We also discuss the estimation of return periods for specific events and the effects of spatial selection, which can occur when the trigger event might have taken place in any of several time series in related locations. The extent to which some selection biases influence EEA is well-documented in the recent literature. The selection of the "trigger event" itself produces a bias, since we are mainly interested in extreme events that happened and for which an increased probability due to climate change is expected (Philip et al., 2020; van Oldenborgh et al., 2021). The bias introduced by reducing the spatial area of interest (Stott et al., 2004; Hammerling et al., 2019) or the possible weather conditions (Philip et al., 2020; van Oldenborgh et al., 2021) to those of a specific observed event is also documented in recent EEA studies. However, to our knowledge, there is no comparable study of such biases in the literature on climate event attribution.

Sections 1.2 and 1.3 introduce timing and spatial selection biases, and simulation results are then provided to show how such biases affect return level estimation. Three fast event attribution studies are then re-analysed, accounting for those selection biases. The paper closes with some recommendations for future rapid return-level-based attribution analyses.

1.2 Accounting for timing bias

1.2.1 Preliminaries

We precede our discussion of remedies for timing bias by recalling the cumulative distribution functions of the generalized extreme value and generalized Pareto distributions, and of the joint cumulative distribution function associated with a logistic copula in S dimensions, viz

$$\mathrm{GEV}(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, \quad -\infty < x < \infty, \tag{1.1}$$

$$\mathrm{GPD}^u(x) = 1-\left(1+\xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x \in [u,\infty), \tag{1.2}$$

$$C(w_1,\ldots,w_S) = \exp\left[-\left\{\sum_{s=1}^{S}\left(-\log w_s\right)^{1/\alpha}\right\}^{\alpha}\right], \quad 0 < w_1,\ldots,w_S < 1, \quad 0 < \alpha \le 1, \tag{1.3}$$

where $a_+ = \max(a, 0)$ for real numbers a. Expressions (1.1) and (1.2) respectively provide standard models for block (e.g., annual or seasonal) maxima and for the exceedances of a high threshold u. Both models depend on a shape parameter ξ that determines the weight of the distribution tails; the first also depends on location and scale parameters μ and σ, and the second depends on a scale parameter σ_u. The logistic copula (1.3) is a one-parameter dependence model in which the variables $w_1, \ldots, w_S$ are independent when $\alpha = 1$ and become totally dependent when $\alpha \to 0$. Such a simple dependence model rarely fits real data well, but it is adequate for our purposes. Below we denote the unknown parameters for each of these expressions by θ. The development in the univariate case below is closely based on Barlow et al. (2020). Belzile and Davison (2022) give an alternative derivation of the main results and investigate improved inference based on them.

1.2.2 Stopping and estimation

When statistical analysis is performed immediately after the occurrence of a trigger event, the joint probability density of the data should be modified. If for simplicity we suppose that the successive observations are independent replicates of a random variable with probability density and distribution functions f and F, and denote the trigger event by $\mathscr{E}$, at which time the available data are $x_1, \ldots, x_T$, then the joint density of the data, conditional on the occurrence of $\mathscr{E}$, is

$$f(x_1, \ldots, x_T \mid \mathscr{E}) = \frac{\Pr(\mathscr{E} \mid x_1, \ldots, x_T) f(x_1, \ldots, x_T)}{\Pr(\mathscr{E})} = \frac{I(\mathscr{E} \cap \{x_1, \ldots, x_T\}) f(x_1) \cdots f(x_T)}{\Pr(\mathscr{E})},$$

where the first equality follows from Bayes' theorem and the second from the assumption that $x_1, \ldots, x_T$ are independent. The indicator function appearing in the last expression ensures that the joint density equals zero unless the configuration of the data $x_1, \ldots, x_T$ has led to the the trigger event; it can be dropped below, because we assume throughout that this is the case. If $\mathscr{E}$ occurs at time T, when the data series first exceeds some pre-determined high level η, for example, then $\Pr(\mathscr{E}) = F(\eta)^{T-1}\{1 - F(\eta)\}$, leading to

$$f(x_1, \ldots, x_T \mid \mathscr{E}) = \prod_{t=1}^{T-1} \frac{f(x_t)}{F(\eta)} \times \frac{f(x_T)}{1 - F(\eta)}, \quad x_1, \ldots, x_{T-1} \leq \eta < x_T. \tag{1.4}$$

The first $T - 1$ terms on the right-hand side of (1.4) correspond to those observations that did not exceed η, and the last term to the value $x_T > \eta$ that caused the trigger event.

The observations $x_1, \dots, x_{T-1}$ are right-truncated at η, whereas x_T is left-truncated at η.

The above formulation assumes that the trigger event is generated by the same physical mechanisms as earlier data. In some cases this may be untrue, because of changes in the background climatology or novel conjunctions of circumstances, but in any case one aspect of attribution analysis is to gauge the appropriate degree of surprise at the trigger event, and this involves comparison with the past. Moreover if this event is so unprecedented that relevant data are very limited or even unavailable, statistical analysis is difficult to justify. We therefore maintain this assumption, though rather gingerly.

For simplicity above we have suppressed the parameter vector θ on which an expression such as (1.4) depends, but in applications the conditional density is used to fit the model, so we henceforth include θ in the notation. Estimation by maximising the standard log-likelihood function

$$\mathscr{L}^{\text{STD}}(\theta) = \sum_{t=1}^{T} \log f(x_t;\theta) \tag{1.5}$$

does not account for the fact that T is determined by the data. A naïve correction excludes the final observation from the data, giving the 'exclusion' log-likelihood function

$$\mathscr{L}^{\text{EX}}(\theta) = \sum_{t=1}^{T-1} \log f(x_t;\theta), \tag{1.6}$$

but although x_T itself does not appear here, it influences the fit because it helps to determine T. Neither (1.5) nor (1.6) allows for the fact that T is random, and, as mentioned above, fitting using them can be expected to lead to respective under- and over-estimation of the return period for x_T.

Two difficulties in the statistical formulation of the trigger event and its associated stopping rule is that these are typically only known after this event has occurred and that the event itself may be somewhat vaguely defined, so entirely watertight inferences appear unattainable. However, sensitivity analysis based on plausible stopping rules is certainly feasible, and below we shall see that it can provide useful insights.

One natural formulation of the stopping rule is to define the trigger event so that the preceding data are not regarded as particularly unusual. A simple way to do this is to set $T = \min\{t : x_t > \eta_t\}$, where $\eta_1, \eta_2, \ldots$ is a series of thresholds and x_T is the first observation to exceed the corresponding threshold. Thus $x_t < \eta_t$ for $t = 1, \ldots, T-1$, and then $x_T > \eta_T$. In many cases, η_t might be constant over time, but this is not essential to the argument. The resulting full conditional log-likelihood function (Barlow et al., 2020) is a generalisation of expression (1.4),

$$\mathscr{L}^{\text{COND}}(\theta) = \sum_{t=1}^{T} \log\left\{\frac{f(x_t;\theta)}{F(\eta_t;\theta)}\right\} + \log\left\{\frac{f(x_T;\theta)}{1-F(\eta_T;\theta)}\right\}, \tag{1.7}$$

which incorporates this stopping rule and thus allows for the timing bias. We do not consider the partial conditioning approach suggested by Barlow et al. (2020), but by analogy to $\mathscr{L}^{\text{EX}}(\theta)$ we introduce

$$\mathscr{L}^{\text{CONDEX}}(\theta) = \sum_{t=1}^{T-1} \log\left\{\frac{f(x_t;\theta)}{F(\eta_t;\theta)}\right\}, \tag{1.8}$$

which excludes the trigger event from $\mathscr{L}^{\text{COND}}$. Equations (1.5), (1.6), (1.7) and (1.8) easily adapt to the non-stationary case by replacing the parameter vector θ by a time-varying parameter vector θ_t.

The use of varying thresholds would be natural in many applications, but allowing them to depend on recent extremes raises computational issues; see the Supplementary Material.

Analyzing correlated time series to predict return levels in a specific area is common in climate studies. For example, van der Wiel et al. (2017) selected 19 out of 324 stations in the state of Louisiana (US), with at least 0.5° of spatial separation among those selected, in order to reduce spatial dependence between time series of annual maximum 3-day precipitation averages. In van Oldenborgh et al. (2018), maximum annual temperature return levels for two correlated time series close to Phalodi (India) are derived from separate event attribution studies. Thus it is useful to extend our discussion above to the multivariate setting. A simple approach uses a copula to model dependence among S-dimensional variables $x_1, \ldots, x_T$, where $x_t = (x_{t,1}, \ldots, x_{t,S})$ now consists of observations at S spatial locations that we denote collectively by $\mathscr{S}$. Then the log-likelihood functions (1.5), (1.6), (1.7) and (1.8) for independent $x_1, \ldots, x_T$ generalise to

$$\mathscr{L}^{\mathrm{STD}}(\theta) = \sum_{t=1}^{T} \log\left[f(x_t;\theta)c\{F(x_t;\theta);\theta\}\right], \tag{1.9}$$

$$\mathscr{L}^{\mathrm{EX}}(\theta) = \sum_{t=1}^{T-1} \log\left[f(x_t;\theta)c\{F(x_t;\theta);\theta\}\right], \tag{1.10}$$

$$\mathscr{L}^{\mathrm{COND}}(\theta) = \sum_{t=1}^{T-1} \log\left[\frac{f(x_t;\theta)c\{F(x_t;\theta);\theta\}}{C\left\{F(\eta_t;\theta);\theta\right\}}\right] + \log\left[\frac{f(x_T;\theta)c\{F(x_T;\theta);\theta\})}{1-C\left\{F(\eta_T;\theta);\theta\right\}}\right], \tag{1.11}$$

$$\mathscr{L}^{\mathrm{CONDEX}}(\theta) = \sum_{t=1}^{T-1} \log\left[\frac{f(x_t;\theta)c\{F(x_t;\theta);\theta\}}{C\left\{F(\eta_t;\theta);\theta\right\}}\right], \tag{1.12}$$

where $F(x_t;\theta) = \left\{F_1\left(x_{t,1};\theta\right),\ldots,F_S\left(x_{t,S};\theta\right)\right\}$ represents the vector of marginal cumulative distribution functions, $f(x_t) = \prod_{s=1}^{S} f_s(x_{t,s};\theta)$ is the product of their marginal density functions, and C is a copula with uniform margins defined by

$$\mathbb{P}(X_1 \leq x_1,\ldots,X_S \leq x_S;\theta) = C\{F(x;\theta);\theta\} \tag{1.13}$$

with associated density function $c(u;\theta) = \partial^S C(u;\theta)/\partial u_1 \cdots \partial u_S$, with $u = (u_1,\ldots,u_S) \in [0,1]^S$.

1.2.3 Discussion of the stopping rule

In an ideal world the stopping rule would be clearly specified in advance of potential trigger events by listing circumstances exceptional enough to warrant an attribution study. In many practical situations, such a task is impossible, too time-consuming or too restrictive, so attribution analysis is often performed without the clear prior specification of a trigger event. In many cases contextual information about what events can be treated as extreme in a specific geographical region can be used to "guess" the stopping rule and thus to determine terms appearing in equations (1.7) and (1.8). For example, the authors of the attribution study for the August 2016 Louisiana floods give the following quantitative definition of extreme flooding:

> "In places, the 3-day precipitation totals in Louisiana exceeded [...] 3 times the average annual 3-day precipitation maximum"

Quote 3: Extract from van der Wiel et al. (2017).

If this definition was not influenced by the level recorded in August 2016, then a flooding event would be considered as extreme when, somewhere in the region, a 3-day average precipitation annual maximum three times bigger than its historical average was recorded. The data analysed in van der Wiel et al. (2017) involve $S = 19$ different spatial locations. Assuming that there is no spatial selection, the stopping rule could be defined as the first time at which one or more of these spatial locations records a 3-day average precipitation X_t^s that exceeds three times the historical annual average 3-day maximum. If data for the years 1950–2000 are used to compute the historical average $\bar{X}^s$ at each location s in the set $\mathscr{S}$ containing the locations and stopping can only occur in subsequent years, we might take T to be the first time from the year 2001 onwards that such an event occurs at one or more locations in $\mathscr{S}$, i.e.,

$$T = \min\left\{ t \geq 2001 : \bigcup_{s \in \mathscr{S}} \left(X_t^s \geq 3\bar{X}^s \right) \right\}. \tag{1.14}$$

In van Oldenborgh et al. (2018), an extreme heatwave is declared when TXx, the annual maximum daily temperature between May and June, is at least 4 or 5 degrees above its average for 1981–2010. Equation (1.15) transcribes this contextual vision of an extreme temperature to a quantitative stopping rule for use in fitting the observed series, i.e.,

$$T = \min\left\{ t \geq 2010 : \mathrm{TXx}_t - \overline{\mathrm{TXx}}_{[1981,2010]} \geq 4 \right\}. \tag{1.15}$$

When the precise definition of the extreme event is unclear, various plausible stopping rules could be formulated and used as the basis for sensitivity analyses.

1.3 Accounting for spatial selection

Thus far we have discussed how analysis immediately after a trigger event can influence the estimation of an underlying extremal probability model and thus affect the probability and/or return period associated with that event. Bias can also arise when the trigger event occurs in a single time series that is selected among several related series, and no allowance is made for the selection. We now give a stylised discussion of how this affects estimated return periods for the event in question.

Suppose that S independent time series are monitored and that extreme events occur in the sth series with distribution $\mathrm{GEV}_s(x)$, where the subscript indicates that the parameters that determine the distribution depend on the series. Suppose that analysis takes place when the largest of the corresponding variables $X_1, \ldots, X_S$ exceeds a given return level, and that this selection is ignored. Without loss of generality we further suppose that this largest value occurs in series $s = 1$, and that its value x_1 is associated with a return period of m years based on the distribution $\mathrm{GEV}_1(x)$, i.e.,

$$\mathrm{GEV}_1(x_1) = 1 - 1/m.$$

This calculation ignores the fact that corresponding values $x_2, \ldots, x_S$ in time series $2, \ldots, S$, each such that $\mathrm{GEV}_s(x_s) = 1 - 1/m$, would also have led to the same return period estimate. Taking the selection into account, the true return period m_S is therefore given by

$$1 - 1/m_S = \Pr(X_1 \le x_1, \ldots, X_S \le x_S) = \prod_{s=1}^{S} \Pr(X_s \le x_s) = (1 - 1/m)^S,$$

i.e.,

$$m_S = \left\{1 - (1 - 1/m)^S\right\}^{-1}. \tag{1.16}$$

If $S = 1$, i.e., there is no selection, then $m_S = m$, and if m is large then $m_S \approx m/S$: m-year events will occur S times more frequently in S independent series.

At first sight, these calculations for independent series might appear irrelevant to the analysis of dependent series. For so-called asymptotically dependent series, however, and with the same notation, one can write

$$\Pr(X_1 \le x_1, \ldots, X_S \le x_S) = (1 - 1/m)^{\chi} = 1 - 1/m_{\chi}, \tag{1.17}$$

where the so-called extremal coefficient χ satisfies $1 \le \chi \le S$ and can be interpreted as the “number of independent series” contributing to the overall maximum. If $X_1, \ldots, X_S$ are totally dependent, then $\chi = 1$, and if they are fully independent, then $\chi = S$; see expression (31.12) of Davison et al. (2019), for example. Asymptotically dependent models for spatial extremes can be expected to provide reasonable approximations to phenomena such as maxima of temperature time series at S sites in a relatively small spatial region, and such models will then provide upper bounds on m_S. An alternative class of models, often found appropriate for phenomena such as rainfall at spatially separated sites, has the property of asymptotic independence: increasingly rare observations become closer to independence, i.e., $\chi \lesssim S$ for very rare events. The corresponding m_S given by (1.16) will then provide a lower bound on the true return period.

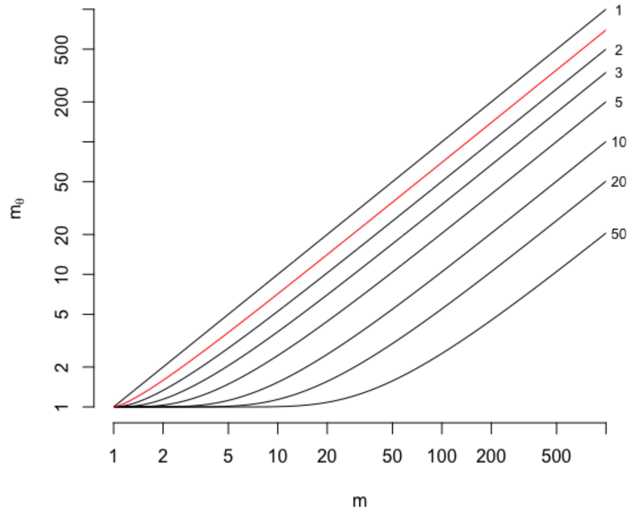


Figure 1.1: Dependence of true return period m_χ on naïve return period m when the selection of an extreme event in one series among χ "equivalent independent" series is ignored. The figures at the right of the black lines show χ. The red line corresponds to $\chi \approx 1.43$ for the Phalodi analysis in Section 1.5.2.

Figure 1.1 shows how m_χ is related to m for various values of χ. Each function is roughly linear for $m \geq \chi$, so the approximation $m_\chi \approx m/\chi$ seems adequate in most cases.

1.4 Simulation studies

1.4.1 Setup

We now use stochastic simulation of extremal data to compare how fitting based on the various log-likelihood functions described above affects the estimation of a p-year return level, i.e., the level expected to be crossed by the variable of interest every p years, taking $p = 200$ for illustration. We shall see that not accounting for timing bias can lead to poor estimation in both univariate and bivariate settings. We also consider the association of a return period with a specific observation.

Our Monte Carlo settings were chosen to resemble real uses of extreme event attribution. Many climate variables studied are positive, unbounded and somewhat heavy-tailed, and their annual maxima are commonly fitted with a generalized extreme-value

(GEV) distribution. As in Barlow et al. (2020), we defined quantitative stopping rules using a simulated historical sample of $n_h = 10$ maxima, then generated further independent variables from the same GEV distribution, applied different stopping rules, and used the resulting samples of maxima to estimate the three GEV parameters and the p-year return level.

For each run, stopping rules in which the chosen thresholds were return levels η_τ for a GEV (see Equation (1.1)) with parameters $\mu = 0$, $\sigma = 1$ and $\xi = -0.2, 0, 0.2$ with different return periods τ were applied, giving stopping times

$$T^\tau = \min\left\{t > n_h : X_t \geq \eta_\tau\right\}. \tag{1.18}$$

The goal was to evaluate the impact of increasing the unlikeliness of the trigger event on the estimation of the 200-year return level based on the log-likelihood functions (1.5)–(1.8). Fits from 1000 simulated datasets were compared to the true 200-year return level in terms of the bias and relative root mean squared error (RRMSE). Confidence interval coverage (CIC) and width (CIW) are derived from confidence bounds for the 200-year return level estimator.

1.4.2 Timing bias with univariate extremes

We first discuss the effect of timing bias when estimating return levels based on a univariate time series. The results from parameter estimation using the log-likelihood functions (1.5), (1.6), (1.7), and (1.8) are respectively designated by “Standard”, “Excluding Extreme”, “Cond. Including Extreme” and “Cond. Excluding Extreme” in the figures and the text.

Barlow et al. (2020) sampled GEV random variables until they first exceeded a threshold η_τ (see Equation (1.18)) or until the maximum sample size N was reached, estimated the parameters θ and then estimated the 200-year return level and its 95% confidence bounds. Their simulation studies result in lower relative bias and root mean squared error using the full conditional log-likelihood than using the standard likelihood, whether or not the trigger event is included, with comparable confidence interval coverage and width; see the Supplementary Material. The relative bias decreases when the full conditional log-likelihood includes the extreme event for return periods of $\tau \geq 500$, but this is due to the imposition of a maximum sample size. As τ increases, exceedances of η_τ become less likely, and when no realization exceeds η_τ, sampling stops when the maximum sample size is reached, and it is inappropriate to condition the log-likelihood with regard to a stopping rule that has not been applied.

To avoid the aforementioned problem we performed simulations with the sample size fixed to be $n_C = 200$ and return periods τ exceeding $n_C - n_h$, so that the last event observed is unlikely enough for the stopping rule to make sense. We first generated a "historical sample" of n_h GEV variables, and then, for each return period τ considered (see Equation (1.18)), we generated $n_C - n_h - 1$ GEV variables right-truncated at η_τ, followed by a final GEV variable left-truncated at η_τ; these correspond to the conditional densities appearing in (1.7). We then concatenated the historical sample, the data under the stopping threshold η_τ and the last observation above η_τ to yield a sample of n_C observations, of which the only observation to exceed η_τ was the last, provided η lies above all n_h historical values. Figure 1.2 shows the results of this experiment with GEV shape parameter $\xi = 0.2$. The standard fit shows an upward relative bias that increases with the size of the trigger event, and the resulting 200-year return level is less and less reliable (the confidence interval coverage decreases with τ). The differences between results for the other three log-likelihoods are smaller, especially for large η_τ, partly because the conditioning term has little effect on Equations (1.8) and (1.7) when $F(\eta_\tau) \approx 1$. The coverage of two-sided confidence intervals is most stable for the conditional fits, but this disguises a difference in the one-tailed errors: the intervals tend to be too short in the upper tail and too long in the lower tail. There is little to choose between the results using the conditional fits, though that with all the available information, based on (1.7), seems slightly preferable for smaller τ.

The corresponding results with $\xi = 0$ and -0.2 reported in the Supplementary Material lead to similar conclusions: conditioning while either including (Equation (1.7)) or excluding (Equation (1.8)) the "trigger" provides less biased 200-year return level estimates than the standard fit, whether or not the trigger is included in the latter. However, the coverage of the conditioned fit that includes the "trigger" deteriorates when $\xi = -0.2$, and its upper coverage error also significantly increases, as the upper bound for the 95% confidence interval is under-estimated for negative ξ. In this case, excluding the trigger without conditioning leads to underestimation of the upper confidence bound for $\tau < 1000$ and of the lower confidence bound for any τ considered.

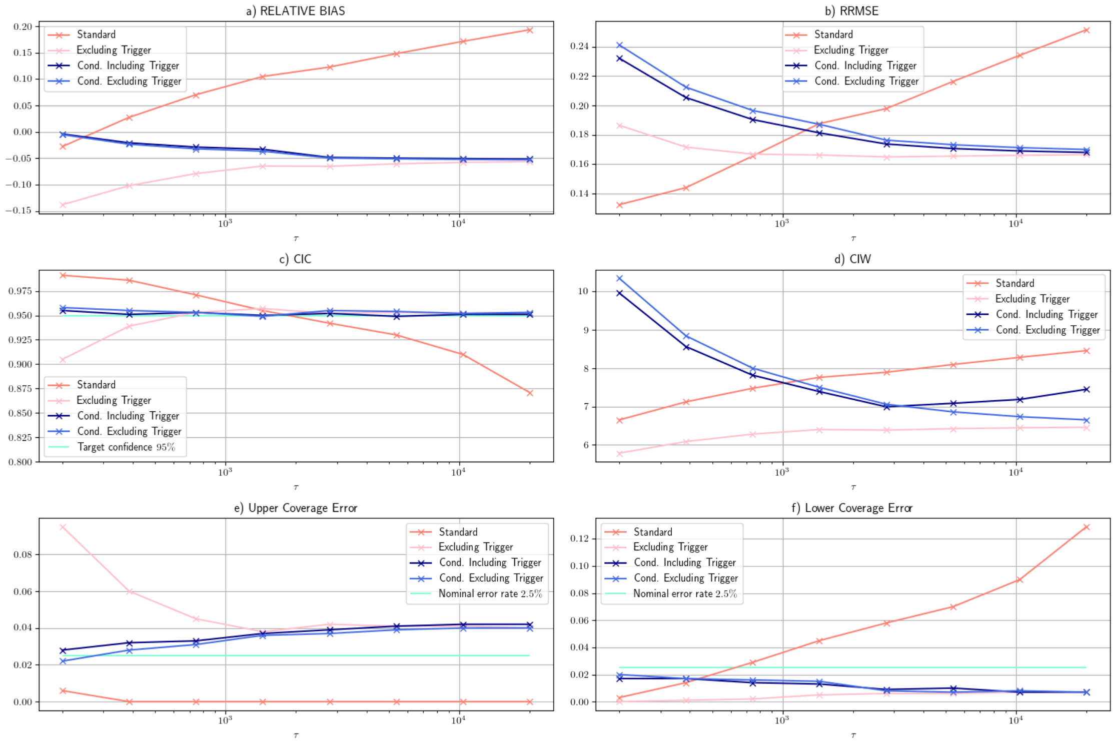


Figure 1.2: Summary results for the estimation of a 200-year return level based on simulated GEV random variables with shape parameter 0.2, with stopping thresholds defined from the return periods τ shown on the x-axis as in Equation (1.18). The relative bias and mean squared error are shown in Panels a and b, and coverage of 95% confidence intervals and their average widths are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors. The time series are generated so that the first exceedance of the stopping threshold occurs at a specified time.

1.4.3 Timing bias with correlated extremes

We now investigate the impact of timing bias in a bivariate setting. The univariate fit for the variable of interest is labelled “Independent”, while fits using the log-likelihood functions (1.9), (1.10), (1.11), and (1.12) are respectively labelled “Including Extreme”, “Excluding Extreme”, “Cond. Including Extreme” and “Cond. Excluding Extreme”.

We suppose that the stopping rule is applied to one variable but the other is merely part of the analysis. This situation can arise when, for instance, a location s_1 is very close to that of the trigger event, but its time series for the variable of interest lacks the data for that event itself. Often a more complete time series is available, and though its location s_2 lies further from that of the trigger event, it can serve as a monitoring reference for data at s_1. If there is strong dependence between time series at the two locations, then observation of an extreme event at s_2 may aid in event attribution for an extreme at s_1.

To explore this setting we generated replicates of two GEV variables X and Y with shape parameters 0.2 and dependence given by the logistic copula (1.3) with its parameter $\alpha = 0.5$ taken to be known. The maximum sample size was set to N, as in Barlow et al. (2020), and the univariate stopping rule of Equation (1.18) was applied to Y: sampling of both series stopped when $Y \geq \eta_\tau$ for some return period τ. Although the stopping rule is applied only to Y, it influences the estimation of quantiles of X because the series are dependent. For every return period considered, the marginal distributions of X and Y were estimated from the time series stopped at T^τ, say, using the log-likelihood functions (1.9)–(1.12). The cumulative distribution function for Y is denoted by F_Y. In Equations (1.11) and (1.12), the bivariate conditional terms $C\left\{F(\eta_t;\theta)\right\}$ for $t \leq T^\tau$ and $1 - C\left\{F(\eta_{T^\tau})\right\}$ are respectively replaced by $F_Y(\eta_\tau;\theta)$ and $1 - F_Y(\eta_\tau;\theta)$. Finally, the 200-year return level for X and its confidence bounds were derived and the summaries used in the univariate case were computed. The standard univariate fit for X from Equation (1.5) was used as a benchmark.

When $\alpha = 0.5$, the probability that X is extreme given that Y is extreme can be shown to equal $2 - 2^\alpha \simeq 0.59$. This leads to the following two cases:

$\mathscr{A}$. both X and Y are extreme when the trigger event occurs. In this case, we expect the return levels of X to be overestimated when fitting the time series for X assuming X and Y are independent, like in the standard univariate case. The corresponding simulation results, displayed in Figure 1.3, suggest that for every stopping threshold η_τ the relative bias and root mean squared error from the full conditional fit are much lower than for other fits, while confidence intervals have similar coverages and widths. Upper and lower coverage errors are very similar across methods accounting for the dependence between the series (Figure 1.3e and f), though for high stopping thresholds, excluding the trigger with the appropriate conditioning provides the closest upper coverage error to the nominal rate, i.e., the most reliable upper confidence limit for the 200-year return level, while other methods tend to underestimate the upper confidence bound (Figure 1.3e);

$\mathscr{B}$. X is not extreme when the trigger event occurs, and we then expect the univariate fit for X to underestimate the return levels for X. Indeed, realizations of X sampled until the trigger event will tend to be low because they are related to those of Y, which lie below η_τ until sampling stops. Figure 1.4 shows very reduced relative bias and RRMSE with conditioned bivariate fits, both including

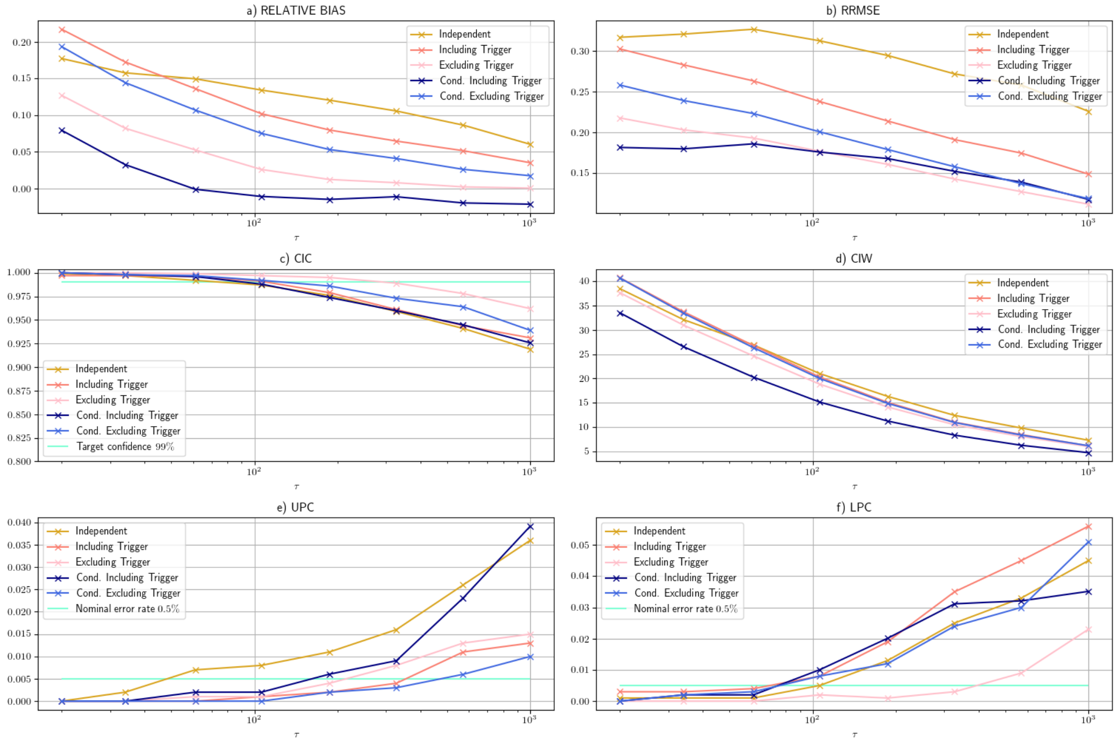


Figure 1.3: Summary results for simulated bivariate extremes, case $\mathscr{A}$: X is also extreme when Y is stopped. Simulations of two correlated random variables X and Y following a logistic ($\alpha = 0.5$) copula with GEV ($\mu = 0, \sigma = 1, \xi = 0.2$) margins. The stopping rule is defined for Y as the return level of period τ as in Equation (1.18). The return periods τ are shown on the x-axis. The relative bias and relative mean squared error from the theoretical 200-year return level for X are shown in Panels a and b, and 99% confidence interval coverage and width are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors.

> and excluding the extreme at s_2, compared with the independent fit for X, which strongly underestimates the 200-year return level at s_1, and with the standard fit including the extreme event, which gives positively biased estimators of the 200-year return level. Excluding the extreme leads to slight downward bias of the estimated return level for X for every stopping threshold τ considered. Of all methods, excluding the trigger with the appropriate conditioning provides the upper coverage error closest to the nominal error rate, especially for high stopping thresholds (Figure 1.4e).

The figures show how both cases affect the return level estimates. The improvement due to accounting for the timing bias is much clearer in case $\mathscr{A}$, but case $\mathscr{B}$ better illustrates the situation in which the data are incomplete at the location of interest s_1 but the trigger event is seen only at s_2.

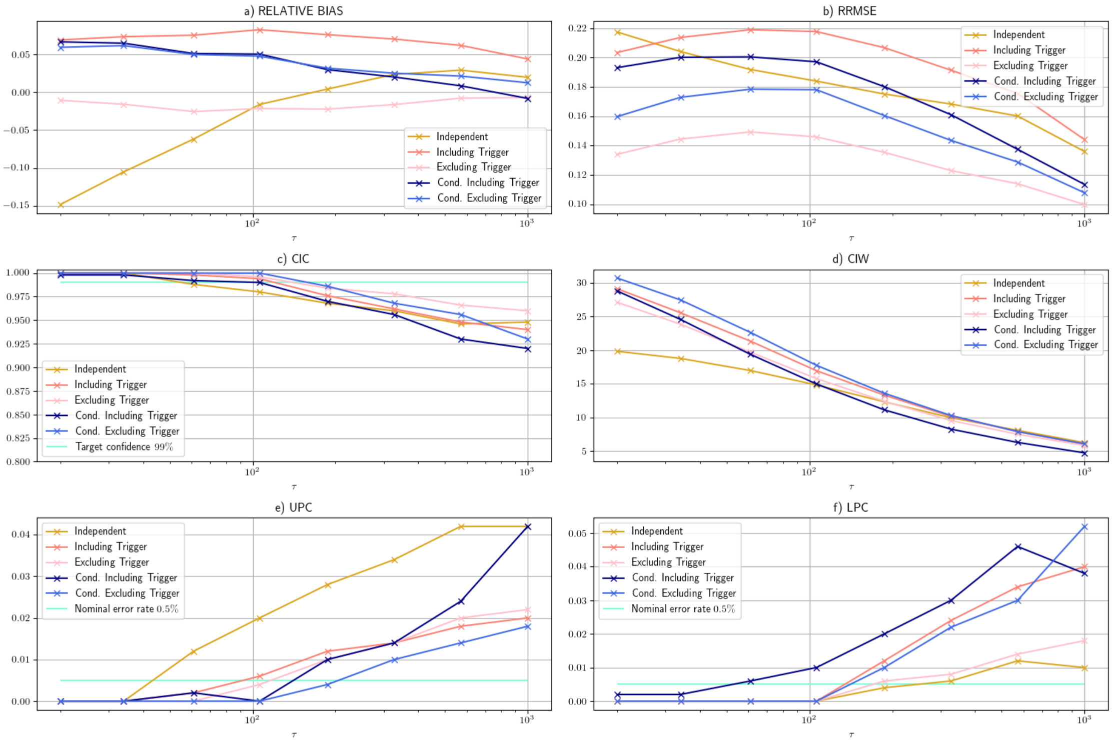


Figure 1.4: Summary results for bivariate extremes, case $\mathscr{B}$: X is not extreme when Y is stopped. The simulation setup and stopping rule are the same as in Figure 1.3. The relative bias and relative mean squared error from the theoretical 200-year return level for X are shown in Panels a and b, and the 99% confidence interval coverage and width are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors.

1.4.4 Bias in return period estimation

Fits of extreme-value models can be highly sensitive to the largest or smallest observations in a sample (Davison and Smith, 1990), so it is natural to wonder whether estimated return periods for particular observations corresponding to rare events might be biased. For concreteness, suppose that the generalized extreme-value distribution (1.1) has been fitted to a sample whose largest value is $X_{\max}$ and that the return period of $X_{\max}$ is to be estimated from the fit. The fitted distribution is $\widehat{\mathrm{GEV}}(x)$, so the true return period M and its estimate $\widehat{M}$ may be written as

$$M = \{1 - \mathrm{GEV}(X_{\max})\}^{-1}, \quad \widehat{M} = \left\{1 - \widehat{\mathrm{GEV}}(X_{\max})\right\}^{-1}.$$

The observation $X_{\max}$ is often described as an "M-year event", but this term applies in relation to $\mathrm{GEV}(x)$. In practice the estimate $\widehat{\mathrm{GEV}}(x)$ is often based on data that include $X_{\max}$, and the latter may strongly influence the estimated distribution. It seems plausible that $\widehat{M} < M$ if $X_{\max}$ is included in the fit, and that $\widehat{M} > M$ if not. The situation here differs from those in previous sections, which concerned the estimation of a return level, i.e., a parameter of the distribution, as the return period M depends on $X_{\max}$ and thus is itself random. To investigate the relation between M and $\widehat{M}$ we performed a further simulation study, which we now describe.

We generated 1000 independent datasets using a simplification of the approach described in Section 1.4.1, by simulating $n_C - 1$ observations from (1.1) right-truncated at a fixed threshold η, supplemented by a final observation with return period m. In order to measure the bias as accurately as possible, this final observation is determined by the equation $\mathrm{GEV}(x) = 1 - 1/m$ and thus is fixed. For each such dataset we computed return period estimates $\widehat{m}$ using fitted distributions $\widehat{\mathrm{GEV}}(x)$ found using the log-likelihood functions (1.5)–(1.8). This process was repeated for different configurations of values of n_C, η and m, with shape parameter $\xi = 0.2$ throughout.

Figure 1.5 shows boxplots of the resulting ratios $\widehat{m}/m$. The results are extremely variable, with many simulated datasets in which $\widehat{m} \gg m$, but some patterns emerge. When GEV is based on the standard log-likelihood and the largest observation is included, we tend to see $\widehat{m} < m$, especially when the effect of not allowing for the right-truncation is reinforced by increasing n_C; $\widehat{m}$ is systematically too large when the largest observation is excluded. The situation is more variable with the conditional likelihood, which gives the most consistent results when the largest observation is excluded. In itself this is not surprising, as use of log-likelihood (1.8) is then appropriate and the fitted distribution is independent of the largest observation; thus in this case we expect that $\widehat{m}/m \to 1$ as n_C increases, but clearly such convergence is unlikely to be visible for values of n_C seen in applications. Perhaps the most striking feature of the results is that $\widehat{m}/m$ is very variable and/or systematically biased in all cases: an event with $m = 1000$, say, might easily have $\widehat{m}$ anywhere in the range 250 to 4000. This suggests that extreme caution is required when attributing return periods to particular events; indeed, this should not be attempted without a statement of uncertainty.

