

---

# ON THE ASSESSMENT OF PRECIPITATION EXTREMES IN REANALYSIS AND ENSEMBLE FORECAST DATASETS

---

Inaugural dissertation  
of the Faculty of Science,  
University of Bern

*presented by*

PAULINE MARIE CLÉMENCE RIVOIRE

*from* FRANCE

*Supervisor of the doctoral thesis:*

Prof. Dr. Olivia Romppainen-Martius  
Geographisches Institut, Universität Bern

*Co-supervisor of the doctoral thesis:*

Dr. Philippe Naveau  
Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ



This work is licensed under a Creative Commons Attribution 4.0 International License.





---

---

# ON THE ASSESSMENT OF PRECIPITATION EXTREMES IN REANALYSIS AND ENSEMBLE FORECAST DATASETS

---

---

Inaugural dissertation  
of the Faculty of Science,  
University of Bern

*presented by*  
PAULINE MARIE CLÉMENCE RIVOIRE  
*from* FRANCE

*Supervisor of the doctoral thesis:*  
Prof. Dr. Olivia Romppainen-Martius  
Geographisches Institut, Universität Bern

*Co-supervisor of the doctoral thesis::*  
Dr. Philippe Naveau  
Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ

Accepted by the Faculty of Science.

Bern, 14th July 2022

The Dean:  
Prof. Dr. Z. Balogh



*Sur la mousse des nuages  
Sur les sueurs de l'orage  
Sur la pluie épaisse et fade  
J'écris ton nom*

LIBERTÉ

Paul Eluard  
*(Extrait)*



## Abstract

Precipitation extremes can trigger natural hazards with large impacts. The accurate quantification of the probability and the prediction of the occurrence of heavy precipitation events is crucial for the mitigation of precipitation-related hazards. This PhD thesis provides methods for the assessment of precipitation extremes. The methods are applied to different gridded datasets. The framework of extreme value theory, and more precisely the extended generalized Pareto distribution (EGPD), is used to quantify precipitation distributions.

Chapter 2 compares ERA-5 precipitation dataset with observation-based datasets and identifies the regions of low or high agreement of ERA-5 precipitation with observations. ERA-5 is a reanalysis dataset, i.e. a reconstruction of the past weather obtained by combining past observations with weather forecast models. The strengths of reanalysis precipitation fields are the regular spatio-temporal coverage and the consistence with the data on the atmospheric circulation from the reanalysis. However, precipitation in ERA-5 stem from short-term forecasts and the precipitation data calculation does not include observed precipitation. Therefore a comparison with observational datasets is needed to assess the quality of the precipitation data. We compare ERA-5 precipitation with two observation-based gridded datasets: EOBS (station-based) over Europe and CMORPH (satellite-based) globally. Both intensity and occurrence of precipitation extremes are compared.

We measure the co-occurrence of extremes between ERA-5 and the observational datasets with a hit rate of binary extreme events. We find a decrease in the hit rate with increasing rarity of events. Over Europe, the hit rate is rather homogeneous except near arid regions where it has a larger variability. In the global comparison, the midlatitude oceans are the regions with the largest agreement for the occurrence of extremes between the satellite observations and the reanalysis dataset. The areas with the largest disagreement are the tropics, especially over Africa. We compare the precipitation intensity extremes between ERA-5 and the observational datasets using confidence intervals on the estimation of extreme quantiles and a test based on the Kullback-Leibler divergence. Both the confidence intervals and the Kullback-Leibler divergence calculations are based on the fitting of the precipitation distribution with the EGPD. The quantile comparison indicates an overlap of the confidence intervals on extreme quantiles (with a probability of non-exceedance of 0.9) for about 85% of the grid points over Europe and 72% globally. The regions with non-overlapping confidence intervals between ERA-5 and EOBS correspond to regions where the observation coverage is sparse and therefore where EOBS is more uncertain. The two datasets have a good agreement over countries with dense observational coverage. ERA-5 and CMORPH precipitation intensities agree well over the midlatitudes. The tropics are a region of disagreement: ERA-5 underestimates quantiles for heavy precipitation compared to CMORPH.

In Chapter 3, we provide return levels of heavy precipitation events with regional fittings of the EGPD. The goal of this chapter is to develop a regional fitting method being a good trade-off between a robust estimation of the distribution and parsimony of the model, with a focus on precipitation extremes. We apply the method to ERA-5 precipitation data over Europe. This area of the dataset contains more than 20,000 grid points. A local fit of EGPD distributions for

all grid points in Europe would therefore imply estimating a large number of parameters. To reduce the number of estimated parameters, we identify homogeneous regions in terms of extreme precipitation behaviors. Locations with a similar distribution of extremes (up to a normalizing factor) are first clustered with a partitioning-around-medoid (PAM) procedure. The distance used in the clustering procedure is based on a scale-invariant ratio of probability-weighted moments focusing on the upper tail of the distribution. We then fit an EGPD with a constraint: only one parameter (out of three) is allowed to vary within a homogeneous region.

The outputs of Chapter 3 are 1) a step-by-step blueprint that leverages a recently developed and fast clustering algorithm to infer return level estimates over large spatial domains and 2) maps of return levels over Europe for different return periods and seasons. The relatively parsimonious model with only one spatially varying parameter can compete well against statistical models of higher complexity.

The last part of this thesis (Chapter 4) evaluates the prediction skill of operational forecasts on a subseasonal (S2S) time scale. Good forecasts of extreme precipitation are crucial for warnings and subsequent mitigation of natural hazards impacts. The skill of extreme precipitation forecasts is assessed over Europe in the S2S forecast model produced by the European Centre for Medium-Range Weather Forecasts. ERA-5 precipitation is used as a reference.

Extreme events are defined as daily precipitation exceeding the 95<sup>th</sup> seasonal percentile. The precipitation data is transformed into a binary dataset (threshold exceedance vs. no threshold exceedance). The percentiles are calculated independently for the forecast and the reference dataset: the direct comparison of dataset-specific quantiles removes potential biases in the data. The Brier score is computed as a reference metric to quantify the skill of the forecast model. In addition to the Brier score, a binary loss function is used to focus the verification on the occurrence of the extreme, discarding the days when the daily precipitation is not extreme, in both the forecast and the verification datasets. A daily and local verification of extremes is conducted; the analysis is extended further by aggregating the data in space and time. Results consistently show higher skill in winter compared to summer. Portugal, Norway and the South of the Alps are the regions with the highest skill in general. The Mediterranean region also presents a relatively good skill in winter. The spatial and temporal aggregation increases the skill.

Each part of this thesis provides methods to model and evaluate precipitation extremes. The outcome of Chapter 2 is an evaluation of ERA-5 precipitation. Europe is found to be a region of good performance in this dataset. ERA-5 is therefore used to apply the regionalized estimation of return levels developed in Chapter 3. Furthermore, the reanalysis dataset is used as a reference for the estimation of the S2S forecast skill for precipitation extremes, in Chapter 4.

The appendix contains the additional articles in which I was involved during my PhD project, as a lead author or as a coauthor.

# CONTENTS

<b>Abstract</b>	<b>ii</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.1.1 What are precipitation extremes? . . . . .	1
1.1.2 Rare events and extreme value theory . . . . .	2
1.1.3 Gridded precipitation datasets . . . . .	4
1.2 Aims and outline of this thesis . . . . .	6
<b>2 A comparison of moderate and extreme ERA-5 daily precipitation with two observational datasets</b>	<b>9</b>
2.1 Introduction . . . . .	11
2.2 Data . . . . .	12
2.2.1 ERA-5 Precipitation . . . . .	12
2.2.2 Observation-based datasets . . . . .	12
2.2.3 Data Processing . . . . .	13
2.3 Methods . . . . .	13
2.3.1 Co-occurrence of Precipitation Events . . . . .	13
2.3.2 Intensity Assessment . . . . .	14
2.3.3 Difference in Number of Wet Days . . . . .	17

2.4	Results . . . . .	17
2.4.1	Number of Wet Days . . . . .	17
2.4.2	Co-occurrence of Precipitation Events . . . . .	18
2.4.3	Intensity Verification . . . . .	19
2.5	Summary and Discussion . . . . .	23
2.6	Conclusion . . . . .	27
2.7	Appendix . . . . .	29
2.7.1	Mean precipitation per day . . . . .	29
2.7.2	Number of Wet Days . . . . .	30
2.7.3	Hit Rate . . . . .	31
<b>3</b>	<b>High return level estimates of daily ERA-5 precipitation in Europe estimated using regionalized extreme value distributions</b>	<b>33</b>
3.1	Introduction . . . . .	35
3.2	Data . . . . .	37
3.3	Methods . . . . .	38
3.3.1	A scale-invariant ratio of PWM . . . . .	38
3.3.2	Clustering algorithm: partitioning around medoids (PAM) . . . . .	39
3.3.3	Regional fitting . . . . .	40
3.3.4	Assessment of the fitting . . . . .	42
3.4	Results . . . . .	43
3.4.1	Partition of ERA-5 over Europe . . . . .	43
3.4.2	Assessment of the fitting . . . . .	44
3.4.3	Return levels . . . . .	45
3.5	Discussion . . . . .	49
3.6	Conclusions . . . . .	50
3.7	Appendix . . . . .	52
3.7.1	Difference between the regional and the local fittings . . . . .	52
3.7.2	Return levels in Switzerland . . . . .	54
<b>4</b>	<b>A verification of extreme precipitation occurrence in S2S forecasts over Europe</b>	<b>57</b>
4.1	Introduction . . . . .	59
4.2	Data and Methods . . . . .	60
4.2.1	Data . . . . .	60
4.2.2	Definition of extreme events . . . . .	61
4.2.3	Metrics . . . . .	61
4.3	Results . . . . .	68
4.3.1	Daily and local comparison . . . . .	68
4.3.2	Temporal aggregation . . . . .	68
4.3.3	Spatial aggregation . . . . .	68
4.3.4	Dependence on weather regimes . . . . .	69



4.4	Discussion and Conclusion . . . . .	70
4.5	Appendix . . . . .	73
4.5.1	95 <sup>th</sup> percentile . . . . .	73
4.5.2	Local and daily comparison of extremes . . . . .	74
4.5.3	Temporally accumulated extremes . . . . .	75
4.5.4	Spatially accumulated extremes . . . . .	76
<b>5</b>	<b>Summary, concluding remarks and outlook</b>	<b>79</b>
5.1	Summary and concluding remarks . . . . .	79
5.2	Outlook . . . . .	82
	<b>Bibliography</b>	<b>85</b>
<b>A</b>	<b>Identifying meteorological drivers of extreme impacts: an application to simulated crop yields</b>	<b>111</b>
A.1	Introduction . . . . .	113
A.2	Data and Methods . . . . .	114
A.2.1	Climate and crop model simulations . . . . .	114
A.2.2	Data processing . . . . .	115
A.2.3	Explanatory data analysis . . . . .	116
A.2.4	Lasso regression . . . . .	118
A.2.5	Other models . . . . .	119
A.2.6	Segregation threshold adjustment . . . . .	119
A.2.7	Model performance assessment and sensitivity analysis . . . . .	120
A.3	Results . . . . .	123
A.3.1	Overall performance . . . . .	123
A.3.2	Explanatory variables . . . . .	123
A.4	Discussion . . . . .	127
A.4.1	Predicting bad yield years . . . . .	127
A.4.2	Important predictors . . . . .	129
A.5	Conclusion . . . . .	130
A.5.1	APSIM-Wheat model settings . . . . .	131
A.5.2	Additional figures . . . . .	133
<b>B</b>	<b>A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods</b>	<b>139</b>
<b>C</b>	<b>Guidelines for Studying Diverse Types of Compound Weather and Climate Events</b>	<b>141</b>
	<b>Acknowledgment</b>	<b>143</b>



# LIST OF FIGURES

1.1	Distribution of positive precipitation in Fall in Maringes . . . . .	3
1.2	S2S predictions between weather forecasts and seasonal outlooks . . . . .	6
2.1	95 <sup>th</sup> percentile of ERA-5 fall precipitation . . . . .	14
2.2	Relative position of the confidence intervals . . . . .	20
2.3	Number of seasons with overlapping confidence intervals . . . . .	21
2.4	p-value of the Kullback-Leibler divergence test . . . . .	22
2.5	Number of seasons not rejecting of null hypothesis of Kullback-Leibler test . . . . .	23
2.6	Global climatology ERA-5 precipitation . . . . .	29
2.7	Ratio of the number of wet days . . . . .	30
2.8	Hit rate ERA-5 (Europe) . . . . .	31
2.9	Hit rate ERA-5 (global) . . . . .	32
3.1	Partition of the PAM algorithm on ERA-5 data (Europe) . . . . .	43
3.2	QQ plots for regional and local fittings . . . . .	46
3.3	10-year return levels ERA-5 . . . . .	46
3.4	50-year return levels ERA-5 . . . . .	47
3.5	100-year return levels ERA-5 . . . . .	48
3.6	Difference 50-year return levels . . . . .	48
3.7	Difference 10-year return levels . . . . .	52
3.8	Difference 100-year return levels . . . . .	53
3.9	10-year return levels MeteoSwiss . . . . .	54
3.10	50-year return levels MeteoSwiss . . . . .	55
3.11	100-year return levels MeteoSwiss . . . . .	56

4.1	Bias of the forecast for the 95 <sup>th</sup> percentile . . . . .	62
4.2	Definition of the last skilful day, BLF . . . . .	65
4.3	Illustration of the temporal aggregation . . . . .	66
4.4	Illustration of the spatial aggregation . . . . .	66
4.5	Brier score for local daily comparison . . . . .	68
4.6	Brier score with temporal aggregation . . . . .	69
4.7	Brier score with spatial aggregation . . . . .	70
4.8	95 <sup>th</sup> percentile ERA-5 . . . . .	73
4.9	BLF for daily local comparison . . . . .	74
4.10	BLF with temporal aggregation . . . . .	75
4.11	Illustration of the size of grid points . . . . .	76
4.12	BLF with spatial aggregation . . . . .	77
4.13	Brier score depending on NAO . . . . .	77
4.14	BLF depending on NAO . . . . .	78
4.15	Zonal mean BLF depending on NAO . . . . .	78
A.1	Mean annual yield over the 1600 model years . . . . .	116
A.2	Composite time series of meteorological variables . . . . .	121
A.3	Correlations between potential meteorological predictors and yield . . . . .	122
A.4	Confusion matrix for normal and bad years . . . . .	122
A.5	CSI of the Lasso regression . . . . .	124
A.6	Correlation between CSI and yield . . . . .	124
A.7	Map of selected predictors by the Lasso model . . . . .	125
A.8	Data processing . . . . .	126
A.9	Comparison between simulated yield and yield statistics . . . . .	133
A.10	Composite times series for USA . . . . .	134
A.11	Composite times series for China . . . . .	135
A.12	Number of months in the growing season . . . . .	136
A.13	Selected climate extreme indicators . . . . .	137

# LIST OF TABLES

1.1	Gridded precipitation datasets used in this thesis . . . . .	4
2.1	Difference in the number of wet days . . . . .	17
2.2	Mean hit rate ERA-5 vs observation . . . . .	18
2.3	Summary of precipitation distribution comparison of ERA-5 with the observation .	24
3.1	Description of the four EGPD models . . . . .	41
3.2	Anderson-Darling test, local and regional fittings . . . . .	42
3.3	AIC criterion, local and regional fittings . . . . .	45
A.1	Meteorological drivers used in the analysis . . . . .	117



## CHAPTER

# 1

# INTRODUCTION

## 1.1 Motivation

### 1.1.1 What are precipitation extremes?

Precipitation extremes are one of the most impactful natural hazards, triggering for example floods with various ranges of spatial extents. Intense daily or sub-daily precipitation can result in flash floods (Alfieri and Thielen, 2015) and can overwhelm sewer capacities in urban areas, such as the extreme precipitation event that resulted in an urban flood in Lausanne (Switzerland) in June 2018. On a temporal scale of days to weeks, large amounts of precipitation can result in regional scale floods. One recent example in Europe is the flood of the River Ahr in the Rhineland-Palatinate region in July 2021 (Ibebuchi, 2022; Munich Re, 2022). Heavy precipitation can also trigger landslides (see Dikshit et al., 2020; Bevacqua et al., 2021, for the Himalayan Region and Italy, respectively). Compound events of precipitation extremes together with other hazards represent an additional risk, e.g. compound wildfires and intense rainfall increase the risk of debris flows mudslides and flash floods in a region (Touma et al., 2022). Natural hazards caused by precipitation extremes have large repercussions for our society, with casualties (Bocheva and Pophristov, 2019; Zhang et al., 2021b; Munich Re, 2022), damage to infrastructure (Munich Re, 2022), impact on the economy (Kotz et al., 2022) and on crops (Mahmood et al., 2012; Rosenzweig et al., 2002). A good knowledge of natural hazards is a key step in the process of risk assessment (see Box 1). Understanding the formation, quantifying the probability and predicting the occurrence of extreme precipitation is therefore of crucial importance.

Climate change fostered by anthropogenic emissions brings further uncertainty to the estimation of precipitation. Even if the Clausius Clapeyron law provides information about water content in the air (approximately 7% per °C, Martinkova and Kysely, 2020), projecting changes in heavy

precipitation is not straightforward (Kim et al., 2020). Many regions experience an increase in drought frequency and intensity, linked to global warming exists for many regions. However, uncertainty exists in the modification of precipitation occurrence (Cook et al., 2018; Dai et al., 2018). The change in precipitation intensity depends on the location, the season and the intensity itself (Kim et al., 2020; Donat et al., 2014; Liang and Zhang, 2021). Precipitation extremes have a significant signal of intensification (Min et al., 2011; Westra et al., 2013; Gillett et al., 2022). The development of methods for the assessment of precipitation datasets, with a focus on precipitation extremes, is therefore essential.



This schematic illustrates the concept of disaster risk as defined by the Intergovernmental Panel on Climate Change (IPCC, 2012). The risk is defined as the likelihood of severe alterations in the normal functioning of a society, due to hazardous physical events interacting with vulnerability and exposure of the given society. The risk can be measured with the quantification of climate and weather extremes, vulnerability and exposure. This PhD dissertation focuses on the analysis of one type of extreme weather event, the precipitation extremes (blue pin).

*Box 1: Contextualisation of the analysis of precipitation extremes in the framework of disaster risk, as defined in the report of the IPCC (2012).*

It is even more critical to put effort into the analysis of precipitation extremes, because precipitation is a complex variable. Unlike other common atmospheric variables, like temperature or pressure, characterized by only one continuous process, precipitation requires to be described by two characteristics: 1. “Is it raining or not?” which is a binary process, 2. “if it is raining, what is the intensity of precipitation?” which is a continuous process. The objects of study of this thesis are both the occurrence and the intensity of precipitation extremes.

### 1.1.2 Rare events and extreme value theory

Since precipitation is characterized by its intensity and its occurrence, this PhD thesis investigates two approaches for extreme precipitation analysis. The emphasis is on methods to quantify the precipitation intensity and on methods to quantify the occurrence of precipitation events.

The probability density function is a tool that summarizes the information on precipitation intensity: it associates a probability to every possible amount of positive precipitation (per rainy day, in our case). For an analysis focused on heavy precipitation, the extreme value theory (EVT) is a suitable framework to fit the precipitation distribution, i.e. to determine a functional notation



of the probability density function. With the EVT methodology, distribution of extremes for intensity ranges even beyond observation can be estimated. The two classical EVT methods are the generalized extreme value (modelling block maxima, Jenkinson, 1955) and the generalized Pareto distribution (modelling exceedances over thresholds, Pickands III et al., 1975). The extended generalized Pareto distribution (EGPD, Tencaliec et al., 2020) is an extension of the generalized Pareto distribution to the whole distribution of precipitation intensity. The EGPD is very flexible and avoids the problem of the threshold selection; we therefore use this approach in this thesis to estimate the distribution of precipitation intensity. Figure 1.1 is an example of a precipitation distribution fitting: the whole range of positive precipitation for a given season can be modeled and summarized by the EGPD. This distribution modelling is the basis for the global assessment of ERA-5 precipitation (see Chapter 2) and the identification of coherent regions over Europe to estimate high return levels (see Chapter 3). Previous studies identified coherent regions in terms of extremes and performing regional fitting (e.g. Forestieri et al., 2018; Evin et al., 2016; Darwish et al., 2021), but these studies are conducted at a spatial scale much smaller than a continent. In Chapter 3, we perform a spatial clustering and a regional fitting of the EGPD for precipitation over the entire European continent.

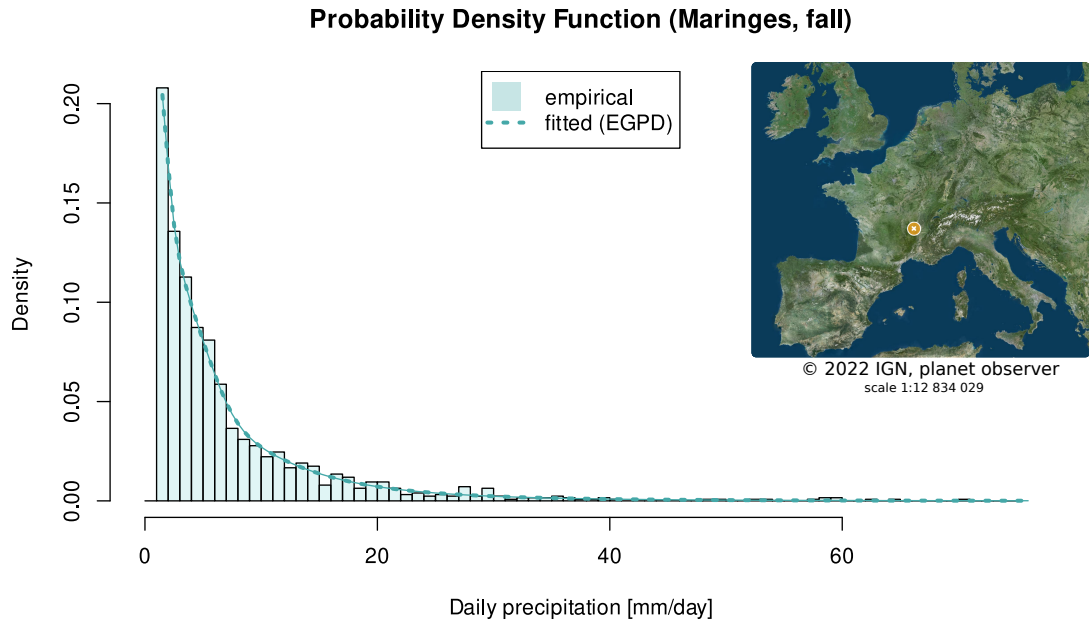


FIGURE 1.1: *Distribution of positive precipitation in Fall (September-October-November between 1979 and 2021, daily precipitation  $> 1\text{mm/day}$ ). The histogram represents the empirical distribution and the line represent the EGPD fit. Data: ERA-5, at the grid point containing Maringes (France).*

To characterize the binary occurrence of precipitation events, and more specifically extreme precipitation events, two metrics are used: the hit rate and the binary loss function. The former measures the event coincidences between ERA-5 and observation-based datasets, for moderate

Name (type)	Provided by	Spatial resolution	Temporal coverage	More details in
ERA-5 (reanalysis)	ECMWF	$0.25^\circ \times 0.25^\circ$ $0.5^\circ \times 0.5^\circ$ global and Europe	1979–2018, 2001–2020	Chapters 2, 3 and 4 Hersbach et al. (2019)
EOBS (station observations)	ECA&D	$0.25^\circ \times 0.25^\circ$ Europe lands	1979–2018	Chapter 2, Haylock et al. (2008)
CMORPH (satellite data)	NCAR	$0.25^\circ \times 0.25^\circ$ global	2003–2016	Chapter 2, Joyce et al. (2004)
ECMWF S2S data (hindcast)	ECMWF	$0.5^\circ \times 0.5^\circ$ Europe	2001–2020	Chapter 4, Vitart (2020)

TABLE 1.1: *Brief description of the gridded precipitation datasets used in this thesis. The different spatial resolutions and temporal coverage for ERA-5 depend on the dataset with which we compare ERA-5, e.g. a  $0.5^\circ \times 0.5^\circ$  grid resolution to fit with the hindcast resolution (see more details in the corresponding chapters).*

and extreme events (Chapter 2); the latter evaluates the ability of the subseasonal-to-seasonal forecast to capture extremes, penalizing false alarms and missed events (Chapter 4). Both metrics are extended with spatial and temporal flexibility.

### 1.1.3 Gridded precipitation datasets

*Precipitation* refers to any kind of liquid or solid particles of water originating in the atmosphere and falling to the earth’s surface (AMS, 2003). The main forms of precipitation include rain, snow, ice pellets and hail (ECMWF, 2018b). For analysis purposes, precipitation can be measured (with e.g. raingage, radar, satellite data, NASA and WorldBankGroup, 2022) or modelled (e.g. reanalysis data, climate models, ECMWF, 2022; NASA, 2022)

In this thesis, we use gridded precipitation datasets, i.e. geo-spatial datasets summarizing the precipitation occurring in squares of fixed size (fixed longitude width and latitude height). The advantage of such datasets is the regular spatial resolution, facilitating statistical analysis. Table 1.1 summarizes the datasets used in the different chapters.

For the statistical analysis of present-day precipitation (here second half of the XXth century to present), one can choose among different types of gridded datasets, with different characteristics. As stated above, the precipitation data can be observation-based or model-based. Both types of datasets have their uncertainties (Tapiador et al., 2012; Sun et al., 2018). These uncertainties can stem from the spatial aggregation to a grid or from measurement errors (Herrera et al., 2019; Yu et al., 2009) or from the model itself for model data (Sun et al., 2018). Reanalysis is a type of model dataset combining past observations with current numerical weather models. Physical and dynamical processes are computed to provide a consistent picture of the past weather, with a regular spatio-temporal coverage (ECMWF, 2022). In this thesis, when analysing past precipitation, we use ERA-5, the latest reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF, Hersbach et al., 2019). Previous studies focused on the verifica-

tion of ERA-5 precipitation, however these analyses are limited to a country or restricted region (Hénin et al., 2018; Wang et al., 2018; Mahto and Mishra, 2019; Tarek et al., 2020; Amjad et al., 2020; Jiang et al., 2021). To our knowledge, precipitation extremes have not been verified in ERA-5 on a European or global scale. Chapter 2 provides a European and global comparison of ERA-5 moderate and extreme precipitation with observational datasets.

Numerical weather forecasts predicting future precipitation are key for mitigation and adaptation to extreme events. One can assess the quality of forecast data only once the event actually happened, i.e. when the future is not the future anymore, but present or past. Hindcast data consist of forecast models run on past data. Hindcasts are necessary to verify forecast models over long time periods, against a past reference for “observation” (observation-based or reanalysis data, for a consistent and regular picture of the past). Because a forecast is constrained only by initial conditions, after several days, a simple deterministic forecast has limited skill. Indeed, initial condition errors can grow uncontrollably (WCS, 2021). An ensemble forecast consists of several forecast runs, so-called members: one member with the exact initial conditions, and several other members, with slightly perturbed initial conditions (Wilks, 2011). The ensemble members provide an estimation of the uncertainty of the forecast. The subseasonal-to-seasonal (S2S) forecasting timescale refers to forecasting timescales from two weeks to a season. S2S prediction aim to fill the gap between weather forecasts and monthly or seasonal outlooks (see figure 1.2 and White et al., 2017), with a large range of applications (White et al., 2021, and see the introduction to Chapter 4 for more details). Among the existing S2S precipitation hindcasts, the one provided by ECMWF (Vitart, 2020) is one of the most skilful (de Andrade et al., 2019). Existing studies on S2S extreme precipitation forecast are mainly focused on North America (Zhang et al., 2021a; DeFlorio et al., 2019), Africa (Olaniyan et al., 2018) and Asia (Yan et al., 2021; Li et al., 2019). Little is known about the performance of S2S models to predict heavy precipitation over Europe. In Chapter 4, we provide a skill assessment of ECMWF S2S forecast for precipitation extremes.

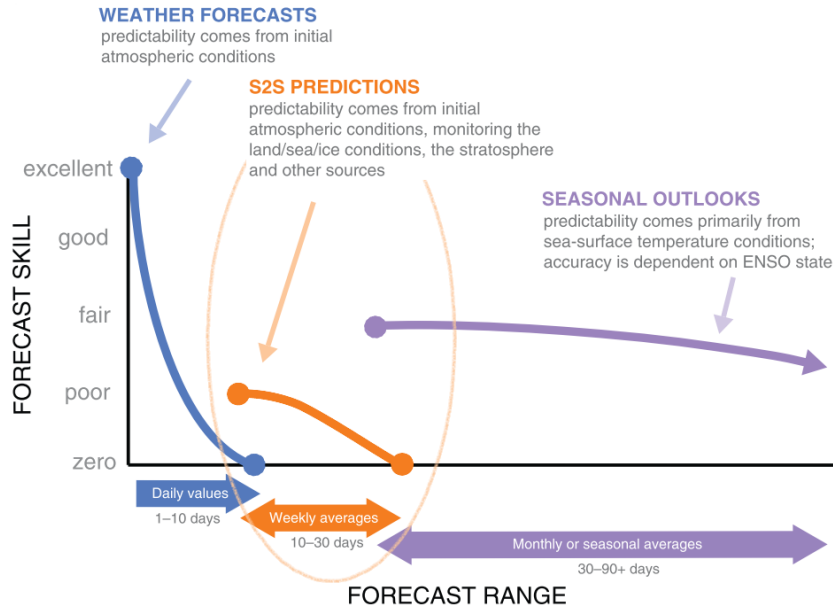


FIGURE 1.2: *S2S prediction aim to fill the gap between weather forecasts and monthly or seasonal outlooks (inspired from White et al. (2017)).*

## 1.2 Aims and outline of this thesis

The goal of this doctoral project is to provide a verification of precipitation extremes in the ERA-5 reanalysis precipitation dataset and in the ECMWF S2S precipitation hindcast dataset. This evaluation includes spatio-temporal considerations, with a regional estimation of ERA-5 extremes precipitation intensity and an assessment of temporally clustered extremes in the S2S hindcast.

To evaluate precipitation extremes, we fit distributions, using extreme value theory locally and regionally and we quantify the co-occurrence of events, with temporal and spatial flexibility. Additionally, we estimate the uncertainty and significance of results with bootstraps and statistical tests.

The developed methods provide a framework for the evaluation of precipitation and a quantification of the performance of ERA-5 and ECMWF S2S forecasts, two widely used datasets. This quantification includes the estimation of return levels, uncertainty, biases and ability to capture extreme events.

This thesis is structured in three main parts (Chapters 2 to 4), followed by a conclusion and appendices:

- In Chapter 2, we compare the reanalysis precipitation dataset ERA-5 with two observation-based datasets (Rivoire et al., 2021b). The comparison against the satellite dataset CMORPH

is global, and the comparison against the station-based dataset EOBS is conducted over the European continent. This chapter answers the research question: What is the performance of ERA-5 in modelling moderate and heavy precipitation compared to observation-based datasets? The novelty of this analysis is the global assessment of ERA-5 and the application of the EGPD fit to a European and a global level.

- Chapter 3 contains an estimation of high return levels of daily ERA-5 precipitation in Europe, using regionalized extreme value distributions. The goal is to answer the question: How can we characterize regionally extreme precipitation in Europe? We first define homogeneous regions in ERA-5 European precipitation, gathering grid points consistent in terms of extreme intensity behavior. Then we use this spatial clustering as a basis for a regional fitting of the EGPD. The spatial clustering of ERA-5 according to the extreme behavior and the estimation of high return levels with a regionalized EGPD fitting are the novelty brought by this chapter.
- We assess precipitation extremes in S2S forecast data in Chapter 4. ECMWF hindcast extreme events are verified against ERA-5 data. This chapter answers the question: What is the skill of the ECMWF S2S forecast in capturing precipitation extremes over Europe? We define extreme events as exceedances over thresholds, providing debiased binary data. The novelties of this analysis are a) the application of the binary loss function, in addition to the classic Brier score, with a temporal and spatial accumulation consideration and b) the evaluation of the ECMWF S2S forecast over Europe with a focus on extremes.
- Chapter 5 draws the conclusions of this thesis and brings an outlook and some perspective to this PhD project.
- The appendix collects three additional research articles in which I was involved during my PhD. Chapter A corresponds to the article “Identifying meteorological drivers of extreme impacts: an application to simulated crop yields” (Vogel et al., 2021), to which I contributed as a lead author. The abstract of the article “A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods” (Kopp et al., 2021) for which I was a co-author is included in Chapter B. Chapter C contains the abstract of the article “Guidelines for Studying Diverse Types of Compound Weather and Climate Events” (Bevacqua et al., 2021), to which I contributed as a co-author.



## CHAPTER

# 2

# A COMPARISON OF MODERATE AND EXTREME ERA-5 DAILY PRECIPITATION WITH TWO OBSERVATIONAL DATASETS

This chapter contains an article written together with Philippe Naveau and Olivia Martius. It was published in 2021 with the title “A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets” in the journal *Earth and Space Science* (Rivoire et al., 2021b).

## Abstract

A comparison of moderate to extreme daily precipitation from the ERA-5 reanalysis by the European Centre for Medium-Range Weather Forecasts (ECMWF) against two observational gridded datasets, EOBS and CMORPH, is presented. We assess the co-occurrence of precipitation days and compare the full precipitation distributions. The co-occurrence is quantified by the hit rate. An extended generalized Pareto distribution is fitted to the positive precipitation distribution at every grid point and confidence intervals of quantiles compared. The Kullback-Leibler divergence is used to quantify the distance between the entire extended generalized Pareto distributions obtained from ERA-5 and the observations. For days exceeding the local 90<sup>th</sup> percentile, the mean hit rate is 65% between ERA-5 and EOBS (over Europe) and 60% between ERA-5 and CMORPH (globally). Generally, we find a decrease of the co-occurrence with increasing precipitation intensity. The agreement between ERA-5 and EOBS is weaker over the southern Mediterranean region and Iceland compared to the rest of Europe. Differences between ERA-5 and CMORPH are smallest over the oceans. Differences are largest over North-West America, Central Asia and land areas between 15°S and 15°N. The confidence intervals on quantiles are overlapping between ERA-5 and the observational datasets for more than 80% of the grid points on average. The intensity comparisons indicate an excellent agreement between ERA-5 and EOBS over Germany, Ireland, Sweden and Finland, and a disagreement over areas where EOBS uses sparse input stations. ERA-5 and CMORPH precipitation intensity agree well over the mid-latitudes and disagree over the tropics.



## 2.1 Introduction

Natural hazards related to extreme precipitation (river floods, flash floods, landslides, debris flows and avalanches) cause casualties, damages to infrastructures and buildings and have direct and indirect economic impacts (MunichRE, 2018). For infrastructure planning and prevention measures, information about rare events, e.g., events that occur on average only once in a hundred years, is important. Such information can be obtained from precipitation data with statistical tools. Assessing the accuracy in high quantiles depends on spatial domain sizes and temporal availability. Different types of global precipitation datasets are available (Sun et al., 2018): global precipitation datasets are based on ground observations, satellite observations, combinations of ground observations and satellite observations and on short term weather model forecasts in reanalyses datasets. Reanalyses combine past observations with weather forecast models to reconstruct past weather. The main advantage of this type of precipitation dataset is its regular spatial and temporal coverage. Reanalyses ensure consistency of the precipitation data with the atmospheric conditions, which is important for weather and climate process studies. Here, we focus on ERA-5 precipitation (C3S, 2017). ERA-5 is the latest reanalysis product from the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA-5 precipitation is computed in short-term forecast started from reanalysis initial conditions (Hennermann, 2020). The ERA-5 precipitation production process does not include precipitation observation inputs. Hence comparison with observational data makes sense, keeping in mind that observation data have (partly substantial) uncertainties as well (Sun et al., 2018; Kulie et al., 2010; Prein and Gobiet, 2017). ERA-5 precipitation has already been widely used since its release in 2018, but very few assessments of this dataset have been conducted over large regions. Only precipitation over restricted areas and precipitation associated with specific type of events have been assessed (Wang et al., 2018; Hénin et al., 2018; Mahto and Mishra, 2019; Tarek et al., 2020; Amjad et al., 2020). The goal of this article is to assess daily precipitation in ERA-5 against observational datasets over large regions: Europe, comparing with the station-based dataset EOBS (Haylock et al., 2008), and the entire globe, comparing with the satellite-based dataset CMORPH (Joyce et al., 2004). The verification of daily precipitation will focus on the intensity distribution and on the temporal consistence, i.e., the co-occurrence of events. We use the extended generalized Pareto distribution (Tencaliec et al., 2020) to evaluate the intensity distribution and we calculate co-occurrence hit rates to assess the joint occurrence of precipitation events.

This paper is structured as follows. Section 2.2 describes the data used for this study. We introduce methods used for the comparison of co-occurrence and intensity in section 2.3. The results of our analysis are presented in section 2.4. Finally, the results are summarized and discussed in section 2.5.