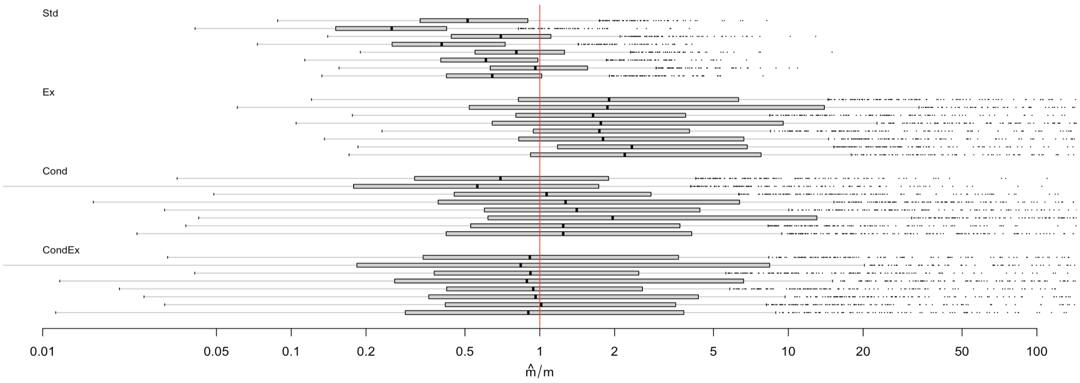


Figure 1.5: Ratios of estimated and true return periods $\hat{m}/m$ for various fits to samples of size n_C, with threshold η and true return period m, with parameters estimated from standard and conditional log-likelihoods including and excluding the largest value, labeled Std, Ex, Cond and CondEx. From top in each block of boxes: (η, m, n_C) = (100,200,50), (100,200,80), (150,200,50), (150,200,80), (200,400,50), (200,400,100), (200,1000,50), (200,1000,100).

1.5 Real data analyses

1.5.1 1999 flooding in Vargas state, Venezuela

We now consider an extreme flooding event in the Venezuelan state of Vargas in December 1999. According to Méndez et al. (2015), the form of the San Julián basin implies that major rainfall events are extremely rare in this area, but when they do happen the consequences can be very serious. Indeed, these authors observe that this basin has a very wide range of slopes, provoking increased erosion over time and that its small area (20.68 km^2) implies rapid concentration of surface runoff, so water can very quickly arrive in residential zones. In December 1999, such flooding, combined with a landslide, massive debris transportation and poor infrastructure, caused disastrous damage in the Caraballeda area.

To predict the likelihood of such an event, we estimate return levels for daily maximum hourly precipitation (mm) in Vargas from 1961 to 1999. The San Julián basin is not subject to much seasonality (Méndez et al., 2015), and no long-term trend is perceptible in these data.

We use a generalized Pareto distribution (1.2) to model daily rainfall amounts over $u = 12$ mm, a choice of threshold justified in the Supplementary Material using the approaches of Northrop and Coleman (2014) and Varty et al. (2021). Let η_p be defined

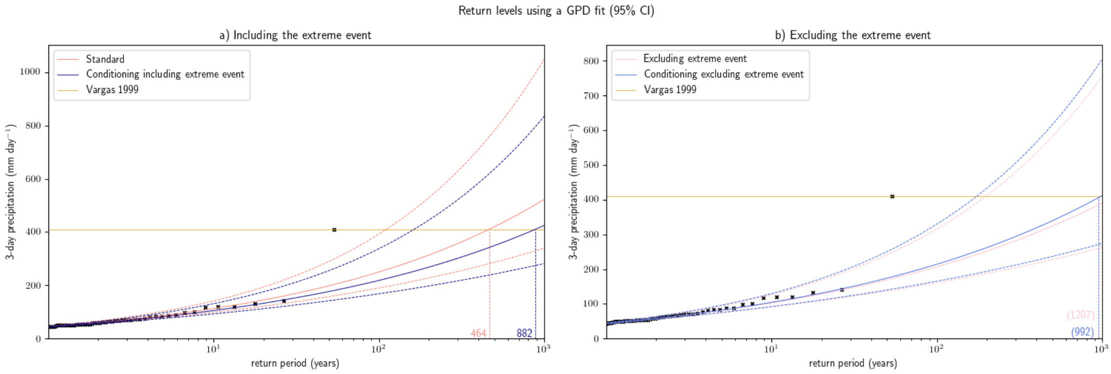


Figure 1.6: Vargas data analysis. Estimated p-year return level η_p (see Equation (1.19)) and its 95% confidence interval (dashed lines) from a GPD fit. The x-axis shows the return period p (years). Panel a) shows standard and conditioned fits when the trigger event is included (see equations (1.5) and (1.7)), and Panel b) shows results from standard and conditioned fits when it is excluded (see equations (1.6) and (1.8)). Vertical dotted lines show the estimated return periods for the event in Vargas in December 1999.

such that

$$\mathbb{P}(X > \eta_p) = \frac{1}{p\lambda}, \tag{1.19}$$

where λ is the average number of exceedances per year, so $p\lambda$ is the average number in a p-year period. This allows us to interpret the GPD quantile η_p as the p-year return level.

The stopping rule here is ill-defined, so for illustration we took the trigger event to be the first crossing of the historical 100-year return level computed using the first two decades of data. The standard and conditioned fits with and without the extremes from December 1999 are displayed in Figure 1.6. When the trigger event is included, the return time for the 1999 event is 464 (95% CI [352, 647]) years for the standard log-likelihood fit but 882 [597, 1447] for the full conditional log-likelihood fit (Figure 1.6a). The standard fit changes considerably when the trigger event is excluded, and the return period for the Vargas 1999 event is multiplied by 2.6 to become 1207 [766, 2182] years, but the full conditional results change little except for the upper confidence bound, which increases faster with the return period (Figure 1.6b). The uncertainty range for every return period computed is very wide.

An alternative analysis fits the GEV distribution to the annual maxima of daily precipitation values using the same stopping rule. There are fewer annual maxima than exceedances of 12 mm, so each has a larger influence on the fitted model, as we see in Figure 1.7, where the standard fit including the extreme event has a heavier upper tail

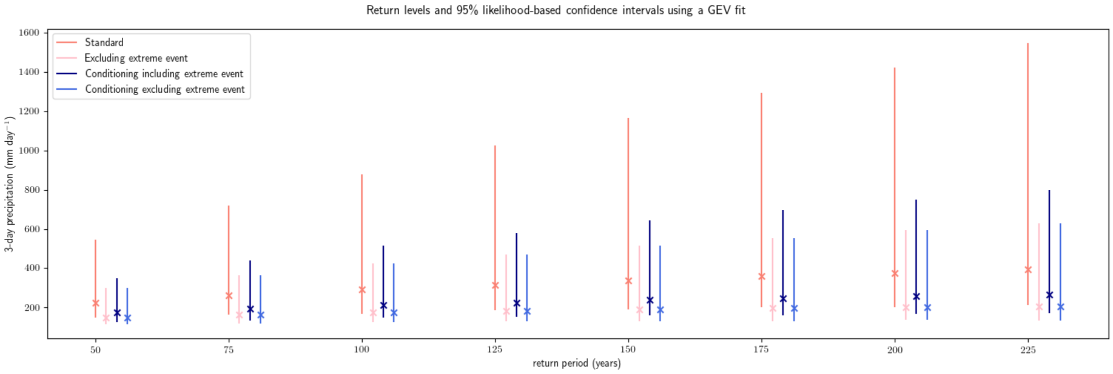


Figure 1.7: Vargas data analysis. Estimated p-year event η_p using a GEV fit. The x-axis represents the return period in years p. Marks represent the estimated p-year event, vertical bars denote 95% likelihood-based confidence intervals. Different colors represent results from different fits.

than the other fits. The return level for the 1999 extreme event using a GEV fit is comparable to using a GPD fit when the likelihood is conditioned, but the unconditioned GEV fit predicts a 250-year return level for the Vargas event, around 200 years shorter than the prediction using a GPD; see Coles and Pericchi (2003), Coles et al. (2003) and the Supplementary Material.

Return levels and return periods for a flood as extreme as the trigger estimated with a conditioned likelihood that includes the trigger event are quite different from their unconditioned analog, especially using a block maxima approach. Return level estimates based on the usual likelihood including and excluding the extreme event are very different, whereas inferences based on the conditional likelihoods with or without the trigger event are more stable.

1.5.2 2016 heatwave in Phalodi, India

The importance of accounting for timing bias can be seen by reconsidering the attribution analysis for the 2016 heatwave in Phalodi, India, which had disastrous public health consequences. Data sources and methods are detailed in van Oldenborgh et al. (2018), though here we compute likelihood-based confidence intervals rather than use a bootstrap. The Phalodi series is not available in the GHCN-D dataset, but sufficiently complete annual maximum temperature time series are available at two nearby stations, Jodhpur and Bikaner, and we analyse these as a proxy for data at Phalodi. Our findings for a standard fit of the Jodhpur series with a time-related trend, shown in Figure 1.8, are similar to those in van Oldenborgh et al. (2018). The location parameter of the fitted GEV distribution slightly decreases over time (Figure 1.8a) and a risk ratio (see Naveau et al., 2020, for a definition) of 0.511 is found for the occurrence of the trigger event in 2016 relative to 1973 (Figure 1.8b). The heterogeneity in the Jodhpur time series could make the slightly negative trend in the location parameter very sensitive to the three observations between 1940 and 1960 (Figure 1.8a). However, our aim is to reproduce as closely as possible the work of van Oldenborgh et al. (2018) in order to compare findings when accounting or not for timing and spatial selection bias. When fitting the Jodhpur time series with the fully conditioned log-likelihood (1.7) and a trend in the location parameter, the estimated risk ratio decreases from 0.511 to 0.4. Return periods for the heatwave as in 1973 and 2016 are given in Figure 1.8b for the standard fit and in Figure 1.8c for the conditioned fit, although they are very uncertain. The return period for a similar heatwave with the standard fit is 26 (95% CI [14, 150]) years in 1973 and 51 [26, 91] years in 2016. Conditioning slightly increases both return periods and the width of the 95% confidence interval, to reach respectively 32 [15, 302] and 80 [31, 147] years in 1973 and 2016.

Our analysis was performed in two steps, using the fact that the temperature time series at Jodhpur is more complete than that at Bikaner and contains the 2016 extreme event (van Oldenborgh et al., 2018), whereas Bikaner is closer to Phalodi. The stopping rule is defined as in Equation (1.15). The first step was an extremal analysis using only the Jodhpur series of annual temperature maxima, TXx. The return levels estimated from univariate fits based on (1.5)–(1.8) are shown in Figure 1.9. Those obtained using the standard likelihood (1.5) and including the trigger event are higher than for the other fits, with much higher upper confidence limits. To illustrate how allowing for timing bias can stabilise estimation, we extend the Jodhpur time series with later data and recompute return levels using standard and conditional likelihoods. The latter involves conditioning up to the trigger event year, while standard likelihood contributions are used for data after 2016. The full conditional fits before 2016 use the log-likelihood function (1.8), since the stopping rule has not yet been applied. Figure 1.10 shows that using the standard log likelihood (1.5) results in a jump in the predicted return levels after the extreme 2016 heatwave, followed by a slow decrease, whereas those from the conditional fits are more stable.

In a second step, we attempt to estimate the return level in Bikaner, where the extreme is not directly observed, by using a logistic copula (1.3) to model the dependence between the annual maximum temperatures there and at Jodhpur. Figure 1.18 of the Supplementary Material compares contours of the fitted joint density and cumulative

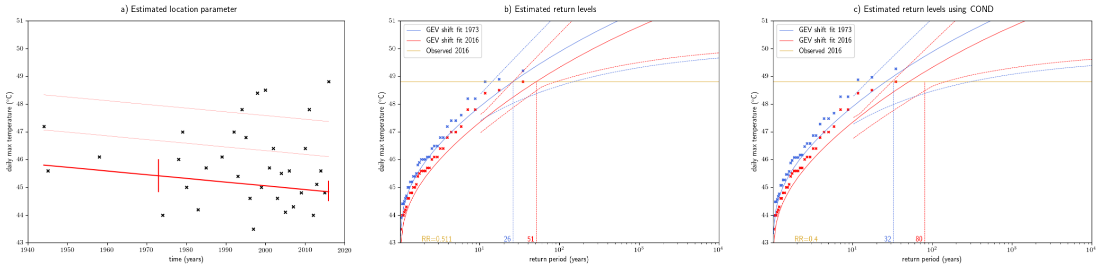


Figure 1.8: Phalodi heatwave analysis. Panel a) shows the estimated location parameter $\hat{\mu}_t$ of the GEV fit (red line) for annual temperature maxima (blue points) at Jodhpur for 1944–2016. The vertical red lines show 95% profile likelihood confidence bounds for μ_t in 1973 and 2016, and the thin red lines denote $\hat{\mu}_t + \hat{\sigma}$ and $\hat{\mu}_t + 2\hat{\sigma}$, where $\hat{\sigma}$ is the estimated scale parameter. Panel b) displays return level estimates for 1973 (solid blue) and 2016 (solid red) and their 95% confidence intervals (dotted). The observations are shown twice, scaled with the time-related trend (blue and red points). The golden horizontal line represents the extreme temperature observed in Jodhpur in 2016 (48.8°C), which which return periods in 1973 (blue) and 2016 (red) are shown by vertical dotted lines. Panel c) reproduces plots from Panel b using the COND log-likelihood (1.7).

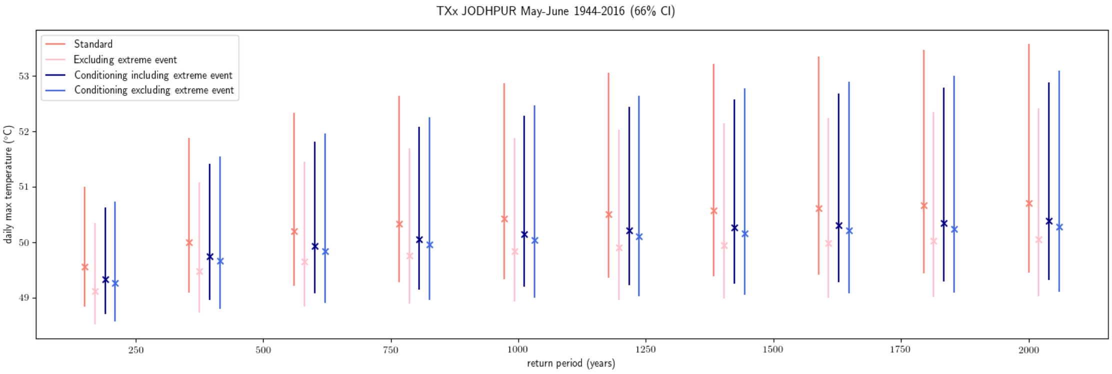


Figure 1.9: Phalodi heatwave analysis. Estimated p-year event η_p using GEV fits. The x-axis represents the return period in years. Marks represent the estimated p-year event, for $p = 200, 400, \ldots, 2000$, and vertical bars denote 66% likelihood-based confidence intervals.

distribution functions with the maxima.

A stopping rule for the Jodhpur series is then defined using the principle described in Section 1.4. The parameter α for the logistic copula (1.3), assumed constant, is estimated. Figure 1.11 shows how the estimated return levels in Bikaner change over time. Assuming independent data yields lower return level estimates, while using the dependence with the Jodhpur series to incorporate the extreme event into the prediction is stable over time only when the full conditional likelihood is used, as the jump seen in 2016 in the Jodhpur series in Figure 1.10 is also visible in the Bikaner return levels estimated with the standard likelihood; see Figure 1.11.

The analysis of the Jodhpur TXx series shows that not accounting for timing bias leads

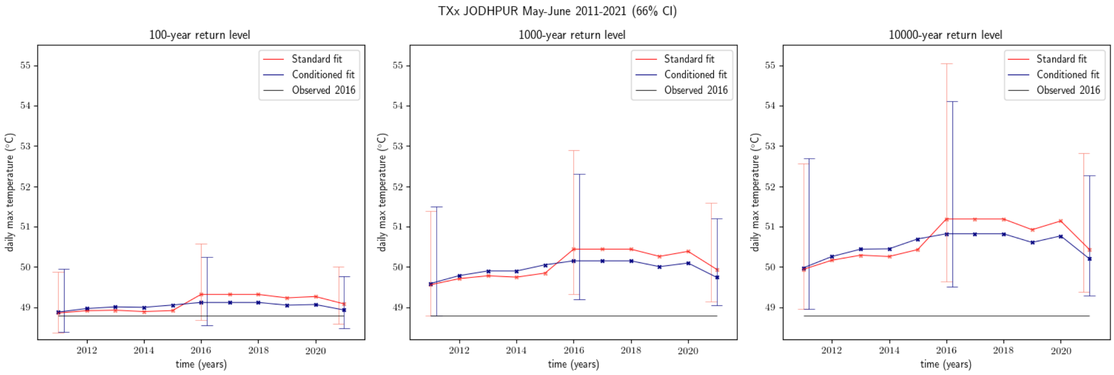


Figure 1.10: Phalodi heatwave analysis. Estimated return levels and their 66% confidence intervals (vertical lines finishing with ticks) with the standard and conditioned univariate fits for three return periods for 2011–2021 for TXx in Jodhpur. The horizontal black line indicates the extreme observed in Jodhpur in 2016 ($48.8°C$).

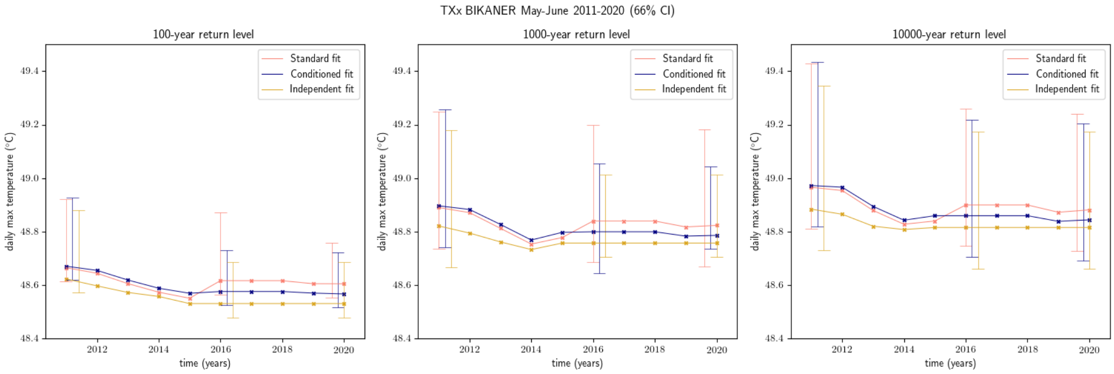


Figure 1.11: Phalodi heatwave analysis. Estimated return levels with an independent standard fit (golden line) and with the standard (navy) and conditioned (salmon) fit with a logistic correlation structure for three return periods throughout the 2011-2021 period for TXx in Bikaner.

to a jump in return level estimates that dissipates slowly for several years after a trigger event, whereas appropriate conditioning avoids this. For bivariate time series, using even a very basic correlation model instead of assuming independence has a huge impact on return level estimates, and accounting for timing bias prevents the bias transfer in return level estimation from the stopped series to the nearby series.

We now discuss the impact of spatial selection (Section 1.3). The logistic copula (1.3) has $\chi = S^{\alpha}$, and if we assume that we would have performed a similar analysis had an equally extreme event been observed in 2016 at Bikaner rather than at Jodhpur, then $S = 2$ and $\hat{\chi} = S^{\hat{\alpha}} \approx 1.43$. This lies between $\chi = 1$, which would correspond to total dependence between extremes at Bikaner and Jodhpur, and $\chi = 2$, which would correspond to independence. Under this argument the return period of 51 years found in Figure 1.8 for the event at Jodhpur, with this location specified before

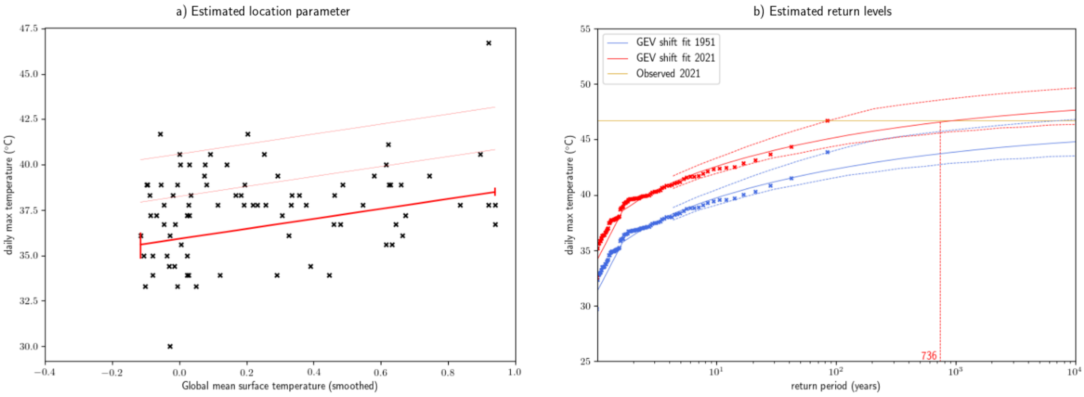


Figure 1.12: Portland heatwave analysis. Panel a) shows the location parameter $\hat{\mu}_t$ of the GEV fit (red line) to annual temperature maxima (blue points) at Portland for 1938–2021, with 95% profile likelihood confidence bounds (vertical red lines). The thin red lines denote $\hat{\mu}_t + \hat{\sigma}$ and $\hat{\mu}_t + \hat{\sigma}$, where $\hat{\sigma}$ is the estimated scale parameter. Panel b) displays return level estimates for the years 1951 (solid blue) and 2021 (solid red) and their 95% profile likelihood confidence intervals (dotted). The observations are shown twice, scaled with the time-related trend (blue and red points). The golden horizontal line represents the extreme temperature observed in Portland in 2021 (46.7°C). The return period estimate for this event in 1951 (blue) is shown by a vertical dotted line.

the event occurred, reduces to around 36 years for such an event at one of the two locations, using either the exact formula $m_\chi = 1/\left\{1-(1-1/m)^\chi\right\}$ given by (1.17) or the approximation $m_\chi \approx m/\chi$.

1.5.3 2021 heatwave in the Pacific Northwest

Our third re-analysis concerns the unprecedented “heat dome” event in the United States and Canada in June 2021, which led to wildfires that resulted in the inhabitants of the Canadian town of Lytton becoming climate refugees within a couple of days.

We use the Portland International Airport series of observed daily temperatures from the GHCN-D dataset to reproduce part of the attribution study of Philip et al. (2022), wherein data sources and methods are detailed, though we compute likelihood-based confidence intervals rather than use a bootstrap. The chosen stopping rule is the same as in the Phalodi case study; see (1.15). An increasing trend in the location parameter as a function of the global temperature anomaly (data from NASA-GISS) found in Philip et al. (2022) is shown in Figure 1.12a, and Figure 1.12b shows return levels using the standard log-likelihood and their 95% confidence intervals. The return period for the 2021 event is displayed in Figure 1.12b, though the prediction is very uncertain (95% confidence bounds are available in Table 1.1).

Table 1.1 shows how timing bias affects risk ratio estimation. The historical and current probabilities of crossing the previous TXx record of 41.7°C and its confidence interval are computed by fixing the location parameter of the GEV to the historical/current value for the linear trend in the global temperature anomaly, which is the approach of the WWA (Hammerling et al., 2019). We then parametrize the GEV in terms of the probability of exceeding this level and fit using the different log-likelihoods. Confidence intervals for the risk ratio were obtained using the delta method on its logarithm; see the Supplementary Material. A similar computation applies to the return period for the 2021 Portland event, with the GEV parametrized in terms of its return period and confidence intervals obtained using the profile likelihood.

Including the extreme event without conditioning yields a much shorter return period for the 2021 Portland temperature of 46.7° than when using conditioning, but the latter somewhat increases the risk ratio (Table 1.1); note that the confidence intervals for the risk ratio based on the standard and conditional fits do not overlap. Excluding the trigger event makes it impossible to estimate its return period, and the estimated risk ratio is less than half that computed by including this event; the same applies for conditional analysis without the trigger event.

Table 1.1: Comparison of estimated risk ratios p_1/p_0 and current return periods for the extreme 2021 temperature $\mathbb{P}(\text{TXx}_{2021} > 46.7°\text{C})$ for different log-likelihoods used to fit the Portland TXx series. The factual probability is defined as $p_1 = \mathbb{P}(\text{TXx}_{2021} > u)$ and the counterfactual (or pre-industrial) probability as $p_0 = \mathbb{P}(\text{TXx}_{1951} > u)$ for an extreme threshold u, here taken to be the previous record of 41.7°C. Also given are 95% confidence intervals for the risk ratio and the return period for the 2021 event.

	Risk ratio	**Return period for Portland 2021 (years)**
Standard	3.31 [3.20, 3.44]	736 [147, 5744]
Excluding	1.41 [0.94, 2.10]	∞ [∞, ∞]
Cond	3.77 [3.68, 3.86]	1830 [183, 16987]
CondEx	1.51 [1.06, 2.14]	∞ [∞, ∞]

1.6 Discussion

Our results in Sections 1.4.2 and 1.4.3 imply that in both univariate and bivariate settings it is generally better to exclude the trigger event if a conditioned fit is not used. In the univariate simulation framework with fixed sample size, the relative bias and relative root mean squared error reduce for $\tau \geq 80$ if the trigger is excluded (Figure 1.2). Fitting using a conditional log-likelihood always gives less biased return level estimates, even if the trigger event is not very extreme: simulations for both univariate and bivariate data show much lower bias using the conditioned log-likelihood function for $\tau \leq 200$, and it is increasingly important to use an appropriate likelihood when the trigger event becomes more extreme (see the results for $\tau > 500$ in Section 1.4). Table 1.1 suggests that although it depends heavily on the trigger event, the estimated risk ratio is much more stable, presumably because it contrasts two probabilities that are typically positively correlated; the same can be expected for functions of the risk ratio, such as the fraction of attributable risk.

The results of Section 1.4.4 suggest that attributing a return period to a specific observation should if possible be avoided, but if this is essential then the observation itself should be excluded from the fit, which should be performed using a conditional log-likelihood; an uncertainty statement should be included. In any case, the ratio of the estimated and true return periods for a single large observation is extremely uncertain. When the estimated shape parameter $\hat{\xi}$ of the extremal distribution is negative, as often arises for temperature data (see Sections 1.5.2 and 1.5.3), the return periods for certain future events may be infinite (see Table 1.1). This highlights another limitation of the statistical method: when $\hat{\xi} < 0$, excluding the trigger event may make this event effectively impossible. Including the extreme event is then preferable to excluding it, and applying appropriate conditioning will provide roughly unbiased (but very variable) results.

We now summarise the issues that our work raises for the choice of the statistical model for event attribution under an implicit stopping rule.

1. Potential timing bias may be suggested by time series in which the last value is rather unusual.
2. The stopping rule may be difficult to formulate precisely: if obtaining a suitable quantitative definition of an extreme event is impossible, it will be necessary to assemble contextual evidence about what is seen as extreme in the given context and to use that to guess a stopping rule for use in sensitivity analyses.
3. Accounting for timing bias by fitting the data with a conditional log-likelihood is generally desirable, but if for some reason a standard log likelihood must be used, then it is better to exclude the trigger event.
4. A multivariate extremal model allows the analyst to assess the potential effects of spatial selection in the analysis of several related series.

5. Return period estimation for the trigger event can be biased and very uncertain and thus should be avoided, but if it is required then some indication of its uncertainty is essential.

Further work could explore sensitivity analysis on a set of plausible stopping rules with varying thresholds and historical sample sizes. This paper concerns EEA studies that use observations in combination with possibly non-stationary extreme value distributions to estimate return levels, and our simulation studies and examples are specific to the timing bias problem using extreme value models. However, the general framework described in Section 1.2 could be used for conditioning any type of event with any distribution. Although conceptually straightforward, numerical aspects may become problematic when the computing the probability of the stopping event is complex; see the Supplementary Material. Further work could address selection biases relevant to other EEA methodologies.

1.7 Conclusion

Existing work on overcoming timing or spatial selection bias in extreme-value statistics has implications for return-level-based extreme event attribution analysis. Indeed, when such a bias exists, not taking it into account in the event attribution can lead to poor, unstable, return level estimates, seriously biased estimates of return periods for extreme observations, and hence to potentially misleading conclusions. Conditioning of the likelihood term uses contextual information more appropriately and hence leads to more reliable findings.

1.8 Data statement

The simulations and real case studies in Vargas, Phalodi, and Portland are implemented in Python code available on GitHub (https://github.com/OpheliaMiralles/timing–bias–extremes). The open-source package `pykelihood` was used for implementation of inference using stopping rule (https://github.com/OpheliaMiralles/pykelihood). A pipeline for downloading and processing the Phalodi data can be found in the same GitHub repository. Daily maximum temperatures were obtained from the NOAA publicly accessible dataset. The Vargas precipitation data may be found in the R package `mev`.

1.9 Supplementary material

1.9.1 Data-dependent thresholds

Suppose that $X_1,\dots,X_T$ are independent with common density and distribution functions f and F, that a series of varying thresholds $\eta_1,\dots,\eta_T$ is used, and let $\mathscr{A}_t$ denote the event $X_t \le \eta_t$ $(t=1,\dots,T)$. Then the stopping event $\mathscr{E}$ can be expressed as $\mathscr{A}_1 \cap \cdots \cap \mathscr{A}_{T-1} \cap \mathscr{A}_T^c$, with $\mathscr{A}_T^c$ the complement of $\mathscr{A}_T$, and thus

$$\Pr(\mathscr{E}) = \Pr(\mathscr{A}_1) \times \prod_{t=2}^{T-1} \Pr(A_t \mid \mathscr{A}_1 \cap \cdots \cap \mathscr{A}_{t-1}) \times \Pr(\mathscr{A}_T^c \mid \mathscr{A}_1 \cap \cdots \cap \mathscr{A}_{T-1}). \qquad (1.20)$$

If the events $\mathscr{A}_1,\dots,\mathscr{A}_T$ are independent, then this expression reduces to

$$\Pr(\mathscr{E}) = \prod_{t=1}^{T-1} \Pr(A_t) \times \Pr(\mathscr{A}_T^c),$$

and the conditional log-likelihood of $x_1,\dots,x_T$ given $\mathscr{E}$ is then (1.7). This simplification applies if the η_t do not depend on previous values $x_1,\dots,x_{t-1}$, but are estimated from unrelated data, for instance if the η_t vary seasonally according to a time series model fitted to the bulk of the available observations. If on the other hand the events $\mathscr{A}_t$ are dependent, as would be the case if the $\eta_t \equiv \eta_t(X_1,\dots,X_{t-1})$ depend on recent extremes, then the more general expression (1.20) should in principle be used. Consider $\Pr(\mathscr{A}_2 \mid \mathscr{A}_1)$, for example, which equals

$$\int_{-\infty}^{\eta_1} \Pr\{X_2 \le \eta_2(x_1) \mid X_1 = x_1\}\{f(x_1)/F(\eta_1)\}\,\mathrm{d}x_1 = F(\eta_1)^{-1} \int_{-\infty}^{\eta_1} F\{\eta_2(x_1)\} f(x_1)\,\mathrm{d}x_1;$$

more complex expressions apply for further terms.

When the η depend on recent extremes it is tempting to replace the T-dimensional integral in (1.20) by

$$\Pr(\mathscr{A}_1) \times \prod_{t=2}^{T-1} \Pr(A_t \mid X_1 = x_1,\dots,X_{t-1} = x_{t-1}) \times \Pr(\mathscr{A}_T^c \mid X_1 = x_1,\dots,X_{T-1} = x_{T-1}),$$

and thus replace $\Pr(\mathscr{E})$ by

$$F(\eta_1) \times \prod_{t=2}^{T-1} F\{\eta_t(x_1,\dots,x_{t-1})\} \times \left[1 - F\{\eta_T(x_1,\dots,x_{T-1})\}\right].$$

The consequences for the estimation of the parameter vector θ from the resulting approximate conditional log-likelihood are unknown.

1.9.2 Further simulation results

Figure 1.13 gives further simulation results that reproduce those in Barlow et al. (2020), in which the sampling stops at random times, rather than at a fixed time as in the present paper.

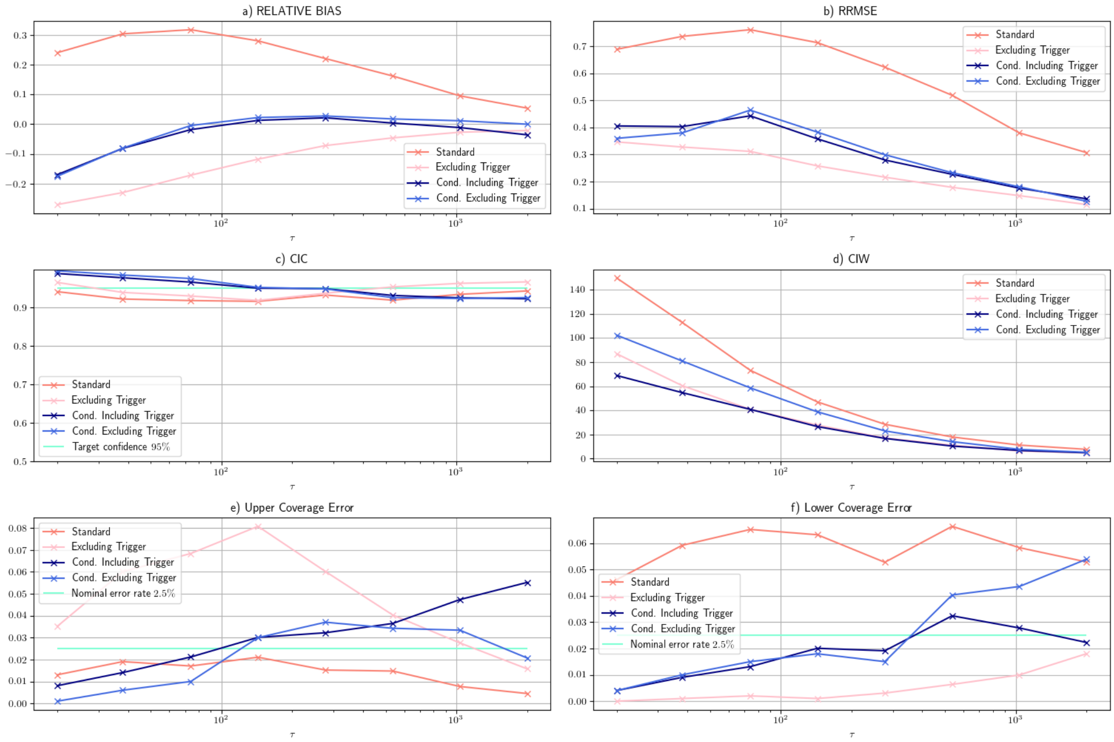


Figure 1.13: Summary results for the estimation of a 200-year return level based on simulated GEV random variables with shape parameter 0.2, with stopping thresholds defined from the return periods τ shown on the x-axis as in Equation (1.18). The relative bias and mean squared error are shown in Panels a and b, and coverage of 95% confidence intervals and their average widths are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors. Here the time series stops at a maximum sample size of $N = 800$.

As a complement to Figure 1.2 of the paper, Figures 1.14 and 1.15 give simulation results for truncated sampling with shape parameters $\xi = 0$ and $\xi = -0.2$.

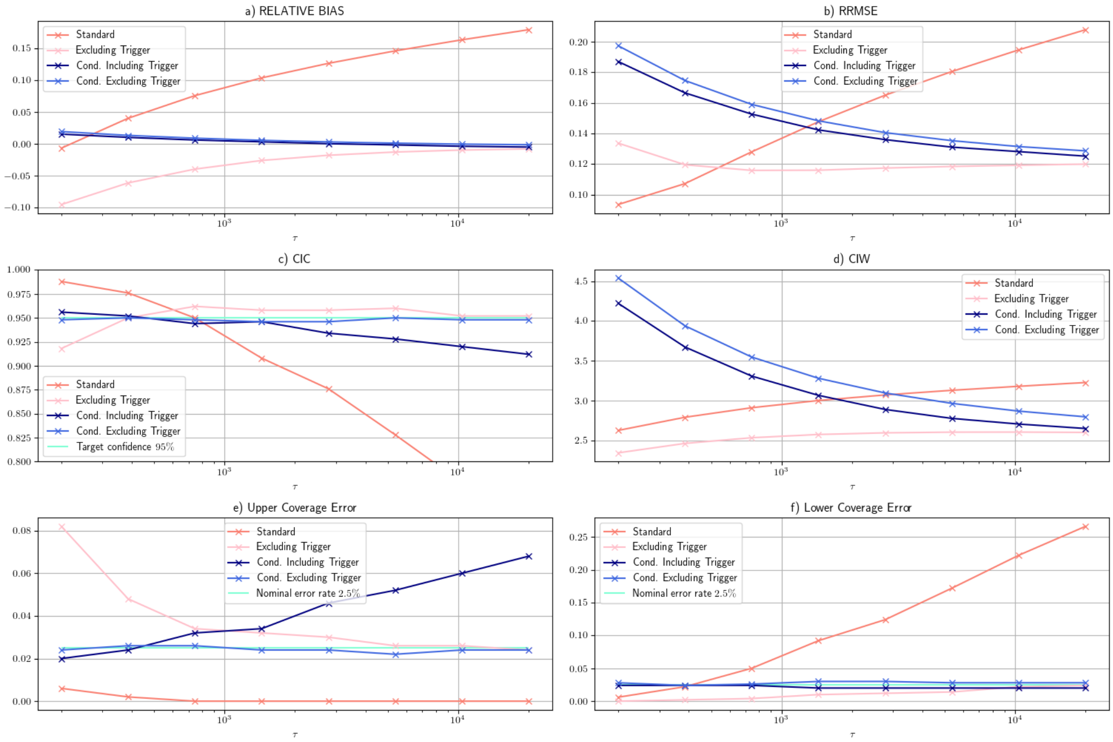


Figure 1.14: Summary results for the estimation of a 200-year return level based on simulated GEV random variables with shape parameter $\xi = 0$, with stopping thresholds defined from the return periods τ shown on the x-axis as in Equation (1.18). The relative bias and mean squared error are shown in Panels a and b, and coverage of 95% confidence intervals and their average widths are shown in Panels c and d. Panels e and f represent the upper and lower coverage errors. In this situation, the time series is generated so that the first exceedance of the stopping threshold occurs at a specified time.