## 2.2 Data

### 2.2.1 ERA-5 Precipitation

Reanalysis precipitation in this study are extracted from ERA-5 reanalysis dataset. ERA-5 is the latest global reanalysis dataset provided by the European Center for Medium-Range Weather Forecasts (C3S, 2017; Hersbach et al., 2020). In this dataset, precipitation stem from short-term forecasts and are available at an hourly resolution that we aggregate to daily precipitation. The precipitation data calculation does not rely on observed precipitation (ECMWF, 2018a). The data is interpolated to a regular grid with  $0.25^\circ$  resolution.

### 2.2.2 Observation-based datasets

The two gridded observation-based precipitation datasets used in this study are EOBS (Haylock et al., 2008) that is based on European station observations and CMORPH that is based on satellite observations (Joyce et al., 2004).

The EOBS dataset is provided by the European Climate Assessment & Dataset and is a daily gridded dataset based on spatially interpolated station data. The version used is 19.0e, with a  $0.25^\circ$  by  $0.25^\circ$  grid. The interpolation to a  $0.25^\circ$  by  $0.25^\circ$  grid is a combination of monthly precipitation totals and daily anomalies products. Figure 1a in Cornes et al. (2018) displays the station coverage for version 16.0. This coverage is heterogeneous, with a very dense network in Ireland, the Netherlands, Germany, Switzerland, and northern Italy, for example, and very few stations in northern Africa, in the Middle East, in Iceland, in Norway, and in Sweden. The station density strongly modulates the influence of the spatial interpolation procedure with the seasonal climatology becoming more dominant in data-sparse regions. Cross-validation with station data showed that EOBS exhibits the highest seasonal RMSE in summer and the absolute bias is highest for the uppermost decile (Cornes et al., 2018). EOBS covers land precipitation only, and the comparison with ERA-5 is conducted for the time period between January 1979 and December 2018.

The second observational data, CMORPH, is provided by the National Center for Atmospheric Research (NCAR) (Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2011). This gridded precipitation product combines passive microwave satellite scans and geostationary satellite infrared data and provides 3 hour accumulations that we aggregate in daily accumulation. CMORPH stands for climate prediction center morphing method, the name of this combination technique. The precipitation estimation algorithm of this dataset is not able to capture snow (Joyce et al., 2004). A detailed evaluation of the CMORPH dataset can be found in Sun et al. (2018). CMORPH tends to produce more light and fewer heavy rain events in East Asia due to the bilinear interpolation routine (Yu et al., 2009). The spatial resolution of the gridbox is also  $0.25^\circ$ . The comparison with ERA-5 is conducted for the period 2003-2016, for latitudes between  $60^\circ$  S and  $60^\circ$  N.

The two observation-based datasets have the same grid resolution as ERA-5 but a shift of  $0.125^\circ$

in latitude and longitude is present for the coordinates of the grid points compared to ERA-5.

### 2.2.3 Data Processing

The study evaluates seasonal precipitation for September, October, November (SON); December, January, February (DJF); March, April, May (MAM); and June, July, August (JJA). Figure 2.6 in appendix displays the seasonal mean precipitation of the ERA-5 dataset. Separation between seasons ensures stationarity of the time series. The intensity distribution analyses are based on wet days, defined as days with precipitation accumulations exceeding 1 mm. The 1 mm threshold corresponds to standard recommendations for station data (Hofstra et al., 2009) and eliminates potential drizzle effect in reanalysis data (Maraun, 2013). The co-occurrence analysis is conducted on the entire seasonal time series, including days with precipitation lower than 1 mm.

The precipitation time series are not de-trended in the present study as the response of precipitation to increasing atmospheric CO<sub>2</sub> varies with the precipitation intensity (Pendergrass and Hartmann, 2014) and trends depend on the length of time series (Scherrer et al., 2016). Moreover, Donat et al. (2014) identified mostly small or insignificant trends for the past thirty years.

## 2.3 Methods

For the sake of simplicity, in the method section OBSER denotes the observation-based datasets. OBSER can be either EOBS or CMORPH, as the comparison procedures between ERA-5 and EOBS and between ERA-5 and CMORPH are identical.

### 2.3.1 Co-occurrence of Precipitation Events

Binary events are defined here as occurrences of daily precipitation above the  $P^{th}$  seasonal percentile with  $P \in \{75, 90, 95, 99\}$ . Percentile values can be different between ERA-5 and OBSER. In Figure 2.1, the 95<sup>th</sup> precipitation percentiles in SON is displayed for: (a) ERA-5 over the EOBS domain (1979-2018), (b) EOBS over its entire domain (1979-2018), (c) ERA-5 over the CMORPH domain (2003-2016) and (d) CMORPH over the entire domain (2003-2016).

We define a co-occurrence between two datasets when two exceedances occur either at the same grid point on the same day, at the same grid point with one day lag, or at one of the eight surrounding grid points on the same day. During the spatial and temporal shift, a single event is never used more than once when looking for co-occurrences. We allow for one day shift to bypass uncertainties that arise having a fixed 24h time window (Haylock et al., 2008). The extension to the eight grid points around the centre point addresses potential issues arising from the precipitation interpolation to the different grids.

The hit rate is the ratio between the number of joint events and the total number of events (Rhodes et al., 2015). The total number of events is the same if computed from ERA-5 or from OBSER.

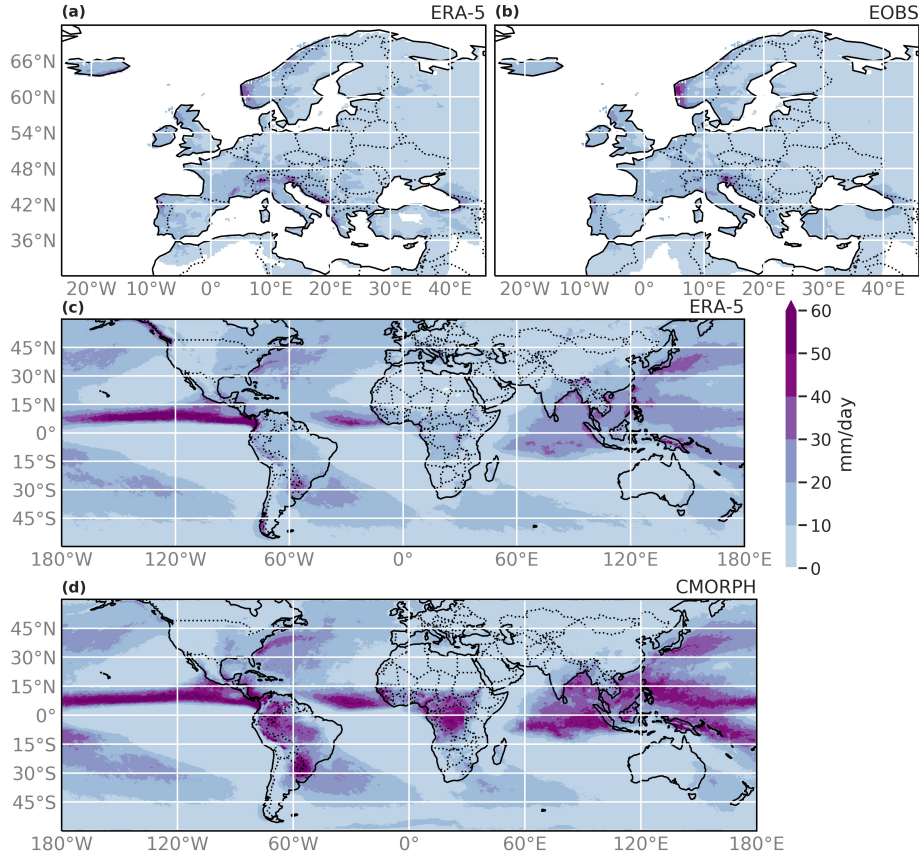


FIGURE 2.1: 95<sup>th</sup> precipitation percentile (mm) for all days in SON for (a) ERA-5 over Europe 1979-2018 (b) EOBS 1979-2018 (c) ERA-5 globally 2003-2016 (d) CMORPH 2003-2016.

### 2.3.2 Intensity Assessment

Extreme value theory is often used in hydrology and climate sciences (e.g. Lamb and Kay, 2004; Cooley et al., 2007; Trambly et al., 2013; Kang and Song, 2017). This approach states that peaks over high thresholds, i.e., amounts of rain exceeding a given threshold  $u$ , may be approximated by a generalized Pareto distribution, provided the threshold and the number of observations are large enough and some additional mild conditions are satisfied (see section 2.3.2 for generalized Pareto distribution definition). However, the generalized Pareto distribution fitting has drawbacks. First, it only captures the upper tail behavior. A distribution combining gamma behavior for small and moderate precipitation amounts with generalized Pareto distribution behavior for high amounts can be a solution. Second, a threshold has to be determined for every station or grid point to separate the upper tail from the rest of the distribution (Dupuis, 1999). To overcome these challenges, here we use the extended generalized Pareto distribution (EGPD) (Naveau et al., 2016; Tencaliec et al., 2020).

To study wet day precipitation intensity distributions, this section presents a comparison of quantiles and a homogeneity test based on the Kullback-Leibler divergence. Both parts rely on our EGPD fit.

For the intensity comparison, we discard grid points where the number of wet days is smaller

than 500 days for the comparison with EOBS and smaller than 200 days for the comparison with CMORPH. Moreover, auto-correlation can be present in daily time series, for example when two consecutive wet days are fostered by the same weather system (e.g. Lenggenhager et al., 2019a). To address the auto-correlation in time, we consider that two precipitation events separated by two days are independent (Barton et al., 2016; Lenggenhager and Martius, 2019b; Fukutome et al., 2015). To ensure independence of the time series, the intensity assessment is conducted on one-third of the data that is randomly drawn. This approach is a trade-off between keeping enough data in the sub-samples to ensure robust fitting and best removing the auto-correlation in the data.

### 2.3.2.1 Extended General Pareto Distribution

In extreme value theory, one way to model the extremal tail behavior is the so-called peak-over-threshold approach (Pickands III et al., 1975). Under this framework, rainfall exceedances above a large threshold  $u$  are assumed to follow a Generalized Pareto distribution defined as

$$H_{\xi}(z) = \begin{cases} 1 - (1 + \xi z)_{+}^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - e^{-z} & \text{otherwise,} \end{cases} \quad (2.1)$$

where the positive scalar  $\sigma$  represents a scale parameter and the real  $\xi$  drives the upper tail behavior. A negative, null and positive  $\xi$  corresponds respectively to the “bounded”, “light” and “heavy” tail case, i.e. an upper tail that is bounded for  $\xi < 0$ , with exponential decay for  $\xi = 0$  or polynomial decay for  $\xi > 0$ . The selection of the threshold  $u$  is not trivial for large datasets, as each grid point may need a different optimal threshold (e.g. Deidda, 2010). A large threshold implies a small sample size of extremes and consequently, large uncertainties in the estimation of  $\sigma$  and  $\xi$ . Conversely, a moderate threshold leads to a possible incorrect approximation by a Generalized Pareto distribution, i.e. a large model error. To bypass this complex threshold selection step, Naveau et al. (2016) proposed a simple scheme to smoothly transition between the main body of the distribution and its upper tail, while keeping the constraint of modeling extremes with a Generalized Pareto distribution. The proposed model can be written as

$$F(x) = G\{H_{\xi}(x/\sigma)\}, \quad \text{for all } x > 0, \quad (2.2)$$

where  $G$ , the transition function, is a continuous cumulative distribution function (cdf) in the unit interval. By imposing the two constraints,  $\lim_{u \downarrow 0} \frac{1-G(1-u)}{u}$  is finite and positive and  $\lim_{u \downarrow 0} \frac{G(u)}{u^s}$  is finite and positive for some  $s > 0$ , the new cdf  $F(\cdot)$  is bound to be in compliance with extreme value theory for its lower and upper tails. This class of cdf is called extended generalized Pareto distribution (EGPD) family. In this study, the cdf  $G(\cdot)$  is estimated using a specific Bernstein polynomial approximation, more information can be found in Tencaliec et al. (2020). The R code is available upon request.

### 2.3.2.2 Quantile Confidence Intervals

Confidence intervals for the quantiles of ERA-5 and OBSER precipitation are computed using a semi-parametric bootstrap on EGPD fitting. For each grid point, the following bootstrap procedure is conducted. Two subsamples containing one-third of the time series each are randomly drawn from the initial wet day time series. Each of these two subsamples is bootstrapped 100 times each, and each bootstrapped sample is fitted by a EGPD.

Having our disposal 200 bootstrapped estimates of  $G(\cdot)$ ,  $\sigma$  and  $\xi$ , quantiles for any given non-exceedance probability can be computed from Eq. (2.2). In particular, 95% confidence intervals of the quantiles are obtained by calculating the empirical 2.5% and 97.5% quantiles of the 200 bootstrapped quantiles values. An important feature to assess the proximity of our different datasets is to check if the confidence intervals from ERA-5 overlap (or not) with the observational datasets.

### 2.3.2.3 Kullback-Leibler Divergence Test

The well-known Kullback-Leibler divergence used in various fields “measures” the distance between two probability density functions, say  $f_1$  and  $f_2$ . It is given by equation

$$\mathbb{E}_{f_1} \left[ \log \left\{ \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})} \right\} \right] + \mathbb{E}_{f_2} \left[ \log \left\{ \frac{f_2(\mathbf{Y})}{f_1(\mathbf{Y})} \right\} \right] \quad (2.3)$$

with  $\mathbf{X}$  and  $\mathbf{Y}$  being random variables following respectively the probability density functions  $f_1$  and  $f_2$ .

Let  $X_{ERA-5} = (X_i)_{i=1, \dots, n}$  and  $Y_{OBSER} = (Y_j)_{j=1, \dots, m}$  be the time series (after removing the autocorrelation) of wet day precipitation in ERA-5 and OBSER. With  $\hat{f}_1$  and  $\hat{f}_2$  estimated by the EGPD fitting, we obtain the empirical value of the Kullback-Leibler divergence with equation

$$\frac{1}{n} \sum_{i=1}^n \log \frac{\hat{f}_1(X_i)}{\hat{f}_2(X_i)} + \frac{1}{m} \sum_{j=1}^m \log \frac{\hat{f}_2(Y_j)}{\hat{f}_1(Y_j)}. \quad (2.4)$$

The null hypothesis of our test is “ $X_{ERA-5}$  and  $Y_{OBSER}$  have the same distribution”, i.e.  $\hat{f}_1 = \hat{f}_2$ . The alternative hypothesis is  $\hat{f}_1 \neq \hat{f}_2$ .

The distribution of the Kullback-Leibler divergence under the null hypothesis is estimated using 300 values of the divergence between two vectors randomly drawn from a concatenation of  $X_{ERA-5}$  and  $Y_{OBSER}$ . The probability of “The Kullback-Leibler divergence between  $X_{ERA-5}$  and  $Y_{OBSER}$  is greater than the Kullback-Leibler divergence under the null hypothesis” is the  $p$ -value of the test. This  $p$ -value is empirically determined from the 300 values of the Kullback-Leibler divergence under the null hypothesis. The null hypothesis is rejected with a confidence level of 5% if the  $p$ -value is greater than 0.05.

### 2.3.3 Difference in Number of Wet Days

The intensity comparison is based on the EGPD, which is fitted to wet days only. Discrepancies in the number of wet days between ERA-5 and the observational datasets could have an impact when comparing the EGPD fitted to ERA-5 and the observational datasets. To quantify these discrepancies at a fixed grid point, we use two simple measures. The first measure is the ratio of the seasonal number of wet days, defined by :

$$\frac{N_{ERA-5}^{wd}}{N_{OBSER}^{wd}} \quad (2.5)$$

where, for a fixed season,  $N_{ERA-5}^{wd}$  and  $N_{OBSER}^{wd}$  are the number of wet days in ERA-5 and OBSER, respectively.

The second measure is the absolute value of the difference in the number of wet days between ERA-5 and OBSER, normalized by the time series length of OBSER, given by:

$$100 \times \left| \frac{N_{ERA-5}^{wd} - N_{OBSER}^{wd}}{N_{OBSER}^{wd}} \right|. \quad (2.6)$$

Note that the absolute difference quantifies the distance between the ratio of the number of wet days and 1.

## 2.4 Results

### 2.4.1 Number of Wet Days

Table 2.1 presents the mean absolute value of the difference in the number of wet days. The differences are computed only over grid points retained for the intensity comparison, i.e. with time series longer than 200 days for CMORPH and 500 days for EOBS.

TABLE 2.1: *Mean absolute value of the difference in the number of wet days.*

EOBS				CMORPH			
DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
14.5%	20.7%	18.5%	10.9%	66.1%	73.7%	75.2%	76.0%

Note: the mean absolute value of the difference in the number of wet days is defined in equation 2.6, and is computed here for grid points with more than 500 wet days for ERA-5 vs EOBS and for grid points with more than 200 wet days for ERA-5 vs CMORPH.

Over Europe, the mean absolute difference is between 11% (SON) and 21% (MAM) of EOBS number of wet days. In SON and DJF, the number of wet days is lower in ERA-5 than in EOBS in northern Europe and higher in southern Europe. In MAM and JJA the number of wet days is almost always higher in ERA-5 than in EOBS (see Figure 2.7a in the appendix for a map of the

ratio of the number of wet days in SON). The difference in number of wet days between ERA-5 and EOBS can be considered as low and will not have impact on the EGPD fitting.

The global comparison with CMORPH reveals larger discrepancies in the seasonal number of wet days than with EOBS. The mean difference in the number of wet days corresponds to between 66% (DJF) and 76% (SON) of the CMORPH number of wet days. This difference is mainly due to the ERA-5 wet days being more numerous than in CMORPH. The number of wet days in ERA-5 is twice as large as in CMORPH for 19% (DJF) to 25% (JJA) of grid points. Figure 2.7b in appendix displays the map of the ratio of number of wet days in SON. This ratio is greater than 2 over bands at fixed latitudes e.g. over bands close 60° S, 20° S, 20° N and 60° N.

### 2.4.2 Co-occurrence of Precipitation Events

In the comparison between ERA-5 and both observation-based datasets, the hit rate decreases with increasing intensity of the events and is similar across the seasons (Table 2.2). Grid points with a given percentile of less than 1 mm are not considered.

TABLE 2.2: *Mean hit rate ERA-5 vs EOBS and ERA-5 vs CMORPH.*

Percentile	EOBS				CMORPH			
	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
75 <sup>th</sup>	76%	75%	73%	77%	74%	73%	72%	73%
90 <sup>th</sup>	66%	65%	61%	67%	62%	61%	60%	60%
95 <sup>th</sup>	59%	58%	53%	60%	53%	52%	51%	52%
99 <sup>th</sup>	44%	45%	39%	45%	37%	35%	35%	36%

Note: for a given percentile, the mean is computed over all grid points where the precipitation percentile is larger than 1 mm. See section 2.3.1 for the definition of the hit rate.

Over Europe the average hit rate between ERA-5 and EOBS for the 75<sup>th</sup> percentile is between 73 % (in JJA) and 77 % (in SON), i.e. about three quarters of the events exceeding the 75<sup>th</sup> percentiles coincide. For the 95<sup>th</sup> percentile, the mean hit rate is between 53% (JJA) and 60% (SON).

For the 99<sup>th</sup> percentile, the hit rate varies between 39 % and 45 % depending on the season. The global mean hit rate is of the same order of magnitude as for Europe. The mean hit rate between ERA-5 and CMORPH for the 95<sup>th</sup> percentiles is above 50%. The mean hit rate associated with the 99<sup>th</sup> percentile is between 35% and 37% depending on the season.

Maps of the hit rate for the 95<sup>th</sup> percentile can be found in the appendix (Figure 2.8 and Figure 2.9). The spatial pattern does not strongly depend on the season or the percentile. For Europe, the hit rate has a large variability near arid regions (Maghreb and Turkey). The rest of Europe is quite homogeneous. A lower hit rate is observed in Iceland and southern Italy. For the global comparison, the best hit rate is reached over the oceans in the mid-latitudes. The hit rate is substantially lower in Eastern China, along the equator, in South America and in tropical Africa.



### 2.4.3 Intensity Verification

#### 2.4.3.1 Confidence Intervals on Quantiles

The confidence interval overlap between ERA-5 and EOBS is independent of the probability of non-exceedance, i.e., the intensity of the events. Figure 2.2a shows the relative position of the 95% confidence intervals for quantiles associated with probability of non-exceedance 0.9 in SON, between ERA-5 and EOBS. A grid point is displayed in yellow if the confidence intervals are overlapping, in orange if the upper boundary of ERA-5 confidence interval is lower than the lower boundary of the EOBS confidence interval, and in blue if the lower boundary of ERA-5 confidence interval is larger than the upper boundary of the EOBS confidence interval. Figure 2.3a shows the number of seasons with a confidence intervals overlap for quantiles with non-exceedance probability 0.9 between ERA-5 and EOBS. The confidence intervals overlap during all the seasons for a major part of Europe. The exceptions are Iceland, Norway and Western Russia, Romania, the Adriatic sea coast and some grid points in the Alps. Non-overlapping confidence intervals correspond primarily to an underestimation of the quantiles by ERA-5 for low precipitation intensities (not shown), and overestimation for large intensities (Figure 2.2a.). Quantiles with probability of 0.3 have a larger number of grid points with disagreement in JJA. ERA-5 quantiles for probability 0.3 during JJA are underestimated compared to EOBS quantiles for a major part of Europe (not shown).

The global comparison of ERA-5 with CMORPH shows less overlap of the confidence intervals with increasing precipitation intensity, as we will see in section 2.5. Figure 2.2b displays the relative position of the 95% confidence intervals for quantiles associated with probability of non-exceedance 0.9 in SON, between ERA-5 and CMORPH and Figure 2.3b shows the number of seasons with an overlap of the confidence intervals for quantiles with non-exceedance probability 0.9 between ERA-5 and CMORPH. The confidence intervals for non-exceedance probabilities of 0.3 and 0.5 overlap for more than 85% of the grid points. For probabilities 0.75 to 0.95, in a band along the equator the confidence intervals do not overlap for all the seasons for many grid points, see e.g. Figure 2.2b. For all seasons and between 15° S and 15° N, ERA-5 quantiles are smaller compared to CMORPH. Another disagreement that deserves to be highlighted is the higher ERA-5 quantiles compared to CMORPH over the mountainous regions of the west coast of North America, the south of Chile, in Papua New Guinea, the Himalayas and the Alps, regardless of the non-exceedance probabilities and seasons.

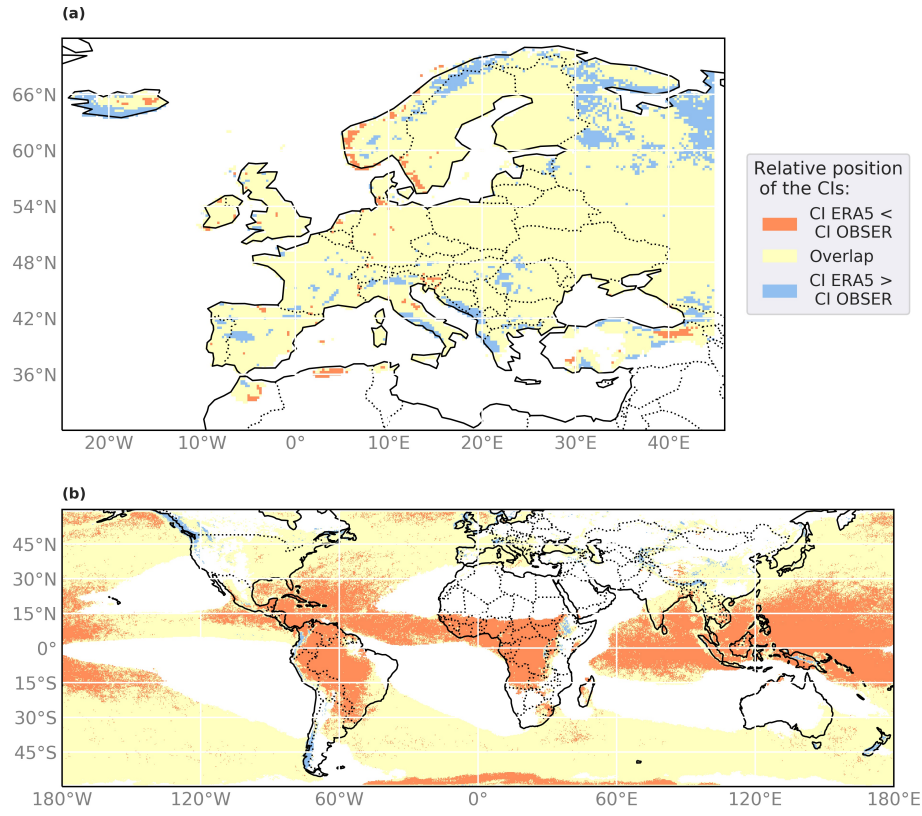


FIGURE 2.2: *Relative position of the confidence intervals (CIs) for SON quantiles associated with non-exceedance probability 0.9 between (a) ERA-5 and EOBS (1979-2018) and between (b) ERA-5 and CMORPH (2003-2016). See section 2.3.2.2 for computational details. Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.*

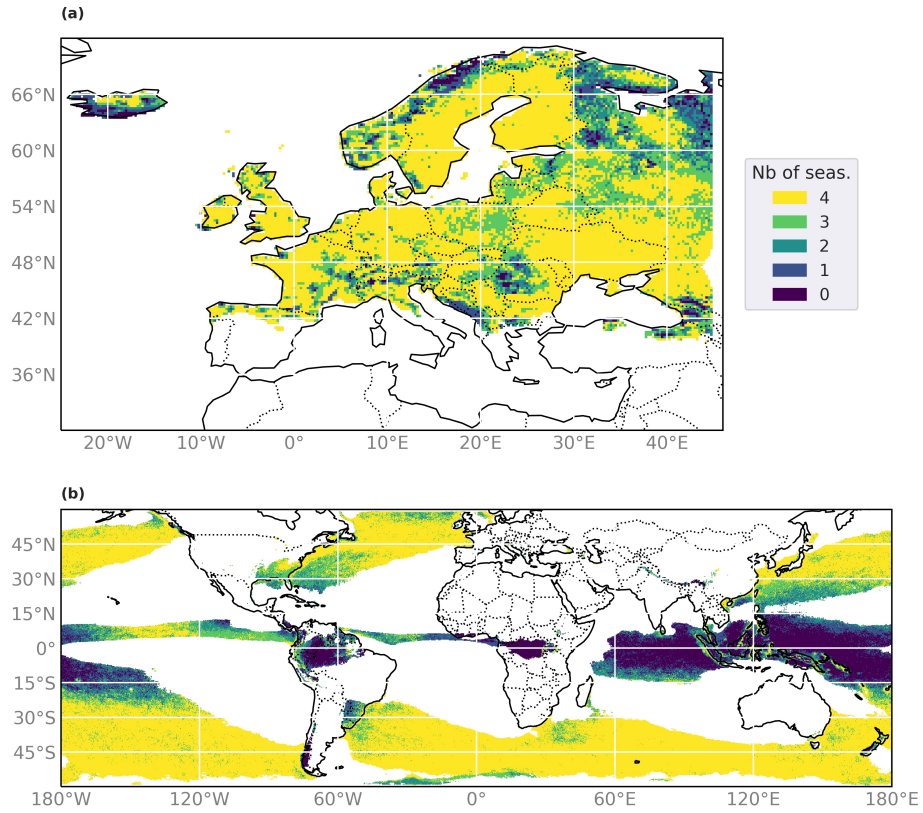


FIGURE 2.3: *Number of seasons with overlapping confidence intervals for quantiles associated with non-exceedance probability 0.9 between (a) ERA-5 and EOBS (1979-2018) and between (b) ERA-5 and CMORPH (2003-2016). See section 2.3.2.2 for computational details. Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.*

#### 2.4.3.2 Comparison of the Full Distributions Using the Kullback-Leibler Test

The Kullback-Leibler divergence of the full precipitation distribution points to regions of agreement and disagreement independent of the seasons, for both the comparison between ERA-5 and EOBS, and ERA-5 and CMORPH, see Figure 2.4. Figures 2.5a and 2.5b display the number of seasons for which the  $p$ -value of the Kullback-Leibler test is greater than 0.05, i.e. where the EGPD distributions fitted to ERA5 and to EOBS and CMORPH do not differ significantly.

ERA-5 and EOBS wet day precipitation intensities agree best over Germany, Ireland, Sweden and Finland. Wet day precipitation intensity follows the same distribution in ERA-5 and EOBS for most grid points in these countries. Regions with the least agreement, i.e. where the null hypothesis is rejected for all the seasons, are Iceland, Norway, Hungary and the Balkans. The area with at least one season where the null hypothesis is rejected is rather large. The Kullback-Leibler test gives weight to differences in the entire distribution, thus the low intensity precipitation disagreement in JJA mentioned previously in section 2.4.3.1 has an impact on the Kullback-Leibler divergence (see Figure 2.4a and Figure 2.4c for the plots of the  $p$ -value in JJA and SON). Note the pattern following the border between Norway and Sweden, and Finland and Russia: a very good agreement is observed in Sweden and Finland, whereas the null hypothesis is rejected

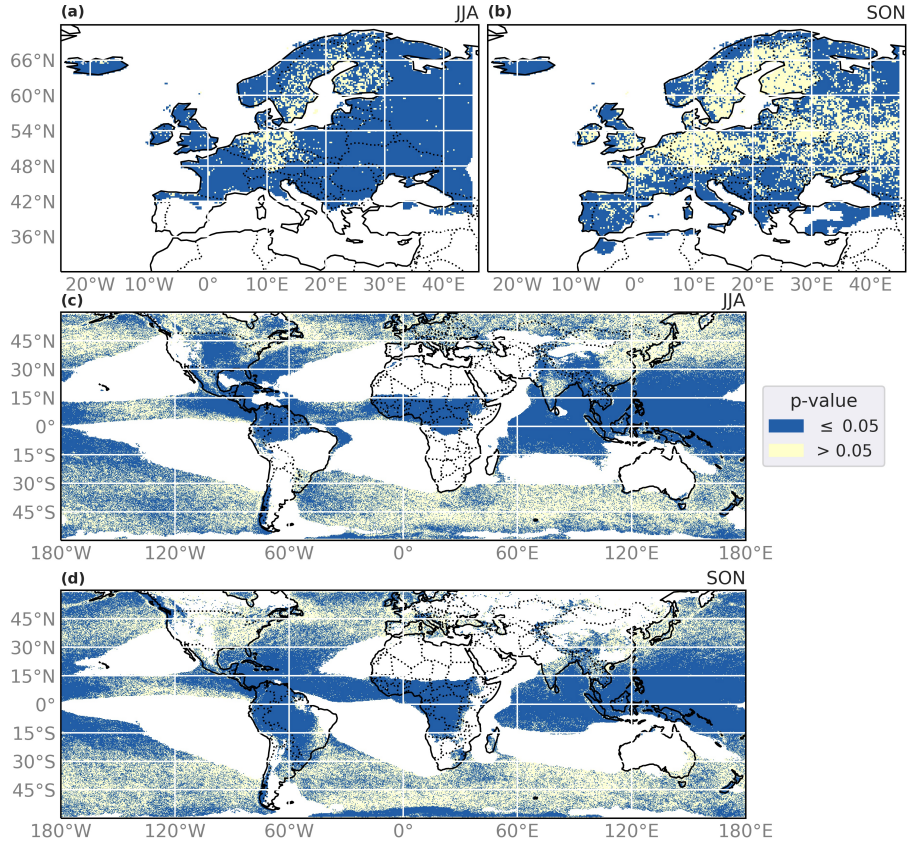


FIGURE 2.4:  $p$ -value of the Kullback-Leibler divergence test (as defined in section 2.3.2.3) between ERA-5 and EOBS (1979-2018) in JJA (a) and SON (b) and between ERA-5 and CMORPH (2003-2016) in JJA (c) and SON (d).  $p$ -values  $\leq 0.05$  indicate that the distributions differ significantly. Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.

for almost all seasons in Norway and Karelia.

The Kullback-Leibler test between ERA-5 and CMORPH has a clear signal of agreement in the mid-latitudes and disagreement in the tropics, for all seasons. The summary over all seasons mainly informs about intensity agreement over the oceans, because of the time series length constrain removes most land grid points. Figures 2.4b and 2.4d present the  $p$ -value for JJA and SON, and show over land the same general pattern of disagreement in the tropics and agreement in the mid-latitudes. One exception to this pattern is the disagreement over mountainous regions of the mid-latitudes (western North America, Himalayas, South Chile), in agreement with the results in section 2.4.3.1.

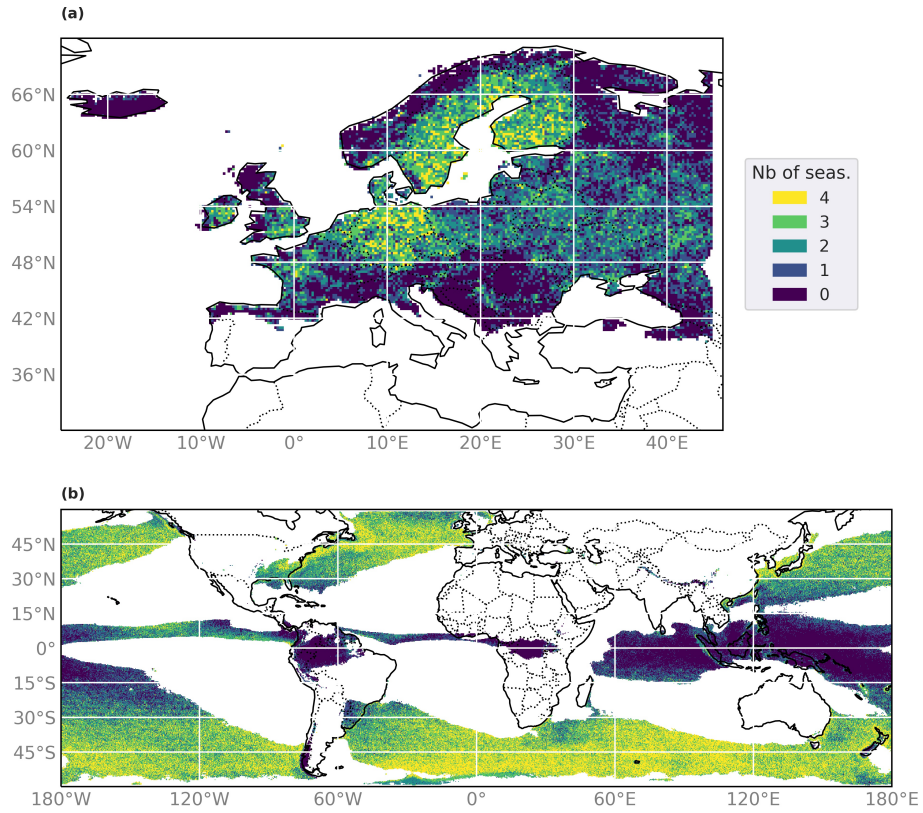


FIGURE 2.5: *Number of seasons where the distributions are similar, i.e., without rejection of the null hypothesis of the Kullback-Leibler test (as defined in section 2.3.2.3) (a) between ERA-5 and EOBS (1979-2018) and (b) between ERA-5 and CMORPH (2003-2016). Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.*

## 2.5 Summary and Discussion

The analysis of precipitation event co-occurrence between ERA-5 and EOBS and ERA-5 and CMORPH reveals a decreasing agreement with increasing intensity of events, independently of the season.

Key results of the intensity comparison of ERA-5 with EOBS over Europe depend on the season (Table 2.3). Quantiles in MAM and SON show a good agreement for all non-exceedance probabilities  $p$ . Indeed, between 81% (for  $p = 0.3$  in MAM and for  $p = 0.5$  in DJF) and 90% (for  $p = 0.5$  in MAM and SON, and for  $p = 0.75$  in SON) of the grid points have overlapping confidence intervals. The percentages of grid points where the distributions agree (Kullback-Leibler test) are highest for MAM (34%) and SON (39%). In DJF, the agreement between quantiles is between 69% ( $p = 0.9$ ) and 82% ( $p = 0.3$ ). In JJA agreement is much better for high quantiles (up to 94% for  $p = 0.95$ ) than low ones, the confidence intervals for quantiles with probability  $p = 0.3$  overlap for only 39% of grid points. This discrepancy for low precipitation intensity has an impact on the Kullback-Leibler test: the null-hypothesis could not be rejected for only 10% of grid points in JJA. The other seasons show a higher fraction of grid points for which the

TABLE 2.3: *Summary of the wet day precipitation distribution comparison of ERA-5 with the observational datasets.*

Precipitation intensity	EOBS				CMORPH			
	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
Low $p = 0.3$	82%	81%	39%	82%	94%	93%	92%	93%
Median $p = 0.5$	81%	90%	72%	90%	90%	86%	87%	88%
Moderate $p = 0.75$	72%	89%	93%	90%	79%	76%	76%	75%
High $p = 0.9$	69%	85%	93%	87%	75%	72%	73%	70%
Extreme $p = 0.95$	73%	87%	94%	87%	76%	73%	74%	71%
whole distrib.	29%	34%	10%	39%	57%	55%	53%	52%

Note: for a given non-exceedance probability  $p$ , the percentage denotes the proportion of grid points for which the confidence intervals are overlapping. For the whole distribution, the percentage denotes the proportion of grid points where the null-hypothesis can not be rejected, i.e. where the distributions are similar.

null-hypothesis was not rejected (between 29% in DJF and 39% in SON).