1.9.3 Supplementary material related to case studies

Additional figures for the Vargas and Phalodi analyses

Figure 1.16 and Figure 1.17 supplement the study of the 1999 Vargas event (Section 1.5.1), respectively by summarising methods used to select the threshold for the GPD and by providing similar analysis results than in Figure 1.6 fitting a GEV distribution to annual maxima. Figure 1.18 provides a visual representation of the bivariate fit used in Section 1.5.2 for the study of the Indian heatwave.

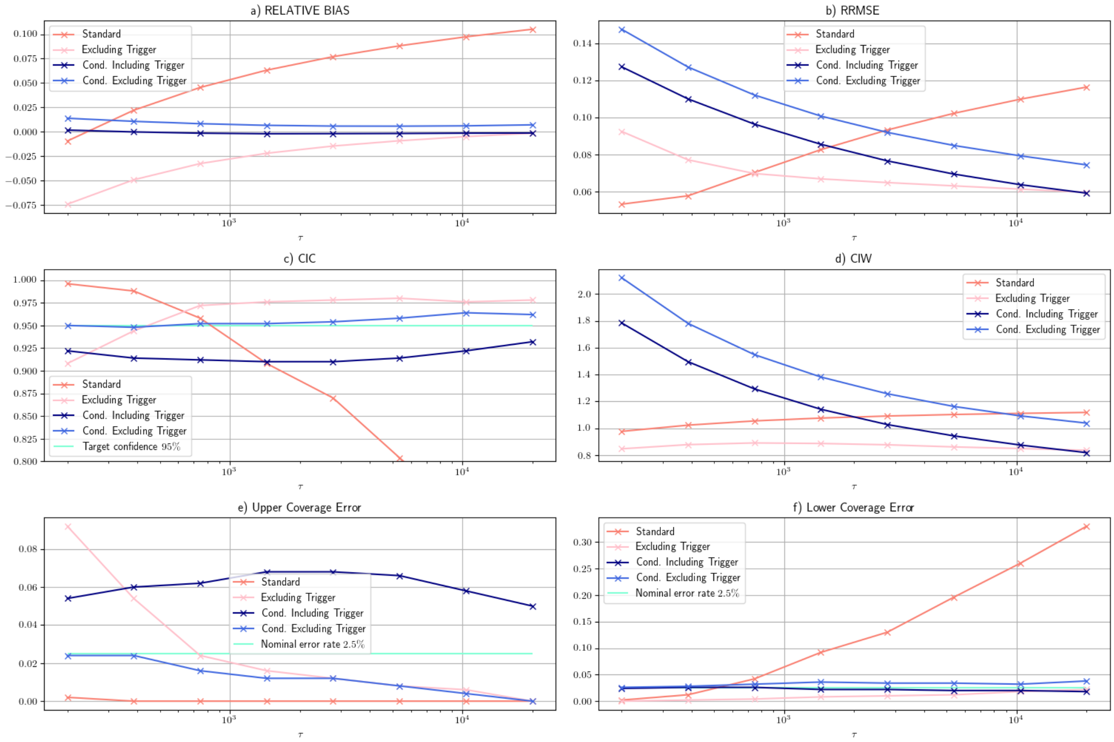


Figure 1.15: Summary results for the estimation of a 200-year return level based on simulated GEV random variables with negative shape parameter −0.2. See the caption to Figure 1.14.

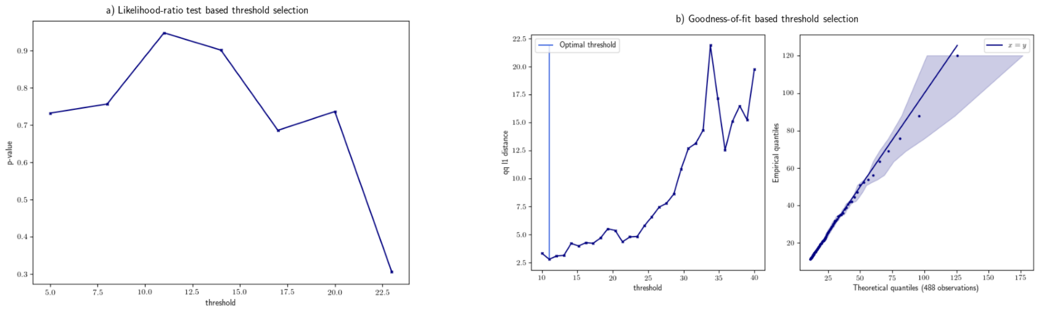


Figure 1.16: Vargas data analysis. Threshold selection: a) graphical threshold selection (Northrop and Coleman, 2014) for daily precipitation maxima between 1961 and 1999, based on thresholds 5, 8, . . . , 20, 23. b) Goodness-of-fit based threshold selection (Varty et al., 2021); the left-hand panel shows the d^Q distances for different thresholds, with the optimal threshold shown by a vertical bar and the right-hand panel shows a QQ-plot comparing the data above the optimal threshold and the fitted generalized Pareto distribution.

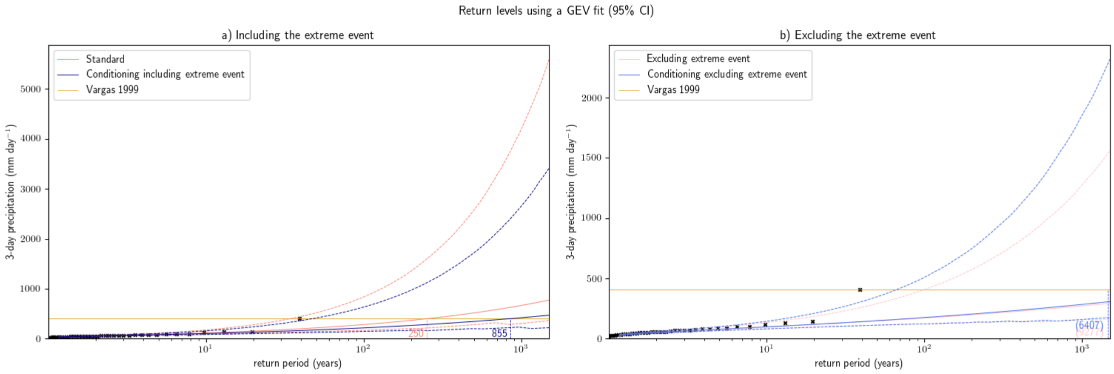


Figure 1.17: Vargas data analysis. Estimated p-year event η_p (see Equation (1.19)) and its 95% confidence interval (dashed lines) using a GEV fit. The x-axis represents the return period in years p. Different colors are used to represent different conditioning methods: Panel a) shows standard and conditioned fits for series that include the extreme event (see equations (1.5) and (1.7)), while Panel b) shows standard and conditioned fits for series that exclude the extreme event (see equations (1.6) and (1.8)). Vertical dotted lines show the return period for the extreme event in Vargas in December 1999.

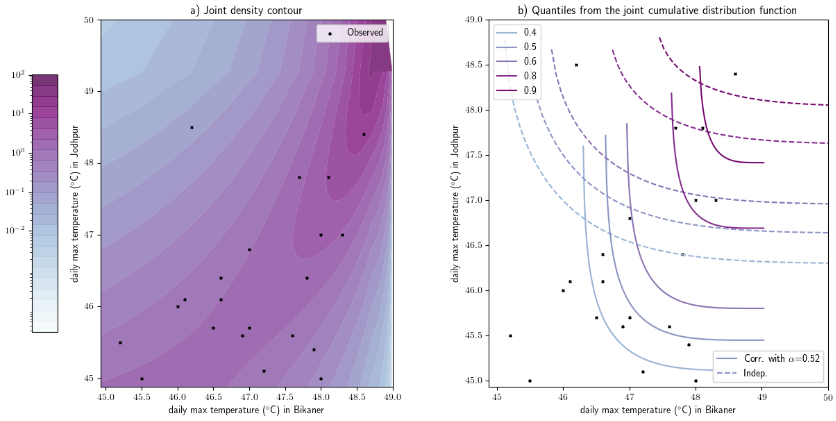


Figure 1.18: Phalodi heatwave analysis. Panel a) shows the density contour for the fitted joint logistic copula ($\hat{\alpha} = 0.52$) of the Jodhpur and Bikaner time series. Panel b) shows the joint cumulative distribution function, comparing an independent fit (dashed lines) and the logistic copula fit (thick lines). Data are represented by black marks.

Delta method for the Pacific Northwest heat dome event

To compute the confidence intervals for the risk ratio given in Table 1.1, we write the log risk ratio estimator as

$$\log \hat{p}_1 - \log \hat{p}_0,$$

where the circumflexes indicate maximum likelihood estimators. This difference has variance

$$\operatorname{Var}(\log \hat{p}_1) + \operatorname{Var}(\log \hat{p}_0) - 2\operatorname{Cov}(\log \hat{p}_1, \log \hat{p}_0).$$

In general $\hat{p}_1 = p_1(\hat{\theta})$, where θ denotes the parameters, so the chain rule and an application of the delta method give

$$\operatorname{Var}(\log \hat{p}_1) = \operatorname{Var}\{\log p_1(\hat{\theta})\} = (\nabla p_1)^T C (\nabla p_1)/p_1^2,$$

where C denotes the covariance matrix of $\hat{\theta}$ and $\nabla p_1 = \partial p_1(\theta)/\partial\theta$ is the gradient vector for p_1 with respect to θ. A similar expression holds for $\log \hat{p}_0$, and in the same notation we have

$$\operatorname{Cov}(\log \hat{p}_1, \log \hat{p}_0) \approx (\nabla p_1)^T C (\nabla p_0)/(p_0 p_1).$$

In this specific case an annual temperature maximum TXx series is fitted by a GEV with a trend $\mu = a + bg$ in the location parameter, where g is the global annual temperature anomaly. Estimates for p_0 and p_1 are derived from GEV distributions with location parameter fixed to the fitted location parameters in 1951, $\hat{a} + \hat{b}g_0$, and in 2021, $\hat{a} + \hat{b}g_1$. Hence $\theta = (a, b, \sigma, \xi)$, and the estimate $\hat{\theta}$ and covariance matrix $C \approx \operatorname{Cov}(\hat{\theta})$ are provided by the maximum likelihood fit to the data, while ∇p_1 contains the derivatives of $p_1 = 1 - \{1 + \xi(u - a - bg_1)/\sigma\}^{-1/\xi}$ with respect to the components of θ, and similarly for ∇p_0.

The variance of $\log \hat{p}_1 - \log \hat{p}_0$ is obtained by putting the above terms together, all evaluated at the maximum likelihood estimates. A confidence interval for $\log p_1 - \log p_0$, obtained using a normal approximation, is then exponentiated to provide the corresponding interval for p_1/p_0.

2 Downscaling of Historical Wind Fields over Switzerland using Generative Adversarial Networks

Ophélia Miralles[1], Daniel Steinfeld[2], Olivia Martius[2], Anthony C. Davison[1]

1 – Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

2 – Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Switzerland

This paper represents joint work between the École Polytechnique Fédérale de Lausanne (Anthony Davison and myself) and the Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Switzerland (Daniel Steinfeld and Olivia Martius). Daniel Steinfeld contributed to the writing of Section 3.2 (in particular Sections 2.3.2, and 2.3.1). Olivia Martius and Anthony Davison provided relevant and valuable feedback on the work and paper. I provided the rest of the work, i.e., the literature review, data processing, model design, training of the model, model validation, and diagnostics on the test set.

This chapter was published as a paper by the journal Artificial Intelligence for the Earth Systems in July 2022. It can be accessed via the link doi.org/10.1175/AIES-D-22-0018.1 and cited using the following bibliographical reference.

Miralles, O., Steinfeld, D., Martius, O., and Davison, A. C. (2022). Downscaling of historical wind fields over Switzerland using generative adversarial networks. *Artificial Intelligence for the Earth Systems*, 1(4):e220018

Abstract

Near-surface wind is difficult to estimate using global numerical weather and climate models, as airflow is strongly modified by underlying topography, especially that of a country such as Switzerland. In this article, we use a statistical approach based on deep learning and a high-resolution Digital Elevation Model to spatially downscale hourly near-surface wind fields at coarse resolution from ERA5 reanalysis from their original 25 km to a 1.1 km grid. A 1.1 km resolution wind dataset for 2016–2020 from the operational numerical weather prediction model COSMO-1 of the national weather service, MeteoSwiss, is used to train and validate our model, a generative adversarial network (GAN) with gradient penalized Wasserstein loss aided by transfer learning. The results are realistic-looking high-resolution historical maps of gridded hourly wind fields over Switzerland and very good and robust predictions of the aggregated wind speed distribution. Regionally averaged image-specific metrics show a clear improvement in prediction compared to ERA5, with skill measures generally better for locations over the flatter Swiss Plateau than for Alpine regions. The downscaled wind fields demonstrate higher-resolution, physically plausible orographic effects, such as ridge acceleration and sheltering, which are not resolved in the original ERA5 fields.

2.1 Statement of interest

Statistical downscaling, which increases the resolution of atmospheric fields, is widely used to refine the outputs of global reanalysis and climate models, most commonly for temperature and precipitation. Near-surface winds are strongly modified by the underlying topography, generating local flow conditions that can be very difficult to estimate. This study develops a deep learning model that uses local topographic information to spatially downscale hourly near-surface winds from their original 25 km resolution to a 1.1 km grid over Switzerland. Our model produces realistic high-resolution gridded wind fields with expected orographic effects but performs better in flatter regions than in mountains. These downscaled fields are useful for impact assessment and decision-making in regions where global reanalysis data at coarse resolution may be the only products available.

2.2 Introduction

Near-surface wind fields are of interest in applications such as wind energy projects (Emeis, 2014; Staffell and Pfenninger, 2016; Dujardin et al., 2021), risk and damage assessment for intense windstorms (Schwierz et al., 2010; Stucki et al., 2014; Welker et al., 2016; Stucki et al., 2016), snow distribution and avalanche forecasting (Lehning et al., 2008) and modeling the spread of wildfires (e.g., Sharples et al., 2012). Detailed wind information at high spatial and temporal resolution and for long time periods is needed to study wind-related impacts, space-time variability, and long-term trends, but the accurate representation of surface winds in complex terrain is challenging because winds fluctuate over a wide range of time and spatial scales. Surface weather stations provide accurate and long-term local wind measurements, but are sparsely distributed and the spatial interpolation of wind between them is difficult (Kruyt et al., 2017; Harris et al., 2020). Climate and weather prediction models provide spatially and temporally continuous gridded wind data that are physically consistent, but the observed wind field varies at much smaller spatial scales than those in global versions of such models (Koller and Humar, 2016; Molina et al., 2021), whose grid resolutions range from tens to hundreds of kilometers and at best resolve only major topographical features. Models at these resolutions do not capture local flow effects such as wind speed-up over ridges, flow channeling in valleys, flow deflection around and over mountain ranges, and thermally induced winds that alter the local flow field.

Reanalysis datasets produced by global weather prediction models, such as the state-of-the-art ERA5 reanalysis from the European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020), provide long-term gridded wind fields on a global scale, but their coarse spatial resolution (~25 km) limits their use for impact assessment in complex terrain. Although large-scale atmospheric flow conditions associated with surface winds are broadly well-represented in reanalysis datasets (Molina et al., 2021), especially over flat regions (Ramon et al., 2019), such data are too coarse to accurately represent local surface wind conditions in regions with complex terrain, such as the Swiss mountains (Graf et al., 2019; Dörenkämper et al., 2020). The horizontal grid resolution in global reanalyses is relatively coarse, in part due to their high computational demands. On the other hand, high-resolution numerical model data are available, but typically only for short time periods. The regional operational weather prediction model COSMO-1 (MeteoSwiss, 2016) of the Swiss weather service has been successfully run at a grid resolution of 1.1 km over Switzerland since 2016, producing realistic representations of local wind conditions, but no long-term (>5 years) gridded climatology for wind exists (MeteoSwiss, 2018). Hence there is a trade-off between geographic coverage and time span on the one hand and spatial detail on the other. This data gap can be filled by applying downscaling methods to long-term historical reanalysis and climate model outputs (Gutowski Jr. et al., 2016). This motivates the development of a downscaling technique to produce a gridded near-surface wind climatology at higher spatial and temporal resolution.

Statistical downscaling must address the question of what is considered to be the *ground truth*. Most statisticians would agree that field observations are a noisy version of the truth, whereas physicists tend to attribute value to re-analyzing such data, correcting for measurement errors, and smoothing it to fit physical theory. A consequence of these considerations for downscaling is that researchers favor either point-by-point modeling and forecasting based on a limited number of observation stations (Winstral et al., 2017; Nerini, 2020) or mapping of low-resolution grids directly to high-resolution ones (Höhlein et al., 2020; Leinonen et al., 2021; Ramon et al., 2021).

Spatio-temporal regression models have been proposed for statistical downscaling (Winstral et al., 2017; Ramon et al., 2021), though they generally assume linear dependence and Gaussianity and often do not account for unobserved spatial phenomena. More complex statistical models have been avoided in the past because of the computational burden of dealing with very large datasets, which precludes applying the simulation-based methods widely used in other contexts. Latent variable models attempt to account for hidden or unobserved effects in high-dimensional data, and Gaussian processes can flexibly capture local correlations and uncertainties. Latent Gaussian models combine these concepts (Lawrence, 2003; Rue et al., 2009) and provide a large class of statistical tools. The R-INLA package (Rue et al., 2017) can estimate posterior distributions for latent Gaussian models, but the size of the latent field affects the complexity of precision matrix computation. Rue et al. (2017) argue that assuming Markov properties for the target process can greatly reduce the computational burden, and efficient solutions now exist for fitting multi-layer statistical models to large numbers of data points and have been used for environmental applications. For example, Castro-Camilo et al. (2019) use R-INLA to fit a hierarchical Bayesian model involving a biphasic distribution for extreme and non-extreme wind speeds at 260 stations across the United States. However, there is little to no literature on downscaling climate time series using such models. In our study, we use grid-to-grid downscaling to produce entire maps of wind fields. Although we considered using a spatio-temporal Bayesian hierarchical model, the very large number of data points

(~ 10 billion in total) was impossible to handle using R-INLA.

Instead, we implement deep learning methods that deal with very large amounts of data by introducing a network hierarchy that allows a computer to build complicated structures from simple ones (Goodfellow et al., 2016). This hierarchy is commonly described as a series of layers; the deeper the network, the more layers there are and the more specific the role of each layer. Downscaling atmospheric fields using neural networks is a very recent development (Vandal et al., 2018; Reichstein et al., 2019; Baño Medina et al., 2020; Sha et al., 2020). Machine learning methods for downscaling environmental variables can provide good results, avoid information loss, and require reasonable computational effort if the structure has enough hidden layers (Höhlein et al., 2020). However, neural networks are mainly used to produce deterministic outcomes, which is an issue if one wants to know the distribution of the target process. This can be overcome with a recurrent generative adversarial network (GAN) that adds noise to the original input in order to make predictions more robust, as proposed by Leinonen et al. (2021) for rainfall data. A more probabilistic approach is to use neural networks to estimate the parameters of a given statistical model, for instance by estimating the parameters of gamma distribution for wind speed data (Nerini, 2020). As far as we know, no existing neural network can efficiently downscale wind fields on complex terrain from different low- and high-resolution sources. In this paper, we propose a stochastic deep learning approach using a GAN to downscale historical maps of hourly near-surface wind fields over Switzerland from open-source ERA5 data and local topography. The target high-resolution maps are wind fields from the COSMO-1 model, provided by MeteoSwiss, which represent the local surface winds well. The resulting time series of downscaled wind fields can be used for detailed case studies of past weather events or climatological analyses.

This study is structured as follows. The data used for the downscaling and associated challenges are described in Section 2.3, and the specific deep learning model and its training are explained in detail in Sections 2.4 and 2.5. Quantitative analysis of the obtained predictions is performed in Section 2.6, and the main findings are given in Section 2.7.

2.3 Data

2.3.1 Geographical setting and typical wind systems in Switzerland

Switzerland has a complex and diverse topography with three main subregions (cf. Figure 2.1): the Alps in the central and southern part of the country with high mountain ranges and deep valleys, covering ~ 60% of Switzerland; the Jura in the north-western part with lower and narrow mountain ranges, covering ~ 10%; and, between them, the hilly and densely-populated Plateau, covering ~ 30%. The elevation ranges from below 300 m to above 4,500 m. Figure 2.1 shows how this topography is represented in the ERA5 reanalysis and in the COSMO-1 model, with respective horizontal grid resolutions of 25 km and 1.1 km. The ERA5 grid cannot resolve the complex mountain terrain, but the high mountain ranges and deep inner alpine valleys are well resolved in COSMO-1. This terrain interacts with and modifies the synoptic-scale flow at different scales, generating region-specific surface winds (Barry, 2008). At the larger (alpine) spatial scale, the frequent westerly winds are modified by the high mountains, for example, by horizontal and vertical deflection creating mountain waves, and by channeling of the flow (Jackson et al., 2013). A well-known example in Switzerland is the north-south Foehn flow, which crosses the main Alpine ridge and leads to a warm and dry downslope windstorm in the lee, affecting many Alpine valleys (Richner and Hächler, 2013; Sprenger et al., 2016). Another example is the Bise, an easterly wind that is enhanced in the Plateau region by channeling between the Jura and the Alps (MeteoSwiss, 2015). At the more local scale, thermally-driven diurnal mountain-valley winds are generated by temperature contrasts that form within the mountains and valleys due to radiative heating during the day and cooling at night (Weissmann et al., 2005; Zardi and Whiteman, 2013).

2.3.2 Low-resolution input fields: ERA5 reanalysis

We use the ERA5 reanalysis, the fifth generation of global reanalysis datasets from the European Centre for Medium-Range Weather Forecasts (ECMWF), which has a spatial resolution of 0.25° (≃ 25 km) and is available hourly from 1979 onwards. Long-term climate data sets such as this are built by assimilating observations from multiple data sources and solving the main atmospheric evolution equations, with

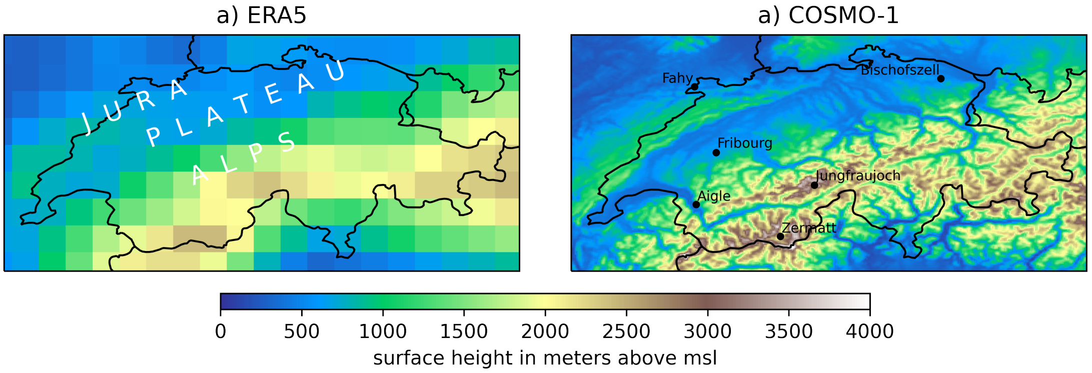


Figure 2.1: Maps of Switzerland showing topography in meters above sea level for ERA5 with 25 km resolution (left) and COSMO-1 with 1.1 km resolution (right). The three sub-regions of Switzerland are indicated on the left map, and the location of validation sites (see Figure B1 in Appendix) is indicated on the right map.

the aim of representing past or current climates on a regular grid (Hersbach et al., 2020). Cycle 41r2 of the Integrated Forecast System (IFS), the global numerical forecast model of the ECMWF, and 4D variational data assimilation of past observations were used to produce the ERA5 reanalysis, which is freely available through the EU-funded Copernicus Climate Change Service (C3S). It will eventually be extended back to 1950.

Low-resolution surface (10-meter) wind fields covering Switzerland are retrieved from the ERA5 reanalysis as predictors. These consist of gridded u (east-west) and v (south-north) hourly wind speed components on a horizontal grid of 0.25° ($\simeq$25 km) from 2016 to 2020. We tested additional predictors from ERA5, also used by Höhlein et al. (2020), in the hope of obtaining information about the local wind systems and driving processes described in Section 2.3.1: boundary layer height, surface pressure, forecast surface roughness and geopotential height at 500hPa. However, they did not improve the performance of the GAN and were not included in the final model.

2.3.3 Topographic descriptors

The terrain of Switzerland is complex: local topographic features strongly modify surface wind speeds, and to allow the GAN to learn this relationship we use the topography from the freely-available 90-meter resolution SRTM3 digital elevation model (DEM) constructed by NASA and NGA (Jarvis et al., 2008). We also tested a comprehensive set of DEM-derived descriptors, calculated using the Python package `topo-descriptors` (Nerini and Zanetta, 2021) provided by MeteoSwiss: directional (south-north and east-west) derivatives, slope and aspect, the ridge/valley norm and direction, and the Topographic Position Index (TPI), which evaluates a gridpoint's elevation relative to its surroundings (Winstral et al., 2017). However, best performance was reached with the raw DEM.

2.3.4 High-resolution target fields: COSMO-1

High-resolution 10-meter target fields are from the COSMO-1 (Consortium for Small-Scale Modelling) model. COSMO-1 is a non-hydrostatic deterministic limited-area numeric weather prediction model that is based on primitive, thermo-hydrodynamical equations describing compressible flow in a moist atmosphere (COnsortium for Small-scale MOdeling, 2017). MeteoSwiss has run COSMO-1 at a grid resolution of 1.1 km with the domain centered over Switzerland operationally since March 2016 (MeteoSwiss, 2016), which provides a little more than four years of hourly (reanalysis) data. Boundary conditions are provided by the ECMWF Integrated Forecasting System, which is also the global weather model underlying the ERA5 reanalysis. The performance of COSMO-1 was assessed against weather stations in Kruyt et al. (2018) and found to give good overall wind speed results. We use surface wind estimates from the COSMO-1 analysis provided by MeteoSwiss at hourly resolution from March 2016 to October 2020. The ERA5 and COMSO-1 10-meter wind components u at 00 UTC on 13 January 2017 and at 00 UTC on 4 March 2017 are compared in Figure 2.2.

2.4 Generative Adversarial Network (GAN)

Below we use the term "tensor" to refer to data provided as input to a neural network, data resulting from the transformation associated with a hidden layer of the network, or predictions made by the network. The network is fed with square frames, or "patches", that are randomly selected from the input map. To ensure stability and speed of training, we do not update the parameters for each observation, but process a "batch" of observations at a time. In this study, all tensors are of dimension five: the first dimension is the batch size, the second is the time coordinate, the third and fourth are the spatial coordinates, and the last, the "channels", refers to individual scalar variables.

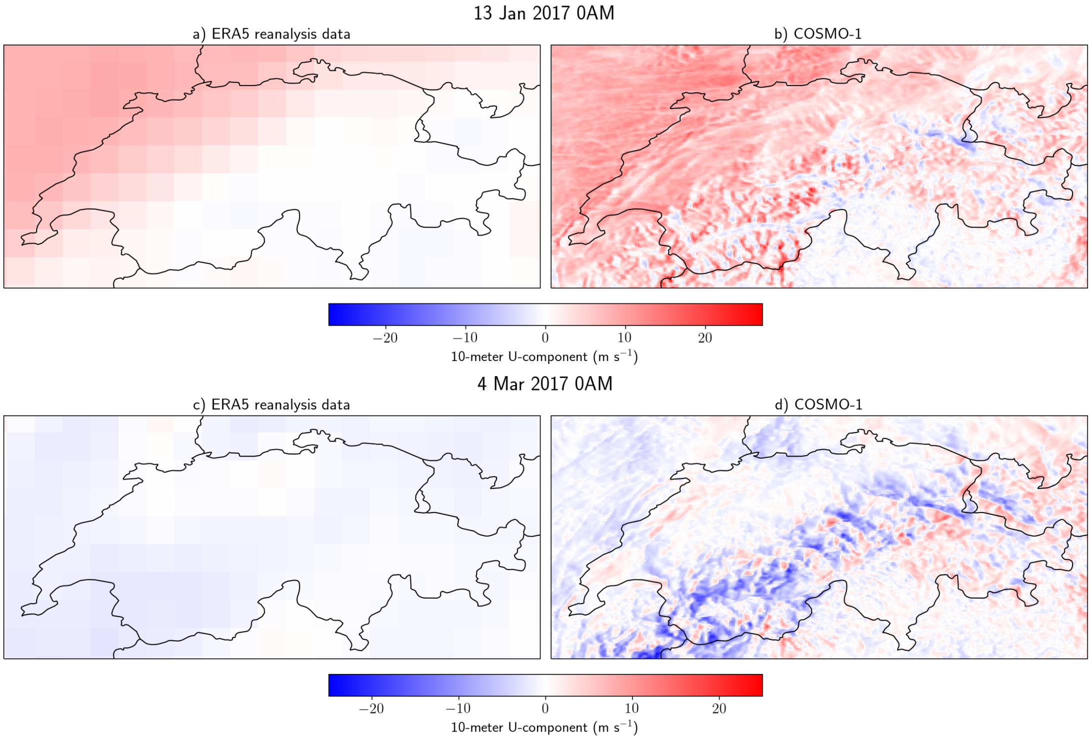


Figure 2.2: Examples of ERA5 reanalysis input 10-meter u wind component with resolution 25 km (a and c) and target 10-meter u wind component from COSMO-1 with resolution 1.1 km (b and d).

2.4.1 General architecture

The generative adversarial network we use has a standard Wasserstein GAN with gradient penalty (WGAN-GP) architecture, comparable to that used for precipitation data by Leinonen et al. (2021). Such a network comprises two different deep neural networks with specific roles. The generator network, or “artist”, takes low-resolution sequences of wind and other covariates as input, convolves and upsamples them through sequential layers, and produces two output images that are fitted to the high-resolution wind fields during training. The discriminator network, or “critic”, attributes a score that measures the match between the low-resolution input data and the high-resolution wind field prediction. Thus the purpose of the discriminator is not so much that predicted winds look exactly like COSMO-1 winds, but to attribute a score assessing the consistency of a pair of low/high-resolution winds. The score function is obtained after compressing the information in the low- and high-resolution wind fields into a scalar value through convolutional layers. The critic is optimized throughout the training to make its output score as discriminating as possible. The goal is to clearly distinguish between fake wind fields created by the generator and their real counterparts. Its optimal parameters are found by minimizing a gradient-penalized version of the Wasserstein loss (Gulrajani et al., 2017),

$$\mathrm{Loss}_D(x,y,z) = D\left(x,y\right) - D\left\{x, G\left(x,z\right)\right\} + \gamma\left\{\|\nabla_{\tilde{y}} D\left(x,\tilde{y}\right)\|_2 - 1\right\}^2, \tag{2.1}$$

where x is the low-resolution input tensor, y is the true high-resolution wind field, z is a noise field, D is the score given by the discriminator to a pair of low and high-resolution fields, and $G(x,z)$ is the prediction of the generator (the fake high-resolution wind field). The score is obtained by minimizing the loss function. The final term of equation (2.1) is the gradient penalty, whose influence is determined by the positive scalar γ, and which attracts the norm of the gradient toward unity. This term contains a random combination $\tilde{y} = \epsilon y + (1-\epsilon)G\left(x,z\right)$ of true and predicted wind fields, with ϵ a standard uniform random variable.

Scores attributed to both the true high-resolution and predicted winds should be robust in order that the discriminative ability of the network is reliable. The gradient term in equation (2.1) was introduced by Gulrajani et al. (2017) to enforce the 1-Lipschitz constraint on the discriminator's score relative to its inputs, but it also prevents gradient explosion at the start of the training, which is otherwise common when using deep structures (Huang et al., 2016). The artist's loss is simply the score given by the discriminator to the fake high-resolution output, i.e.,

$$\mathrm{Loss}_G(x) = D\left\{x, G\left(x,z\right)\right\}.$$

On the one hand, the critic should score unrealistic predictions as highly as possible so that the artist can improve, while reducing the score attributed to COSMO-1 high-resolution wind fields as much as possible. On the other hand, the optimum of the artist is reached when its loss is minimal, which means that the networks act on $D\left(x, G\{x,z\}\right)$ in opposite ways.

2.4.2 Modelling the wind time series

The wind time series of the two 10-meter wind components u and v from COSMO-1 present strong short-term autocorrelation (Figure 2.3a), which reduces to about 0.2 only after about 30 hours. To allow the artist to accurately reproduce this, we augment the generator network with a long short-term memory (LSTM) layer (Hochreiter

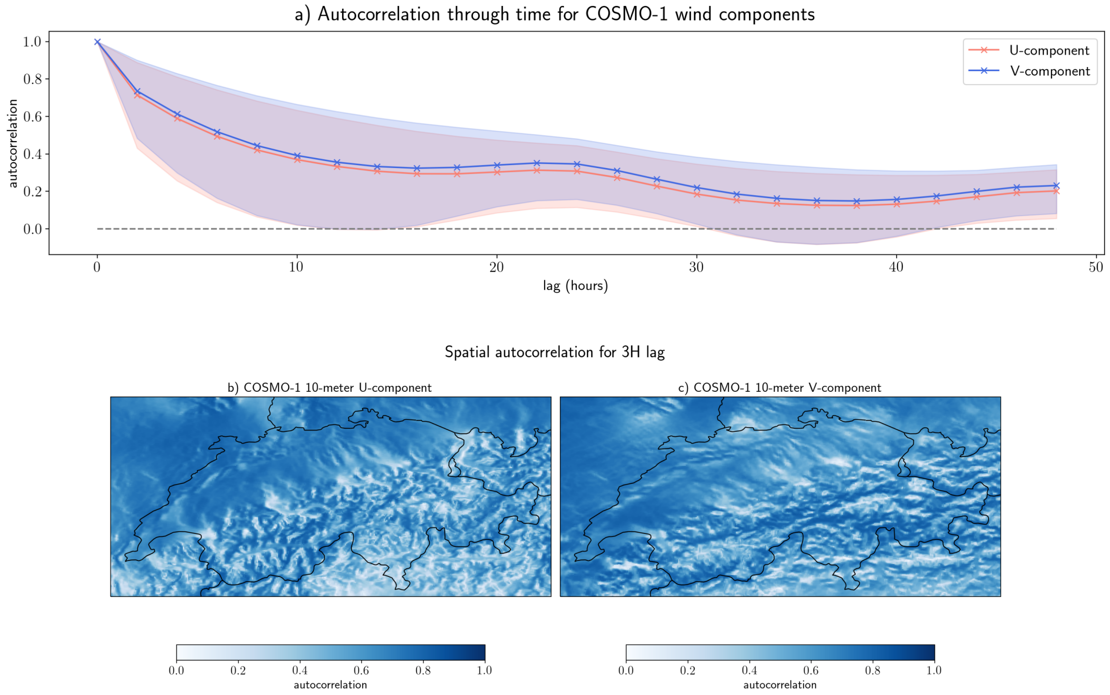


Figure 2.3: Mean autocorrelation as a function of lag in hours for COSMO-1 wind components u and v (fig.a). The shaded area corresponds to 5% and 95% quantiles of the spatio-temporal distribution for u and v. Spatial distribution of u (fig.b) and v (fig.c) autocorrelation for a 3-hour lag.

and Schmidhuber, 1997) that uses a hidden state to recall information about the past. The critic is also given such a layer so that scores are computed and optimized based on a wind sequence rather than on individual wind fields. As Figure 2.3 shows, spatial autocorrelation depends on local topography: for both u (Figure 2.3b) and v (Figure 2.3c) components, autocorrelation is stronger in the plains of the Swiss Plateau and on top of the high mountain ridges than on steep slopes and in the valleys. Hence it is crucial that the topography is fed to the network before the activation of the LSTM layer in order to account for its effect on autocorrelation. The complete architecture of the network is displayed in the Appendix (Figure 2.15).

2.4.3 Generator network

The entry layer of the generator is a concatenation of the input low-resolution wind fields (of size $N_B \times N_T \times S \times S \times N_P$, where N_B is the batch size, N_T is the number of consecutive time steps used for building the sequences, S is the patch size and N_P is the number of predictors) and random Gaussian noise (of size $N_B \times N_T \times S \times S \times N_N$, where N_N is the number of noise channels), which is used to robustify learning by making it less dependent on the precise data used. Introducing noise also allows for stochasticity in the model by sampling from the latent field distribution. The noise standard deviation of 0.1 m/s is chosen to represent small deviations from the input wind field.

After concatenating the input data and noise, we progressively increase the resolution of the input random vector to attain the desired resolution in the output. We decompose this step into two simultaneous sub-steps. The number of channels, $N_P + N_N$, is first increased using padded convolutions to leave room for the information contained in the spatial dimension of the tensor, and convolutional layers with strides are simultaneously applied to the tensor to decrease the spatial dimension, triggering the transfer of information to the channels. This sub-step is shown in Figure 2.15a: it starts after the concatenation of input and noise channels and is terminated by an LSTM layer and a first split connection. At the end of the sub-step, the image size has been reduced by a factor of four. Layers in which the same operation is applied to different time steps are referred to as “TimeDistributed” in Figure 2.15a. This sub-step can be seen as an organized and efficient destructuring by the generator of the information contained in the input layer in order to recreate a higher-resolution version of the image. The second sub-step increases the resolution by transferring information from the channels back to the spatial dimensions. The last upsampling layers of the generator use spatial bilinear interpolation rather than transposed convolutions, as this produces smoother outputs. All convolution layers from the upsampling step are activated with the Leaky ReLU function $x \mapsto x^+ - 0.2x^-$, where x^+ and x^- are the positive and negative parts of x.

Finally, wind fields (of size $N_B \times N_T \times S \times S \times 2$) are predicted using padded convolution with linear activation (see the last convolution layer in Figure 2.15a). Using bounded activation functions is known to increase the stability of training, especially on visual feature recognition problems (Liew et al., 2016). The idea of constraining the generated wind fields using a normalization constant and a tanh activation function for the last layer was considered but not applied, primarily to avoid underestimating extreme winds.

To assess the functioning of the generator network, we blur COSMO-1 high-resolution

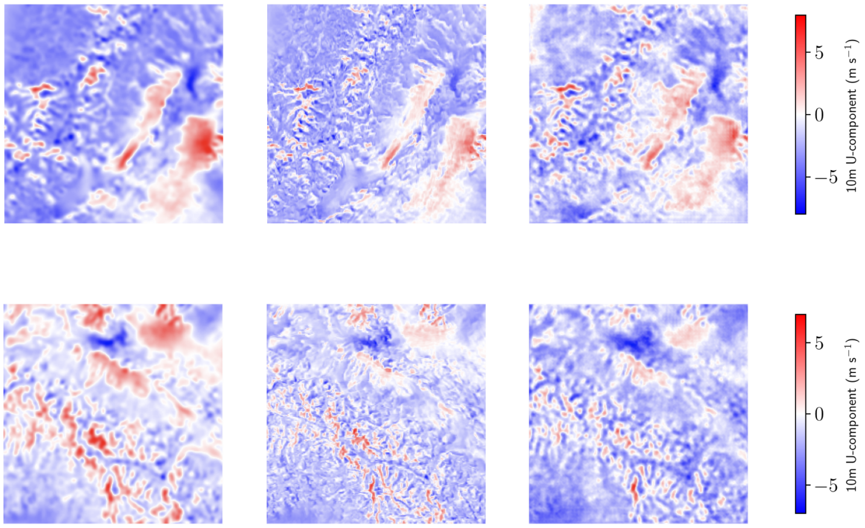


Figure 2.4: Prediction of the u component of 10-meter wind field by the generator model presented in Section 2.4. The rows denote different 80 km patches at different times. The columns represent inputs from the COSMO-1 model at resolution 1.1 km blurred with a Gaussian filter with standard deviation 2 (left), the original raw high-resolution wind fields (middle), and the model prediction using RMSE loss (right).

fields using a Gaussian filter with a standard deviation of 2 and try to predict the unblurred high-resolution fields by minimizing the root mean squared error between generated and realized fields. No other predictor is added to this optimization problem, in order to check whether the generator alone can perform well on a very simple task. Figure 2.4 shows that the network produces good results when trained on a small number of steps, or “epochs”: The blurring pattern seems to be rapidly understood by the generator. The training validation metrics detailed in Section 2.6 confirm that it performs well.

2.4.4 Discriminator network

The discriminator network, or critic, is used to determine whether a pair of low and high-resolution wind fields (both of size $N_B \times N_T \times S \times S \times 2$) are a good match and to decide if the high-resolution wind field is from COSMO-1 or predicted by the generator. Accordingly, the first layer of the discriminator inputs concatenated low and high-resolution images in order to evaluate how well they match. In Figure 2.15b,

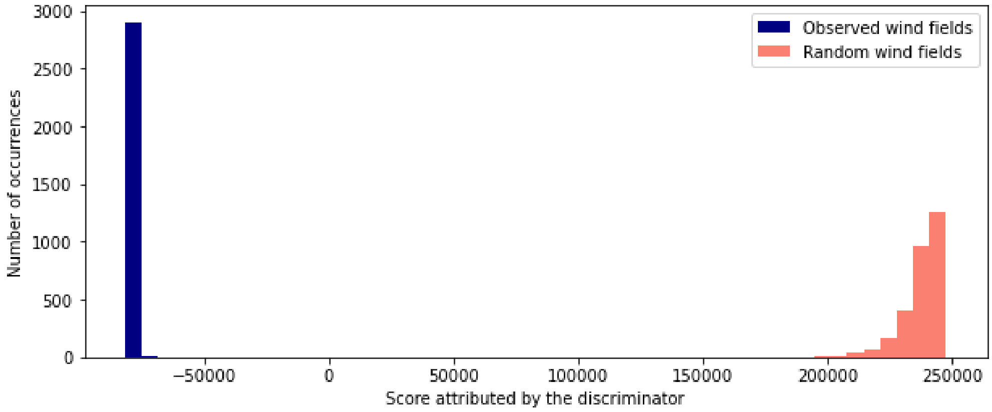


Figure 2.5: Scores for 10-meter wind fields predicted by the discriminator network presented in Section 2.4. The score is a unit-free relative quality measure internal to the GAN and thus has no meaning in absolute terms.

this step is represented by the temporary creation of two branches, one in which the match between low and high-resolution images is processed as a time-varying tensor, and one in which only the high-resolution image goes through this process. The two branches allow the generator to learn time series specificities for both the pair of low/high-resolution winds and the high-resolution field. The tensors containing information about the match and the high-resolution wind field alone are then concatenated and undergo a progressive information transfer from the spatial dimensions to the channels, as described for the first step of the generator network (the successive application of TimeDistributed convolution layers is shown in Figure 2.15b). Finally, a dense layer with linear activation is averaged on the time dimension to produce the final score with size $N_B \times 1$ for the two wind fields.

To check whether the critic can attribute different scores to realized and generated inputs, we train it alone by minimizing the loss introduced in equation (2.1). Inputs generated by the artist are not available when we train the critic alone, so we replace them with more obvious fake images, Gaussian random fields with a standard variation of 10. This task is similar to binary classification, although the scores here can take any value in $\mathbb{R}$. Figure 2.5 shows that the scores attributed by the trained critic to fake and real wind fields are clearly separated, so the critic performs correctly. The scores vary more for random wind fields, which could be interpreted as the network introducing uncertainty regarding the reliability of the classification in the presence of potentially fake winds.

2.4.5 Input data

The ERA5 inputs are grids of 12 × 24 pixels (25 km resolution), COSMO-1 targets are on 294 × 429 grids (1.1 km resolution) and topographic descriptors are constant on grids of size 3312 × 6912 (90 m resolution). To better control information reduction and expansion throughout the convolutional layers of the GAN, we created sequences of square patches as input for the model using the same geographical areas to create the inputs and the targets, while keeping the patch size constant. To do so we project all the inputs and outputs onto the COSMO-1 target grid. In particular, ERA5 inputs, topographical descriptors, and outputs are processed to reach 1.1 km grid resolution. The resolution of the ERA5 fields is artificially increased by filling the gaps with the nearest available value from the reanalysis dataset. Topographic predictors that have a higher spatial resolution than COSMO-1 fields are slightly blurred to meet the resolution standards.

Sequences of square patches of size S are created around random points in space and time for low-resolution inputs and corresponding high-resolution fields and are then randomly flipped or rotated before being input to the model. Data augmentation such as this has been found to enhance network efficiency and out-of-sample performance in other applications (Perez and Wang, 2017).

2.5 Training

2.5.1 Adversarial training

Fitting a GAN can be difficult because the generator and discriminator networks may train at different speeds. For this study, the training was stabilized using spectral normalization that enforces a Lipschitz constraint on parameters in the convolution layers of both networks (Miyato et al., 2018), and different learning rates were chosen (Heusel et al., 2017). Indeed, a generator training against a poor discriminator does not generate better images because the discriminator cannot score them well. Typically, the learning rate of the discriminator is set 4 to 5 times above that of the generator to allow the scoring function to improve enough between successive updates of the generator. We further aided discriminator training by updating its network three times for each update of the generator to give the discriminator more time to process the change in the generator network (Gulrajani et al., 2017, Algorithm 1). To avoid vanishing gradients the generator network includes not only split connections (Ronneberger et al., 2015; Srivastava et al., 2015), i.e., shortcuts between deep layers, but also batch normalization layers (Santurkar et al., 2018) to normalize data across samples of a batch. For the same reason, the discriminator includes one split connection, and layer normalization of data aggregated on channels (Ba et al., 2016) was applied to the discriminator's convolutional layers. Normalization here should be understood as standardization, i.e., transforming the data to have zero mean and unit standard deviation. The Adam optimizer, a stochastic gradient descent method based on adaptive estimation of first and second-order moments (Kingma and Ba, 2014), was used with learning rates of 1×10^{-4} for the generator and 4×10^{-4} for the discriminator. The values for the first and second moment estimates $\beta_1 = 0.$ and $\beta_2 = 0.9$ were derived from the calibration of the Adam optimizer for WGAN-GP by Gulrajani et al. (2017). Reconstruction loss was considered in this study to improve the training stability, by using an auto-encoder to extract features from wind maps, but the results were more satisfactory with other techniques, such as layer normalization, split connections, and adjusting the optimizer hyper-parameters. The small effect of inserting a reconstruction loss could be due to the very basic structure of the auto-encoder, which we built ourselves for this study: efficiently extracting relevant features from the wind fields is a research project on its own. Moreover, the implementation of the GAN with a reconstruction loss had the undesired impact of keeping the prediction close to the original ERA5 pixellated style. Lowering the weight of the reconstruction loss in the overall loss turned out to be equivalent to doing without reconstruction loss entirely.

2.5.2 Transfer learning

The GAN is trained using transfer learning (Bozinovski and Fulgosi, 1976): after it is trained for one task, the learning curve for a similar task should be less steep and the training more efficient. Our downscaling problem is difficult for two main reasons. First, the difference in resolution between inputs and targets is large, as wind fields from ERA5 reanalysis data are available on 25 km grids, while COSMO-1 is on 1.1 km grids. Second, input and target winds come from two different sources, and discrepancies in modeling techniques make it more difficult for a network to understand how a high-resolution COSMO-1 field is linked to an ERA5 field than to an artificially blurred version of itself. In our case, no known transform of the high-resolution output data links it to the low-resolution predictors, so the network is first trained to downscale winds from artificially blurred COSMO-1 data to the high-

Algorithm 1 WGAN-GP with different update rates for the generator and discriminator networks, as proposed by Gulrajani et al. (2017). θ and w represent respectively the generator and discriminator's parameters throughout training.

```
Require: For data processing: batch size $N_B$, time steps $N_T$, patch size $S$ and number of predictors $N_P$. The number of noise channels $N_N$ is also required.
Require: Learning rates $\alpha_G$ and $\alpha_D$ for the generator and discriminator networks, optimizer hyperparameters $\beta_1 = 0$ and $\beta_2 = 0.9$ and the number of consecutive discriminator updates $n_{\text{critic}} = 3$ for one generator update.
Require: Initial discriminator parameters $w_0$ and initial generator parameters $\theta_0$.
  while $\theta$ has not converged do
    for $t = 1, \ldots, n_{\text{critic}}$ do
      Create inputs batches $x$ (of size $N_B \times N_T \times S \times S \times N_P$) and corresponding target batches $y$ ($N_B \times N_T \times S \times S \times 2$)
      Sample noise $z \sim \mathcal{N}(0, 10^{-2})$ of size $N_B \times N_T \times S \times S \times N_N$
      Sample a random number $\epsilon \sim \mathcal{U}(0, 1)$
      $\hat{y} \leftarrow G_\theta(x, z)$
      $\tilde{y} \leftarrow \epsilon y + (1 - \epsilon)\hat{y}$
      $\text{Loss}_D \leftarrow D\left(x, y\right) - D\left(x, \hat{y}\right) + \gamma\left(\|\nabla_{\tilde{y}} D\left\{x, \tilde{y}\right\}\|_2 - 1\right)^2$
      $w \leftarrow \text{Adam}\left(\nabla_w \text{Loss}_D, \alpha_D, \beta_1, \beta_2\right)$
    end for
    Sample noise $z \sim \mathcal{N}(0, 10^{-2})$ of size $N_B \times N_T \times S \times S \times N_N$
    $\theta \leftarrow \text{Adam}\left(\nabla_\theta D\{x, G_\theta(x, z)\}, \alpha_G, \beta_1, \beta_2\right)$
  end while
```

resolution target wind fields, and then the training continues with low-resolution winds from the ERA5 reanalysis data.

2.5.3 Technical challenges

The MeteoSwiss data cover a period from April 2016 to October 2020, yielding 1673 days of hourly observations on grids of resolution 429 × 324 pixels. Our interest is in surface wind vectors with components u and v, so the number of individual data points to be predicted is about 10 billion. To capture daily patterns, we chose to train the GAN on 24-hour sequences ($N_T = 24$) of square patches using two years of data (about 4.8 billion individual points). In the end, the only predictors inputted to the GAN are the low-resolution wind fields from COSMO-1 blurred data for the first training phase and the wind fields from ERA5 for the second training phase, mainly because the ratio of performance improvement to additional computational burden was too low when other predictors from ERA5 (see Section 2.3.2) were added. Using raw DEM as the only topographic predictor gave the best results. Small batches ($N_B = 8$) were chosen because we found that such micro-batches stabilize the training. The generator contains 1.7 million parameters and the discriminator contains 3.3 million parameters, so the total number of parameters to be estimated is about 5 million. The training was done over about 200 epochs (training steps) on an Nvidia GPU with Volta microarchitecture provided by the EPFL Scientific IT and Application Support (SCITAS) system.

2.6 Metrics

The discriminator network scores the wind fields predicted by the generator according to its discrimination criteria by generating a model internal score function that is continuously improved during training. This function provides relative comparisons of the model at different training stages and cannot be interpreted in absolute terms. To gain a detailed understanding of the network's performance, we also use other metrics to monitor the training and the final results. The Fréchet inception distance is the most commonly used metric to assess the performance of GANs (Heusel et al., 2017), but its implementation relies on the use of another neural network trained to recognize features that are friendly to human perception on static RGB images. This is irrelevant in our case because the two-dimensional field we aim to predict has only one channel per variable. Hence we use standard metrics to assess the performance of image prediction, such as modified versions of the root mean squared error (RMSE), log spectral distance (LSD), angular cosine distance (ACD), and a spatially convolved version of the Kolmogorov–Smirnov statistic, which are detailed below.

2.6.1 Root mean squared error (RMSE)

We use two versions of the original RMSE (Hyndman and Koehler, 2006). Although weighting the metrics based on realized (COSMO-1) values rather than on predictions could bias model selection and validation (Lerch et al., 2017), these metrics were chosen to best meet the goal of the study, which is to provide accurate historical covariates for analysis of past weather and climate in which extreme winds might play a role as an aggravating factor. Although special attention was paid to extreme winds, other metrics and visual methods were also used to assess prediction reliability and thereby counterbalance any bias (Lerch et al., 2017).

The wind speed weighted RMSE (Dujardin, 2021) is defined as

$$\text{WSRMSE} = \sqrt{\frac{1}{N_T \times P} \sum_{t \le N_T, i \le P} \tau \left\{ \left(u_{it} - \beta \hat{u}_{it}\right)^2 + \left(v_{it} - \beta \hat{v}_{it}\right)^2 \right\}},$$

where N_T is the number of time steps, P the number of pixels in a single image, u and v are the 10-meter high-resolution eastern and northern components of wind, $\hat{u}$ and $\hat{v}$ are their respective estimates and

$$\beta = \frac{\epsilon + w}{\epsilon + \hat{w}}, \quad \tau = \begin{cases} t, & \hat{w} \ge w, \\ 1 - t, & \text{otherwise,} \end{cases}$$

where w is the 10-meter high-resolution wind speed and $\hat{w}$ its estimate. The calibration of hyperparameters by Dujardin (2021) sets $\epsilon = 4$ and $t = 0.425$.

Another RMSE variant developed for this specific problem,

$$\text{ExtremeRMSE} = \sqrt{\frac{1}{N_T \times P} \sum_{t \le N_T, i \le P} \frac{u_{it}^2}{\sum_{j,k} u_{jk}^2} (u_{it} - \hat{u}_{it})^2 + \frac{v_{it}^2}{\sum_{j,k} v_{jk}^2} (v_{it} - \hat{v}_{it})^2},$$

tries to condemn bad predictions of extreme components.

Both of these metrics put more weight on extreme winds, which explains their similar evolution during training (see Figure 2.6). However, WSRMSE penalizes extremes in a relative sense (β becomes large when the wind speed is underestimated), whereas ExtremeRMSE directly puts component-wise weights that increase with the realized extremeness of each direction, whether or not the estimated component is also extreme.

2.6.2 Angular cosine distance (ACD)

The angular cosine distance (Foreman, 2013) computes the average angle between the target and generated vectors,

$$\mathrm{ACD} = \frac{1}{N_T \times N_P} \sum_{t \le N_T, i \le N_P} \arccos\left(\frac{u_{it}\hat{u}_{it} + v_{it}\hat{v}_{it}}{\sqrt{u_{it}^2 + v_{it}^2}\sqrt{\hat{u}_{it}^2 + \hat{v}_{it}^2}} \right),$$

and thus quantifies agreement between predicted and observed wind field directions. The ACD and the RMSE metrics complement each other, as ACD measures the performance of the network in terms of wind direction, whereas RMSE evaluates the distance between realized and predicted wind speed, which is the wind vector's norm. Both are needed for an accurate performance assessment.

2.6.3 Spatially convolved Kolmogorov–Smirnov statistic (SKSS)

This new metric, developed in the scope of this research, represents the disagreement between the distributions of the generated and observed wind fields. It is computed as the maximum absolute difference of empirical cumulative distribution functions for the generated and realized fields, summed over 10×10 patches of the image of interest. The aim is to obtain a metric with properties close to those of the Fréchet inception distance (Heusel et al., 2017) for RGB images by assessing the match between input wind fields and images produced by the generator, as a human eye would. Indeed, a GAN's performance can be hard to assess and visual checks of the generator's output may be preferred by users. The SKSS assesses whether the output is visually pleasing by checking whether the u and v fields on small spatial patches look similar to those in the original image. First, M spatial patches of constant size are extracted from the target and predicted images. We then set

$$\text{SKSS} = \sum_{t\leq N_T, j\leq M} \sum_{c\in\{u,v\}} \max_{x\in\mathbb{R}} |F^c_{jt}(x) - \hat{F}^c_{jt}(x)|,$$

where F^c_{jt} represents the empirical cumulative distribution function for a single spatial patch j, point in time t and channel c (u or v component of wind) and $\hat{F}^c_{jt}$ is its analog for generated data. This metric is intended to evaluate the agreement between two local distributions rather than focusing on individual pixels.

2.6.4 Log spectral distance (LSD)

The LSD metric (Rabiner and Juang, 1993) is expressed as the log-difference of power spectra between the generated and realized samples,

$$\text{LSD} = \sqrt{\frac{1}{2N_T \times P} \sum_{t\leq N_T, i\leq P} \sum_{c\in\{u,v\}} \left[10\log_{10}\left(\frac{|f(c_{it})|^2}{|f(\hat{c}_{it})|^2}\right)\right]^2},$$

where f is the Fourier transform, $|f(\cdot)|^2$ the power spectrum, c is the wind component and $\hat{c}$ its estimate. The LSD evaluates whether the generated images reproduce the spatial structures noticeable in the target images.

2.7 Results

The GAN described above is a stochastic model: predictions may vary with different samples of the input noise. In the following analysis of the results, the average prediction for the test set over 200 different noise samples is used to construct graphs and maps. As explained in Section 2.5.2, the network is trained in two phases that are evaluated separately below. Unless specified otherwise, the years 2016 to 2018 were used for training, the year 2019 was used as the validation set, and the year 2020 was used as the test set for all results and plots.

2.7.1 Training phase 1: downscaling COSMO-1 blurred wind fields

Model selection

The first step of the quantitative analysis entails model selection, as the network produces a different set of parameters for every training epoch, i.e., every complete training round of the GAN on all wind fields. The best model must be selected to perform predictive analyses and make diagnostic plots. Metrics are expected to be non-monotonic throughout training, as the generator and discriminator improve in an adversarial way. As the generator improves, it is more difficult for the discriminator to determine whether a given image is observed or predicted data, and is, therefore, more likely to attribute an incorrect score. As the discriminator's loss decreases, the classification becomes more accurate and the images produced by the generator are more likely to receive very positive scores, increasing the generator's loss. Figure 2.6 shows that the six metrics considered partially agree on the best-performing epoch, i.e., the training step with the minimum value for a given metric. All three RMSE metrics (Figure 2.6a, b, and c) and the LSD (Figure 2.6d) indicate a local minimum at epoch 55, while the local minima for the KS-statistic (Figure 2.6e) occur at the very beginning of the training. The angular cosine distance (Figure 2.6f) shows no clear minima. Epoch 55 is used to produce the diagnostic plots and computations that follow.

Quantitative analysis of wind time series

The network is built to capture time series features in the target data. To evaluate the performance of the model, input, target and predicted autocorrelation are compared for lags from 2 hours to 2 days in Figure 2.7. The artificial blurring of the COSMO-1 wind fields (input) introduces additional autocorrelation that the network can successfully remove, as seen in Figure 2.7. However, the predicted wind components show lower autocorrelation for lags below 48 hours than do the COSMO-1 data (target), with a marked increase for multiples of 24-hour lags, perhaps because we process the wind field time series in 24-hour sequences. Wind exhibits very specific sub-daily patterns that vary with the topography. To evaluate whether these are well captured by the network, Figure 2.16 shows the sub-daily wind variability averaged over the validation set for locations in valleys, plains, and on mountaintops. The GAN can

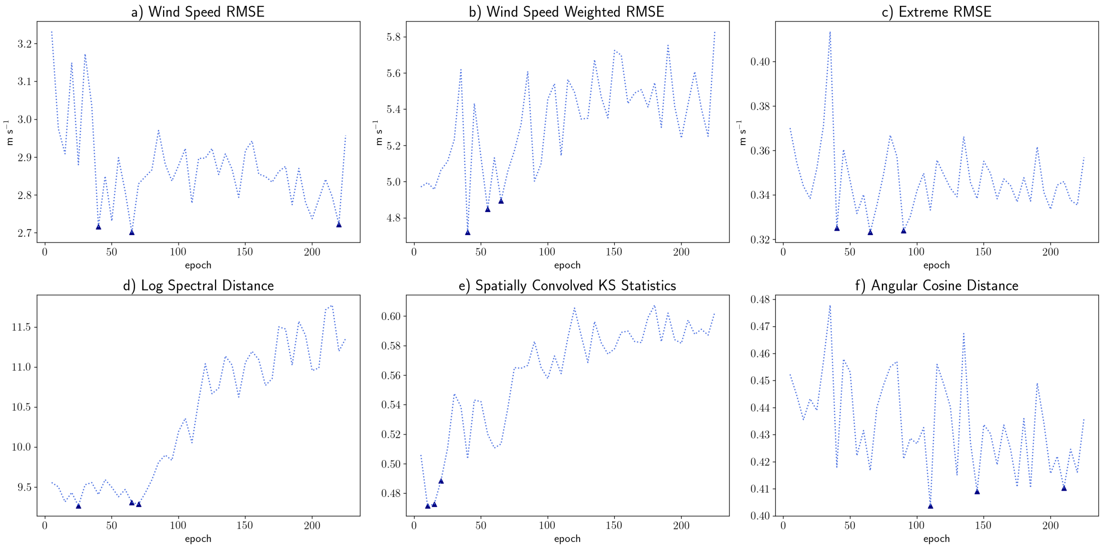


Figure 2.6: Behaviour of the different verification metrics throughout the training of the GAN. The x-axis represents the epochs, with one epoch sampling patches of wind fields for all available dates. The three smallest values attained by each metric during training are denoted by triangle dots. Metrics are evaluated every day over random square patches in Switzerland on a 2-months validation set going from September 2019 to November 2019. Results are averaged over space and time.

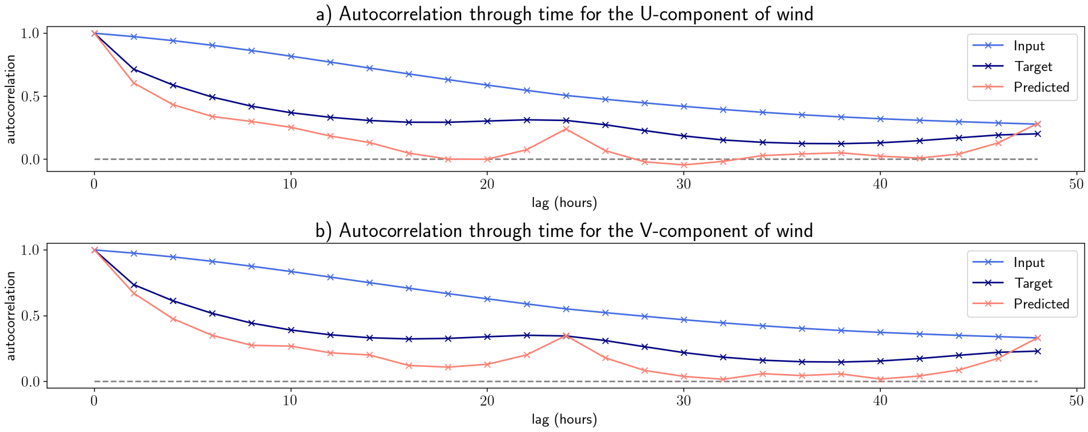


Figure 2.7: Spatially and temporally averaged estimated autocorrelation for input, predicted, and target u (fig.a) and v (fig.b) wind components.

accurately reproduce average daily patterns for both u and v components in relatively flat zones, e.g., Fahy lies in the Jura and Aigle in a valley in the Alps (Figure 2.16a and b), and Bischofszell and Fribourg are in the Swiss Plateau (Figure 2.16c and d). In extremely complex terrain (Zermatt and Jungfraujoch are located on mountain passes), the model does not capture the sub-daily pattern of the u component (Figure 2.16e and f) well.

Visualisation of historical wind maps

The GAN is trained on square patches of size $S = 96$ pixels, while COSMO-1 maps of Switzerland have size 429×324 (Subsection 2.5.3 of the Section 2.5). We combine the patches to predict entire wind fields. One possibility would be to crop the initial COSMO-1 map into a 384×288 map focusing on Switzerland only, predict a grid of 4×3 patches of size S and accept that there will be a discontinuity at their borders, but as our goal is to create realistic-looking historical wind fields we prefer to predict a grid of overlapping 5×4 patches and average them to give smoother borders. Maps of target and predicted u and v components of wind averaged over the 1-year test period are displayed in Figure 2.8. Specific patterns, such as local acceleration at exposed sites (ridges) and sheltering in valleys, are very well reproduced by the network for both u and v components in all three sub-regions of Switzerland. Both COSMO-1 (Figure 2.8c) and the predictions (Figure 2.8d) show strong regional patterns of the mean v wind direction depending on the topography, with southerly winds on north-facing slopes and northerly winds on south-facing slopes. This is probably the fingerprint of the foehn, an intense, warm, and dry downslope windstorm that occurs frequently on both the northern and southern sides of the Alps (Richner and Hächler, 2013). The appendix contains examples of hourly maps produced from blurred COSMO-1 after the first training phase (Figure 2.17) and from ERA5 low-resolution inputs after the second training phase (Figure 2.18).

The spatial quality of the predicted wind fields is evaluated by plotting the median values for WSRMSE and ACD, computed for the test year 2020 (Figure 2.9). The former was made unitless by applying a hyperbolic tangent transform to facilitate interpretation. Predictions for wind speed (Figure 2.9a) and direction (Figure 2.9b) are good in the Swiss Plateau and Jura. Differences in wind direction occur in valleys and at the feet of mountains, while wind direction is well predicted on upper slopes and ridges (Figure 2.9b). In the Swiss Alps, where the terrain is more complex, the wind speed is predicted less well at the high-wind exposed mountain sites (ridges) than at the sheltered valley sites (Figure 2.9a).

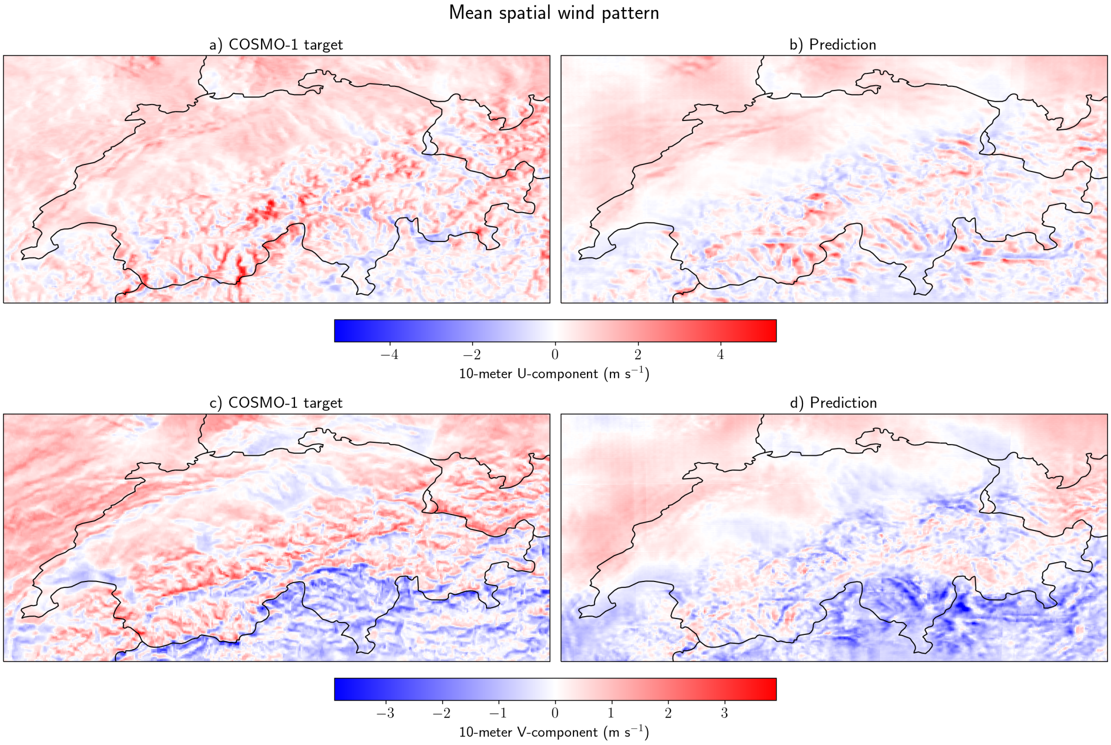

Mean spatial wind pattern
a) COSMO-1 target
b) Prediction
−4
−2
0
2
4
10-meter U-component (m s^{-1})
c) COSMO-1 target
d) Prediction
−3
−2
−1
0
1
2
3
10-meter V-component (m s^{-1})

Figure 2.8: Target (left) and predicted (right) wind components u (top) and v (bottom) averaged over the test period of 1 year in 2020.

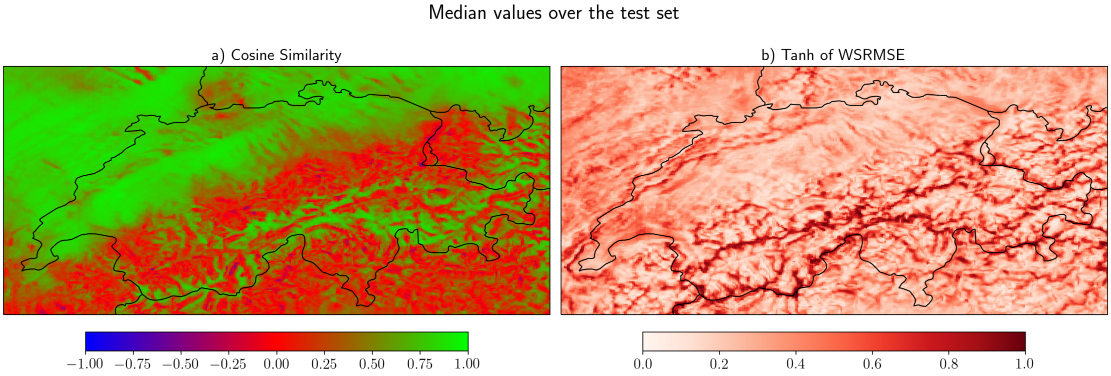


Figure 2.9: Visualisation of GAN performance after the first phase of training. Median cosine similarity is shown in fig.a (1 is perfect and −1 is bad) and median wind-speed weighted RMSE (WSRMSE) after a bounded transform is shown in fig.b (right, 0 is perfect and 1 is bad).

2.7.2 Training phase 2: downscaling ERA5 wind fields

In the second phase, the training continues and the parameters of the best model of the first training phase are used as initial parameter values. When the second training phase is completed, the steps of the first training phase are repeated to find the epoch with the best performance. Maps of wind fields from target and predicted from ERA5 averaged over the 1-year test period are shown in Figure 2.10 and median values of the metrics WSRMSE and ACD are shown in Figure 2.11. The prediction (Figure 2.10b and d) reproduces the high-resolution mean wind pattern of the COSMO -1 target (Figure 2.10a and c) with stronger westerly winds at the exposed mountain sites in the Alps and Jura and weaker or easterly winds at the sheltered valley sites. Looking at the mean wind direction, we see that more regions have northerly and easterly winds compared to COSMO-1. These differences in wind direction between prediction and COSMO-1 are also seen in the median cosine similarity in Figure 2.11b and occur over the entire Alps. The wind speed is predicted less well at the high-wind exposed mountain sites (ridges) than at the sheltered valley sites (Figure 2.11a). Results from the first (Figure 2.9) and the second (Figure 2.11) training phase using ACD and WSRMSE show better predictive performance when downscaling from COSMO-1 blurred inputs. Indeed, the Gaussian filter used to create low-resolution winds from the COSMO-1 high-resolution target produces smoother maps than those created with ERA5 winds, which may facilitate pattern detection.

The ultimate goal of this project is to build accurate historical wind fields for the analysis of specific extreme events, such as windstorms and the role of foehn in forest fires, so the main question addressed in the diagnostics for the second training phase is whether the network accurately represents extreme wind speeds. Strong winds during storms are known to cause damage in populated regions at lower altitudes (Swiss Plateau and valleys) (Schwierz et al., 2010; Welker et al., 2016) and are important in spreading forest fires (Sun et al., 2009; Sharples et al., 2012; Cruz et al., 2020), so we desire them to be accurately downscaled from the low-resolution fields of ERA5.

Speed and direction distributions for predicted (from ERA5) and target winds are displayed and compared in Figure 2.12. We consider the wind speed computed from

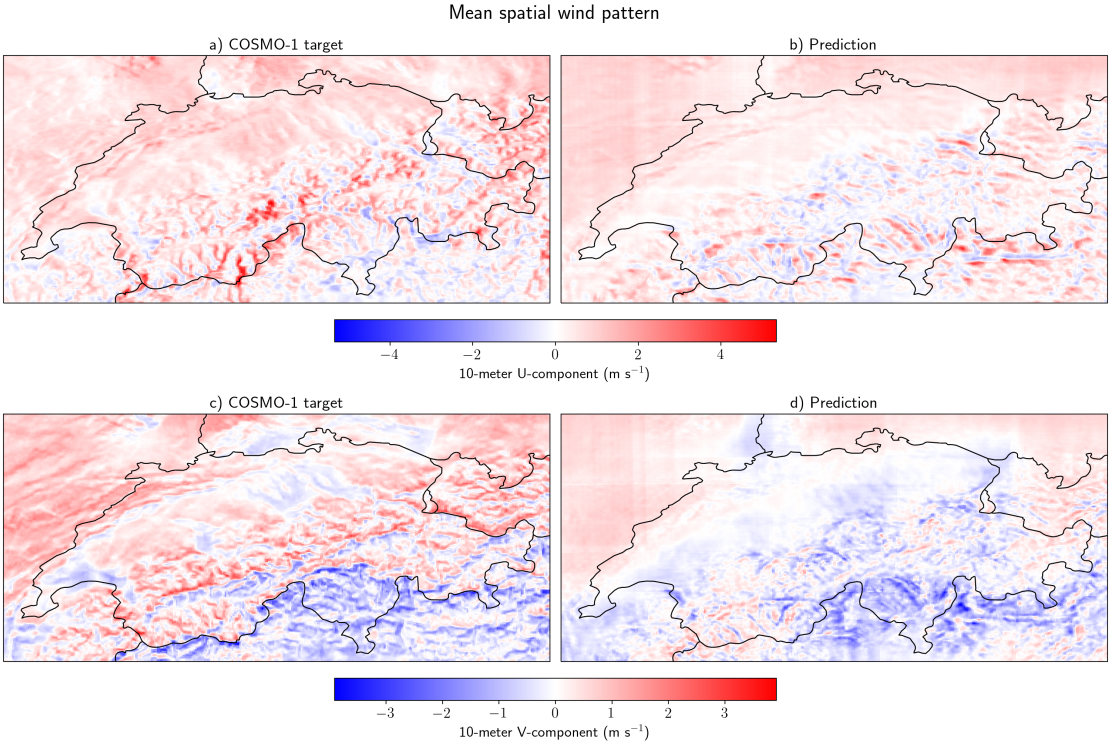

Mean spatial wind pattern
a) COSMO-1 target
b) Prediction
−4
−2
0
2
4
10-meter U-component (m s^{-1})
c) COSMO-1 target
d) Prediction
−3
−2
−1
0
1
2
3
10-meter V-component (m s^{-1})

Figure 2.10: Target (left) and predicted from ERA5 (right) u (top) and v (bottom) components of wind fields averaged over the test period (1 year).