The study of the wet day precipitation distribution (Kullback-Leibler test) between ERA-5 and EOBS over Europe reveals a robust agreement over regions where the station coverage of EOBS is dense. Areas with the largest differences in the distribution are areas with thin station coverage, e.g. in southern Europe and Russia. Cornes et al. (2018) highlighted that “station coverage is the most important factor in determining the success of the gridded data”. In areas with sparse station data, the precipitation is interpolated from distant stations (Hofstra et al., 2009). Additionally, extreme precipitation is smoothed by the spatial interpolation (Hofstra et al., 2010), which justifies our quantiles larger in ERA-5 than in EOBS for extreme precipitation. For all seasons and a majority of non-exceedance probabilities, we find the same dry bias in south Austria and wet bias in North Austria as in Sharifi et al. (2019). In southern Italy, we observe an overestimation along the Tyrrhenian sea, as reported in Reder and Rianna (2021).

The intensity comparison between ERA5 and CMORPH indicates a decreasing agreement between the two datasets with increasing precipitation intensity (Table 2.3). The percentage of grid points with confidence intervals overlapping is between 92% (JJA) and 94% for  $p = 0.3$ , and between 70% (SON) and 75% (DJF) for  $p = 0.9$ . One exception is the slightly better agreement of quantiles for extreme precipitation (non-exceedance probability  $p = 0.95$ ) than for moderately extreme precipitation ( $p = 0.9$ ), with an overlap rate 1% higher, for all seasons. This can be explained by confidence intervals becoming larger for larger quantiles, and there is thus a higher



chance of overlap between ERA-5 and the observational dataset. This remark holds also for the EOBS results. The Kullback-Leibler test presents little variation with season, like for the quantile study. Between 52% (SON) and 57% (DJF) of the grid points studied did not reject the null-hypothesis of the test.

The analysis of the entire precipitation distribution reveals proportionally more grid points agreeing between ERA-5 and CMORPH than between ERA-5 and EOBS (see last row of Table 2.3). This can be due to the longer time series in EOBS leading to a stricter test. Another explanation can be that there are proportionally more challenging regions for a model over Europe, with the Alps for example, whereas globally the largest regions compared are oceans, where the agreement is good in general.

The global comparison of the wet day precipitation distributions between ERA-5 and CMORPH over the period 2003-2016 shows a rather good agreement in the mid-latitudes, and a strong disagreement over the tropics. This result is robust over the seasons and does not depend on the method used. The band along the equator where precipitation intensities are lower in ERA-5 compared to CMORPH corresponds to a region with a ratio of the number of wet days rather close to 1 (see Figure 2.7). The number of wet days does not differ substantially in this region and therefore does not play a role in the robust disagreement between ERA-5 and CMORPH over the tropics. In this region, ERA-5 quantiles are lower than CMORPH ones, especially for non-exceedance probabilities larger than 0.75. This feature has already been observed by Pfahl and Wernli (2012) with ERA-interim, another reanalysis product from ECMWF. They computed the empirical 99<sup>th</sup> percentiles of 6-hourly precipitation in ERA-interim and CMORPH for the period 2003-2016 (Figure 2 in their article). The authors showed a strong underestimation of ERA-interim precipitation over the tropics compared to CMORPH. They concluded that the deep convection, a central process in tropical extreme precipitation, was not properly captured by the reanalysis dataset. Even though Nogueira (2020) found an improvement of the precipitation simulation over the tropics in ERA-5 compared to ERA-interim the previous reanalysis dataset from ECMWF, we assume ERA-5 to still contain an underestimation of the tropical extreme rainfall.

One limit of CMORPH that it is important to emphasize is the limitations of this dataset to capture snow (Joyce et al., 2004) and low-intensity precipitation during winter in the mid-latitudes (Sun et al., 2018; Tian et al., 2007). This property leads to smaller number of wet days in winter (DJF or JJA depending on the hemisphere) as seen in section 2.4.1, hence the large areas where our intensity analysis can not be conducted. Some regions at high latitudes (close to 60° S and 60° N) and some mountainous regions have a number of wet days large enough for the intensity comparison to be conducted even if snow can be expected. However a substantial difference in the number of wet days is observed. The higher ERA-5 precipitation compared to CMORPH over mountainous regions might be related to snow. Timmermans et al. (2019) revealed a disagreement between CMORPH and gauge-based product for extreme precipitation in mountainous regions of western USA in DJF and interpreted it as a consequence of the post-processing performed in CMORPH leading to lots of missing data in winter (Xie et al., 2017). The wet bias observed in JJA and SON on the southern slope of central Himalaya is nevertheless confirmed by comparisons

with gauge station data (Chen et al., 2021).

Hénin et al. (2018) assessed ERA-5 daily accumulated precipitation during extreme precipitation events over the Iberian Peninsula for the period 2000-2008 against precipitation from a ground based gridded dataset. They found an overestimation of daily sums for moderate extreme events and underestimation for the most extreme events. Our study of quantiles in ERA-5 and EOBS for non-exceedance probabilities greater than 0.75 reveals a moderate signal of overestimation of ERA-5 precipitation in the same region. One exception is the southern Basque Country where an underestimation of ERA-5 quantiles for these probabilities in DJF. Our comparison of moderate and large extremes over the period 1979-2018 is therefore only partially in agreement with their study over the period 2000-2008.

For the period 2014-2018, Amjad et al. (2020) showed that ERA-5 overestimates the precipitation observed over Turkey compared to ground based stations, independently of the wetness and slope classes. Our comparison with EOBS indicates the same signal for high quantiles, but also a dry bias for lower quantiles. This can be due to the fact that our study period is longer or that EOBS has a poorer station coverage in this region.

In their study, Tarek et al. (2020) compared the mean seasonal precipitation in ERA-5 with station observations over North America between 1979 and 2018. In JJA, they found an underestimation of precipitation in ERA-5 over Florida, and an overestimation along the west coast of Canada, which is in agreement with our comparison of quantiles between ERA-5 and CMORPH for this season and for all probabilities of non-exceedance. In DJF, they showed an underestimation precipitation in ERA-5 over the west coast of USA and Florida, and an overestimation over the west coast of Canada. Our analysis highlighted that ERA-5 presents larger quantiles than CMORPH over the west coast of North America for all non-exceedance probabilities. Our results are thus in agreement for Canada but not for USA. This can be due to the fact that the time periods studied are different and that CMORPH underestimates precipitation during the cold months (Sun et al., 2018).

Mahto and Mishra (2019) assessed ERA-5 precipitation in India against observation comparing precipitation sums during the monsoon season (June-September) between 1980 and 2018. They found a wet bias over Indo-Gangetic Plain and foothills of Himalaya and a dry bias in semi arid regions of western India. These results are in agreement with our quantile analysis in ERA-5 and CMORPH in JJA for the period 2003-2016.

In Africa, the local overestimation of precipitation over Lake Victoria, Lake Tanganyika and the Ethiopian highlands is in agreement with existing literature (Gleixner et al., 2020). Over the Northern Great plains in North America in summer, we observe a slight dry bias as in Xu et al. (2019), in the western part of the region. We do not observe the slight wet bias signal, likely because of the resolution of CMORPH is coarser than the station data network used in their study. In the central region of China, quantiles are overestimated in JJA and DJF, agreeing with Jiang et al. (2021).



## 2.6 Conclusion

We compare daily precipitation from the ERA-5 reanalysis dataset with daily precipitation from two observation-based datasets, EOBS and CMORPH. The comparison addresses three aspects: i) the temporal co-occurrence of moderate to high extreme events in two datasets, ii) the agreement of return values for moderate to extreme non-exceedance probabilities derived from the extended generalized Pareto distribution (EGPD), and iii) a comparison of the full precipitation distribution captured by the EGPD using the Kullback-Leibler divergence. We quantify the co-occurrence of precipitation events with the hit rate. We compare the EGPD distributions between ERA-5 and the observational datasets with confidence intervals for several non-exceedance probabilities and with a test based on the Kullback-Leibler divergence.

Between ERA-5 and EOBS over Europe the hit rate is above 65% for moderate precipitation and approximately 50% for extreme precipitation. Between ERA-5 and CMORPH globally the hit rate is above 60% for moderate precipitation and around 40% for extreme precipitation. Over Europe areas with the least agreement are the southern Mediterranean region and Iceland and for the global comparison areas with the least agreement are land areas between 15°S and 15°N, North-West America and Central Asia.

For a majority of grid points confidence intervals for non-exceedance probabilities of 0.3 to 0.95 overlap between ERA-5 and EOBS. We find a disagreement between ERA-5 and EOBS in areas where EOBS uses fewer input stations. We therefore hypothesize that the reanalysis dataset might better capture moderate to extreme precipitation in regions where the station coverage is sparse. The analysis also showed that ERA-5 underestimates extreme precipitation compared to CMORPH in the tropics. In general, the magnitudes of the non-exceedance probabilities agree between ERA-5 and the observation-based datasets in the mid-latitudes.

The Kullback-Leibler test on the entire precipitation distributions over Europe shows an agreement of the EGPD distributions in ERA-5 and EOBS over Germany, Ireland, Sweden and Finland. The precipitation distributions differ significantly between in ERA-5 and EOBS in all four seasons in Iceland, Norway, Karelia, Hungary and the Balkan. The Kullback-Leibler test between ERA-5 and CMORPH shows that precipitation distributions are generally in agreement over the mid-latitudes and differ significantly over the tropics for all seasons, confirming the results of the quantile comparison. ERA-5 should only be used with great care to study extreme precipitation over the tropics.

The strengths of ERA-5 daily precipitation data are the regular spatial and temporal resolution and the consistency with the large-scale circulation and there is generally a good agreement with observation-based datasets in the extra-tropics. The reanalysis dataset provides valuable complementary information to observational data in regions where observational datasets are sparse, e.g. in areas where the EOBS station coverage is poor or for CMORPH in regions and seasons where snow is prevalent. In the tropics, an observational dataset should be preferred over ERA-5.

## Acknowledgment

P.R. and O.M. acknowledge funding from the the Swiss National Science Foundation (grant number 178751). Part of P.N. work was supported by the French national program FRAISE-LEFE/INSU, ANR-Melody and ANR-Trex. The support of DAMOCLES-COST-ACTION on compound events is also acknowledged.

The authors thank the assistance of Andrey Martynov (Institute of Geography, Bern), who prepared the precipitation datasets.

The article processing charges for this open-access publication were covered by the university of Bern.

The authors declare that they have no conflict of interest.

The authors acknowledge the E-OBS dataset from the EU-FP6 project UERRA (<http://www.uerra.eu>) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (<https://www.ecad.eu>). EOBS daily precipitation data at resolution  $0.25^\circ$  over Europe lands are available on this link: [https://surfobs.climate.copernicus.eu/dataaccess/access\\_eobs.php](https://surfobs.climate.copernicus.eu/dataaccess/access_eobs.php) Variable rr, ensemble mean. One must be a registered EOBS user and be signed in to download this dataset.

For ERA-5 we use global total daily precipitation data at resolution  $0.25^\circ$ . Hourly values were downloaded from the ECMWF MARS server (a valid ECMWF account required): <https://apps.ecmwf.int/data-catalogues/era5/?type=fc&class=ea&stream=oper&expver=1> Forecast steps 6 to 17, variable: Total precipitation (228.128). The MARS / EMOSLIB interpolation library has been used.

CMORPH global daily precipitation rate data at resolution  $0.25^\circ$  have been downloaded from <https://rda.ucar.edu/datasets/ds502.0/> (registration on <https://rda.ucar.edu/> required) Variable: CMORPH precipitation estimate.

The datasets to reproduce the figures and tables can be found in Zenodo (<https://doi.org/10.5281/zenodo.4443804>).

The codes for the co-occurrence verification and the intensity assessment are available from GitHub ([https://github.com/PauRiv/characterization\\_ERA-5\\_daily\\_precipitation](https://github.com/PauRiv/characterization_ERA-5_daily_precipitation)), as well as the figures for all seasons and probabilities/percentiles.

## 2.7 Appendix

### 2.7.1 Mean precipitation per day

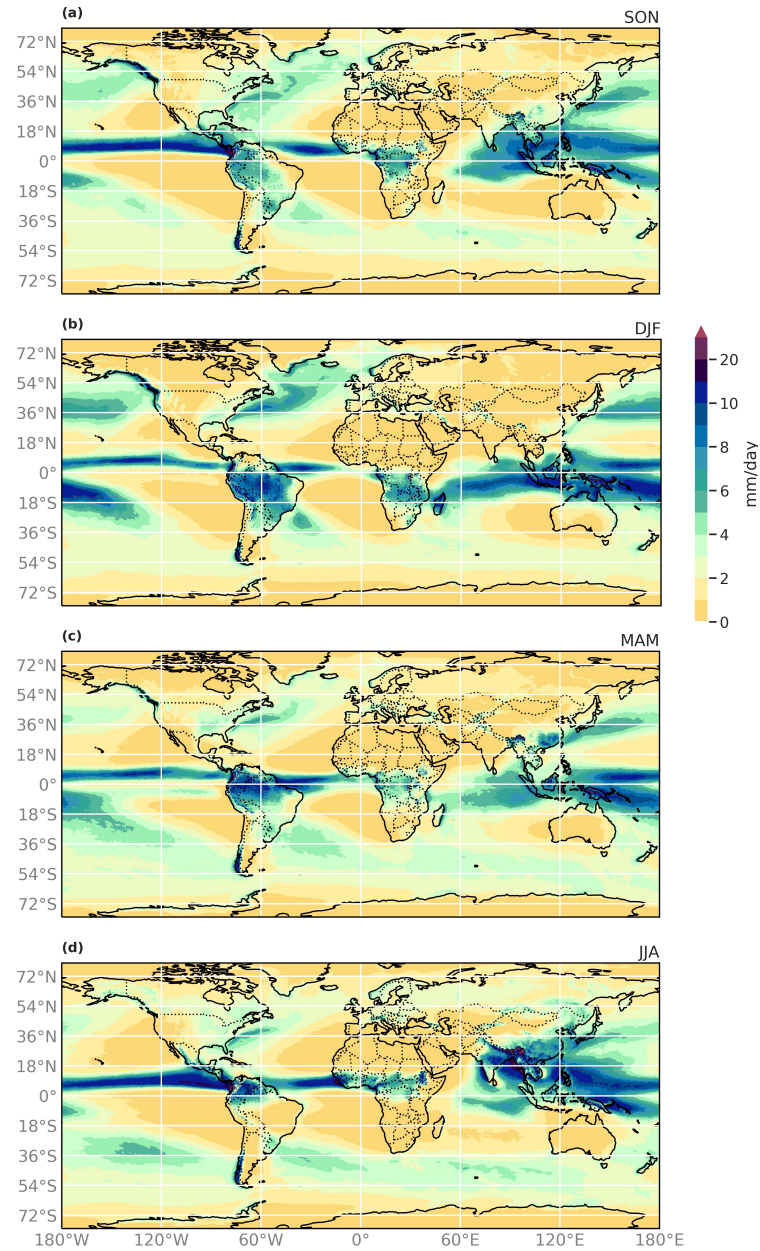


FIGURE 2.6: Mean precipitation per day in ERA-5 globally 1979-2018 in (a) SON (b) DJF (c) MAM and (d) JJA.

### 2.7.2 Number of Wet Days

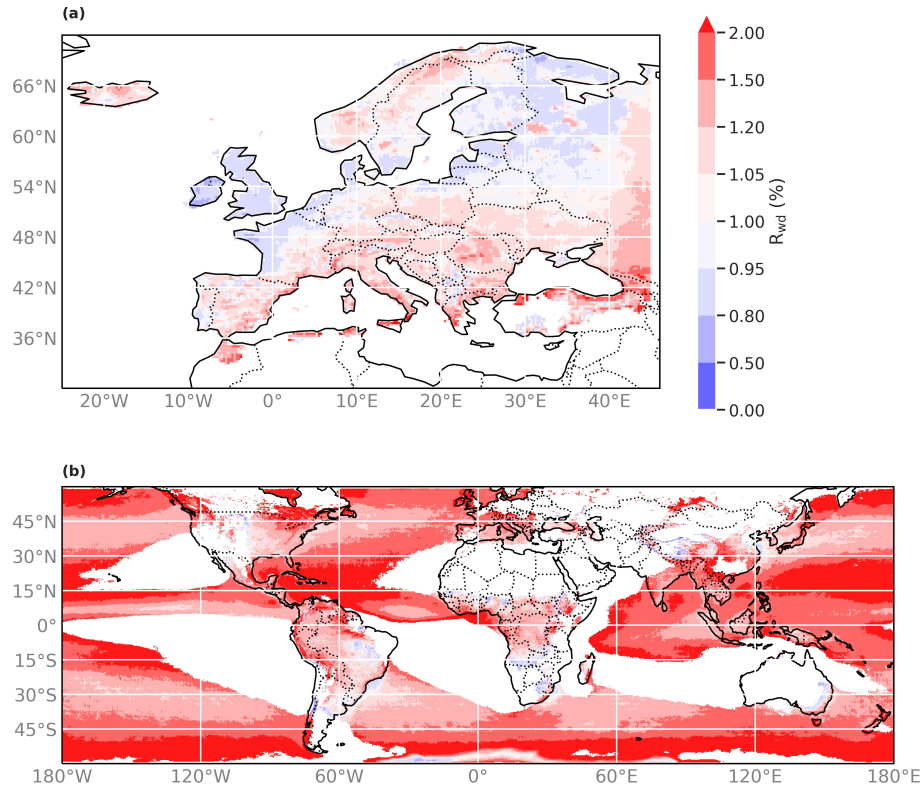


FIGURE 2.7: Ratio of the number of wet days as defined with Eq. (2.5) in SON between (a) ERA-5 and EOBS (1979-2018) and between (b) ERA-5 and CMORPH (2003-2016). Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.