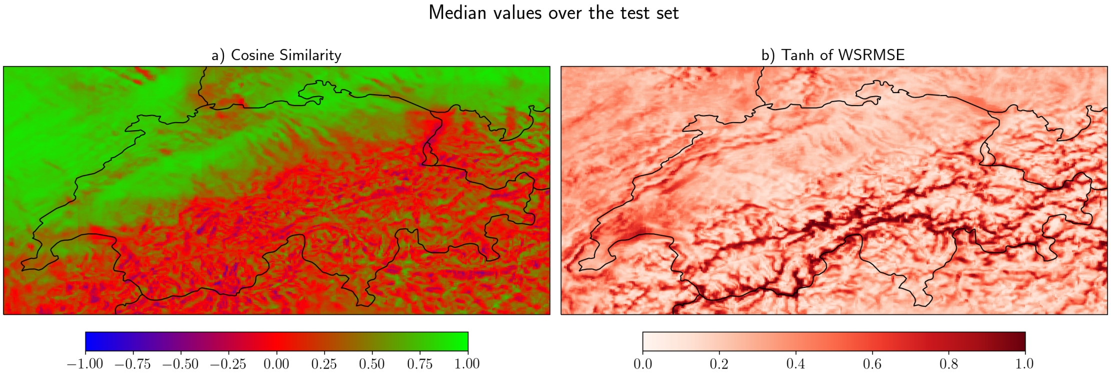


Figure 2.11: Visualisation of the GAN performance after the second phase of training. Median cosine similarity is represented in fig.a (1 is perfect and −1 is bad) and median wind-speed weighted RMSE (WSRMSE) after a bounded transform is represented in fig.b (right, 0 is perfect and 1 is bad).

the mean predicted maps of u and v components rather than the mean wind speed over the noise samples. The predictions of the average wind speed distribution are very close to the target distribution, although the predicted wind speed is slightly less long-tailed than that for COSMO-1 (Figure 2.12a). The network underestimates very high wind speeds and predicts more southwestern (0 to 90 degrees) and northeastern

(180 to 270 degrees) winds than are observed in COSMO-1 (Figure 2.12b). The wind speed distribution of ERA5 winds shows a much thinner tail than the target and predicted wind speed (Figure 2.12a), while the ERA5 wind direction seems completely out of sync compared to the target and predicted distributions (Figure 2.12b).

Further analysis of wind speed predictions within 10 km around the largest Swiss cities is encouraging. Figure 2.13 shows that the extremes of wind speed are well captured by the model, especially for cities in the Swiss Plateau: Zürich (Figure 2.13a, ~ 400K inhabitants), Lausanne (Figure 2.13d, ~ 135K inhabitants), Winterthur (Figure 2.13f, ~ 108K inhabitants) and Luzern (Figure 2.13g, ~ 81K inhabitants).

How much do the downscaled wind fields improve on the original ERA5 fields? We compare the metrics WSRMSE, LSD, and SKSS, averaged over the test set, between predicted winds and ERA5 winds in Figure 2.14b,c,d. Each point represents the average at a specific grid cell and is colored according to the sub-region shown in Figure 2.14a to highlight regional differences. The diagonal line ($x = y$) represents identical values of the metric before and after prediction by the network. Average values of the metrics for each geographical sub-region, given in Table 2.1, show that for all regions the LSD is much smaller when comparing predicted winds to the target COSMO-1 than for the ERA5 inputs. The SKSS is smaller for predicted winds than for ERA5 inputs in the Alps but slightly higher at some points on the Swiss Plateau, maybe because ERA5 predicts the winds sufficiently well on homogeneous and flat zones, while the GAN could add artifacts, i.e., undesired signals with non-physical origins, at these locations. There is no reduction in WSRMSE, which is essentially preserved by the GAN. This is expected because pointwise comparisons do not need to detect the visual improvements highlighted by LSD and SKSS. Table 2.1 shows clear improvements using the GAN prediction instead of ERA5 winds in Alpine regions (Alpes Valaisannes, Alpes Vaudoises, Alpi Lepontine, Alpi Retiche, Berner Alpen, Glarner Alpen, and Urner Alpen), especially for LSD and SKSS. On the Swiss Plateau (Mittelland) and in the Jura, a sharp decrease of the LSD can be noted, while the SKSS is preserved.

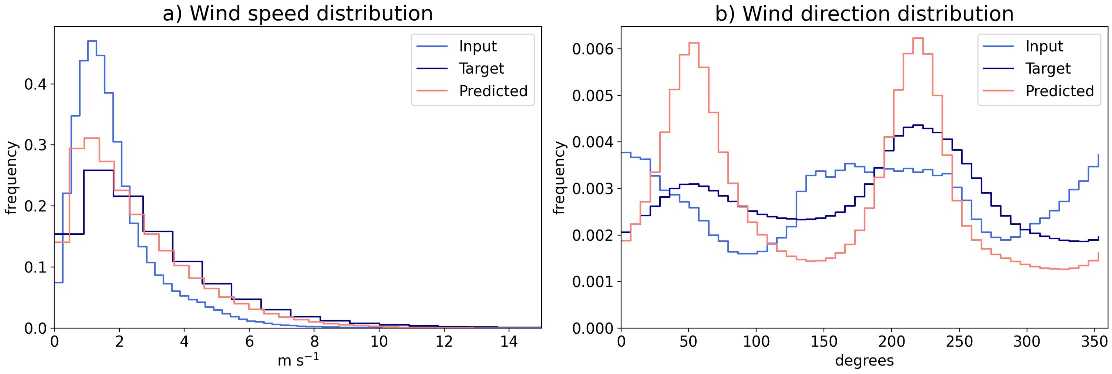


Figure 2.12: Estimated distribution of input (ERA5), target (COSMO-1) and predicted wind speed (fig.a) and wind direction (fig.b).

Table 2.1: Before-to-after comparison of the regional log spectral distance, the hyperbolic tangent of WSRMSE and SKSS averaged over the time dimension of the test set.

Geographical region	**LSD**			**Tanh WSRMSE**			**SKSS**		
	Input	*Predicted*	*%Variation*	*Input*	*Predicted*	*%Variation*	*Input*	*Predicted*	*%Variation*
Alpes Valaisannes	28.04	13.98	−50.2	0.69	0.68	−2.3	0.77	0.69	−10.5
Alpes Vaudoises	27.98	13.96	−50.1	0.69	0.68	−1.9	0.78	0.68	−13.1
Alpi Lepontine	26.86	14.97	−44.3	0.66	0.66	−0.9	0.77	0.69	−9.9
Alpi Retiche	26.60	13.12	−50.7	0.67	0.66	−1.5	0.77	0.71	−7.8
Berner Alpen	25.48	16.82	−34.0	0.69	0.67	−2.7	0.78	0.70	−10.6
Glarner Alpen	25.74	16.35	−36.5	0.66	0.65	−0.8	0.76	0.68	−11.0
Jura	27.06	14.21	−47.5	0.59	0.64	10.2	0.80	0.82	3.3
Mittelland	26.41	15.56	−41.1	0.56	0.60	6.8	0.80	0.83	3.0
Préalpes	25.78	16.42	−36.3	0.62	0.63	0.2	0.79	0.73	−7.4
Urner Alpen	24.77	17.76	−28.3	0.68	0.66	−2.6	0.76	0.68	−10.6

2.8 Conclusion

In this paper, we developed and applied a deep learning model for downscaling hourly near-surface gridded wind fields in complex terrain using low-resolution (~25 km) inputs from ERA5 reanalysis and high-resolution (1.1 km) targets from the numerical weather prediction model COSMO-1. Topographic information from high-resolution (90 m) digital elevation data from the SRTM3 was used as a static input to incorporate local orographic effects that modify airflow. A Wasserstein recurrent generative adversarial network (GAN) with a gradient penalty architecture was chosen with an autoencoder-like structure for the upsampling part of the generator. Adapted normalization of layer outputs was introduced in both networks, and weight normalization was used to speed up and smooth the training. Careful attention was paid to coordinating the networks so that neither became too strong for the other to train, and a good balance was achieved by using different learning rates and updating the discriminator more frequently than the generator. Due to the complexity of the problem, an

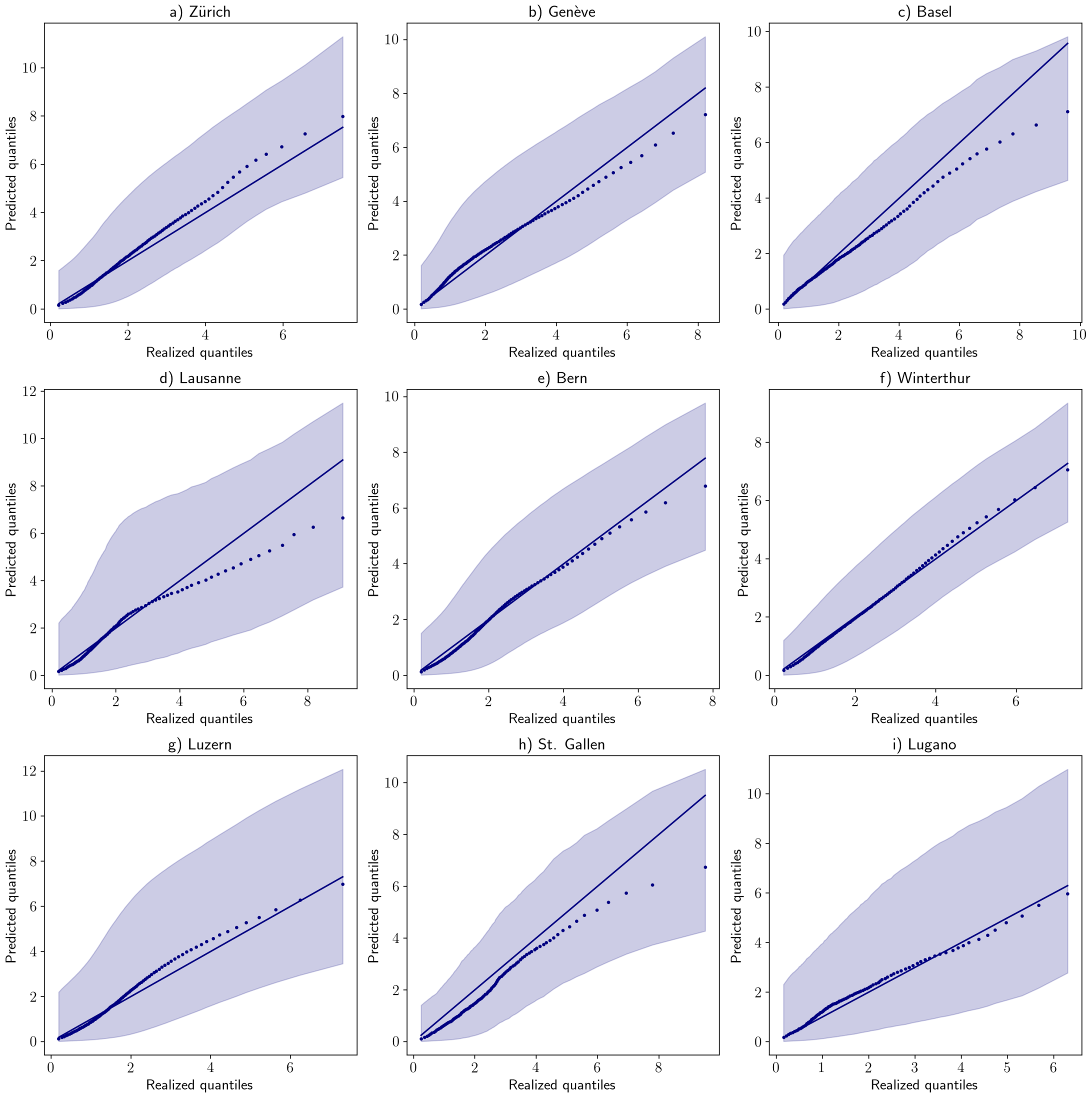
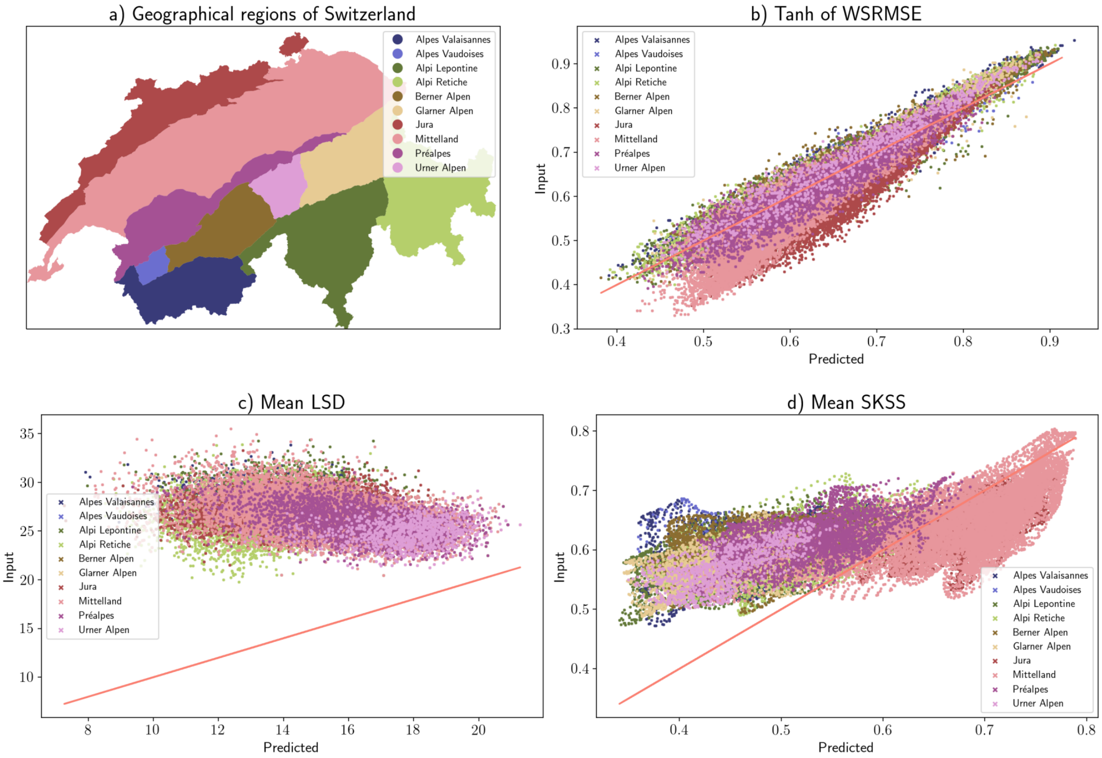


Figure 2.13: QQ plots of predicted (using ERA5 as input) versus target (COSMO-1) quantiles of the wind speed distribution within 10 km of the largest Swiss cities. From left to right: Zürich, Genève and Basel (top row), Lausanne, Bern and Winterthur (middle row), and Luzern, St. Gallen and Lugano (bottom row). Shaded areas are plotted between quantiles computed from the distribution of the minimum and maximum wind speed across 200 different noise samples given as inputs to the GAN. The diagonal line represents $x = y$.

approach based on transfer learning was chosen to train the GAN. Segmentation of the learning curve greatly improved the performance of the network compared to a more direct approach. This appears to be the first deep learning model trained using transfer learning that can efficiently perform such an extreme (25x, from 25 km to 1.1 km) downscaling of wind fields from two different data sources. Its performance was tested over the complex terrain of Switzerland for the period 2016–2020, but if the


a) Geographical regions of Switzerland

b) Tanh of WSRMSE

c) Mean LSD

d) Mean SKSS

Figure 2.14: Before-to-after regional comparison between ERA5 inputs (y-axis) and predicted high-resolution winds (x-axis) averaged over the test set. The WSRMSE (fig.b), LSD (fig.c), and SKSS (fig.d) metrics were used for this plot.

local topography is available, our model could be applied to wind fields elsewhere in order to generate long-term, high-resolution wind climatologies.

Historical maps, created for all of Switzerland using overlapping patches of predicted wind fields, were visually appealing for both training phases, with maps predicted from either blurred COSMO-1 or ERA5 inputs consistently resembling the COSMO-1 target. As our goal was to produce wind field predictions for the analysis of historical extreme weather events and long-term climatology, the second phase of training diagnostics focused on the wind-speed distribution to verify how well extreme values were captured. The findings indicate an excellent prediction of the aggregated wind speed distribution around densely-populated areas. Quantitative analysis of time series and spatial averages after the first training phase showed that the network missed some autocorrelations and that there were differences in the predictive performance between flat and mountainous regions. Wind speed and mean daily patterns were less well predicted in high altitudes of the mountainous terrain than in the hilly plains and valleys. While wind direction was well predicted on mountaintops, the network struggled when predicting wind direction in valleys.

As most issues stem from differences in topography, the global architecture could be improved by building different models for predicting patches that are mainly over mountainous areas or over the Swiss Plateau, rather than using a single network for the entire country. Another deep structure trained for image classification based on topography could use an input sensor to select which of those two networks should be applied. Different parameters would thus be used to predict winds in plains and in complex terrain, perhaps leading to less topographic variation in model performance. The training of such a structure would require more time and resources but might overcome most remaining issues with our model.

2.9 Acknowledgements

We thank the Federal Office of Meteorology and Climatology (MeteoSwiss) for providing the COSMO-1 analysis data. We are grateful to Daniele Nerini (MeteoSwiss) and Jérôme Dujardin (EPFL) for fruitful discussions. Ophélia Miralles acknowledges funding from the Swiss National Science Foundation (project 200021_178824) and Daniel Steinfeld acknowledges funding from the Wyss Academy for Nature and the canton of Bern. This work was performed using computational resources from the EPFL Scientific IT and Application Support system.

2.10 Data statement

The downscaling GAN model is implemented in Python and the code is available on GitHub (https://github.com/OpheliaMiralles/wind–downscaling–gan). A pipeline for map generation using the GAN is publicly available, along with the best-performing model parameters. ERA5 reanalysis data can be downloaded freely from the Copernicus Climate Data Store (https://climate.copernicus.eu/climate–reanalysis) and topographical data can be downloaded freely from the SRTM 90m DEM Digital Elevation Database (http://srtm.csi.cgiar.org). Data from the COSMO-1 model is not open-source but can be obtained from MeteoSwiss on demand.

2.11 Appendix

2.11.1 GAN structure

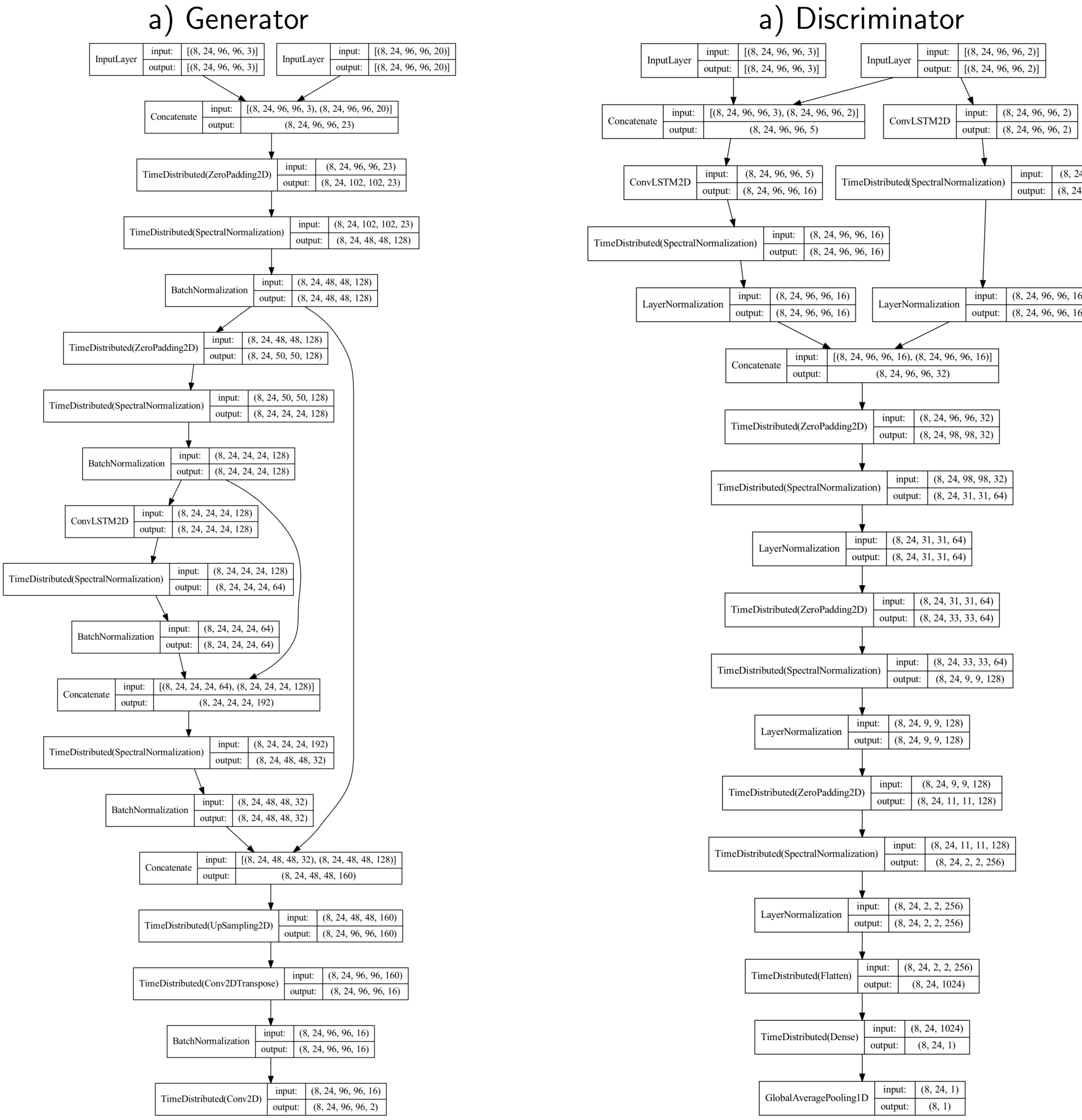
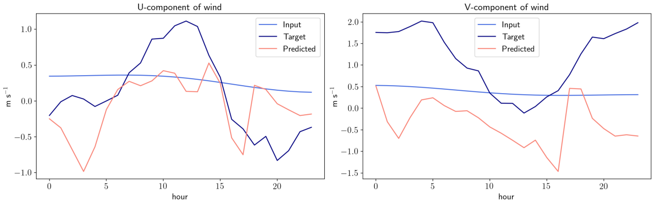

a) Generator
InputLayer input: [(8, 24, 96, 96, 3)] output: [(8, 24, 96, 96, 3)]
InputLayer input: [(8, 24, 96, 96, 20)] output: [(8, 24, 96, 96, 20)]
Concatenate input: [(8, 24, 96, 96, 3), (8, 24, 96, 96, 20)] output: (8, 24, 96, 96, 23)
TimeDistributed(ZeroPadding2D) input: (8, 24, 96, 96, 23) output: (8, 24, 102, 102, 23)
TimeDistributed(SpectralNormalization) input: (8, 24, 102, 102, 23) output: (8, 24, 48, 48, 128)
BatchNormalization input: (8, 24, 48, 48, 128) output: (8, 24, 48, 48, 128)
TimeDistributed(ZeroPadding2D) input: (8, 24, 48, 48, 128) output: (8, 24, 50, 50, 128)
TimeDistributed(SpectralNormalization) input: (8, 24, 50, 50, 128) output: (8, 24, 24, 24, 128)
BatchNormalization input: (8, 24, 24, 24, 128) output: (8, 24, 24, 24, 128)
ConvLSTM2D input: (8, 24, 24, 24, 128) output: (8, 24, 24, 24, 128)
TimeDistributed(SpectralNormalization) input: (8, 24, 24, 24, 128) output: (8, 24, 24, 24, 64)
BatchNormalization input: (8, 24, 24, 24, 64) output: (8, 24, 24, 24, 64)
Concatenate input: [(8, 24, 24, 24, 64), (8, 24, 24, 24, 128)] output: (8, 24, 24, 24, 192)
TimeDistributed(SpectralNormalization) input: (8, 24, 24, 24, 192) output: (8, 24, 48, 48, 32)
BatchNormalization input: (8, 24, 48, 48, 32) output: (8, 24, 48, 48, 32)
Concatenate input: [(8, 24, 48, 48, 32), (8, 24, 48, 48, 128)] output: (8, 24, 48, 48, 160)
TimeDistributed(UpSampling2D) input: (8, 24, 48, 48, 160) output: (8, 24, 96, 96, 160)
TimeDistributed(Conv2DTranspose) input: (8, 24, 96, 96, 160) output: (8, 24, 96, 96, 16)
BatchNormalization input: (8, 24, 96, 96, 16) output: (8, 24, 96, 96, 16)
TimeDistributed(Conv2D) input: (8, 24, 96, 96, 16) output: (8, 24, 96, 96, 2)
a) Discriminator
InputLayer input: [(8, 24, 96, 96, 3)] output: [(8, 24, 96, 96, 3)]
InputLayer input: [(8, 24, 96, 96, 2)] output: [(8, 24, 96, 96, 2)]
Concatenate input: [(8, 24, 96, 96, 3), (8, 24, 96, 96, 2)] output: (8, 24, 96, 96, 5)
ConvLSTM2D input: (8, 24, 96, 96, 2) output: (8, 24, 96, 96, 2)
ConvLSTM2D input: (8, 24, 96, 96, 5) output: (8, 24, 96, 96, 16)
TimeDistributed(SpectralNormalization) input: (8, 24 output: (8, 24
TimeDistributed(SpectralNormalization) input: (8, 24, 96, 96, 16) output: (8, 24, 96, 96, 16)
LayerNormalization input: (8, 24, 96, 96, 16) output: (8, 24, 96, 96, 16)
LayerNormalization input: (8, 24, 96, 96, 16 output: (8, 24, 96, 96, 16
Concatenate input: [(8, 24, 96, 96, 16), (8, 24, 96, 96, 16)] output: (8, 24, 96, 96, 32)
TimeDistributed(ZeroPadding2D) input: (8, 24, 96, 96, 32) output: (8, 24, 98, 98, 32)
TimeDistributed(SpectralNormalization) input: (8, 24, 98, 98, 32) output: (8, 24, 31, 31, 64)
LayerNormalization input: (8, 24, 31, 31, 64) output: (8, 24, 31, 31, 64)
TimeDistributed(ZeroPadding2D) input: (8, 24, 31, 31, 64) output: (8, 24, 33, 33, 64)
TimeDistributed(SpectralNormalization) input: (8, 24, 33, 33, 64) output: (8, 24, 9, 9, 128)
LayerNormalization input: (8, 24, 9, 9, 128) output: (8, 24, 9, 9, 128)
TimeDistributed(ZeroPadding2D) input: (8, 24, 9, 9, 128) output: (8, 24, 11, 11, 128)
TimeDistributed(SpectralNormalization) input: (8, 24, 11, 11, 128) output: (8, 24, 2, 2, 256)
LayerNormalization input: (8, 24, 2, 2, 256) output: (8, 24, 2, 2, 256)
TimeDistributed(Flatten) input: (8, 24, 2, 2, 256) output: (8, 24, 1024)
TimeDistributed(Dense) input: (8, 24, 1024) output: (8, 24, 1)
GlobalAveragePooling1D input: (8, 24, 1) output: (8, 1)

Figure 2.15: The GAN architecture of the generator (fig.a) and discriminator (fig.b) models for downscaling winds from ERA5 reanalysis to COSMO-1 data. Both graphs were made using the publicly available software Keras model plotting utility.

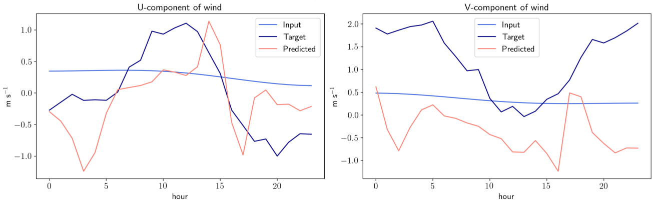
2.11.2 Examples of wind mean daily patterns from COSMO-1 blurred test sample

Mean daily wind pattern

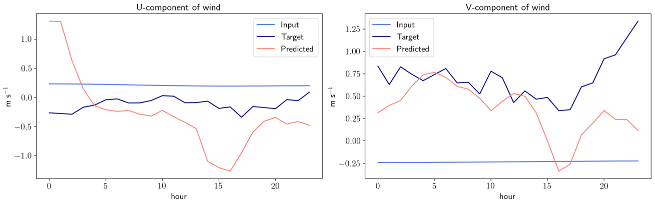
a) Fahy



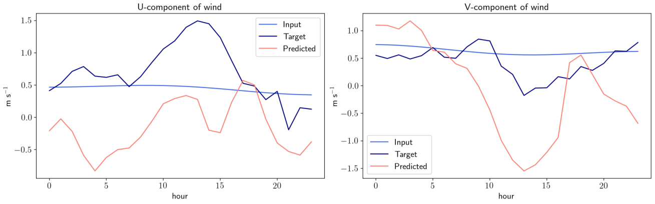
b) Aigle



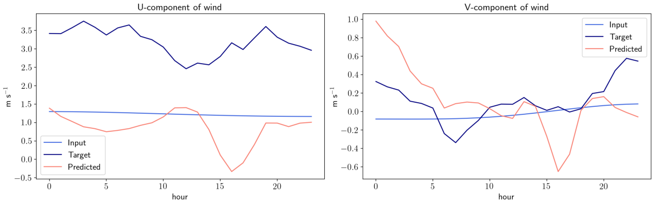
c) Bischofszell



d) Fribourg / Grangeneuve



e) Jungfraujoch



f) Zermatt

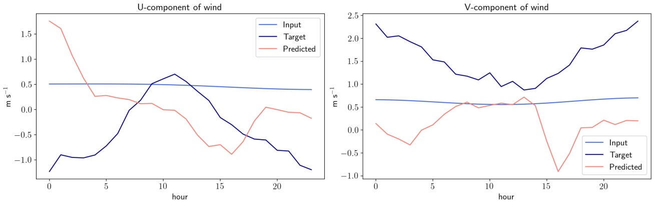


Figure 2.16: Mean daily pattern for u (left) and v (right) wind components in the Jura (fig.a and b), on the Swiss Plateau (fig.c and d), and in mountainous areas (fig.e and f) averaged over time. The locations of the validation sites are shown in Figure 2.1.

2.11.3 Examples of GAN prediction from COSMO-1 blurred test sample

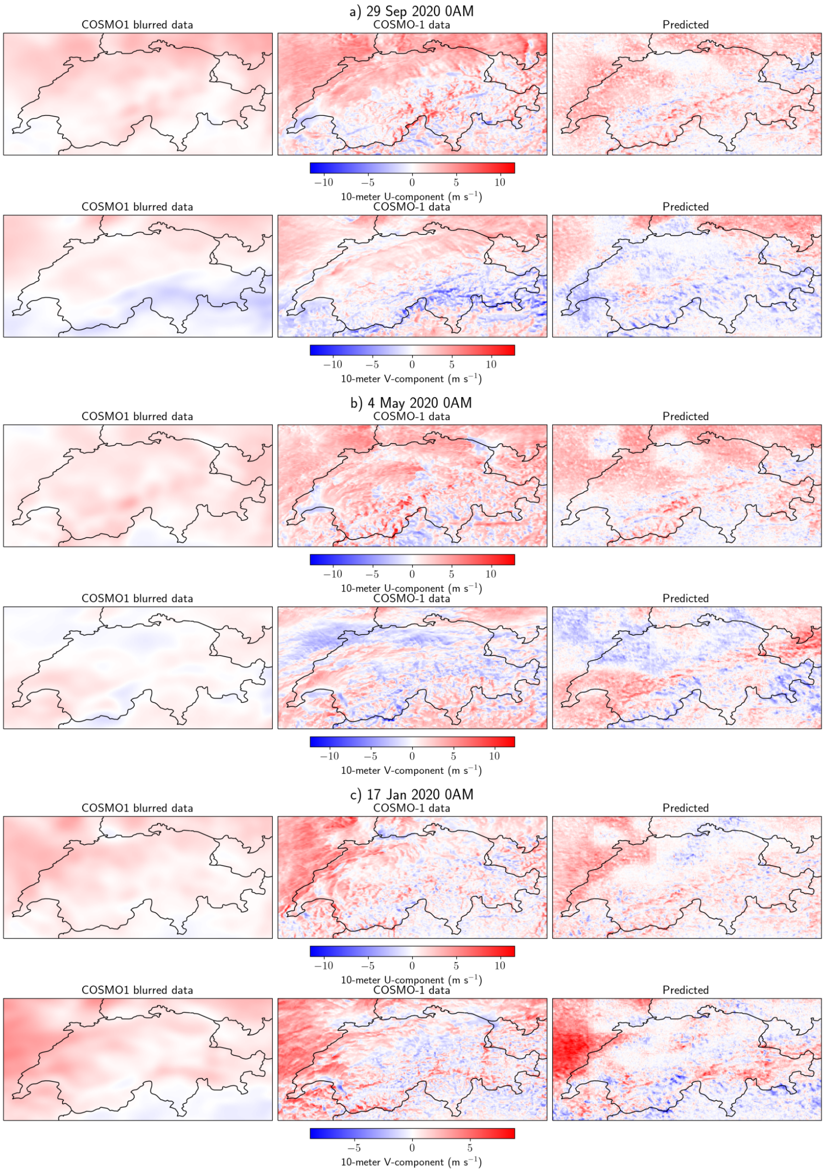


Figure 2.17: Prediction of the u and v components of 10-meter wind field by the GAN presented in Section 2.4. The columns represent inputs (left) from COSMO-1 blurred, the outputs from the COSMO-1 model at 1.1 km resolution (middle), and the model prediction (right).

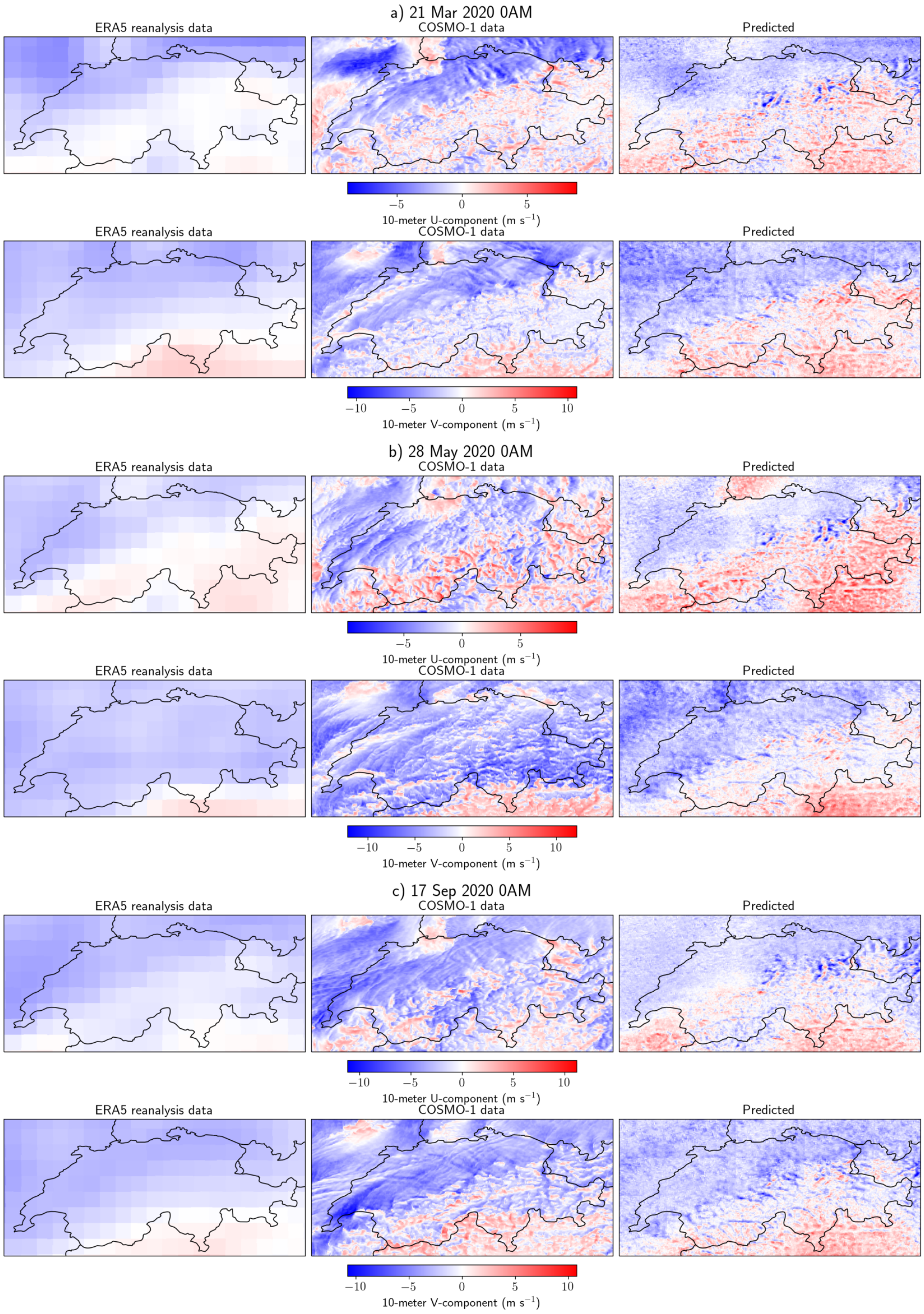


Figure 2.18: Prediction of the u and v components of 10-meter wind field by the GAN presented in Section 2.4. The columns represent inputs (left) from ERA5 25 km resolution grids, the outputs from the COSMO-1 model at 1.1 km resolution (middle), and the model prediction (right).