## 2.7.3 Hit Rate

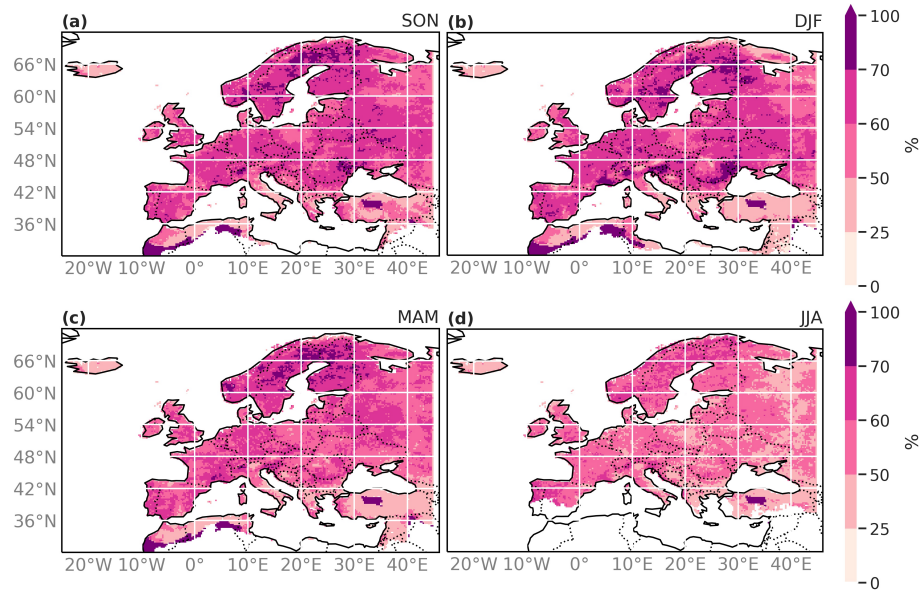


FIGURE 2.8: *Hit rate for events greater than the 95<sup>th</sup> percentile between ERA-5 and EOBS in (a) SON, (b) DJF, (c) MAM and (d) JJA.*



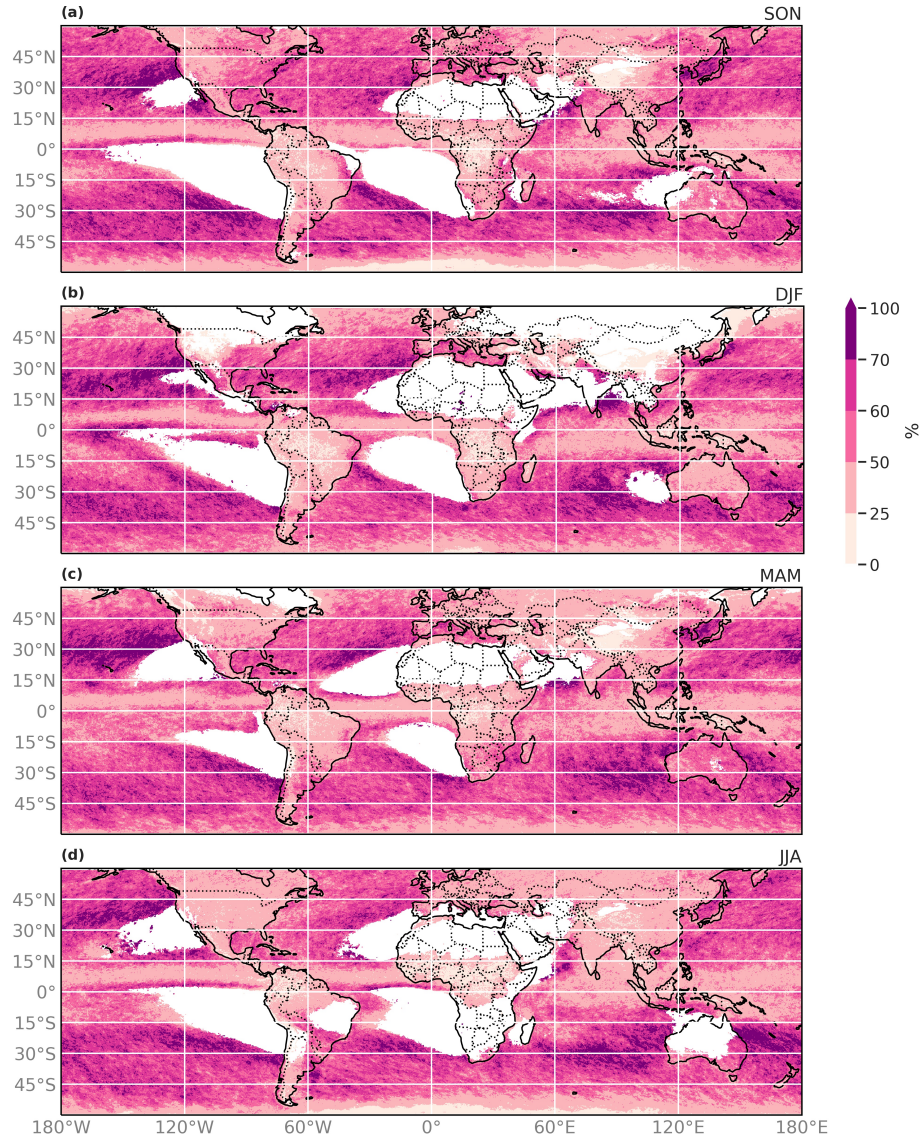


FIGURE 2.9: Hit rate for events greater than the 95<sup>th</sup> percentile between ERA-5 and CMORPH in (a) SON, (b) DJF, (c) MAM and (d) JJA. Grid points with an insufficient number of wet days (see section 2.3.2) are discarded and displayed in white.

## CHAPTER

### 3

# HIGH RETURN LEVEL ESTIMATES OF DAILY ERA-5 PRECIPITATION IN EUROPE ESTIMATED USING REGIONALIZED EXTREME VALUE DISTRIBUTIONS

This chapter contains an article submitted to the journal *Weather and Climate Extremes* in 2021 with the title “High return level estimates of daily ERA-5 precipitation in Europe estimated using regionalized extreme value distributions” (Rivoire et al., 2021a). Philomène Le Gall and I contributed equally to this paper. This article was written together with Anne-Catherine Favre, Philippe Naveau and Olivia Martius.

## Abstract

Accurate estimation of daily rainfall return levels associated with large return periods is needed for a number of hydrological planning purposes, including protective infrastructure, dams, and retention basins. This is especially relevant at small spatial scales. The ERA-5 reanalysis product provides seasonal daily precipitation over Europe on a  $0.25^\circ \times 0.25^\circ$  grid (about  $27 \times 27$  [km]). This translates more than 20,000 land grid points and leads to models with a large number of parameters when estimating return levels. To bypass this abundance of parameters, we build on the regional frequency analysis (RFA), a well-known strategy in statistical hydrology. This approach consists in identifying *homogeneous* regions, by gathering locations with similar distributions of extremes up to a normalizing factor and developing sparse regional models. In particular, we propose a step-by-step blueprint that leverages a recently developed and fast clustering algorithm to infer return level estimates over large spatial domains. This enables us to produce maps of return level estimates of ERA-5 reanalysis daily precipitation over continental Europe for various return periods and seasons. We discuss limitations and practical challenges and also provide a git hub repository. We show that a relatively parsimonious model with only a spatially varying scale parameter can compete well against statistical models of higher complexity.



## 3.1 Introduction

Heavy rainfall can cause natural hazards such as landslides, avalanches and floods (e.g., see Stocker et al., 2013; EEA, 2018). Such hazards can cause casualties and damages, with direct and indirect economic impacts (MunichRE, 2018; Prah et al., 2018). To design protective infrastructure, for instance, a dam, one needs to know the frequency of a given intensity of precipitation (Madsen et al., 2014). The return-period of an event is the duration during which the event occurs once, *on average* (see e.g. Cooley, 2013). Symmetrically, for a given duration, say 100 years, the 100-year return level is defined as the level that is exceeded once every 100 years, *on average*. Given a dataset (e.g. observation or reanalysis), time series are finite and observing an event exactly once in 100 years does not make an event the 100-year return level. One therefore needs a statistical model to predict the intensity of such events, and even unobserved events.

The aim of this paper is to provide return levels for large return periods over Europe. The station coverage being quite heterogeneous over Europe (Cornes et al., 2018) therefore, the use of gridded datasets is appropriate. Various types of gridded precipitation datasets are available (e.g., see Sun et al., 2018, for an overview). Precipitation gridded data can be derived from ground observations, satellite observations, combinations of ground observations and satellite observations and short-term numerical weather forecasts in reanalysis datasets. In reanalyses, past observations are assimilated in numerical weather forecast models to reconstruct past weather. The main advantage of this type of dataset is its regular spatial and temporal coverage. Reanalyses also ensure consistency of the precipitation data with the atmospheric conditions, which is a valuable characteristic for weather and climate process studies. Precipitation in this study is extracted from the ERA-5 reanalysis dataset (C3S, 2017; Hersbach et al., 2020). We study daily precipitation over continental Europe. The region of interest covers more than 20,000 grid points over European land.

Extreme value theory (EVT) provides an asymptotic framework to model the distribution of extremes such as heavy precipitation. Two classical approaches for extreme modelling are the generalized extreme value (GEV) and the generalized Pareto distribution (GPD). The GEV (Jenkinson, 1955) aims at modelling maxima over large blocks (for instance, a year in Posch et al., 2021). The GPD (see e.g. Pickands III et al., 1975, and Section 3.3.3) enables the modeling of exceedances over a given threshold (for instance, the 98-th quantile in Carreau et al., 2017). However, these two approaches only model extremes and our goal is to provide return levels in the full rainfall intensity range. We therefore need a class of distribution that can model the whole spectrum of precipitation intensities. Carreau and Bengio (2009), Papastathopoulos and Tawn (2013), Naveau et al. (2016) and Stein (2020) introduced distributions that model the whole spectrum of rainfall intensities. The methods model the upper tail with a Pareto distribution. Various types of transfer functions then fit the bulk and lower tail distribution. Tencaliec et al. (2020) defined a flexible version of the extended generalized Pareto distribution (EGPD) and Rivoire et al. (2021b) used it to fit the positive daily precipitation of ERA-5. The transfer function is estimated using Bernstein polynomials which bring flexibility to the transfer function estimation but require a large number of parameters (for example 30 for each grid point

in Rivoire et al., 2021b). In this paper, we simply use a monomial transfer function with a single flexibility parameter, see Section 3.3.3 for more details.

Poschlod (2021) fitted GEV distributions at each grid point of ERA-5 in Bavaria (Germany) to estimate 10-year precipitation return levels. However, extending this pointwise analysis of precipitation across Europe is quite onerous. Fitting a GEV and computing return levels for each grid point requires the estimation of more than  $3 \times 20,000$  parameters (location, scale and shape parameters). In addition, estimates of the shape parameter at a specific location are quite sensitive to the length of the time series (e.g., see Zhang et al., 2012; Malekinezhad and Zare-Garizi, 2014; Jalbert et al., 2017). Therefore reducing the dimensionality of the fitted parameters is of great practical importance. In contrast to this local approach, Sang and Gelfand (2009) and Naveau et al. (2014) assumed the shape parameter to be constant over the area of interest (Cape Floristic Region in South Africa and Switzerland, respectively). However, Europe is much larger than these areas, and the diverse climate and complex orography (Beck et al., 2018; Climate-Change-Service, 2020; ECMWF, 2006) strongly influence the spatial distribution of precipitation (e.g., see Evin et al., 2016; Marra et al., 2021). The method used for dimensionality reduction should preserve the diverse spatial patterns of precipitation over Europe. In this paper, we therefore consider an intermediate approach in which the shape parameter is common between grid points within homogeneous regions.

The regional frequency analysis (RFA), a concept from hydrology, attempts to build these homogeneous regions which consist of grid points with similar precipitation distributions (Dalrymple, 1960; Hosking and Wallis, 2005). In a homogeneous region, distributions are all equal to a common regional distribution up to a normalizing factor. In particular, their extreme behaviour should be analogous. Clustering grid points in homogeneous regions reduces the dimensionality of large precipitation datasets while preserving the spatial patterns. We use the definition of homogeneous distributions proposed by St-Hilaire et al. (2003) and Hosking and Wallis (2005): given a region of interest, say  $\mathcal{R}$  (here Europe), a homogeneous cluster ( $\mathcal{C}$ ) is defined as a sub-region where all spatial points  $s$ , have the same marginal distribution up to normalization:

$$\mathcal{C} = \{s \in \mathcal{R} : Q_s = \lambda(s) \times q\}, \quad (3.1)$$

where  $Q_s$  is the quantile function at site  $s$ , the positive scalar  $\lambda(s)$  varies in space, and  $q$  represents a positively-valued and dimensionless quantile function (common to every site in the cluster). As a consequence rescaled quantiles within a homogeneous cluster do not depend on localization  $s$ . Several methods allow regions to be delineated as in Eq. (3.1). They often require climate and/or geographical covariates (see e.g. Fawad et al., 2018; Forestieri et al., 2018, for recent work) and work in three steps: i) selecting explanatory covariates, ii) grouping sites with similar covariates, and iii) testing the homogeneity of the groups obtained. Covariates are selected for their ability to explain the precipitation distribution (Evin et al., 2016; Ouarda et al., 2008). For instance, Darwish et al. (2021) selected explanatory covariates by applying a principal component analysis to available geographical and climate data. They found that longitude, latitude, elevation and seasonality of events explained most hourly precipitation in the UK. With these methods, choosing covariates is an essential step that requires expert knowledge. Moreover, covariate

data must be available and may be complicated to transfer across regions with different climate characteristics. For example, the covariates that best describe precipitation may be different between the UK and Italy. To check the homogeneity of covariate-based groups, Hosking and Wallis (2005) proposed tests to examine the validity of the model corresponding to Eq. (3.1). The tests rely on two components: the moments that characterize the precipitation distribution, and the distributional assumption (Kappa-distributed, see e.g. Hosking, 1994). The tests consist of measuring the dispersion of some estimated L-moments (for all sites in the region) around a theoretical regional value of L-moments. To compute the theoretical value, Hosking and Wallis (2005) assume that the precipitation follows a Kappa distribution. This distributional assumption is not necessarily satisfied in practice. To bypass the selection of covariates, Saf (2009) and Le Gall et al. (2021) proposed methodologies using precipitation data only. They started from the hypothesis that the distributions are partially characterized by their probability weighted moments (PWM, Greenwood et al., 1979).

Le Gall et al. (2021) recently proposed a PWM-based algorithm to identify homogeneous spatial clusters of extreme precipitation and applied the algorithm to Swiss daily precipitation observations. The algorithm provided spatially coherent regions without using any geographical covariate. In this paper, we apply the clustering algorithm from Le Gall et al. (2021) to ERA-5 daily precipitation from all European land areas to group grid points with similar upper tails.

When clusters are delineated, information from all homogeneous grid points can be pooled to accurately estimate the EVT distribution parameters. For the regional distribution, we use an EGPD with three parameters see Section 3.3.3. Only the scale parameter can vary within a homogeneous cluster. The flexibility and shape parameters are constant over the cluster. In a nutshell, the regional approach allows us to go from a model with  $3 \times 20,000$  parameters to a model with  $2 \times n_{\text{clusters}} + 20,000$  parameters,  $n_{\text{clusters}}$  being the number of clusters. We also compare the performance of this regional approach to the performance of a more flexible distribution where the flexibility parameter can vary between sites of the same homogeneous cluster.

This study is the first to provide ERA-5 return levels, which, to our knowledge, have never been provided for the whole of Europe. Second, RFA is traditionally applied to smaller areas such as countries (see e.g. Fowler and Kilsby, 2003; Evin et al., 2016, for RFA on the UK and Switzerland).

Section 3.2 introduces the precipitation dataset and Section 3.3 describes the methods for the non-parametric clustering algorithm, the regional fitting and its assessment. The homogeneous regions, the assessment of the regional fitting and the corresponding 10, 50 and 100-year return levels are presented in the results section, Section 3.4. We discuss our results and compare our clusters to the regions obtained by national-scale studies in Section 3.5. We draw conclusions in Section 3.6.

## 3.2 Data

We use ERA-5 daily precipitation with  $0.25^\circ$  spatial resolution. ERA-5 is the latest global reanalysis dataset provided by the European Center for Medium-Range Weather Forecasts (C3S, 2017;

Hersbach et al., 2020). Precipitation is provided with hourly resolution forecasts that we aggregate to daily precipitation. We study ERA-5 precipitation for the period 1979–2018 in Europe over land, which is a region in which the dataset performs well (Rivoire et al., 2021b). Because practical applications are mainly restricted to the continent, we do not include precipitation data over the oceans. We conduct a seasonal analysis to ensure the stationarity of the time series. We consider the daily positive precipitation for each season. Days are considered as wet when precipitation exceeds 1 mm (Maraun, 2013).

### 3.3 Methods

Here, we introduce the two stages of RFA: i) identify homogeneous regions (Sections 3.3.1 and 3.3.2) and ii) use data from all grid points in the same region to model rainfall intensities (Section 3.3.3). We also introduce the evaluation tools we used to assess the fitted distributions (Section 3.3.4).

#### 3.3.1 A scale-invariant ratio of PWM

Following the notations of Le Gall et al. (2021), we denote  $\alpha_i(Z)$  the  $i$ -th PWM of the positive  $F$ -distributed random variable  $Z$

$$\alpha_i(Z) = \mathbf{E} [ZF(Z)^i].$$

When self-evident, it is denoted simply as  $\alpha_i$ . The first three moments are used to compute the scale-invariant ratio

$$\omega = \frac{3\alpha_2 - 2\alpha_1}{2\alpha_1 - \alpha_0}. \quad (3.2)$$

Le Gall et al. (2021) showed that  $\omega$  can be seen as a ratio of two distances derived from norms. Let  $i$  and  $j$  be two grid point locations, and  $Y_i$  and  $Y_j$  their two associated time series of seasonal positive precipitation. To spatially cluster daily rainfall, we need to compute a dissimilarity measure between two positive time series. Here, we use the  $\omega$ -based distance defined by

$$\hat{d}_{ij} = \left| \widehat{\omega(Y_i)} - \widehat{\omega(Y_j)} \right|, \quad (3.3)$$

where  $\widehat{\omega(Y_i)}$  is the estimate of  $\omega(Y_i)$ . We use this distance for two reasons. First, because the distance is based on PWM, it enables comparison of empirical distribution shapes, including heavy-tailed ones, without fitting a parametric distribution. Second, the key property of  $\omega$  is its scale-invariance. For any precipitation variables  $Y_1, Y_2$  in a homogeneous region, see Eq. (3.1),

$$\omega(Y_1) = \omega(Y_2).$$

The ratio  $\omega$  can be interpreted as the heaviness of the tail within the mathematical framework of EVT. In the block maxima or peak-over-threshold approaches,  $\omega$  only depends on the shape parameter. In the EGPD approach,  $\omega$  depends on the shape and the flexibility parameter, see

Section 3.3.3. The distance between two grid points with homogeneous distributions should be close to zero. The clustering algorithm gather sites with similar  $\omega$  estimates.

### 3.3.2 Clustering algorithm: partitioning around medoids (PAM)

Grouping close  $\omega$  estimates is an unsupervised learning problem: we gather unclassified points that have common characteristics (here, their  $\omega$  value). The grouping of estimates into clusters is based on geometric considerations: estimates are grouped if they are close to each other in the space of variables (here the axis of reals).

Several clustering methods are available (Kaufman and Rousseeuw, 1990; Murty et al., 1999; Schubert and Rousseeuw, 2021), most classic ones fall in two categories: partitioning or hierarchical methods. The  $k$ -medoids, or partitioning around medoids (PAM), and  $k$ -means are iterative algorithms that belong to the first group. They both require the final number of clusters  $k$  as input. The PAM algorithm is preferred to the  $k$ -means because of its ease of interpretation. Indeed, centres of the  $k$ -means clusters are barycentres and therefore virtual points whereas the centres of the PAM clusters are actual points of the dataset (see e.g. Jain et al., 1999; Bernard et al., 2013). For each of these methods, the choice of the dissimilarity measure is paramount. We work with the absolute difference as a distance, also called the Manhattan distance, as recommended in Bernard et al. (2013); Bador et al. (2015), see Eq. (3.3).

The centre of each cluster is the grid point with the smallest dissimilarity to all other grid points in the cluster and is called the medoid. Each non-medoid point of the data-set is associated with its closest medoid. Generally speaking, PAM converges to an ensemble of medoids and clusters that is a local minimum of the total cost, see Eq. (9) in Le Gall et al. (2021).

To solve this optimization problem, PAM starts by selecting  $k$  initial medoids, here in a deterministic way. The first medoid is the medoid of the partition for one cluster: the most centrally located point. The set of  $k$  medoids is then completed by adding the medoids of partitions with an increasing number of clusters one by one until  $k$  is reached.

The second step consists of testing every swap possible between a medoid and any point non-medoid in the whole dataset. If the total cost function (see e.g. Le Gall et al., 2021) decreases, then the point is kept as medoid. Clusters are then updated with respect to their new medoids. The algorithm stops when no swap decreases the total cost.

The computational cost of these two steps increases with the size of the dataset and the number of clusters. Because ERA-5 provides data for about 20,000 grid points in European lands, we use a faster version (Reynolds et al., 2006; Schubert and Rousseeuw, 2021) of the original algorithm. This variation removes some redundant computations in the swap step.

To measure the strength of the link between a point and its cluster, Rousseeuw (1987) introduced the silhouette score. The silhouette score for grid point  $i$  that belongs to the cluster  $k$  is defined as

$$1 - \left( \frac{d_{ik}}{\delta_{i,-k}} \right) \quad (3.4)$$

where  $d_{ik}$  is the average intra-cluster dissimilarity between grid point  $i$  and all other grid points in cluster  $k$ , and  $\delta_{i,-k}$  the smallest of the  $k - 1$  average distance between site  $i$  and all other sites

associated with a cluster different from  $k$ . When a grid point  $i$  is well classified, the intra-cluster average distance is significantly smaller than the distance between clusters. Its silhouette score is then close to 1. By contrast, a silhouette score close to -1 indicates a poorly classified grid point that should be in another cluster. Eventually, a grid point that is not significantly closer to points in the cluster than to other points has a silhouette score close to 0. In other words, it is not strongly linked to any cluster.

Finding the optimal number of clusters in a dataset is a tricky task (Sugar and James, 2003; Pansera et al., 2013). Numerous criteria that aim at identifying tight and well-separated clusters exist (Halkidi et al., 2002; Desgraupes, 2013). We compute five of them (silhouette, Dunn, Davies Bouldin, Xie Beni, S\_Dbw, see e.g. Halkidi et al., 2002; Desgraupes, 2013) to determine the optimal number of clusters, between two and ten. These criteria are based on different distances and provide a different optimal number of clusters. We therefore choose the number of clusters subjectively. We visually compare the maps of the partitions for numbers of clusters. We compromise between a large number of clusters and a partition that is not fragmented.

### 3.3.3 Regional fitting

To model the entire precipitation distribution, Naveau et al. (2016) and Tencaliec et al. (2020) proposed a simple scheme to build a flexible distribution by writing

$$F(z) = G(H_{\sigma,\xi}(z))$$

where the flexibility function  $G$  can be any cumulative distribution function such that there exists  $\kappa > 0$  such that  $\frac{G(u)}{u^\kappa}$  and  $\frac{1-G(1-u)}{u}$  have finite limits when  $u$  goes to zero. These constraints ensure that  $F$  follows EVT for very low and high precipitation accumulations. Here, we use  $G_\kappa(u) = u^\kappa$ ,  $\kappa > 0$ , as flexibility function. Although simple,  $G$  is sufficiently flexible to model daily rainfall distributions while maintaining parsimony in the model (Evin et al., 2016).

We fit the parameters to different levels of regionalization, from  $\sigma$ ,  $\xi$  and  $\kappa$  computed individually at every grid point to  $\sigma$  computed individually and  $\kappa$  and  $\xi$  being common between grid points in a homogeneous region (see Table 3.1).

Models	Flexibility function $G$	$\xi$	$\sigma$
Local Bernstein	$G_i =$ Bernstein polynomials, site specific	site specific	site specific
<b>Local</b>	$G_i(u) = u^{\kappa_i}, \kappa_i > 0$ site-specific	site-specific	site-specific
<b>Semiregional</b>	$G_i(u) = u^{\kappa_i}, \kappa_i > 0$ site-specific	constant on each cluster	site-specific
<b>Regional</b>	$G_i(u) = u^\kappa, \kappa > 0$ constant on each cluster	constant on each cluster	site-specific

TABLE 3.1: Description of the four EGPD models, with various complexity compared in Section 3.43.4.2. The Bernstein EGPD is presented in Tencaliec et al. (2020). The local EGPD is introduced in Naveau et al. (2016) and its regional version in Le Gall et al. (2021). The comparison is mainly conducted between the local, the semiregional and the regional fitting (in bold).

PWM can quickly be estimated non-parametrically and used for estimation of EGPD parameters (see Appendix of Naveau et al., 2016). Estimates of local parameters are taken as initial values for the iterative estimation of regional or semiregional parameters, see Algorithm (1).

---

**Algorithm 1** Regional fit of the EGPD in cluster  $C$ , see last row of Table 3.1.

---

- 1:  $cond = TRUE$ ,  $eps = .001$ , and  $u = 1(\text{mm})$
- 2: **procedure** INPUT(Rainfall Matrix for cluster  $C$ )
- 3:   Remove dry days by only taking  $\{Y(s)|Y(s) > u\}$
- 4:   Fit local Model at each location  $s \in C$
- 5:   Denote  $\kappa_0$  and  $\xi_0$  the cluster means of  $\kappa$  and  $\xi$  from Step 4
- 6:   Compute  $m(s)$  the sample mean at each  $s \in C$
- 7:   **while**  $cond = TRUE$  **do**
- 8:     Compute

$$\sigma_{new}(s) = \frac{\xi_0 m(s)}{\frac{\kappa_0}{F(u)} IB\left(H_{\xi_0}\left(\frac{u}{\sigma_0}\right), 1, \kappa_0, 1 - \xi_0\right) - 1}$$

where  $IB(., ., .)$  is the incomplete Beta function

- 9:     **if**  $|\sigma_{new} - \sigma_0| < eps$  **then**
  - 10:        $cond = FALSE$
  - 11:     **end if**
  - 12:      $\sigma_0 \leftarrow \sigma_{new}$
  - 13:   **end while**
  - 14:   Return  $(\kappa_0, \sigma_0, \xi_0)^T$
  - 15: **end procedure**
- 

The quantile with probability  $p$  can be computed using the explicit formula

$$y_p = F^{-1}(p) = \frac{\sigma}{\xi} [\{1 - G^{-1}(p)\} - 1], \quad \text{if } \xi > 0, \quad (3.5)$$

$0 < p < 1$ . The return level associated with return period of  $T$  years is  $y_p$  for  $p = \frac{1}{T \times n_{\text{wds}}}$ , where  $n_{\text{wds}}$  is the number of wet days per season. We use the mean of the number of wet days per season during the period under study as an approximation for  $n_{\text{wds}}$ .

For every grid point, we assume that the random variable modelling daily positive precipitation is independent and identically distributed. However, precipitation events can last for several consecutive days (Buriticá et al., 2021). To ensure independence in a time series of wet days, we extract one wet day out of three to fit the EGPD models. Despite the climate change, Donat et al. (2014) did not detect any clear trend in the whole precipitation distribution over the period of interest. The absence of a trend and the seasonal analysis are necessary to ensure identical distribution.

### 3.3.4 Assessment of the fitting

We evaluate the goodness-of-fit with standard statistical tools focusing on accuracy, flexibility of estimation, and rewarding of the parsimony (smaller number of parameters).

First, quantile-quantile plots (QQ-plots) provide visual information on the proximity between two distributions. For selected grid points, we present QQ-plots, contrasting the empirical quantiles with the quantiles parametrically estimated with the local, semiregional and regional fits, and EGPD with Bernstein flexibility function (see Table 3.1).

We assess the agreement between the fitting and the empirical distribution with the Anderson-Darling test (see e.g. Anderson and Darling, 1952; Scholz and Stephens, 1987). To ensure independence between the empirical and fitted distribution at a given grid point, we use a third of the positive precipitation time series that was not used in the fitting process as empirical data. Table 3.2 summarizes the results of the Anderson–Darling test over Europe for the regional, the semiregional, and the local fittings. To ensure spatial independence, we perform the tests for 1/8th of the grid points, randomly chosen. This way we avoid repetition of information between neighbouring grid points.

Model	SON	DJF	MAM	JJA
local	91%	89%	90%	87%
semiregional	89%	87%	88%	83%
regional	88%	88%	88%	84%

TABLE 3.2: *Anderson–Darling test at a risk level of 5%: Percentages of grid points for which the hypothesis of equality between the empirical distribution and the fitted distribution is not rejected. Distributions are fitted locally semiregionally and regionally; see second, third and last row of Table 3.1.*

To evaluate the goodness-of-fit, we compute the Akaike information criterion (AIC) (Akaike, 1987) for the local, the semiregional, and the regional models. This criterion combines a measure of the goodness of fit (log-likelihood) with the parsimony and sparsity of the model. The AIC has to be minimized. A smaller number of fitted parameters is a bonus for the model because this reduces the risk of overfitting. For example, the local model requires the estimation of about  $3 \times 20,000$  parameters, whereas the regional model only needs the estimation of about  $20,000 + (\text{number of clusters}) \times 2$  parameters.



## 3.4 Results

### 3.4.1 Partition of ERA-5 over Europe

We apply the clustering algorithm introduced in Section 3.3.2 to ERA-5 positive daily precipitation for each season independently.

The optimal number of clusters is three for September-October-November (SON), December-January-February (DJF), and March-April-May (MAM), and five for June-July-August (JJA, see Section 3.5 for a discussion about this number). Figure 3.1 shows these partitions. The shade of colour indicates the silhouette coefficient of the grid points; light colours indicate low silhouette coefficients and therefore a weak association with the cluster. There are very few isolated grid points. For all the seasons, the borders between clusters follow the orography, for example in the Alps, the Carpathians, and the UK. This orographic link is present in all seasons. Hence, the ratio  $\omega$  captures spatial structures associated with physical features such as orography without requiring additional covariates such as longitude, latitude, or elevation. Silhouette scores are lowest at the borders between clusters, and downward-pointing triangles, which indicate grid points with low and minimum silhouette coefficients, are often located in transition zones between clusters (Fig. 3.1).

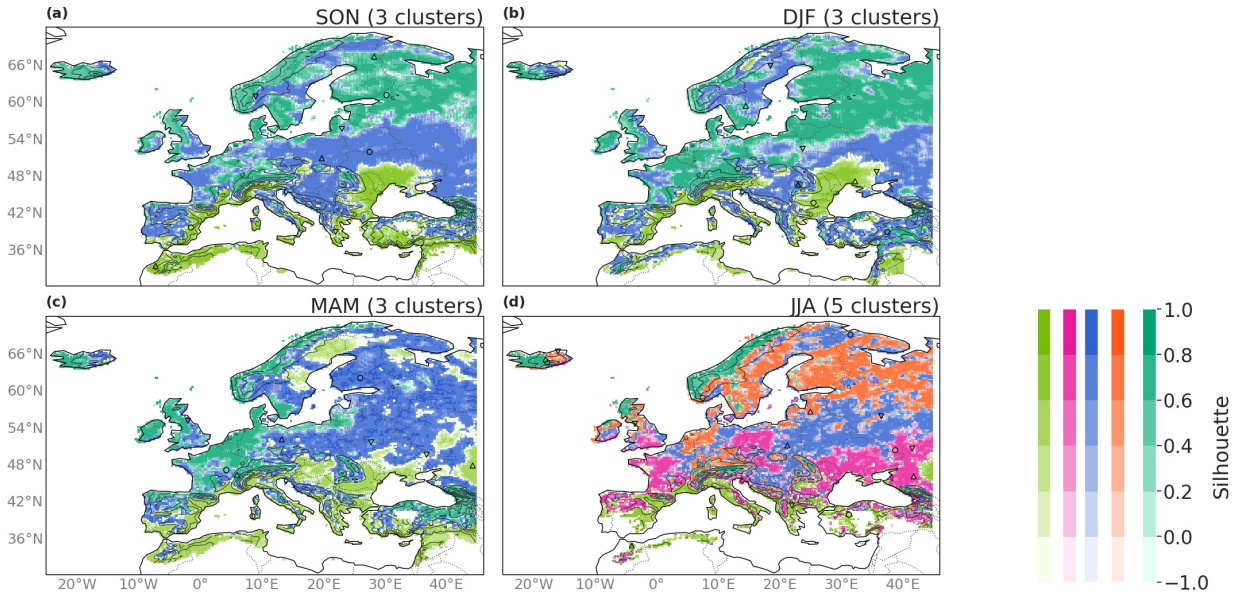


FIGURE 3.1: *Partition with the PAM algorithm applied on ERA-5 daily positive precipitation over Europe for all seasons. Each cluster is identified by a colour. The shades of colour indicate the silhouette coefficient at every grid point. Intense hues indicate a strong association with the cluster. The black lines are 500 [m] altitude isolines of the surface topography in ERA-5. Within a cluster, the circle indicates the location of the medoid, and the triangle pointing up (resp. down) indicates the grid point with the highest (resp. lowest) silhouette coefficient.*

### 3.4.2 Assessment of the fitting

The fitting models are assessed with the Anderson–Darling test, the AIC criterion and QQ-plots. The Anderson–Darling test indicates similar performance for the fitting of the regional and local EGPD models; see Table 3.1. The null hypothesis is that the fitted and the empirical distribution are the same. Table 3.2 displays the non-rejection rates of the null hypothesis for the Anderson–Darling test for each season and model across the entire domain. The null hypothesis is not rejected for 87% of the grid points in JJA and 91% in SON for the local fit. For the regional fit, it is not rejected for 84% of grid points in JJA and 88% in SON, DJF, and MAM. The non-rejection rate for the semiregional fitting is very similar to that the regional fitting. The percentage is lower for the local fit than for the regional fit in all seasons. Nonetheless, the difference between local and regional is smaller than 3% in all seasons. For all seasons and all fittings, the non-rejection rate indicates good performance of the model, the perfect non-rejection rate being 95% on a test with a confidence level of 5%.

The variability of meteorological processes tends to increase with the altitudinal gradient. Around complex topography, local-scale variations in precipitation may occur. Precipitation distributions might differ substantially between grid points, even within a homogeneous region, and the quality of the regional fit might decrease. We therefore distinguished the rejection rate of Anderson–Darling test between grid points below and above 1000 meters above sea level. We did not find any significant difference in the rejection rate of the Anderson–Darling test between grid points at low and high altitudes (not shown). Moreover, the goodness of the classification in the clustering procedure might impact the accuracy of the fit. At a grid point with a poor connection to its cluster, the regional value of  $\xi$  (and  $\kappa$ ) might not be accurate and the distribution fitted regionally might be significantly different from the empirical distribution. We distinguished the Anderson–Darling test between grid points with a silhouette greater or lower than 0.2. Here too, we observe no significant difference between grid points with low and high silhouettes, for either the local or regional fits (not shown). Even if the local model is the most adaptable, the regional model seems to be sufficiently flexible to (i) take into account the local-scale variations caused by complex topography and (ii) compensate for the regionalization of two parameters out of three. The AIC criterion summarizes the information contained in the likelihood and penalizes the number of parameters. It should be as low as possible. The AIC is much lower for the regional model than for the semiregional and local models independent of the season (see Table 3.3). AIC values across all grid points vary between  $-115,106$  in JJA and  $-107,250$  in DJF for the regional fitting, between  $-79,400$  in JJA and  $-67,614$  in DJF for the semiregional fitting and between  $-43,704$  in JJA and  $-27,984$  in SON for the local fitting. These AIC values highlight the trade-off between parsimony and goodness of fit of the regional fitting. Having only one EGP parameter varying within one cluster in the regional model substantially reduces the AIC.

Figure 3.2 displays the QQ plots for cluster medoid, cluster minimum, and cluster maximum silhouette coefficient in each cluster in SON (see partition in Fig. 3.1(a)). For the most centrally located grid point, the medoid, all the fittings perform similarly well. One exception is the upper tail in the northern and southern clusters, which is slightly overestimated with the regional fitting compared to the local one. For the grid point with a minimum silhouette, the regional

Model	SON	DJF	MAM	JJA
local	-30,888	-27,984	-30,016	-43,704
semiregional	-69,792	-67,614	-69,138	-79,400
regional	-108,702	-107,250	-108,266	-115,106

TABLE 3.3: *Akaike information criterion over Europe for each model and season.*

and semiregional fit have a similar performance to the local one or even outperform them in the southern cluster. In the intermediate and southern clusters, for the semiregional and regional fittings only the most extreme precipitation is overestimated compared to the local fit. For the grid point with the maximum silhouette, the regional and semiregional fits outperform the local fit for the whole distribution. Extremes are well captured with the regional method, except in the northern cluster, where the highest precipitation is overestimated for all the fittings. The semiregional and regional fittings seem to significantly improve the quality of estimation for the best-classified points. The semiregional and regional fittings have similar performances. We also compared with the local Bernstein fitting; see Table 3.1. Its performance is similar to the semiregional and regional fittings except in the southern cluster.