2.11.4 Examples of GAN prediction from ERA5 test sample

3 Bayesian Modelling of Insurance Claims for Hail Damage

Ophélia Miralles[1], Anthony C. Davison[1], Timo Schmid[2,3]

1 – Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

2 – Institute for Environmental Decisions, ETH Zürich, Switzerland

3 – Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, Switzerland

My Ph.D. supervisor provided helpful insights about the model and feedback about the manuscript. Timo Schmid provided the downscaled data used as input to the model and proposed some edits concerning the description of this data and the comparison of our model's performance to the benchmark. I provided the rest of the work, i.e., the data processing, methodology for the random line process, Bayesian modelling of insurance claims, sampling, model convergence and performance assessment, visualization of the main findings, and full writing of the manuscript.

This paper was submitted to the journal Annals of Applied Statistics in July 2023. It can be cited using the following bibliographical reference.

Miralles, O., Davison, A. C., and Schmid, T. (2023+). Bayesian modeling of insurance claims for hail damage. *Annals of Applied Statistics.* Submitted

Abstract

Despite its importance for insurance, there is almost no literature on statistical hail damage modeling. Statistical models for hailstorms exist, though they are generally not open-source, but no study appears to have developed a stochastic hail impact function. In this paper, we use hail-related insurance claim data to build a Gaussian line process with extreme marks to model both the geographical footprint of a hailstorm and the damage to buildings that hailstones can cause. We build a model for the claim counts and claim values, and compare it to the use of a benchmark deterministic hail impact function. Our model proves to be better than the benchmark at capturing hail spatial patterns and allows for localized and extreme damage, which is seen in the insurance data. The evaluation of both the claim counts and value predictions shows that performance is improved compared to the benchmark, especially for extreme damage. Our model appears to be the first to provide realistic estimates for hail damage to individual buildings.

3.1 Introduction

Global warming has already begun to affect the behaviour of insurers worldwide, both by increasing premiums and by making companies unwilling to underwrite some risks; a recent example is the May 2023 decision by the US company State Farm to cease offering house insurance to new clients in California. Hail is of particular interest to Swiss insurance companies because of large annual insured losses, averaging several million Swiss francs (CHF) (Botzen et al., 2010), though there is substantial year-to-year variability. Northern Switzerland experienced significant hail events in 2021, leading to estimated insured losses of 2 billion CHF and placing a heavy financial burden on insurers (Müller, 2021). The risk of large losses due to hailstorms is increasing as a result of the construction of more new buildings each year, and climate change may increase the frequency of damaging hailstorms in Europe more broadly (Rädler et al., 2019). Destructive hailstorms are also an important risk for agriculture, buildings, and vehicles elsewhere in the world (Changnon, 2008; Warren et al., 2020), but despite their importance, such storms can be very localised and are hard to model.

The literature on the statistical modeling of the impact of hailstorms on buildings is very limited. Although stochastic models for hailstorm risk (Deepen, 2006; Otto, 2009; Punge et al., 2014; Púčik et al., 2017) or hailstone size (Perera et al., 2018; Liu et al., 2021) exist, most open-source studies on hailstorm impact use deterministic functions to link the intensity of a hail hazard to its monetary damage. The spatial footprint of hail events has been discussed in a few recent modeling studies, in which hailstorms are either represented as ellipses (Otto, 2009) or stretches of constant width (Deepen, 2006). Punge et al. (2014) use a Poisson distribution to estimate the frequency of hail in Europe on a 50 × 30km grid, using a bimodal normal distribution in each grid cell to estimate the pointwise probability of hail based on hail reports and observations of overshooting cloud tops, whose presence for over ten minutes can indicate thunderstorm severity. Deepen (2006) generates a stochastic catalog of hail events in Germany by simulating areas of fixed width, random location, and random length. Hailstorms are also simulated by modeling random hailstones in Liu et al. (2021), while Perera et al. (2018) propose a hailstone size distribution to aid in impact estimation.

On the damage modeling side, Hohl et al. (2002) propose a logistic impact function derived from hail kinetic energy, and Schuster et al. (2006) further explore the link between this energy and impacts. Claim data from insurance companies, though not usually publicly accessible, is a valuable source of information on hail impacts and has been used to complete the radar signal for hail in several recent studies. The hail damage model developed in Schmidberger (2018), for example, derives hail tracks in Germany from radar and insurance data, and Deepen (2006) uses simulated hail footprints to model damage to cars through a Poisson distribution fitted with vehicle insurance data. Brown et al. (2015) use insurance data to explore the link between roof material and hail impact on buildings in Texas.

These studies all involve randomness from the hail event itself. Indeed, radar-based proxies are used to derive the probability and/or the expected intensity of a hailstorm on grids with resolution of several kilometers. Those proxies are often chosen over direct hail measurements with automatic hail sensors (Kopp et al., 2023) because they are more spatially consistent. However, the monetary impacts due to hailstorms appear in very narrow and localized tracks that are usually poorly represented by models with such grids.

The goal of the present study is to propose a spatially consistent model for insurance claims related to hail damage at the building level. This model differs from previous ones, as the probability and intensity of a hail event are supposed to be known, and stochasticity comes from the possible spatial impacts of a hail storm. The model we develop seems to be the first to combine a random line process and an extreme value model in order to represent hail damage tracks accurately. We describe the data available to us in Section 3.2, introduce a line model with extreme marks in Section 3.3, and describe the results we obtain when applying this model in the Swiss canton of Zürich, henceforth “the canton”, in Section 3.6.

3.2 Data and initial analysis

3.2.1 Data

Two variables representing hail risk were provided by the meteorological service MeteoSwiss in the scope of the scClim project, the purpose of which is to combine knowledge from different fields to create a continuous model chain from simulating thunderstorms to quantifying the monetary impacts of hail in Switzerland. Gridded one-kilometer-resolution maps of the probability of hail (POH) and the maximum expected severe hail size (MESHS), derived from volumetric radar reflectivity (Nisi et al., 2016), are available for the canton during the convective season (April–September) for the years 2002–2021. The MESHS offers more spatial granularity than the POH and thus was preferred (Figure 3.1).

The output from a MESHS-based deterministic damage function developed during the scClim project (Schmid et al., 2023) and calibrated with insurance data in the canton is also available for the same period; the results in that paper use the same function, calibrated over several Swiss cantons. The data made available to us contain predictions for the numbers of affected assets and the monetary hail damage in each cell of a 2km square grid. For conciseness below we shall use the terms “grid cell” or sometimes just “cell” in reference to this grid. The PAA and damage functions were developed following the hazard/exposure/vulnerability methodology of the CLIMADA framework described by Aznar-Siguan and Bresch (2019) and will be simply referred to as “CLIMADA” below. We use a per-building version of the CLIMADA output, referred to as “downscaled CLIMADA” below, that will be used as input for our claim value model. This per-building damage is a naive downscaling of the per-cell CLIMADA damage, and attributes weights to each building as a function of its insured value. In practice, this means that every building in the cell is impacted when CLIMADA predicts a positive per-cell value, artificially inflating the number of buildings affected.

In addition to hail-related variables, we use wind direction from the state-of-the-art ERA5 reanalysis from the European Centre for Medium-Range Weather Forecasting (Hersbach et al., 2020), available from 1979 onwards on a 25km square grid over Europe.

Insurance data for hail-related claim damage to buildings in the canton is also available from one of the stakeholders of scClim, the Zürich cantonal insurance company GVZ. These data consist of individual claim values for buildings for the period 2000–2022, during which there were 244 days with positive claims somewhere in the canton and a total of 46254 claims. These are the amount finally paid by the insurance company in Swiss francs (CHF) following hail damage to a building and not estimated values of monetary damage. The construction year, volume, and actualized insured value of every insured building in the canton are also available. We did not explore potential issues linked with preferential sampling, since the owner of every building in the canton is legally obliged to take out natural hazards insurance with GVZ. Consequently, data are very dense, though there are spatial disparities in exposure

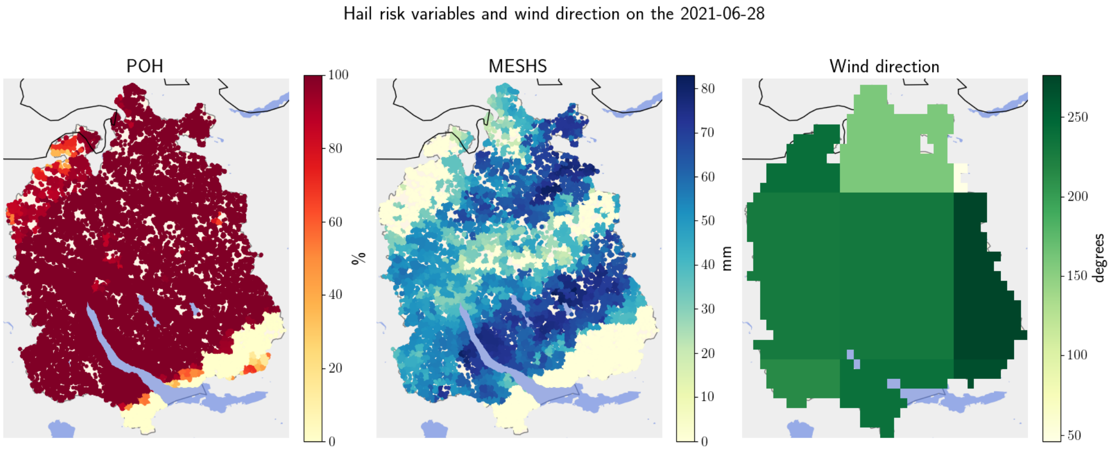


Figure 3.1: Hail risk covariates POH (left), MESHS (center), and wind direction (right) on 28 June 2021 in the canton (47.15–47.70°N, 8.35–8.99°E).

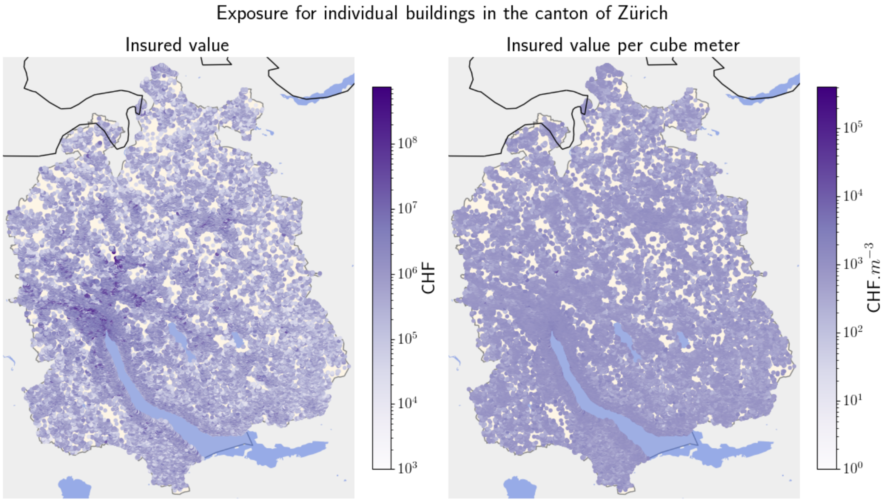


Figure 3.2: Insured monetary value for individual buildings (left) and the exposed monetary value per m^3 (right).

owing to variations in population density. As exposure equals the insured monetary value of buildings, urban areas are much more exposed than suburban areas or the countryside, but the distribution of buildings is spatially rather homogeneous in terms of average exposed value per cubic meter; see Figure 3.2. Our claim value model should be able to predict the difference between the true damage, i.e., the claim values reported by GVZ, and downscaled CLIMADA damage when a positive claim was recorded by GVZ. We call this target variable the “residual damage”.

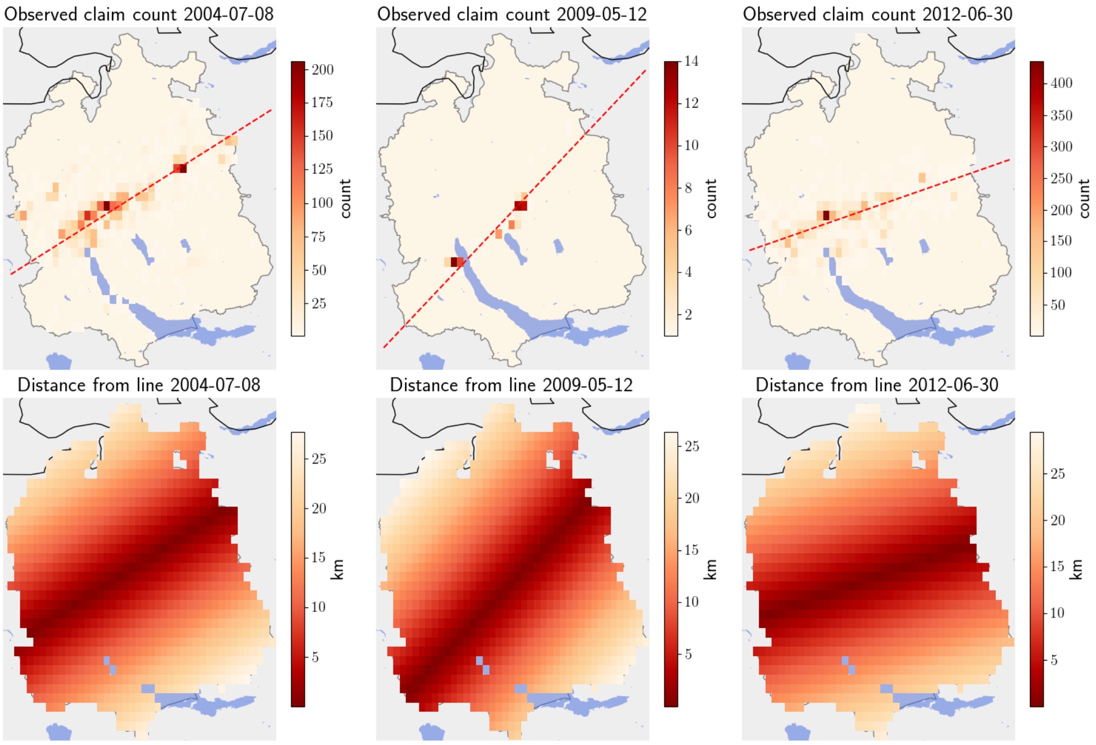

Observed claim count 2004-07-08
Observed claim count 2009-05-12
Observed claim count 2012-06-30
count
Distance from line 2004-07-08
Distance from line 2009-05-12
Distance from line 2012-06-30
km

Figure 3.3: Example days with more than 50 recorded claims. Red-colored squares in the first row correspond to the aggregated claim count per 2km grid cell. The red line has a slope corresponding to the average wind direction on that day. The graphs on the second row represent the distance from the centroid of a 2km grid cell to the red line.

3.2.2 Exploratory analysis

Hail footprints

Literature exploring hailstorm patterns agrees on an ellipsoidal shape (Otto, 2009; Punge et al., 2014), or a sufficiently wide straight sketch (Deepen, 2006) for modeling the spatial extent of a hail footprint. Two of those studies explore hail tracks in Germany (Deepen, 2006; Otto, 2009) while the third treats hailstorm footprints in Central Europe (Punge et al., 2014). Although modeling single hail events on a large territory with locally bounded shapes seems reasonable, the canton of Zürich covers a much smaller area than Germany or Central Europe, and individual hail-related claims in our data suggest that hail storms progress along a line in space with a very narrow lateral dispersion; see Figure 3.3. Exploratory analysis suggested that the line direction is related to the average wind direction on the day of a hail storm.

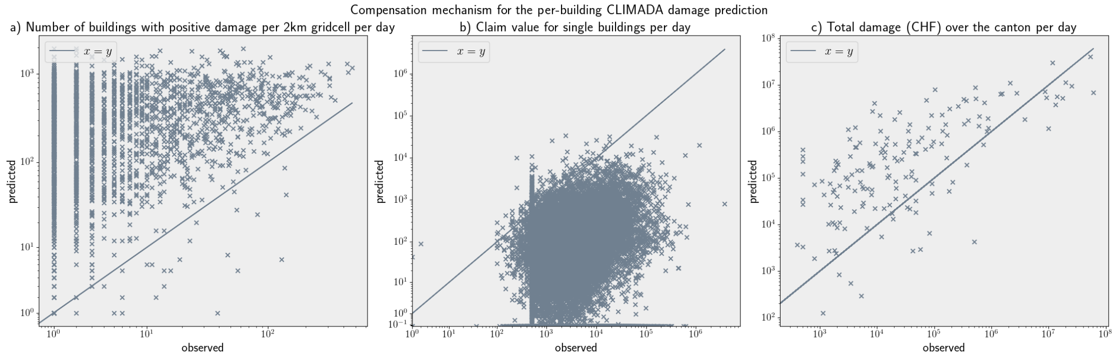


Figure 3.4: Number of buildings with positive predicted damage per grid cell per day using the downscaled CLIMADA impact function: (a) claim value per single building per day: (b) and total damage recorded on a given day aggregated over the canton: (c) all with the line $x = y$.

Local and extreme damage

The deterministic damage function developed through CLIMADA provides good estimates for the amount of monetary damage on a grid and adequately represents spatial patterns over the canton. However, the naively downscaled CLIMADA compensates mispredicted individual claim damage values with a very large number of claims (see Figure 3.4) as damage is distributed over all exposed buildings within a cell. Indeed, CLIMADA cannot distinguish between a few claims of high damage and many claims with low damage in a cell. Furthermore, increases in the frequency or intensity of hail would impact either the count or the value of hail-related damage, causing the compensation mechanism described above to fail. There is thus a need for a model to provide realistic values of the damage per building, which is highly relevant for insurance. The objective of this study is to provide such a model for hail-related monetary damage that respects the count/size ratio observed in the claim dat, by lowering the frequency of positive claims while allowing their values to be locally extreme.

Figure 3.5 shows strong seasonal variation in claims: most damage occurs between June and August, with a peak in June, and claims occurring in April or September look less heavy-tailed than in May–August. Henceforth we only consider claims in April–September, which represent 99.74% of the total data, and define the “hail season” to be May–August.

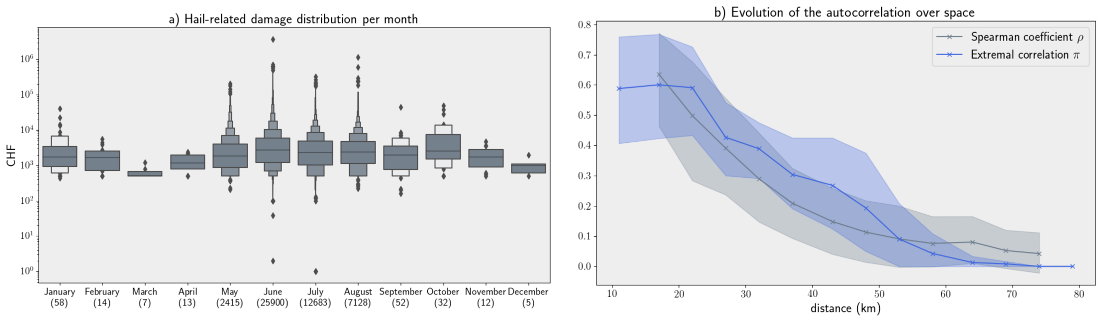


Figure 3.5: Exploratory analysis of claim values: (a) average total claim value over the canton per month: (b) Spearman correlation ρ and extremal correlation π for pairwise time series of claim values per cell as a function of the distance between cells. In (b) the solid lines represent the average pairwise correlation over equally distant cells, and the shaded areas show the 90% confidence range.

3.2.3 Spatiotemporal correlation

Claims for damage from a hail event can be made on that day or with a lag of a few days, so the damage function derived through CLIMADA pre-processes the original insurance data to cluster claims received during a 4-day window around a big hail event detected with the POH values. For any reported claim, POH values in a ±2-day window are scanned. If a POH higher than 50% of that on the day of the reported claim is observed, the claim date is changed to the day with the highest POH (Schmid et al., 2023). This pre-processing step largely succeeds in removing short-term autocorrelation in the claim values, so here we focus on spatial correlation. For this purpose, we introduce the extremal correlation $\pi_h(u) = \mathbb{P}(X_{s+h} \geq u \mid X_s \geq u)$ of the variable X_s, where s denotes the spatial location of a cell, u is a high threshold and h a spatial lag. The threshold u is chosen by applying the threshold selection method described in Varty et al. (2021) to the log total sum of damage X_s (see Supplementary Material). We also study the pairwise Spearman correlation ρ for the daily sum of claim values per cell. Figure 3.5 shows that both π and ρ decrease as the distance between two spatial locations increases.

3.3 Key model elements

In this section we describe the key elements of a model to reproduce the very localized but large damage seen in the data. We first explain how the long and narrow hail footprint (as observed in Section 3.2.2) can be modeled using a Gaussian line process, and then recall the peaks-over-threshold approach from extreme value theory, which is used to account for large impacts on individual buildings (see Section 3.2.2). More details of the modeling are given in Section 3.4.

3.3.1 Random line process

Claim values are usually represented as a spatiotemporal point process $s_t = (t, x, y)$, with $t \geq 0$ and (x, y) the geographical coordinates. In the forthcoming discussion, a sequence of distance-conserving transformations will be applied to map this onto a coordinate system that is better suited for defining the random line model.

For $t > 0$ let $\Theta_t \in [-\pi, \pi]$ and α_t respectively be the time-varying random angle and vertical deviation of a line $\mathscr{L}$ from a chosen origin point $s_0 = (x_0, y_0)$, defined as the set of points

$$\mathscr{L} = \left\{ (s,t) : s^{\mathrm{T}} \begin{pmatrix} -\tan\Theta_t \\ 1 \end{pmatrix} = \alpha_t \right\}, \tag{3.1}$$

where $s = (x, y)$ denotes geographical coordinates and t is the discrete time coordinate. At time t the projection of any pair of spatial coordinates $s = (x, y)$ in the coordinate system in which $\mathscr{L}$ is the horizontal axis and the origin is s_0 may be written as

$$s_t^{\mathscr{L}} = \begin{pmatrix} \cos\Theta_t & \sin\Theta_t \\ -\sin\Theta_t & \cos\Theta_t \end{pmatrix} \begin{pmatrix} x \\ y - \alpha_t \end{pmatrix}, \tag{3.2}$$

and the orthogonal projection of s onto the line would thus be the point $\pi_t^{\mathscr{L}}(s) = (\cos\Theta_t x + \sin\Theta_t y, 0)$ in the new coordinate system. The Euclidean distance between a point in space and the line $\mathscr{L}$ can be computed in any coordinate system, and as $s_t^{\mathscr{L}}$ and $\pi_t^{\mathscr{L}}(s)$ have the same x-coordinate in the new system, at time t this distance can be expressed as

$$d_t(s, \mathscr{L}) = |y'| = |(y - \alpha_t)\cos\Theta_t - x \sin\Theta_t|. \tag{3.3}$$

To allow the intensity of points at time t to be highest close to the random line $\mathscr{L}$, we define a spatiotemporal Gaussian field $X^{\mu}(s,t)$ whose mean is

$$m_t(s) = \frac{\sigma_m}{1 + d_t(s, \mathscr{L})} - 1, \tag{3.4}$$

where σ_m is a dispersion parameter that controls the concentration of points around $\mathscr{L}$. In accordance with the exploratory analysis, we chose a correlation function ρ such that the correlation between $X^{\mu}(s_0, t)$ and $X^{\mu}(s_1, t)$ decreases when the distance

w between s_0 and s_1 increases (see Figure 3.5). A common choice is the Matérn correlation function

$$\rho(w) = \{2^{\nu-1}\Gamma(\nu)\}^{-1}(w/l)^{\nu}K_{\nu}(w/l), \quad w > 0, \tag{3.5}$$

where $\nu > 0$ is a shape parameter controlling the smoothness of the Gaussian process, $l > 0$ is a scale parameter, $\Gamma(\cdot)$ is the Gamma function, and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind. After some experimentation, we took $\nu = 1.5$, which gives fields of similar smoothness to the data, and estimate the parameter l as part of a hierarchical Bayesian model. When $\nu = 1.5$, Equation (3.5) simplifies to

$$\rho(w) = \left(1 + \sqrt{3}w/l\right)\exp\left(-\sqrt{3}w/l\right), \quad w > 0. \tag{3.6}$$

3.3.2 Marginal model for extreme claim values

The exploratory analysis suggests using extreme value theory to model the largest claim values. The generalized Pareto distribution,

$$\mathrm{GPD}^{u}(x) = 1 - \left(1 + \xi\frac{x-u}{\sigma_u}\right)_{+}^{-1/\xi}, \quad x \in [u, \infty), \tag{3.7}$$

where $a_+ = \max(a, 0)$ for real numbers a, provides a standard model for the exceedances of a high threshold u. The model depends on a shape parameter ξ that determines the weight of the distribution tails and on a scale parameter σ_u; both are specified in Section 3.4.2. We select a constant threshold u by applying the method described by Varty et al. (2021) to the log of the total sum of claim values over the canton; see the Supplementary Material.

3.4 Modeling extreme hailstorms

Our model for hail damage uses a discrete zero-inflated count process for the number of claims and a continuous two-part distribution for hail damage values. In the following section, the spatiotemporal matrix of observed covariates will be designated by the letter M. If a variable needs to be specified, it is written M^{NAME} — for the MESHS, for instance, we write M^{MESHS}. Building exposure, MESHS, POH, CLIMADA predicted claim count and downscaled CLIMADA value are respectively designated by Exp, MESHS, POH, NC, and YC. The count of individual claims in a grid cell on a given day is denoted by N, while the value of an individual claim is designated by Y.

3.4.1 Hail damage count

The claim count is modeled on a 2km square grid. Hail can either strike very locally and violently or can be spread out more smoothly, as observed in the data, in which the maximum number of observed claims in a cell on a single day is 469, and the nonzero minimum is 1. In view of this wide range of values, we model N as a negative binomial. Positive counts are scarce, so we model them as realisations of a zero-inflated negative binomial random variable, with probability mass function

$$\mathrm{NB}_{\psi,\mu,\alpha}(x) = \begin{cases} (1-\psi) + \psi\left(\dfrac{\alpha}{\alpha+\mu}\right)^{\alpha}, & x = 0, \\ \psi\dfrac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\dfrac{\alpha}{\mu+\alpha}\right)^{\alpha}\left(\dfrac{\mu}{\mu+\alpha}\right)^{x}, & x = 1,2,3,\ldots, \end{cases} \tag{3.8}$$

where $\psi \in (0,1)$, $\mu > 0$ and $\alpha > 0$ is a shape parameter. We set $N \mid \psi, \mu, \alpha \sim \mathrm{NB}_{\psi,\mu,\alpha}$.

The probability ψ of observing a non-zero claim in grid cell s on day t is modeled as

$$\psi(s,t) = \operatorname{expit}\left\{\psi_0 + \psi_1 \mathbf{1}_{M^{\mathrm{NC}}_{s,t}>0} + \psi_2 M^{\mathrm{NC}}_{s,t} m_t(s)\right\}, \tag{3.9}$$

where $\operatorname{expit}(x) = \{1+\exp(-x)\}^{-1}$ and $\psi_0, \psi_1, \ldots$ are real parameters. We define the mean μ of the negative binomial variable through the equation

$$\log\mu(s,t) = \mu_0 + \sum_{i=1}^{3} \mu_{1i}\left(M^{\mathrm{NC}}_{s,t}\right)^{i} + \mu_2 M^{\mathrm{NC}}_{s,t} m_t(s) + X^{\mu}(s,t) + \epsilon(t), \tag{3.10}$$

where X^{μ} is a spatiotemporal Gaussian field whose mean m_t and covariance function ρ are given respectively in (3.4) and (3.6), and $\mu_0, \ldots$ are real parameters. The Gaussian noise $\epsilon(t)$ has mean zero and one variance for the months April and September and another variance for the months May–August.

3.4.2 Hail damage values

The number of positive claims (46,254) is much smaller than the roughly 350,000 cell-date combinations in which hail events might have occurred, so it is reasonable to model spatial patterns at a coarser resolution than the 2km grid used for the counts. Spatial Gaussian random fields used to model unobserved covariates underlying the hail damage values are thus defined over a grid of resolution roughly 10km in which each cell has at least 100 positive claims over the years 2000–2015.

Downscaled CLIMADA under-predicts the values of 99.3% of reported claims, so we model only positive errors, i.e., if the predicted count N is positive, we only allow a shift upwards from the downscaled CLIMADA value M^{YC}. The resulting residual hail damage variable $Z = Y - M^{\mathrm{YC}}$ is modelled using a beta model for non-extreme values and a generalised Pareto model for extreme values.

We first introduce a binary variable R to model the event that a claim value exceeds the threshold u $(R = 1)$ or not $(R = 0)$, with success probability

$$p(s,t) = \operatorname{expit}\left\{p_0 + p_1 M_{s,t}^{\mathrm{POH}} + p_2 M_{s,t}^{\mathrm{MESHS}} + p_3 M_{s,t}^{\mathrm{MESHS}\cdot\mathrm{POH}} + p_4 M_{s,t}^{\mathrm{Exp}} + \chi(s) + \epsilon_p(t)\right\}, \tag{3.11}$$

where $M_{s,t}^{\mathrm{MESHS}\cdot\mathrm{POH}} = M_{s,t}^{\mathrm{MESHS}} M_{s,t}^{\mathrm{POH}}$ and χ and ϵ_p normally-distributed random effects respectively per grid cell and season. The Beta and Pareto models for non-extreme and extreme claim values are detailed below. In the following sections, f denotes the function $f : x \mapsto \log(1 + x)$.

Non-extreme residual damage

Residual damage Z for which $f(Z) \le u$ is described by letting $Z/f^{-1}(u)$ have a beta density with mean ν and variance $\nu(1-\nu)/\kappa + 1$,

$$\mathrm{Beta}_{\nu,\kappa}(x) = \frac{x^{\nu\kappa-1}(1-x)^{(1-\nu)\kappa-1}}{B\{\nu\kappa,(1-\nu)\kappa\}}, \quad x \in (0,1), \tag{3.12}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. We set

$$\frac{Z}{f^{-1}(u)} \quad \mid \quad \left\{f(Z) \le u\right\}, \mu_B, \sigma_B \sim \mathrm{Beta}_{\nu,\kappa}, \tag{3.13}$$

and model the mean of this variable via the expression

$$\nu_t(s) = \operatorname{expit}\left\{\nu_0 + \nu_1 M_{s,t}^{\mathrm{POH}} + \nu_2 M_{s,t}^{\mathrm{MESHS}} + \nu_3 M_s^{\mathrm{Exp}} + X^{\beta}(s)\right\}, \tag{3.14}$$

where X^{β} is a spatial Gaussian process with zero mean and covariance kernel function

$$\rho_\alpha(w) = \left(1 + \frac{w}{4l^2}\right)^{-2}. \tag{3.15}$$

Extreme residual damage

Damage arising when $f(Z) > u$ is modeled by letting $f(Z) - u$ be a generalized Pareto variable,

$$f(Z) - u \mid \{f(Z) > u\}, \sigma_u, \xi \sim \mathrm{GPD}_{\sigma_u,\xi}; \tag{3.16}$$

the distribution function is given in (3.7). The extremal correlation observed in Figure 3.5 is accommodated by allowing σ_u to depend on POH, MESHS, exposure covariates and unobserved spatial discrepancies and time-related autocorrelation, respectively modeled with a Gaussian process and an auto-regressive process, leading to

$$\log \sigma_{u,t}(s) = \sigma_0 + \sigma_1 M_{s,t}^{\mathrm{MESHS}} + \sigma_2 M_{s,t}^{\mathrm{MESHS}\cdot\mathrm{POH}} + \sigma_3 M_s^{\mathrm{Exp}} + X^{\sigma}(s) \tag{3.17}$$

where X^{σ} is a spatial Gaussian process with zero mean and a Matérn covariance matrix (3.6), in which the Euclidean distance has been replaced by the chordal distance because our Gaussian process occurs on the surface of a sphere, whose curvature should be reflected by our model (Jeong et al., 2017). The chordal distance is the length of a line passing through the three-dimensional Earth to connect two points on its surface. For two locations s_1 and s_2 with respective geographical coordinates (x_1, y_1), (x_2, y_2), the chordal distance between s_1 and s_2 is defined by

$$\mathrm{C}(s_1, s_2) = 2r \arcsin\left[\frac{1}{2}\left\{1 - \cos_p(y_2 - y_1) + \cos_p y_1 \cos_p y_2 (1 - \cos_p(x_2 - x_1))\right\}\right]^{1/2},$$

where $r = 6371$km is the Earth's radius and $\cos_p(z) = \cos(\pi z/180)$. In a small area such as the canton, using the chordal distance instead of the Euclidean distance might not make a huge difference, but it would matter if the model was used for larger regions.

In view of Figure 3.5 we allowed the shape parameter to vary as

$$\xi(t) = \begin{cases} \xi_1, & t \in \{\text{May, June, July, August}\}, \\ \xi_2, & \text{otherwise.} \end{cases} \tag{3.18}$$

3.5 Model fitting and validation

3.5.1 Technical challenges

Fitting the Bayesian hierarchical model described in Section 3.4 is challenging due to its complexity, the size of the parameter space and the large number of data points. Recent advances in spatial statistics allow better computational efficiency for Bayesian models with latent variables (Rue et al., 2017). In our case, it would be desirable to use R-INLA, which has been widely and successfully used for environmental data (e.g., Castro-Camilo et al., 2019; Koh et al., 2023). As our work is part of a collaboration involving several subprojects mostly written in the programming language Python, the model described in Sections 3.3 and 3.4 was also coded in Python so that our collaborators would find it accessible. There is no equivalent of R-INLA in Python, and reproducing it for Python users would have taken far too long, so despite the resulting drop in computational efficiency we resorted to Markov chain Monte Carlo (MCMC) methods.

In contrast to Metropolis–Hastings steps, which make trajectory proposals within a possibly skewed ball (Hastings, 1970; Metropolis et al., 2004), or to Gibbs sampling, which generally only moves in a few dimensions at a time (Gelfand, 2000), Hamiltonian Monte Carlo (HMC) generates proposals based on the shape of the posterior by using its gradient (Betancourt, 2017). In MCMC algorithms, the termination criterion identifies when a trajectory is long enough for adequate exploration of the neighborhood around the current state, but in HMC, this criterion should be chosen to compromise between taking full advantage of the Hamiltonian trajectories and wise use of computational resources (Betancourt, 2017). The No-U-Turn Sampler (NUTS) is a HMC algorithm that proves particularly efficient in converging for high-dimensional posterior distributions (Homan and Gelman, 2014). Indeed, NUTS uses a dynamic termination criterion that considers only the position and momentum of a trajectory's boundaries: when it is met, further sampling typically leads to neighborhoods that have already been explored. In addition to this specific termination criterion, NUTS implements a multiplicative expansion of the trajectory that allows fast exploration of the parameter space within limited computer memory (Betancourt, 2017). We used NUTS for the count model described in Section 3.4.1 and for the extremal model described in Section 3.4.2.

For the non-extreme claims model detailed in Section 3.4.2, a differential evolution

Metropolis (DE-MC) sampling step with a snooker updater was used, as it is more efficient and faster than the classical random walk Metropolis step. DE-MC combines a differential evolution genetic algorithm and MCMC simulation (Ter Braak, 2006). The snooker updater makes it less computationally expensive than classical DE-MC, as it updates different chains in parallel with information from past states (Ter Braak and Vrugt, 2008), and is faster than NUTS for this model, with no significant impact on the results.

We systematically exclude the initial 500 samples drawn, which are reserved for a tuning phase during which the sampler dynamically adjusts the step sizes and scalings to optimize its subsequent performance. We monitor the convergence of the model parameters using informal diagnostic plots; see the Supplementary Material. We check that autocorrelation has decreased to approximately zero during the sampling, and examine trace plots of the sampled parameters for the absence of patterns. Running the claim counts model took about five hours for about a thousand parameters. For the claim values, fitting the model took two hours for the GPD model (38 parameters) and less than an hour for the Beta model (36 parameters). Non-informative priors were found to perform significantly better than weakly informative priors in our case and thus were attributed to all of the model parameters. To make sure the posterior is proper, we check that its distribution percentiles and mean are finite and reasonable.

3.5.2 Metrics

To assess the model's performance in improving spatial patterns we use diagnostic quantities that include the following two specific metrics.

The spatially convolved Kolmogorov–Smirnov statistic (Miralles et al., 2022) represents the disagreement between the spatial distributions of the generated and observed images and is computed as the maximum absolute difference of empirical cumulative distribution functions for the generated and true damage, summed over 10 × 10 patches of the image of interest. The aim is to obtain a metric with properties close to those of the Fréchet inception distance (Heusel et al., 2017) for images by assessing the match between predictions and targets, as a human eye would. After extracting M spatial patches of constant size from the target and predicted images, we set

$$\mathrm{SKSS} = \sum_{t \leq N_T, j \leq M} \max_{x \in \mathbb{R}} |F_{jt}(x) - \hat{F}_{jt}(x)|,$$

where F_{jt} represents the empirical cumulative distribution function of the hail damage for a single spatial patch j and time t and $\hat{F}_{jt}$ is its analog for predicted damage. This metric evaluates the local agreement between two distributions rather than focusing on individual pixels.

The log-spectral distance (Rabiner and Juang, 1993) is expressed as the log-difference of power spectra between the generated and realized samples,

$$\mathrm{LSD} = \left\{ \frac{1}{2N_T \times P} \sum_{t \le N_T, i \le P} \left[10 \log_{10} \left(\frac{|g(c_{it})|^2}{|g(\hat{c}_{it})|^2} \right) \right]^2 \right\}^{1/2},$$

where g is the Fourier transform, $|g(\cdot)|^2$ the power spectrum, c is the target map of damage and $\hat{c}$ its estimate. This evaluates whether the generated images reproduce the spatial structures noticeable in the target images.

3.6 Results

There is no visible long-term trend in either the claim count or value in the insurance data; we can thus split data into sets of consecutive years. The training set is built from years up to 2015, the validation data comprises the years 2016–2017, and later years are used as the test set. In the following analysis of the results, unless specified otherwise, the average prediction for the test set over 1000 different sets of parameters sampled from the posterior distribution is used to construct graphs and maps. We recall that our objectives are to accurately capture spatial patterns for hail damage, to be able to predict localized and extreme damage and to match the distribution of the target data provided by GVZ. We shall see that the fitted random line process with extreme marks achieves this. We start by evaluating the performance of the Gaussian line process, then explain the procedure for combining counts and claim values, and finally discuss predicted claims.

3.6.1 Claim counts

Our benchmark for evaluating the performance of the random line model presented in Section 3.3.1 is the percentage of affected assets (PAA)-based gridded claim count predicted with CLIMADA (Schmid et al., 2023). The PAA, defined as the per-cell proportion of damaged buildings, is expected to increase with the value of MESHS,

Table 3.1: Comparison of the false alarm rate, sensitivity, specificity and positive predictive value (%) averaged over the time dimension for CLIMADA's hail damage model and the random line model. If a, b, c and d denote the true positive, false positive, false negative and true negative numbers, the false alarm rate is computed as $b/(b+d)$, the sensitivity as $a/(a+c)$, the specificity as $d/(b+d)$ and the positive predictive value as $a/(a+b)$.

	False Alarm	**Sensitivity**	**Specificity**	**Positive Predictive Value**
CLIMADA	72.1	64.8	27.9	62.9
Model	29.7	52.1	70.3	77.0

since hailstorms with larger hailstones should cause more damage. Figure 3.6(a) shows that the observed PAA does not increase linearly and is very variable, but that the predicted and observed values are quite similar, while the 95% prediction range captures the observed variation well. The impact function computed through CLIMADA tends to over-predict the percentage of affected assets for any observed MESHS value.

Figure 3.6(b) suggests that small claim counts are over-predicted by our model. For days with more than 1000 recorded claims, the distribution of predicted counts is very close to the observed claim counts; the line process captures days with very many claims particularly well.

Table 3.1 assesses how much our model improves on CLIMADA in terms of predicting the daily claim count. The random line model reduces the false alarm rate by about 40% and increases the positive predictive value by 15% and the specificity by 42%, so it makes fewer mistakes on average in predicting both positive and zero counts. Compared to CLIMADA total predicted counts per day, Table 3.1 shows that the sensitivity has dropped by 12%, i.e., our model might miss days with a positive claim count, but inspection of the data reveals that it only misses days with fewer than ten claims and less than CHF 10K overall damage.

Examples of predicted counts plotted in Figure 3.7 show that the line model helps to concentrate the predicted damage on straight lines, giving results that resemble observed claim counts which are usually concentrated in hail streaks of width just a few km in the Alpine region (Nisi et al., 2018). In contrast, the predictions from CLIMADA are broadly distributed according to the MESHS footprint, which typically covers a whole storm cell core (Nisi et al., 2016). The average predicted count over the canton is also closer to the realized value using the line model than with CLIMADA.

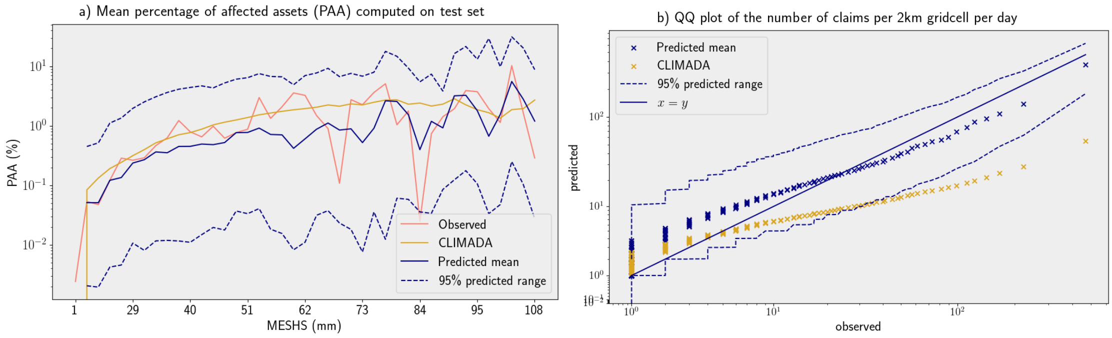


Figure 3.6: Comparison of CLIMADA and model: (a) predicted and observed percentages of affected assets; (b) QQ-plots of realized versus predicted quantiles for the number of claims per grid cell per day.

3.6.2 Combination of claim counts and values

One way to combine counts and values would be to compute a per-cell impact coefficient corresponding to the predicted number of buildings impacted by a hailstorm divided by the total number of buildings in the grid cell. This would then be multiplied by the total possible hail damage in the cell, i.e., the sum of predicted values of claims for all buildings, to obtain the per-cell effective damage. This possibility was considered but not pursued, since our aim is to predict claim values for single buildings and not the aggregated damage in a cell. Our combination of counts and values thus involves choosing which buildings are impacted by hail given the predicted count in a cell. In each cell, buildings are first sorted by their exposure (i.e., insured value), and the first N are selected to compute the total damage. To predict damage and its confidence range, we sample the counts for the cell n times and the claim value for every building m times, apply the procedure described in the previous sentence to the mn samples, and finally compute the average total damage and its 95% prediction range.

3.6.3 Claim values

Evaluation of the full hail damage prediction involves the combination of counts and claim values, as described in Section 3.6.2. We use the two metrics described in Section 3.5.2 to compare the spatial patterns of hail damage predicted with CLIMADA and our model. Figure 3.8(a) shows that both take higher values for CLIMADA than for our model, so the latter better captures the observed hail damage footprints. Figure 3.8 also shows that the distribution of hail damage predicted with our model is close to the observed distribution both on a single building scale (b) and on a 2km square grid (c), though non-extreme damage is slightly over-predicted, which suggests further research on the distribution of non-extreme claim values might be needed. CLIMADA systematically under-predicts the aggregated damage per grid cell, and there is a clear improvement using the random line process. Figure 3.8(b) compares the model's input, downscaled CLIMADA, to the prediction. As expected, downscaled CLIMADA under-predicts damage per building (see Section 3.2.2), while the random line model predicts realistic values, particularly so for claims above CHF 5000.

Figure 3.9 shows some daily hail damage maps for days on which there was over one million Swiss francs of realized hail impacts in the canton (some of these claim dates belong to the train or validation set). The predicted claim values appear to be locally large, matching the spatial pattern of realized damage, whereas CLIMADA damage is more dispersed. The average total predicted damage over the canton for days with extreme realized hail damage is close to the observed value, which the confidence interval usually captures. The line model is thus able to predict well-located extremes while providing reliable estimates of the total damage. The lowest panel in Figure 3.9 shows the most extreme hail event in the two decades of our data, on 28 June 2021, which involved roof-penetrating hail damage with 177 claims above CHF 100,000. Our model manages to capture extreme damage on this day, with average predicted values up to CHF 180,000. The spatial pattern using both CLIMADA and, to a lesser extent, our model, is wider than the observed data, which might be related to overestimation of the MESHS intensity that day in the northern half of the canton (Figure 3.1).

3.7 Conclusion

The model developed in this paper seems to be the first to combine a Gaussian random line process with extreme-value modeling in order to predict the spatial footprint of hail damage. It improves on the use of a benchmark deterministic hail damage function: in particular, it captures extreme damage values for individual buildings well, reproduces the spatial pattern of hail in the insurance data, and its stochasticity enables uncertainty quantification. With appropriate changes, such as for instance the possibility of modeling multiple random lines at the same time, our approach could be generalized to larger areas and would be useful in studying the insurance impacts of climate change. It would be interesting to use thunderstorm cell direction instead of large-scale wind direction as the covariate for the slope of the random line process.

3.8 Authorship contribution statement

Ophélia Miralles: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing (original draft, review, editing), Visualization.

Anthony Davison: Methodology, Supervision, Editing, Funding acquisition.

Timo Schmid: Data Curation, Editing.

3.9 Acknowledgments

We thank the Federal Office of Meteorology and Climatology (MeteoSwiss) for providing the POH and MESHS data, the insurance company GVZ for providing the building damage and exposure data, and Daniel Steinfeld of GVZ for valuable insights.

3.10 Funding

Ophélia Miralles' work was partially funded by the Swiss National Science Foundation (project 200021_178824). Timo Schmid acknowledges funding from the Swiss National Science Foundation (SNSF) Sinergia grant (CRSII5_201792).

3.11 Data statement

The exploratory analysis, model, and diagnostics are implemented in Python and the code is freely available on GitHub (github.com/OpheliaMiralles/hail-damage-modeling). MESHS and POH data are available from MeteoSwiss on demand. The GVZ insurance data are private, and as such are available for use only within the scClim project for research purposes. The CLIMADA impact function is open-source and can be run using the GitHub repository github.com/CLIMADA-project/climada_papers. ERA5 reanalysis data can be downloaded freely from the Copernicus Climate Data Store (climate.copernicus.eu/climate-reanalysis).

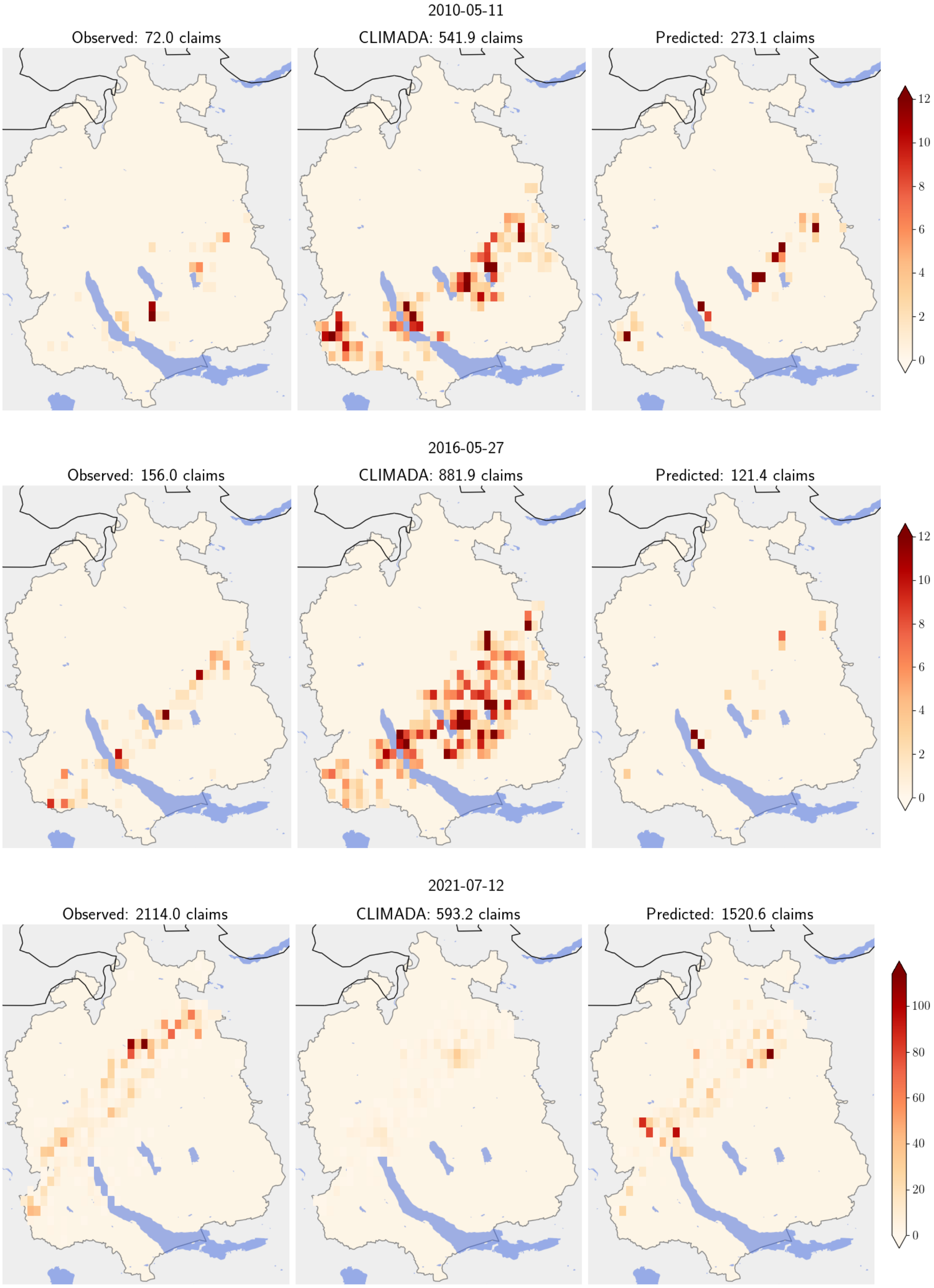


Figure 3.7: Comparison of locations of observed claim counts (left), CLIMADA predicted counts (center), and our predicted counts (right) for three dates selected over all dates with more than 10 observed claims over the canton on the 2000–2021 period. The titles give the observed number of claims, CLIMADA predicted count, and the average count and its 95% predicted range from the random line model (right).

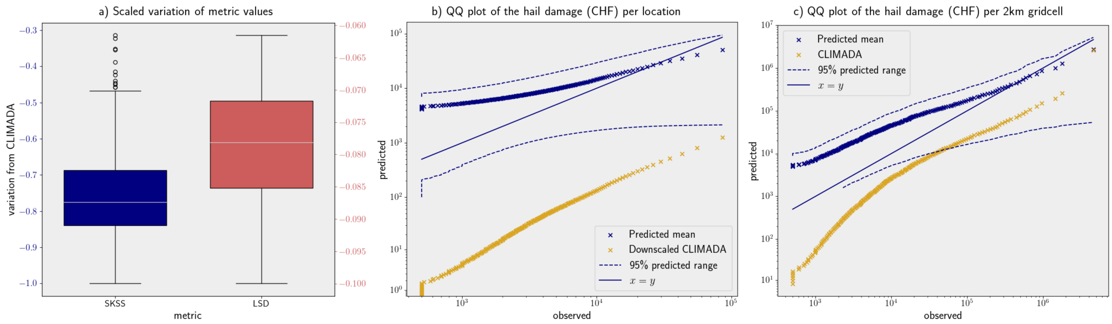


Figure 3.8: Comparison of metrics of predicted damage. (a) values of scaled LSD and SKSS for our model, with those from CLIMADA subtracted. (b) QQ-plots of realized versus predicted quantiles for the damage per building and (c) per 2km grid cell, with dashed blue lines showing the 95% prediction range.

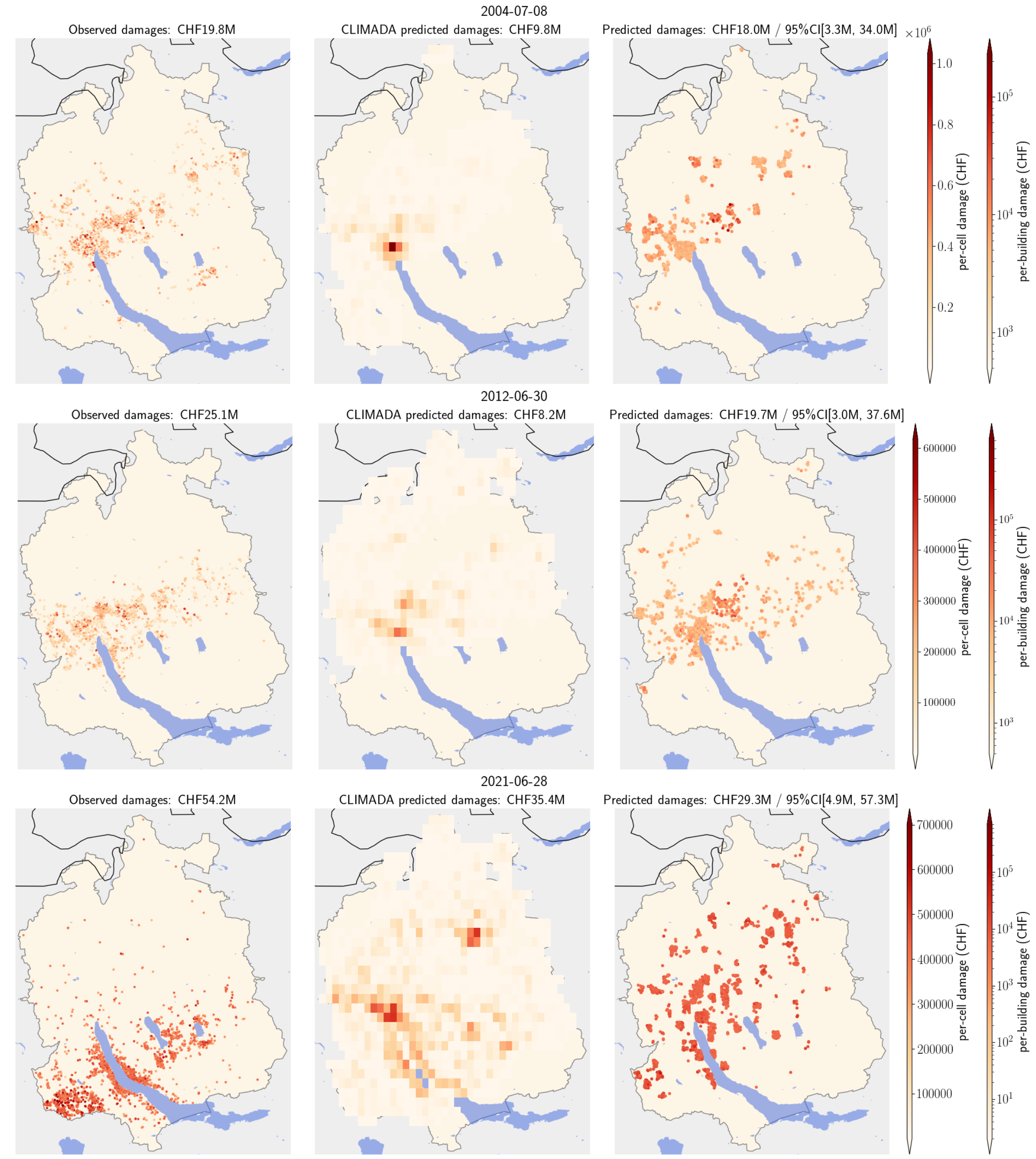

2004-07-08
Observed damages: CHF19.8M
CLIMADA predicted damages: CHF9.8M
Predicted damages: CHF18.0M / 95%CI[3.3M, 34.0M]
$\times10^6$
1.0
0.8
0.6
0.4
0.2
per-cell damage (CHF)
10^5
10^4
10^3
per-building damage (CHF)
2012-06-30
Observed damages: CHF25.1M
CLIMADA predicted damages: CHF8.2M
Predicted damages: CHF19.7M / 95%CI[3.0M, 37.6M]
600000
500000
400000
300000
200000
100000
per-cell damage (CHF)
10^5
10^4
10^3
per-building damage (CHF)
2021-06-28
Observed damages: CHF54.2M
CLIMADA predicted damages: CHF35.4M
Predicted damages: CHF29.3M / 95%CI[4.9M, 57.3M]
700000
600000
500000
400000
300000
200000
100000
per-cell damage (CHF)
10^5
10^4
10^3
10^2
10^1
per-building damage (CHF)

Figure 3.9: Example daily hail impact maps. The columns represent the observed claims (left), CLIMADA-predicted claims (center), and the hail damage prediction using the random line model (right). The left color bar relates to the prediction at the scale of the cell (i.e. relevant for CLIMADA predicted damage), while the right color bar displays a log scale for the per-building claim values (i.e. relevant for realized and predicted damage). The observed monetary cantonal damage (left), CLIMADA predicted total (middle), and average damage from the random line model with its 95% predicted range (right) are displayed in the titles for each selected date.

3.12 Supplementary material

Here we provide additional information and figures about model selection, including material about the choice of fixed hyperparameters such as the threshold for the GPD model, and also MC diagnostics related to the validation of the Bayesian model parameters.

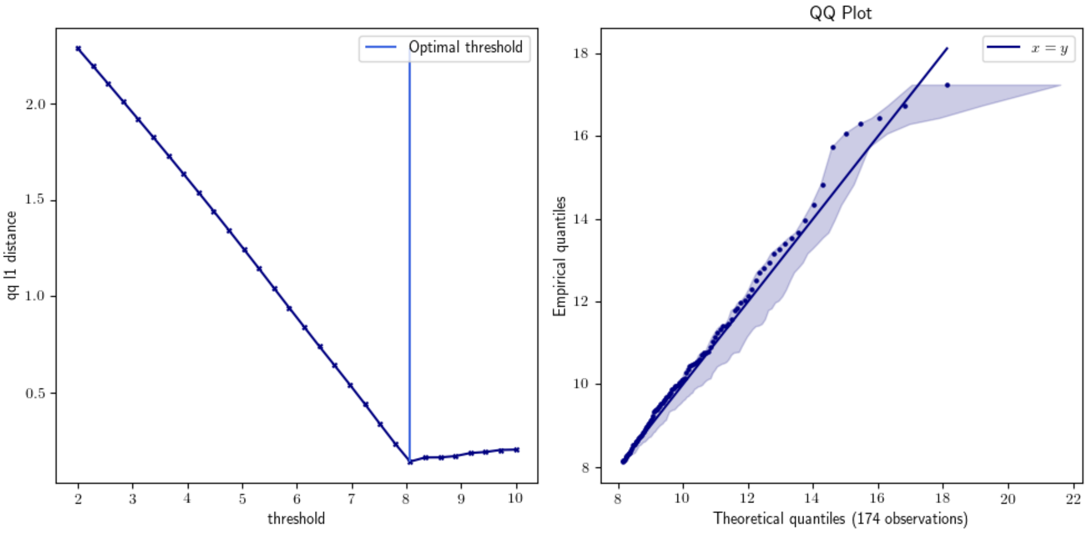


Figure 3.10: Threshold selection method described in Varty et al. (2021) applied to the log total damage per cell. The left panel shows a minimum qq-ℓ_1-distance for a threshold of 8.06. The QQ-plot for the GPD fit of the log total sum of damage above this threshold is displayed in the right panel, in which the profile likelihood-based 95% confidence interval is shown by the blue shaded area.

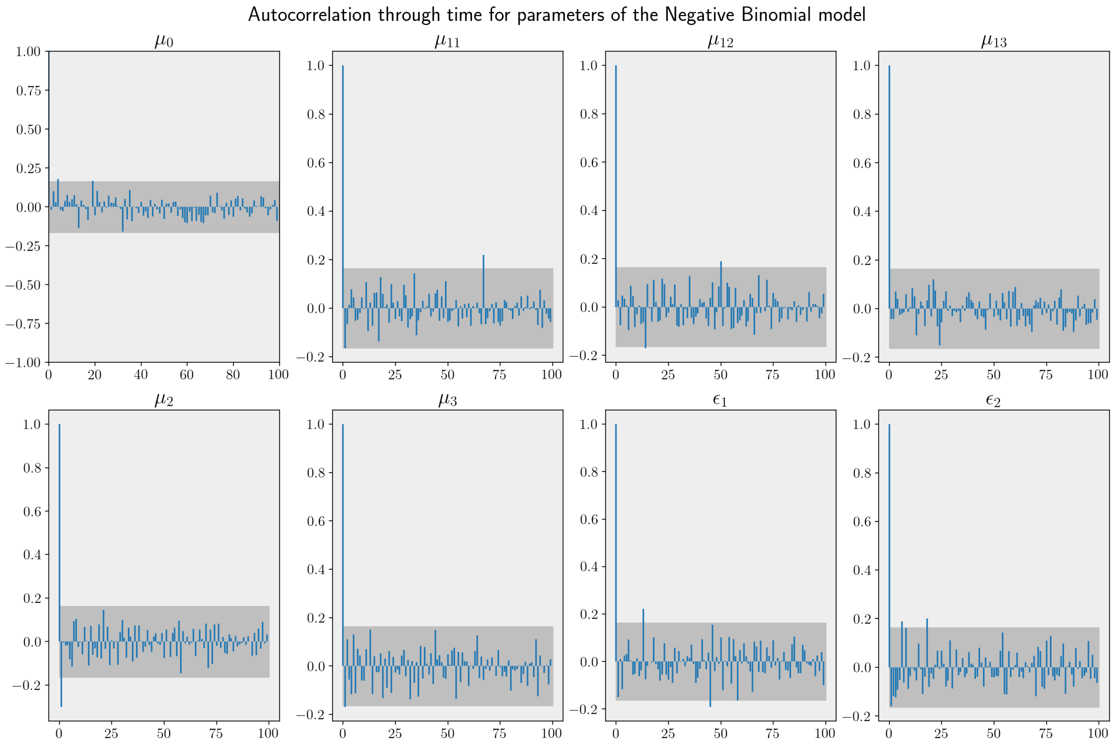


Figure 3.11: Evolution of the autocorrelation through sampling for parameters of the Negative Binomial model presented in Section 3.4.1. The grey area designates the acceptable range for the autocorrelation at the end of sampling to assume convergence of the model. The bounds of the confidence range are computed from the central limit theorem.

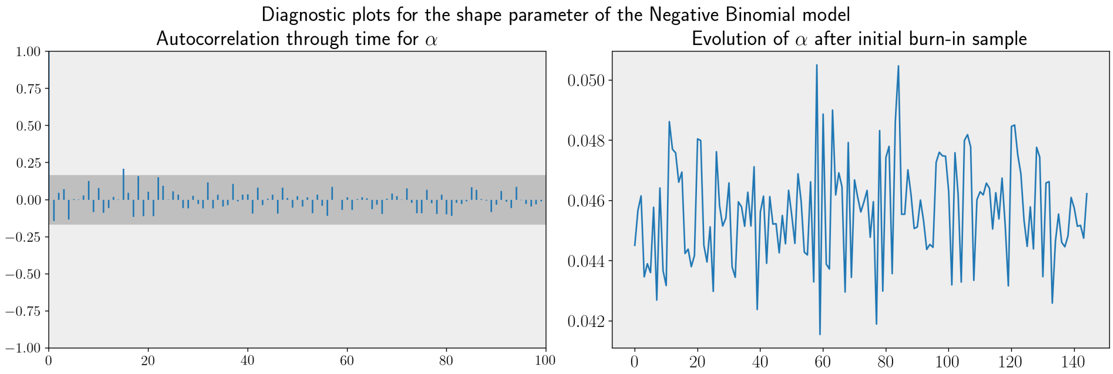


Figure 3.12: Autocorrelation plot (left) and trace plot (right) for the posterior distribution of the shape parameter α in the Negative Binomial model presented in Section 3.4.1. A close-to-zero autocorrelation through sampling and no specific trend or pattern in the trace plot is usually a good sign of model convergence.

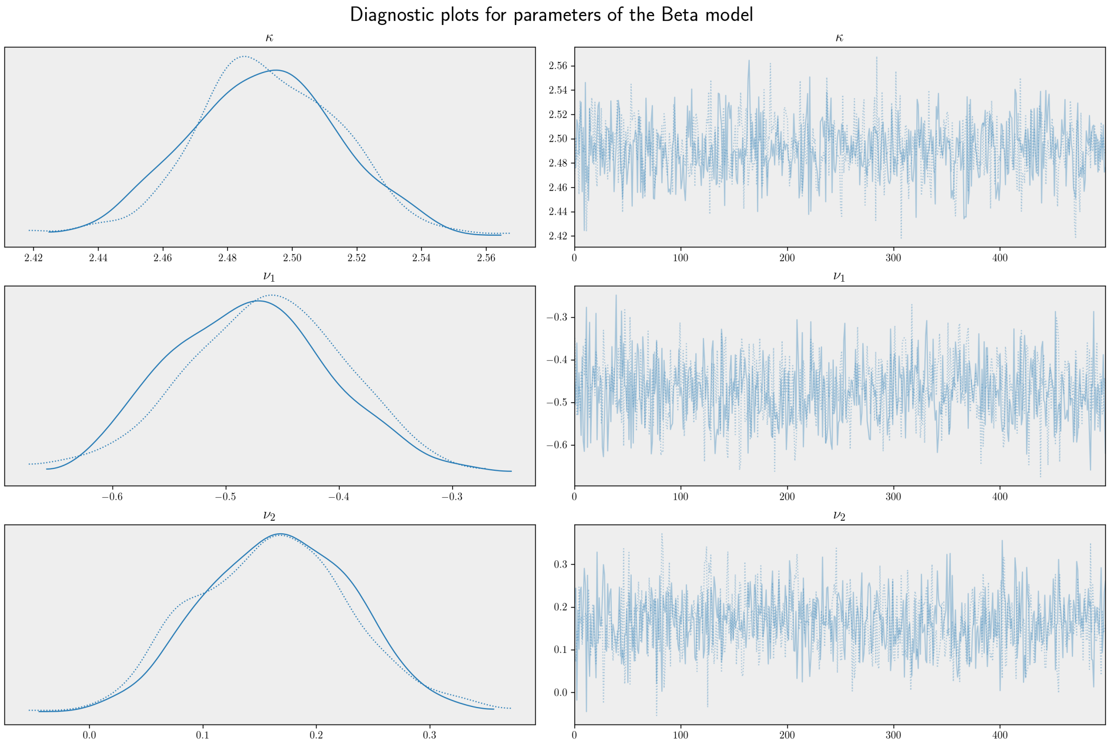


Figure 3.13: Kernel density estimate plot: (left) and trace plot: (right) for parameters of the Beta model presented in Section 3.4.2. No specific trend or pattern in the trace plot is usually a good sign of model convergence.

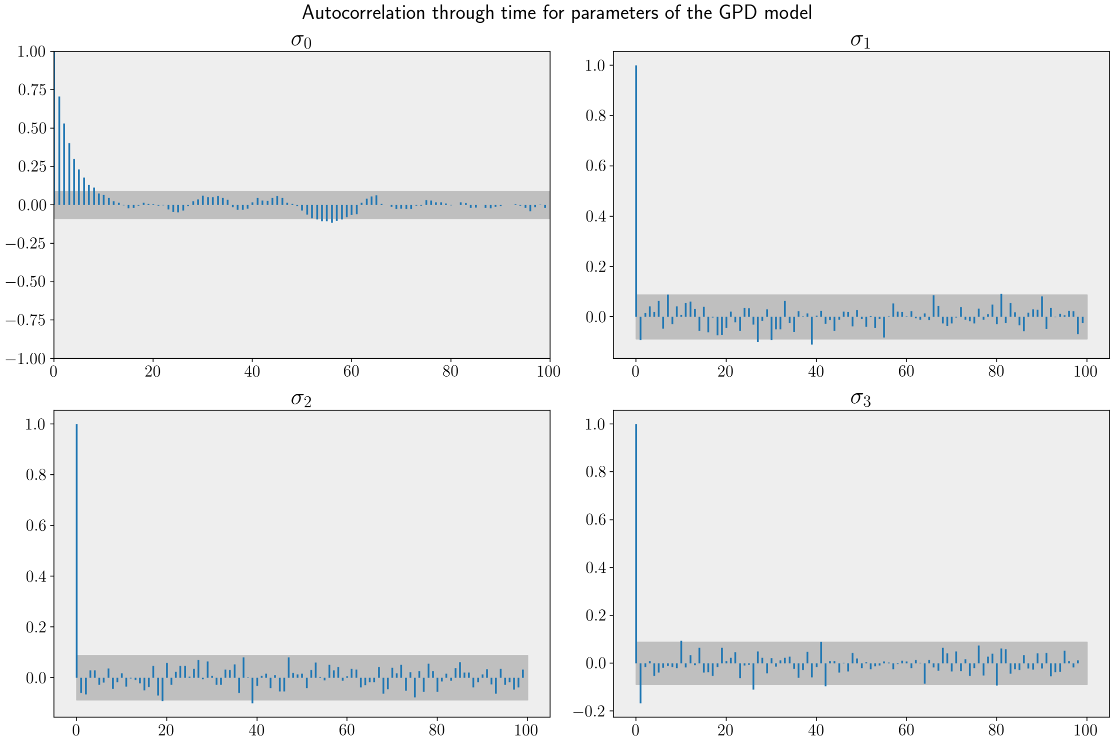


Figure 3.14: Evolution of the autocorrelation through sampling for parameters of the GPD model presented in Section 3.4.2. The grey area designates the acceptable range for the autocorrelation at the end of sampling to assume convergence of the model. The bounds of the confidence range are computed from the central limit theorem.

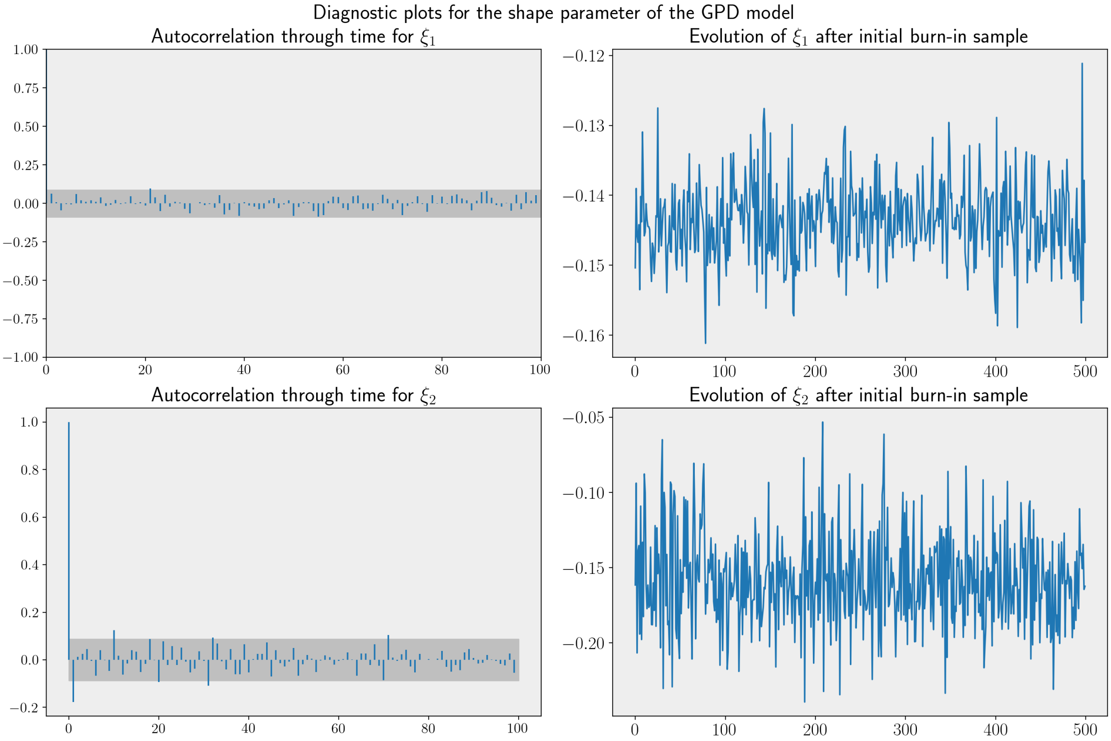


Figure 3.15: Autocorrelation plot (left) and trace plot (right) for the posterior distribution of the shape parameter ξ in the GPD model presented in Section 3.4.2. A close-to-zero autocorrelation through sampling and no specific trend or pattern in the trace plot is usually a good sign of model convergence.

4 Perspectives

This chapter closes the thesis with ideas about further improvements in each specific topic considered during these (almost) three years. Of course, a lot more needs to be done to favor the collaboration between statisticians and climate scientists.

The work on timing and spatial bias in extreme event attribution in Chapter 1 proposes a general framework to study selection bias of any kind. A logical extension to this work would be to consider other selection biases relevant to EEA (National Academies of Sciences, Engineering, and Medicine, 2016), and not only to the method used by the World Weather Attribution group. In the paper, we explain and try to account for spatial selection bias as defined in National Academies of Sciences, Engineering, and Medicine (2016) and Hammerling et al. (2019), i.e., the bias induced by choosing to study the specific location where the extreme happened by considering a spatial area embedding several locations $S = \{s_1, \ldots, s_n\}$ and condition the likelihood on the fact that the extreme happened in location s_m. The simulation framework and case studies only considered bivariate data, but further research could look at larger spatial areas around the observed event with more different locations and see how the results generalize. The bias of mainly performing EEA on events that happened and which we expect to be positively correlated with climate change (Hammerling et al., 2019; van Oldenborgh et al., 2021) might be trickier to study since we would need to generate a catalog of counterfactual scenarios. Another potential area of improvement on the topic is related to the fact that the choice of the stopping rule greatly influences how timing bias can impact the return level estimation. Further work on the topic could perform sensitivity analysis on a set of stopping rules or could study random stopping rules and their effect on the extremal distribution.

In the wind downscaling work in Chapter 2, most issues stem from differences in topography. The global architecture could thus be improved by building different models for predicting patches that are mainly over mountainous areas or over the Swiss Plateau, rather than using a single network for the entire country. Another deep structure trained for image classification based on topography could select which of those two networks should be applied. Different parameters would thus be used to predict winds in plains and in complex terrain, perhaps leading to less topographic variation in model performance. The training of such a structure would require more time and resources but might overcome most of the remaining issues with our model. An alternative is the use of spatial attention mechanisms, such as in transformer networks (Jaderberg et al., 2015), to better capture the topographical context of a specific patch during the training. Indeed, networks containing spatial transformers can select regions of an image that are most relevant (attention), and also encode those regions into a spatiotemporal vector invariant to the input image formatting (cropping, rotating, scaling, ...) to simplify inference in the subsequent layers. The use of transformer networks is widespread in natural language processing and computer vision, and it has started to generalize to the analysis of time series (Ahmed et al., 2023). In our case, the insertion of temporal attention layers instead of LSTM layers in the GAN could help focus on the parts of the day when the wind is strongest, and also reduce the computational cost.

The hail impact model proposed in Chapter 3 could in principle be generalized to all Swiss cantons, given access to cantonal insurance data and CLIMADA's hail impact model output. For larger areas, modelling multiple random lines might be necessary. Another use of the model would be to study how variation in the MESHS prediction (for instance, under the influence of climate change) would impact the total average economic loss. That information would not only be relevant to cantonal insurance companies but could also be used in building climate change mitigation strategies on the cantonal level, helping to highlight the economic and social impacts of climate change. The is a growing interest in statistical modelling of storm tracks in large areas using various geometrical shapes. Statistical models exploring the extremal dependence of several locations on multiple straight lines of pre-defined orientation have been proposed for severe ocean storms (Shooter et al., 2019, 2021). The modelling of a windstorm's spatial extent has been studied in Sharkey et al. (2019) using an ellipse to capture the area impacted by the storm in Europe. Further research could for instance develop a multi-line random process based on the work from Chapter 3 to explore the change in frequency and amplitude of tornadoes in the United States. Indeed, the frequency of tornado hazards in the U.S. has been rising over the last decades (Bryan J. Boruff et al., 2003) with average annual losses of almost $1 billion over the period 1949-2006 (Changnon Stanley A., 2009).

Papers aiming at connecting both fields like the first presented in this thesis are essential to make advances in statistics applicable to climate scientists. Interdisciplinary projects such as those in Chapters 2 and 3 are a direct way of inserting statistical knowledge into scientific projects. The Python package `pykelihood` developed during this Ph.D. (see Appendix) has also made accessible some statistical techniques only available in R before to the Python-coding community, which includes many climate scientists, physicists, and engineers. There are certainly many more ways of bridging the gap between statistics and climate science, and I hope that further ideas will be provided by future researchers.

A Appendix: `pykelihood`

An important outcome of this thesis is the development of the Python package `pykelihood`, which reproduces features from different R packages regarding likelihood-based inference. The development of `pykelihood` represents a non-negligible proportion of my Ph.D. years and is relevant for the overall purpose of my thesis as it facilitates the use of statistical models among the Python-coding community which includes many climate scientists and physicists. This Appendix thus describes the objective and functioning of the package `pykelihood`.

`pykelihood` is a Python package for statistical analysis designed to give more flexibility to likelihood-based inference than is currently possible in Python. Distributions are designed from an Object Oriented Programming (OOP) point of view. In particular, this package allows the fitting of complex distributions to a dataset, add trends of different forms in the parameters of the target distribution, condition the log-likelihood with any form of penalty, and profile parameters of the model based on the chosen likelihood's sensitivity. Installation of the package can be done by following the steps described on the GitHub README file for the package.

The distribution class

The most basic use of `pykelihood` is creating and manipulating distributions as objects. The distribution parameters can be accessed like standard Python attributes. Sampling from the distribution or computing the quantiles can be done using the same semantics as with the Python package `scipy.stats`. To fit the distribution to data, the syntax is simply `distribution.fit(data)`, where fixed parameters can be added as additional arguments to the fit, as in the package `scipy.stats` (for instance, `distribution.fit(data, loc=0)` will fit the distribution to the data while keeping the `loc` parameter null).

Trend fitting

One of the most powerful features of `pykelihood` is the ability to fit arbitrary distributions. For instance, suppose our data has a linear trend in time with very little Gaussian noise we would like to capture.

```python
import numpy as np
data = np.linspace(-1, 1, 365) + np.random.normal(0, 0.001, 365)
```

Fitting a `Normal` distribution with a trend in the `loc` parameter can be done using the following piece of code:

```python
from pykelihood import kernels
Normal.fit(data,loc=kernels.linear(np.arange(365)))
```

and would output the following distribution object.

```
Normal(loc=linear(a=-1.0000458359290572, b=0.005494714384381866),
                                  scale=0.0010055323717468906)
```

The `kernels` module is flexible and can be adapted by users to support any kind of trend. For instance, `kernels.linear(X)` builds a linear model in the form $a + bX$ where a and b are parameters to be optimized for, and X is a covariate used to fit the data. If we assume the data were daily observations, then we find all the values we expected: -1 was the value on the first day, 0.05 was the daily increment ($2/365 = 0.05$), and there was a noise with a standard deviation 0.001.

Fitting with penalties

Another useful feature of `pykelihood` is the ability to customize the log-likelihood function with penalties, conditioning methods, stability conditions, etc. Most statistics-related packages offer to fit data using the standard negative log-likelihood function, or in the best case, preselected models. To our knowledge, `pykelihood` is the only Python package allowing easy customization of the log-likelihood function.

For example, say we want to penalize the target distribution parameters which ℓ_1-norm is too large: we would then apply a Lasso penalty.

```python
def lassolike_score(distribution, data):
    return -np.sum(distribution.logpdf(data)) + np.abs(
                                            distribution.loc())
```

We then compare a fit using the standard negative log-likelihood function to the use of the Lasso-penalized likelihood.

```python
data = np.random.normal(0, 1, 1000)
std_fit = Normal.fit(data)
cond_fit = Normal.fit(data, score=lassolike_score)
```

The outcomes show that the penalty has been taken into account; the `loc` parameter of the distribution applying the penalty is smaller than with the standard opposite log-likelihood function.

```python
>>> std_fit.loc.value
-0.010891307380632494
>>> cond_fit.loc.value
-0.006210406541824357
```

Parameter profiling

Likelihood-based inference requires parameter estimation, so it is important to quantify the sensitivity of a chosen model to each of those parameters. The `profiler` module in `pykelihood` includes the `Profiler` class that allows the linking of a model to a set of observations by providing some goodness of fit metrics and "profiles" for all parameters. Profiles are provided under the form of a dictionary of pandas DataFrame objects. Each key is a parameter to profile, i.e. to fix and vary while the other distribution parameters are optimized, and each associated data frame contains the values of all of the distribution parameters as well as this of the "score" function (usually the opposite log-likelihood) throughout the partial optimization. If the distribution includes a trend in one of the parameters, the parameters of the trend will be profiled. If some parameters were fixed in the distribution provided to the `Profiler`, the associated profiles are not computed. Computing the profile likelihood can be done as follows.

```python
from pykelihood.profiler import Profiler
from pykelihood.distributions import GEV
```

```
fitted_gev = GEV.fit(data, loc=kernels.linear(np.linspace(-1, 1, len(data))))
profiler = Profiler(fitted_gev, data, inference_confidence=0.95) # level of confidence for the likelihood-ratio test
```

```
>>> profiler.AIC
AIC MLE
-359.73533182968777
AIC Standard MLE # a comparison with the standard fit without trend is provided
623.9896838880583
>>> profiler.profiles.keys()
[loc_a, loc_b, scale, shape]
>>> profiler.profiles["shape"].head(5)
    loc_a      loc_b     scale     shape      score
0 -0.000122 1.000812 0.002495 -0.866884 1815.022132
1 -0.000196 1.000795 0.001964 -0.662803 1882.043541
2 -0.000283 1.000477 0.001469 -0.458721 1954.283256
3 -0.000439 1.000012 0.000987 -0.254640 2009.740282
4 -0.000555 1.000016 0.000948 -0.050558 1992.812843
```

A binary search algorithm implemented to compute the parameter confidence intervals allows for very efficient exploration of the parameter space. It can be provided with a “precision” argument, defaulted to 10^{-5}. For example, if the parameter of interest is the location of the GEV distribution, the profile likelihood-based associated confidence interval is computed using the following syntax

```
profiler.confidence_interval_bs("loc", precision=1e-3)
```

from which the output would be an array containing the lower and upper bound for the corresponding confidence interval (using the level defined as a parameter of the `Profiler` object).

```
[-4.160287666875364, 4.7039931595123825]
```

Statement of interest

The toy examples presented above might not seem very useful, but both the trend fitting and the ability to condition a log-likelihood using any type of penalty function have been very useful during this Ph.D. Indeed, the work on the selection bias in extreme event attribution studies involved both the need for a distribution able to accommodate for a trend in one of the covariates in the Phalodi example, and for all of the distributions and datasets considered to be fitted and profiled using conditioned log-likelihoods. A reparametrization of the GEV distribution in terms of the return level also implied the need for flexibility in the parameter definition, which is fortunately provided by the package `pykelihood`. It was also fairly easy to build profile likelihood-based confidence intervals for this study as well with the module `profiling`: this feature was essential for the paper on selection bias in extreme event attribution studies presented in Chapter 1.

Using functionalities already present in some R-packages would have been far easier than writing a Python package to replicate them. Unfortunately, R is a statistician's programming language and as such, it is not widely used in the climate science community. Writing `pykelihood` seemed crucial because it enables researchers from diverse fields outside of statistics to access advanced techniques in extreme value theory, thus facilitating their understanding and application.

Bibliography

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Rasool, G., and Ramachandran, R. P. (2023). Transformers in time-series analysis: A tutorial. *arXiv preprint arXiv:2205.01138*.

Allen, M. (2003). Liability for climate change. *Nature*, 421(6926):891–892.

Aznar-Siguan, G. and Bresch, D. N. (2019). CLIMADA v1: a global weather and climate risk assessment platform. *Geoscientific Model Development*, 12(7):3085–3097.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baño Medina, J., Manzanas, R., and Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124.

Barlow, A. M., Sherlock, C., and Tawn, J. (2020). Inference for extreme values under threshold-based stopping rules. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):765–789.

Barry, R. G. (2008). *Mountain Weather and Climate*. Cambridge University Press, Cambridge, United Kingdom, third edition.

Belzile, L. R. and Davison, A. C. (2022). Improved inference on risk measures for univariate extremes. *Annals of Applied Statistics*, 16:1524–1549.

Berrocal, V. J. (2017). Data assimilation. In Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L., editors, *Handbook of Environmental and Ecological Statistics*, chapter 7, pages 133–146. Chapman & Hall/CRC, Boca Raton, FL, United States.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv e-prints*, page arXiv:1701.02434.

Botzen, W., Bouwer, L., and van den Bergh, J. (2010). Climate change and hailstorm damage: Empirical evidence and implications for agriculture and insurance. *Resource and Energy Economics*, 32(3):341–362.

Bozinovski, S. and Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron B2. In *Proceedings of Symposium Informatica*, pages 3–121–5, Ljubljana, Slovenia. The Slovene Society Informatika.

Brown, T. M., Pogorzelski, W. H., and Giammanco, I. M. (2015). Evaluating hail damage using property insurance claims data. *Weather, Climate, and Society*, 7(3):197–210.

Bryan J. Boruff, Jaime A. Easoz, Steve D. Jones, Heather R. Landry, Jamie D. Mitchem, and Susan L. Cutter (2003). Tornado hazards in the United States. *Climate Research*, 24(2):103–117.

Castro-Camilo, D., Huser, R., and Rue, H. (2019). A spliced gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):517–534.

Changnon, S. A. (2008). Temporal and spatial distributions of damaging hail in the continental United States. *Physical Geography*, 29(4):341–350.

Changnon Stanley A. (2009). Tornado losses in the United States. *Natural Hazards Review*, 10(4):145–150.

Coles, S. and Pericchi, L. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):405–416.

Coles, S. G., Pericchi, L. R., and Sisson, S. A. (2003). A fully probabilistic approach to extreme rainfall modelling. *Journal of Hydrology*, 273:35–50.

COnsortium for Small-scale MOdeling (2017). COSMO model. http://www.cosmo-model.org/. Accessed: 2021–11–25.

Cressie, N. A. C. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley, New York, NY, United States.

Cruz, M. G., Alexander, M. E., Fernandes, P. M., Kilinc, M., and Ângelo Sil (2020). Evaluating the 10% wind speed rule of thumb for estimating a wildfire's forward rate of spread against an extensive independent set of observations. *Environmental Modelling & Software*, 133:104818.

Davison, A. C., Huser, R., and Thibaud, E. (2019). Spatial extremes. In Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L., editors, *Handbook of Environmental and Ecological Statistics*, chapter 31, pages 711–744. Chapman & Hall/CRC, Boca Raton, FL, United States.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52:393–442.

Deepen, J. (2006). Schadenmodellierung extremer Hagelereignisse in Deutschland. Master's thesis, Westfälischen Wilhelms-Universität Münster.

Dörenkämper, M., Olsen, B. T., Witha, B., Hahmann, A. N., Davis, N. N., Barcons, J., Ezber, Y., García-Bustamante, E., González-Rouco, J. F., Navarro, J., Sastre-Marugán, M., Sle, T., Trei, W., Žagar, M., Badger, J., Gottschall, J., Sanz Rodrigo, J., and Mann, J. (2020). The making of the new European wind atlas – part 2: Production and evaluation. *Geoscientific Model Development*, 13(10):5079–5102.

Dujardin, J., Kahl, A., and Lehning, M. (2021). Synergistic optimization of renewable energy installations through evolution strategy. *Environmental Research Letters*, 16(6):064016.

Dujardin, J. F. S. (2021). *The complex winds of the Alps: an unseen asset for the energy transition*. PhD thesis, EPFL, Lausanne, Switzerland.

Efron, B. and Hastie, T. J. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, Cambridge, United Kingdom.

Emeis, S. (2014). Current issues in wind energy meteorology. *Meteorological Applications*, 21(4):803–819.

Fischer, E. M. and Knutti, R. (2015). Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nature Climate Change*, 5(6):560–564.

Fischer, E. M. and Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, 6(11):986–991.

Foreman, J. W. (2013). *Data Smart: Using Data Science to Transform Information into Insight*. John Wiley, Indianapolis, IN, United States.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association*, 95(452):1300–1304.

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

Graf, M., Scherrer, S. C., Schwierz, C., Begert, M., Martius, O., Raible, C. C., and Brönnimann, S. (2019). Near-surface mean wind in Switzerland: Climatology, climate model evaluation and future scenarios. *International Journal of Climatology*, 39(12):4798–4810.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*.

Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F. (2016). WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6. *Geoscientific Model Development*, 9(11):4087–4095.

Hammerling, D., Katzfuss, M., and Smith, R. L. (2019). Climate change detection and attribution. In Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L., editors, *Handbook of Environmental and Ecological Statistics*, chapter 34, pages 789–817. Chapman & Hall/CRC, Boca Raton, FL, United States.

Harris, I., Osborn, T. J., Jones, P., and Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(109).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hernan, M. and Robins, J. (2023). *Causal Inference: What If?* Chapman & Hall/CRC, Boca Raton, FL, United States.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and

Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS), 4–9 December 2017, Long Beach, CA, United States*, volume 30, pages 6626–6637, Red Hook, NY, United States. Curran Associates, Inc.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hohl, R., Schiesser, H.-H., and Aller, D. (2002). Hailfall: the relationship between radar-derived hail kinetic energy and hail damage to buildings. *Atmospheric Research*, 63(3):177–207.

Höhlein, K., Kern, M., Hewson, T., and Westermann, R. (2020). A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, 27(6):e1961.

Homan, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. (2016). Deep networks with stochastic depth. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision (ECCV), 11–14 October 2016, Amsterdam, Netherlands*, volume 14, pages 646–661, Cham, Switzerland. Springer.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Jackson, P. L., Mayr, G., and Vosper, S. (2013). Dynamically-driven winds. In Chow, F. K., De Wekker, S. F., and Snyder, B. J., editors, *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, pages 121–218. Springer, Dordrecht, Netherlands.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, k. (2015). Spatial transformer networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS), 7–12 December 2015, Montréal, Canada*, volume 28, pages 2017–2025, Cambridge, MA, United States. MIT Press.

Jarvis, A., Guevara, E., Reuter, H., and Nelson, A. (2008). Hole-filled SRTM for the globe: version 4: data grid. http://srtm.csi.cgiar.org. Published by CGIAR-CSI on 19 August 2008.

Jeong, J., Jun, M., and Genton, M. G. (2017). Spherical process models for global spatial statistics. *Statistical Science*, 32(4):501–513.

Jones, M. W., Smith, A., Betts, R., Canadell, J. G., Prentice, I. C., and Le Quéré, C. (2020). Climate change increases risk of wildfires. https://sciencebrief.org/briefs/wildfires.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koh, J., Pimont, F., Dupuy, J.-L., and Opitz, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *Annals of Applied Statistics*, 17(1):560–582.

Koller, S. and Humar, T. (2016). Windpotentialanalyse für Windatlas.ch: Jahresmittelwerte der modellierten windgeschwindigkeit und windrichtung (wind potential analysis for windatlas.ch: Annual mean values of the modeled wind speed and wind direction). Technical report, MeteoTest Bundesamt für Energie.

Kopp, J., Schröer, K., Schwierz, C., Hering, A., Germann, U., and Martius, O. (2023). The summer 2021 Switzerland hailstorms: weather situation, major impacts and unique observational data. *Weather*, 78(7):184–191.

Kropp, J. and Schellnhuber, H.-J. (2011). *In Extremis: Disruptive Events and Trends in Climate and Hydrology*. Springer, Heidelberg, Germany.

Kruyt, B., Dujardin, J., and Lehning, M. (2018). Improvement of wind power assessment in complex terrain: The case of COSMO-1 in the Swiss Alps. *Frontiers in Energy Research*, 6:102.

Kruyt, B., Lehning, M., and Kahl, A. (2017). Potential contributions of wind power to a stable and highly renewable Swiss power supply. *Applied Energy*, 192:1–11.

Larsen, A., Yang, S., Reich, B. J., and Rappold, A. G. (2020). A spatial causal analysis of wildland fire-contributed PM2.5 using numerical model output. *arXiv preprint arXiv:2003.06037*.

Lawrence, N. (2003). Gaussian process latent variable models for visualisation of high-dimensional data. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems (NIPS), 9 December 2003, Whistler, Canada*, volume 16, pages 329–336, Cambridge, MA, United States. MIT Press.

Lehning, M., Löwe, H., Ryser, M., and Raderschall, N. (2008). Inhomogeneous precipitation distribution and snow transport in steep terrain. *Water Resources Research*, 44(7):1–19.

Leinonen, J., Nerini, D., and Berne, A. (2021). Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223.

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.

Liew, S. S., Khalil-Hani, M., and Bakhteri, R. (2016). Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. *Neurocomputing*, 216:718–734.

Liu, K., Li, P., and Wang, Z. (2021). Statistical modeling of random hail impact. *Extreme Mechanics Letters*, 48:101374.

Méndez, W., Gil, H. A. P., Ríos, S. C., Montilla, A. M., and León, C. (2015). Caracterización hidroclimatológica y morfométrica de la cuenca del río San Julián (estado Vargas, Venezuela): aportes para la evaluación de la amenaza hidrogeomorfológica. *Cuadernos de Geografía: Revista Colombiana de Geografía*, 24(2):133–156.

MeteoSwiss (2015). Typische Wetterlagen im Alpenraum. https://www.meteoschweiz.admin.ch/home/service-und-publikationen/publikationen.subpage.html/de/data/publications/2015/8/typische-wetterlagen-im-alpenraum.html. Accessed: 2021–11–25.

MeteoSwiss (2016). The new weather orecasting model for the Alpine region. https://www.meteoswiss.admin.ch/home/latest-news/news.subpage.html/en/data/news/2016/3/the-new-weather-forecasting-model-for-the-alpine-region.html. Accessed: 2021–11–25.

MeteoSwiss (2018). Documentation of MeteoSwiss grid-data products. hourly precipitation estimation through rain-gauge and radar: CombiPrecip. https://www.meteoschweiz.admin.ch/content/dam/meteoswiss/en/climate/swiss-climate-in-detail/doc/ProdDoc_CPC.pdf. Accessed: 2021–11–25.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (2004). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miralles, O. (2021). Comptabilité carbone: la clé pour une évaluation adéquate des risques climatiques. https://greentervention.org/2021/06/12/la-parole-aux-jeunes-comptabilite-carbone-la-cle-pour-une-evaluation-adequate-des-risques-climatiques/. Published by Greentervention, a Belgian NGO providing guidance about environmental policies in the European Union. Accessed: 2023–07–19.

Miralles, O. and Davison, A. C. (2023). Timing and spatial selection bias in rapid extreme event attribution. *Weather and Climate Extremes*, 41:100584.

Miralles, O., Davison, A. C., and Schmid, T. (2023+). Bayesian modeling of insurance claims for hail damage. *Annals of Applied Statistics*. Submitted.

Miralles, O., Steinfeld, D., Martius, O., and Davison, A. C. (2022). Downscaling of historical wind fields over Switzerland using generative adversarial networks. *Artificial Intelligence for the Earth Systems*, 1(4):e220018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Molina, M. O., Gutiérrez, C., and Sánchez, E. (2021). Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset. *International Journal of Climatology*, 41(10):4864–4878.

Müller, A. (2021). 2 Milliarden Franken wegen Sturm und Hagel: 2021 wird für die Schweiz eines der teuersten Schadenjahre aller Zeiten. https://www.nzz.ch/wirtschaft/unwetter-schweiz-2021-hagel-und-sturm-kosteten-2-mrd-franken-ld.1652483?reduced=true. Accessed: 2023–06–25.

National Academies of Sciences, Engineering, and Medicine (2016). *Attribution of Extreme Weather Events in the Context of Climate Change.* The National Academies Press, Washington, DC, United States.

Naveau, P., Hannart, A., and Ribes, A. (2020). Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, 7:89–110.

Nerini, D. (2020). Probabilistic deep learning for postprocessing wind forecasts in complex terrain. https://vimeo.com/465719202. Accessed: 2022–01–01.

Nerini, D. and Zanetta, F. (2021). Topo-descriptors. MeteoSwiss. https://github.com/MeteoSwiss/topo-descriptors.

Nisi, L., Hering, A., Germann, U., and Martius, O. (2018). A 15-year hail streak climatology for the Alpine region. *Quarterly Journal of the Royal Meteorological Society*, 144(714):1429–1449.

Nisi, L., Martius, O., Hering, A., Kunz, M., and Germann, U. (2016). Spatial and temporal distribution of hailstorms in the Alpine region: a long-term, high resolution, radar-based analysis. *Quarterly Journal of the Royal Meteorological Society*, 142(697):1590–1604.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

Otto, F. E., Philip, S., Kew, S., Li, S., King, A., and Cullen, H. (2018). Attributing high-impact extreme events across timescalesa case study of four different types of events. *Climatic Change*, 149(3):399–412.

Otto, M. (2009). Modellierung von Hagelschäden in der Pkw-Kaskoversicherung in Deutschland. Master's thesis, Technische Universität Dresden.

Perera, S., Lam, N., Pathirana, M., Zhang, L., Ruan, D., and Gad, E. (2018). Probabilistic modelling of forces of hail. *Natural Hazards*, 91(1):133–153.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. Technical report, Stanford University.

Philip, S., Kew, S., van Oldenborgh, G. J., Otto, F., Vautard, R., van der Wiel, K., King, A., Lott, F., Arrighi, J., Singh, R., and van Aalst, M. (2020). A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2):177–203.

Philip, S., Kew, S. F., van Oldenborgh, G. J., Otto, F., OKeefe, S., Haustein, K., King, A., Zegeye, A., Eshetu, Z., Hailemariam, K., et al. (2018). Attribution analysis of the Ethiopian drought of 2015. *Journal of Climate*, 31(6):2465–2486.

Philip, S. Y., Kew, S. F., 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.

Púčik, T., Groenemeijer, P., Rädler, A. T., Tijssen, L., Nikulin, G., Prein, A. F., van Meijgaard, E., Fealy, R., Jacob, D., and Teichmann, C. (2017). Future changes in European severe convection environments in a regional climate model ensemble. *Journal of Climate*, 30(17):6771–6794.

Punge, H., Bedka, K., Kunz, M., and Werner, A. (2014). A new physically based stochastic event catalog for hail in Europe. *Natural Hazards*, 73:1625–1645.

Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, New Jersey.

Rädler, A. T., Groenemeijer, P. H., Faust, E., Sausen, R., and Púčik, T. (2019). Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Climate and Atmospheric Science*, 2(1):30.

Ramon, J., Lledó, L., Bretonnière, P.-A., Samsó, M., and Doblas-Reyes, F. J. (2021). A perfect prognosis downscaling methodology for seasonal prediction of local-scale wind speeds. *Environmental Research Letters*, 16(5):054010.

Ramon, J., Lledó, L., Torralba, V., Soret, A., and Doblas-Reyes, F. J. (2019). What global reanalysis best represents near-surface winds? *Quarterly Journal of the Royal Meteorological Society*, 145(724):3236–3251.

Reich, B. J., Yang, S., Guan, Y., Giffin, A. B., Miller, M. J., and Rappold, A. (2021). A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3):605–634.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204.

Richner, H. and Hächler, P. (2013). Understanding and forecasting Alpine Foehn. In *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, pages 219–260. Springer, Dordrecht, Netherlands.

Risser, M. D. and Wehner, M. F. (2017). Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Geophysical Research Letters*, 44(24):12–457.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and

Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI), 5–9 October 2015, Munich, Germany*, pages 234–241, Cham, Switzerland. Springer.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). How does batch normalization help optimization? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS), 3–8 December 2018, Montréal, Canada*, volume 32, pages 2483–2493, Red Hook, NY, United States. Curran Associates, Inc.

Schmid, T., Villiger, L., Portmann, R., and Bresch, D. N. (2023+). Open-source hail damage model for buildings and cars. *Natural Hazards and Earth System Sciences*. Submitted.

Schmidberger, M. (2018). *Hagelgefährdung und Hagelrisiko in Deutschland basierend auf einer Kombination von Radardaten und Versicherungsdaten*. PhD Thesis, Karlsruher Institut für Technologie (KIT).

Schuster, S. S., Blong, R. J., and McAneney, K. J. (2006). Relationship between radar-derived hail kinetic energy and damage to insured buildings for severe hailstorms in eastern Australia. *Atmospheric Research*, 81(3):215–235.

Schwierz, C., Köllner-Heck, P., Zenklusen Mutter, E., Bresch, D. N., Vidale, P.-L., Wild, M., and Schär, C. (2010). Modelling European winter wind storm losses in current and future climate. *Climatic Change*, 101(3-4):485–514.

Sha, Y., Ii, D. J. G., West, G., and Stull, R. (2020). Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology*, 59(12):2057–2073.

Sharkey, P., Tawn, J. A., and Brown, S. J. (2019). Modelling the spatial extent and severity of extreme European windstorms. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(2):223–250.

Sharples, J., McRae, R., and Wilkes, S. (2012). Wind-terrain effects on the propagation of wildfires in rugged terrain: Fire channelling. *International Journal of Wildland Fire*, 21:282–296.

Shooter, R., Ross, E., Tawn, J., and Jonathan, P. (2019). On spatial conditional extremes for ocean storm severity. *Environmetrics*, 30(6):e2562.

Shooter, R., Tawn, J., Ross, E., and Jonathan, P. (2021). Basin-wide spatial conditional extremes for severe ocean storms. *Extremes*, 24(2):241–265.

Sprenger, M., Dürr, B., and Richner, H. (2016). Foehn studies in Switzerland. In Willemse Saskia, F. M., editor, *From Weather Observations to Atmospheric and Climate Sciences in Switzerland - Celebrating 100 years of the Swiss Society for Meteorology*, pages 215–247. vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.

Staffell, I. and Pfenninger, S. (2016). Using bias-corrected reanalysis to simulate current and future wind power output. *Energy*, 114:1224–1239.

Stott, P. A., Christidis, N., Otto, F. E. L., Sun, Y., Vanderlinden, J.-P., van Oldenborgh, G. J., Vautard, R., von Storch, H., Walton, P., Yiou, P., and Zwiers, F. W. (2016). Attribution of extreme weather and climate-related events. *WIREs Climate Change*, 7(1):23–41.

Stott, P. A., Stone, D. A., and Allen, M. R. (2004). Human contribution to the European heatwave of 2003. *Nature*, 432(7017):610–614.

Stucki, P., Brönnimann, S., Martius, O., Welker, C., Imhof, M., von Wattenwyl, N., and Philipp, N. (2014). A catalog of high-impact windstorms in Switzerland since 1859. *Natural Hazards and Earth System Sciences*, 14(11):2867–2882.

Stucki, P., Dierer, S., Welker, C., Gómez-Navarro, J. J., Raible, C. C., Martius, O., and Brönnimann, S. (2016). Evaluation of downscaled wind speeds and parameterised gusts for recent and historical windstorms in Switzerland. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):31820.

Sun, R., Krueger, S. K., Jenkins, M. A., Zulauf, M. A., and Charney, J. J. (2009). The importance of fire–atmosphere coupling and boundary-layer turbulence to wildfire spread. *International Journal of Wildland Fire*, 18(1):50–60.

Ter Braak, C. J. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249.

Ter Braak, C. J. and Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446.

van der Wiel, K., Kapnick, S. B., van Oldenborgh, G. J., Whan, K., Philip, S., Vecchi, G. A., Singh, R. K., Arrighi, J., and Cullen, H. (2017). Rapid attribution of the August 2016 flood-inducing extreme precipitation in South Louisiana to climate change. *Hydrology and Earth System Sciences*, 21(2):897–921.

van Oldenborgh, G. J., Otto, F. E., Haustein, K., and Cullen, H. (2015). Climate change increases the probability of heavy rains like those of Storm Desmond in the UK—an event attribution study in near-real time. *Hydrology and Earth System Sciences Discussions*, 12(12):13197–13216.

van Oldenborgh, G. J., Philip, S., Kew, S., van Weele, M., Uhe, P., Otto, F., Singh, R., Pai, I., Cullen, H., and AchutaRao, K. (2018). Extreme heat in India and anthropogenic climate change. *Natural Hazards and Earth System Sciences*, 18(1):365–381.

van Oldenborgh, G. J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R., King, A., Lott, F., Arrighi, J., Singh, R., and van Aalst, M. (2021). Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166(13):1–27.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R. (2018). Generating high resolution climate change projections through single image super-resolution: An abridged version. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI), 13–19 July 2018, Stockholm, Sweden*, volume 37, pages 5389–5393, San Francisco, CA, United States. International Joint Conferences on Artificial Intelligence Organization.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv preprint arXiv:2102.00884*.

Warren, R. A., Ramsay, H. A., Siems, S. T., Manton, M. J., Peter, J. R., Protat, A., and Pillalamarri, A. (2020). Radar-based climatology of damaging hailstorms in Brisbane and Sydney, Australia. *Quarterly Journal of the Royal Meteorological Society*, 146(726):505–530.

Weissmann, M., Braun, F. J., Gantner, L., Mayr, G. J., Rahm, S., and Reitebuch, O. (2005). The Alpine mountainplain circulation: airborne Doppler lidar measurements and numerical simulations. *Monthly Weather Review*, 133(11):3095–3109.

Welker, C., Martius, O., Stucki, P., Bresch, D., Dierer, S., and Brönnimann, S. (2016). Modelling economic losses of historic and present-day high-impact winter windstorms in Switzerland. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):29546.

Winstral, A., Jonas, T., and Helbig, N. (2017). Statistical downscaling of gridded wind speed data using local topography. *Journal of Hydrometeorology*, 18(2):335–348.

Wyatt, T., Miralles, O., Massé, F., Lima, R., da Costa, T. V., and Giovanini, D. (2022). Wildlife trafficking via social media in brazil. *Biological Conservation*, 265:109420.

Zardi, D. and Whiteman, C. D. (2013). Diurnal mountain wind systems. In Chow, F. K., De Wekker, S. F., and Snyder, B. J., editors, *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, pages 35–119. Springer, Dordrecht, Netherlands.

OPHÉLIA MIRALLES

2 September 1994 @ opheliamiralles@gmail.com +33 6 43 92 19 47 Yens, Switzerland in ophelia-miralles

EXPERIENCE

PhD student in Mathematics/Assistant Teacher

EPFL (École Polytechnique Fédérale de Lausanne)

September 2020 – current Lausanne, Switzerland

- Built a deep learning model using generative adversarial networks to downscale wind maps from a 25km grid to a 1km resolution grid. Provided an api to generate high-resolution wind maps (github.com/OpheliaMiralles/wind–downscaling–gan).
- Showed that there is timing and selection bias in climate attribution studies and provided statistical methodology to avoid it and real usecases (github.com/OpheliaMiralles/timing–bias–extremes).
- Created a Python package to use this methodology and fluently perform likelihood-based statistical inference (github.com/OpheliaMiralles/pykelihood)
- Modelled hail damage per buioding in the canton of Zürich with a Bayesian hierarchical model and provided reliable upper bounds for extreme hail damages (github.com/OpheliaMiralles/hail–damage–modeling).
- Techs: Git, Linux, cartopy, xarray, tensorflow, pymc, pykelihood.

Research Assistant

UNEP (United Nations Environment Program)

June 2021 – October 2021 Geneva, Switzerland

- Provided contextual information and meaningful visualisation of geospatial data about environmental risks in Côte d'Ivoire in the scope of a climate risk and conflicts assessment. Helped translating the final report to French.
- Techs: Git, QGIS, cartopy, xarray.

Quantitative Analyst

Goldman Sachs

April 2019 – September 2020 London, United Kingdom

- On-boarded a wide range of clients on the highly competitive execution services market by providing unique and powerful analytics computed using state of the art trade optimisation techniques.
- Redesigned and completely automated the client reporting process from large order databases to a PowerPoint Presentation, dividing the man-hours needed for one request by 10.
- Handled efficiently big order-level databases to retrieve essential data and applied cutting-edge machine learning for stock clustering, providing reliable liquidity profiles for a large universe of assets.
- Led a research project about transaction costs of rebalancing systematic factor investment strategies.
- Techs: Git, Linux, Subversion, Jira, Cvxpy, Pandas, Python-pptx, Imgkit, SQL, QPad, Scikit Learn.

SCIENTIFIC PUBLICATIONS

1. T. Wyatt et al. (2022). "Wildlife trafficking via social media in Brazil". *Biological Conservation*, 265, p. 109420
2. O. Miralles, D. Steinfeld, et al. (2022). "Downscaling of historical wind fields over Switzerland using generative adversarial networks". *Artificial Intelligence for the Earth Systems*, 1(4), e220018
3. O. Miralles and A. C. Davison (2023). "Timing and spatial selection bias in rapid extreme event attribution". *Weather and Climate Extremes*, 41, p. 100584
4. O. Miralles, A. C. Davison, and T. Schmid (2023+). "Bayesian modeling of insurance claims for hail damage". *Annals of Applied Statistics*. Submitted

OTHER PUBLICATIONS

1. O. Miralles (2021b). *The Price of Nature*. `https://blogs.cfainstitute.org/investor/2021/02/10/the-price-of-nature/`. Published on the blog of the CFA Institute. Accessed: 2023–07–19
2. O. Miralles (2021a). *Comptabilité carbone: la clé pour une évaluation adéquate des risques climatiques*. French. `https://greentervention.org/2021/06/12/la-parole-aux-jeunes-comptabilite-carbone-la-cle-pour-une-evaluation-adequate-des-risques-climatiques/`. Published by Greentervention, a Belgian NGO providing guidance about environmental policies in the European Union. Accessed: 2023–07–19

SUPPORTING STATEMENT

I am an applied mathematics graduate with strong corporate experience in the financial markets, passionate about environmental conservation and social justice.
I am looking for a quantitative position where I can make a positive impact in an interesting and challenging environment, focused on sustainable and fair actions.

SOFTSKILLS

- Creative Thinking
- Good Listener
- Inter-Disciplinary Team Work
- Organization
- Flexible
- Multi-Tasking
- Work Under Pressure
- Communication

PROFESSIONAL SKILLS

Python	●●●●●
R	●●●●○
Linux	●●●●○
Excel	●●●○○
C++	●●○○○

STRENGTHS

- Research
  - Environmental statistics
  - Geospatial modelling
  - Statistics
  - Economics
  - Extreme events prediction and risk assessment
  - Machine/Deep Learning
- Development Tools
  - Pycharm
  - Visual Studio
  - R
  - SAS
- Data Analytics
  - pandas
  - numpy
  - matlplotlib
  - seaborn
  - plotly
  - beautifulSoup
  - xarray
  - cartopy
  - scikit learn
  - pymc
  - keras
  - tensorflow
- Data Management
  - Excel
  - Tableau

LANGUAGES

English	●●●●●
French	●●●●●
Spanish	●●●●○

OTHERS

- Driving License – B
- LaTeX, Powerpoint, Outlook

CERTIFICATES

Introduction to Conservation

National Geographic, United for Wildlife

April 2020 – No Expiration Date

Level 3: Regulation, Securities, Derivatives

Chartered Institute for Securities and Investment

Jan 2020 – No Expiration Date

EDUCATION

Master’s degree in Applied Mathematics

École Polytechnique - Sorbonne University (year 2)

September 2018 – December 2019 Paris, France

Ranked 3rd out of 80 with honours (“Mention Bien”).

Master’s degree in Applied Mathematics

Dauphine University (Y1)

September 2017 – June 2018 Paris, France

Ranked 4th out of 121 with the highest honours (“Mention Très Bien”).

Bachelor’s degree in Economics and Applied Mathematics

Toulouse School of Economics

September 2015 – June 2017 Toulouse, France

Ranked 3rd out of 79 with the highest honours (“Mention Très Bien”).

University Degree in Statistics

Paul Sabatier University

September 2016 – June 2017 Toulouse, France

Ranked 2nd out of 61 with the highest honours (“Mention Très Bien”).

VOLUNTEERING

Alliance Manager

RENCTAS International

September 2020 – current Morges, Switzerland

- Built meaningful partnerships with prestigious research laboratories in only a month, including Northumbria University’s Green Criminology department, to find innovative ways of fighting wildlife trafficking in Brazil.
- Supervised research projects involving teams in Brazil and in Europe, providing a structural framework, but also insights and ideas.
- Organized a crowdfunding plan based on micro-participation by individual agents.

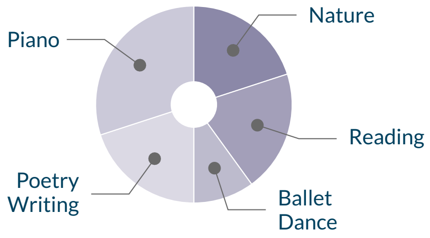
Communication Manager

TSE Junior Etudes

September 2016 – February 2017 Toulouse, France

- Attracted new joiners to the Junior Enterprise by showcasing the association on social networks and within the school with creative posters, events and meetings.
- Designed a descriptive and attractive new website and sent clients and members weekly tailored professional newsletters.
- Techs: Django, HTML, Adobe Muse.

INTERESTS


Nature
Piano
Reading
Poetry Writing
Ballet Dance