### 3.4.3 Return levels

The estimate of the return levels is spatially smooth despite the regionalization of two out of three parameters in the EGPD. Figures 3.3, 3.4 and 3.5 show the 10-, 50-, and 100-year return levels for all seasons. Even though the shape and flexibility parameters  $\xi$  and  $\kappa$  are constant across each cluster, the variability of scale parameter  $\sigma$  (estimated locally) accounts for the high level of spatial detail of the fit. Regions with high return levels are shown in deep blue and purple colours on the map. Specific regions known to experience heavy precipitation are highlighted, such as the Cévennes, South of France (with Cévenols episodes, see e.g. Ducrocq et al., 2008; Vautard et al., 2015) in SON and the Canton of Ticino in southern Switzerland (Isotta et al., 2014; Barton et al., 2016; Panziera et al., 2018) in SON, MAM, and JJA.

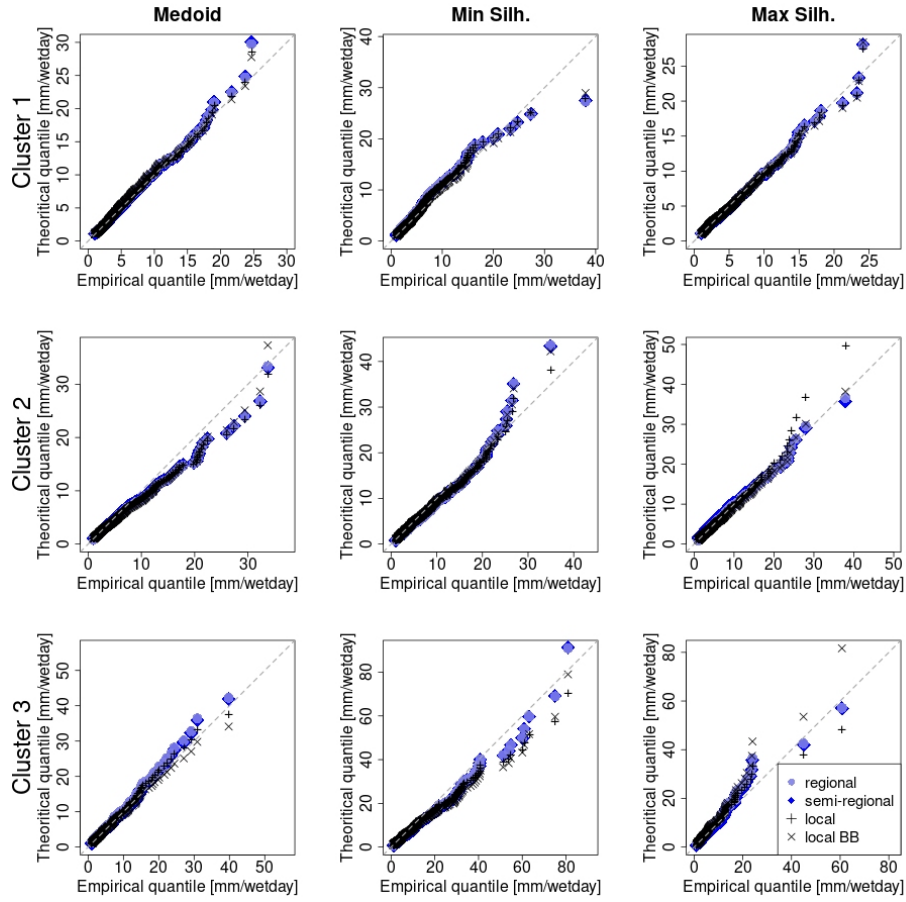


FIGURE 3.2: *Example QQ plots of the regional, semiregional, local, and local Bernstein (local BB) fittings, for the medoid point (left) and the grid points with minimum (middle) and maximum (right) silhouettes in the northern (top row), intermediate (middle row), and southern (bottom row) clusters in SON (blue cluster in Figure 3.1).*

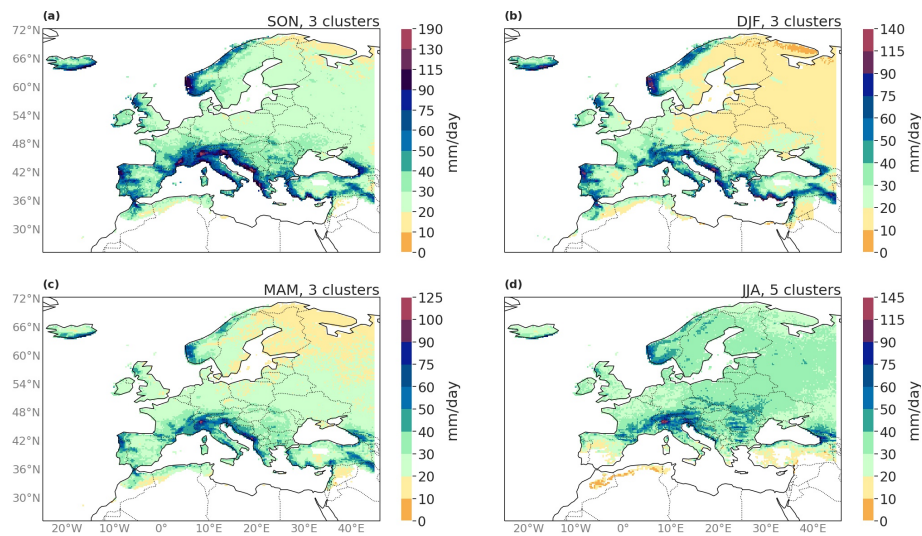


FIGURE 3.3: *10-year return levels computed with the regional fitting, see Table 3.1.*

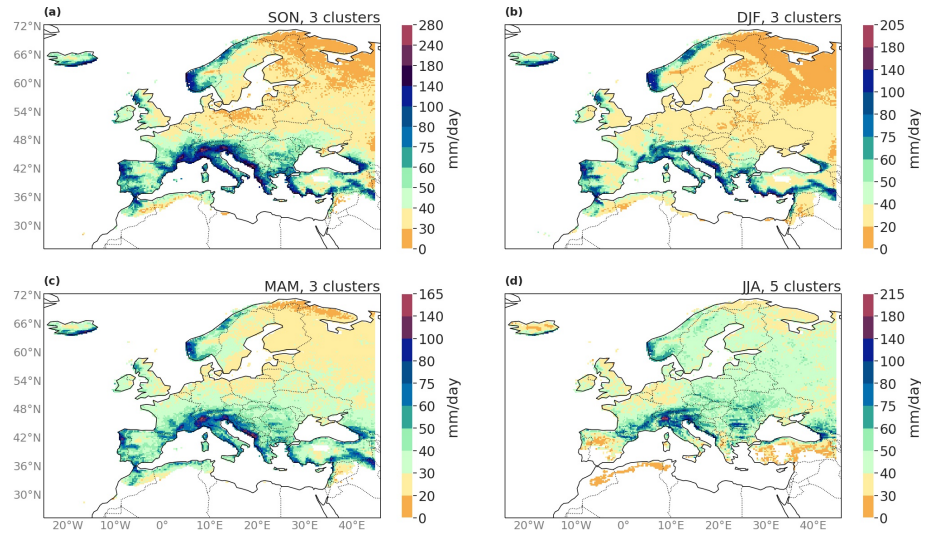


FIGURE 3.4: 50-year return levels computed with the regional fitting, see Table 3.1.

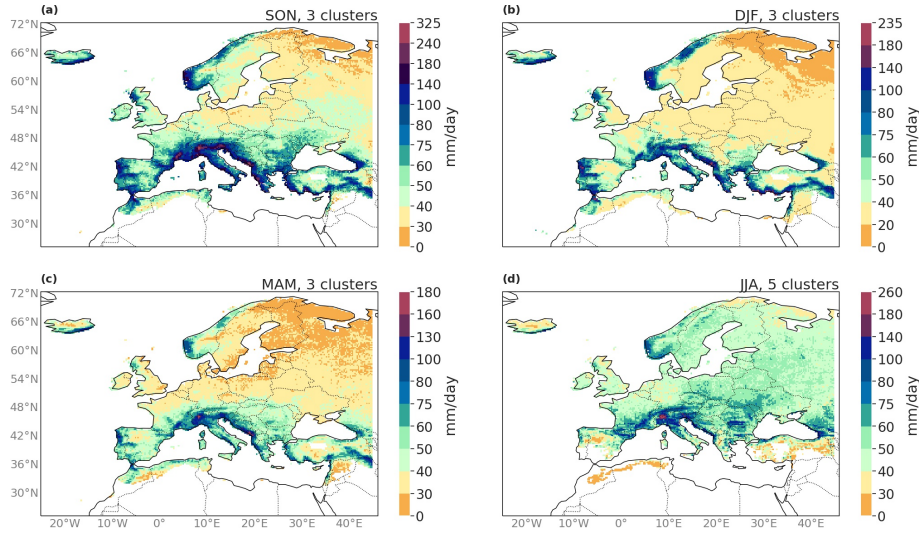


FIGURE 3.5: *100-year return levels computed with the regional fitting, see Table 3.1.*

Finally, we compare the return levels obtained with the local fit and the regional fit. Figure 3.6 displays the relative difference between the 50-year return levels for regional and local fittings. The return levels differ by less than 10% for about 60% of the grid points in SON and for up to 80% of the grid points in DJF. The mean value of the absolute difference lies between 7% (DJF) and 10% (SON). Areas with the highest relative differences are generally located in the cluster with the highest shape parameters: those with more frequent extremes. The same maps for the 10- and 100-year return levels can be found in appendix (Fig. 3.7 and 3.8).

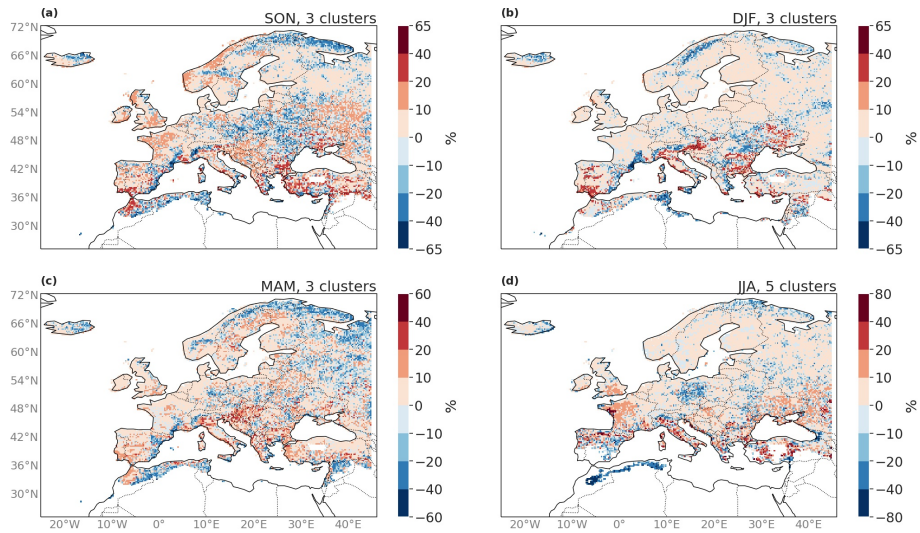


FIGURE 3.6: *Relative difference between the 50-year return levels computed with the regional fitting and the local fitting.*



### 3.5 Discussion

We conduct a PAM clustering procedure based on the PWM ratio  $\omega$ . We find that the optimal number of clusters is three in SON, DJF and MAM, and five in JJA. The higher number of clusters in JJA might be explained by a larger spatial variability of precipitation extremes in Europe in summer (see e.g. Cortesi et al., 2014, in Spain). The same analysis was conducted for hierarchical partitioning, leading to the same results for some parametrization (results available upon request).

The choice of the optimal number of clusters is challenging. The various criteria for the choice proposed in the literature did not agree on the optimal number of clusters (see e.g. Pansera et al., 2013, for the use of two criteria). This can be explained by the large number of grid points in the analysis, resulting in more noise in the criteria than actual information about the goodness of the partitioning. In general, if many grid points are involved, we recommend using more than one criterion and checking the maps visually for the plausibility of the partition obtained.

We analyse the impact of the number of clusters on the regional fit. For this purpose, we compare the difference between the 50-year return levels based on the PAM partition with three clusters and the one with four clusters, in SON (not shown). For a large majority of the grid points, the difference in return levels is lower than 5%. The difference is a bit larger for a few outliers but remains lower than 25%. The outliers are generally located in regions with a very low silhouette score for the partition with 3 clusters.

We compare our partition in Central Europe to that obtained by Gvoždíková et al. (2019) over Germany, Poland, Austria and the Czech Republic. They considered extreme events between 1961 and 2013. Their approach was based on the weather extremity index. They computed Ward's linkage in the hierarchical clustering algorithm. The clusters they found exhibit a West-East pattern. The partition we obtain over these four countries also tends to separate eastern and western regions. Darwish et al. (2021) also found this West-East pattern in the UK. They delineated the regions using the most explanatory covariates (among those that were available) and then assessed their homogeneity by computing tests of Hosking and Wallis (2005) on hourly precipitation. Our results generally agree with the partition they obtained.

The regional model is more parsimonious than the local model; see Table 3.1. It is also more precise on well-classified points (see Figure 3.2). The semiregional and regional models have similar performance; hence, the regional model should be preferred because it is more parsimonious. An alternative to our fitting method would be to select only grid points with a satisfactory silhouette score (e.g. greater than 0.2) to estimate the regional parameters. The quantiles of points with very low silhouette scores would then be estimated locally. This could increase the likelihood of the fitted distribution in some cases but would also increase the number of parameters to fit. However, the performance of the regional fit was not substantially lower than the local fit for the border areas between the clusters, and the rate of rejection in the Anderson–Darling test was not substantially higher at grid points with low silhouettes. For the sake of simplicity and parsimony, we choose to keep the regional approach for all points.

The local Bernstein distributions do not seem to be substantially closer to the empirical distribu-

tion than the regional ones; see Figure 3.2. Hence the flexibility brought by the scale parameter  $\sigma$  in the regional model is sufficient to fit the data well and therefore the most parsimonious model is as precise as the others.

The spatial pattern of our seasonal 10-year return levels (Fig. 3.3) is similar to that of the yearly 10-year return levels obtained by Poschod et al. (2021) with an observational dataset and the Canadian Regional Climate Model.

We also compare the return levels over Switzerland with those provided by MeteoSwiss (2019) for all the seasons (see Fig. 3.9, 3.10 and 3.11 in appendix). This small country provides a good test case for our study because the complex orography leads to a wide variety of precipitation patterns (Schmidli et al., 2002; Umbricht et al., 2013; Isotta et al., 2014; Evin et al., 2018). Return levels obtained by MeteoSwiss (2019) were computed by fitting a GEV to observed seasonal maxima, with a much higher spatial grid resolution than ERA-5 (up to 1km, see MeteoSwiss, 2020). The maps of return levels are close in terms of magnitude and exhibit very similar spatial patterns. Only the small-scale structures are not captured by ERA-5 which is due to the coarser grid resolution of ERA-5. The magnitudes of extremes are slightly underestimated in ERA-5, especially in MAM. This agrees with the study of Hu and Franzke (2020) over Germany. They state that ERA-5 generally underestimates extremes of daily precipitation compared to observation-based gridded datasets (and weather station observations). In our analysis, despite the regionalization (three or five clusters in Switzerland depending on the season) of two parameters out of three, the scale parameter  $\sigma$  presents sufficient variability to have correct return levels. The variability of  $\sigma$  alone is sufficient to provide accurate fitting, even in a country with a complex topography and high local spatial variability of extreme precipitation.

### 3.6 Conclusions

We derive return levels of extreme daily precipitation ( $> 1\text{mm}$ ) over Europe using regionalized parameters for the EGPD fits. The regionalization requires two steps. First, all land grid points are partitioned into a few homogeneous regions with a clustering algorithm. As distance measure, we estimate a scale-invariant ratio of PWM for each grid point, focusing on the tail of the distribution, and then use the PAM clustering algorithm to group these estimates into regions. The second step is the choice and fitting of a model to estimate return levels. We choose to fit an EGP distribution that models the full range of precipitation intensity. Only the scale parameter is allowed to vary within a homogeneous cluster, and the tail and flexibility parameters are common to all grid points in that cluster.

We assessed our regional analysis with classical statistical tools and compared it to previous analyses and return level estimates. Although parsimonious, the regional model is sufficiently flexible to capture the strong spatial variability of rainfall intensities.

This paper provides two main contributions. We provide maps of 10-, 50- and 100-year return levels for European precipitation of ERA-5, and we have made the algorithms for clustering and regional model available in a GitHub repository<sup>1</sup>.

---

<sup>1</sup>[https://github.com/PhilomeneLeGall/RFA\\_regional\\_EGPDk.git](https://github.com/PhilomeneLeGall/RFA_regional_EGPDk.git)



## Author contributions

Pauline Rivoire: Investigation, Conceptualization, Software, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization. Philomène Le Gall: Investigation, Conceptualization, Software, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization. Philippe Naveau: Supervision, Conceptualization, Methodology, Writing - Review & Editing. Olivia Martius: Supervision, Conceptualization, Methodology, Writing - Review & Editing. Anne-Catherine Favre: Supervision, Conceptualization, Methodology, Writing - Writing - Review & Editing.

## Acknowledgments and funding

Within the CDP-Trajectories framework, this work is supported by the French National Research Agency in the framework of the “Investissements d’avenir” program (ANR-15-IDEX-02). P.L. and A.-C.F. gratefully acknowledge financial support for this study provided by the Swiss Federal Office for Environment (FOEN), the Swiss Federal Nuclear Safety Inspectorate (ENSI), the Federal Office for Civil Protection (FOCP), and the Federal Office of Meteorology and Climatology, MeteoSwiss, through the project EXAR (“Evaluation of extreme Flooding Events within the Aare-Rhine hydrological system in Switzerland”).

P.R. and O.M. acknowledge funding from the the Swiss National Science Foundation (grant number 178751).

Part of this work was supported by the DAMOCLES-COST-ACTION on compound events, the French national program (FRAISE-LEFE/INSU and 80 PRIME CNRS-INSU), and the European H2020 XAIDA (Grant agreement ID: 101003469). P.N. also acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project), and the ANR-Melody.

The authors declare that they have no conflict of interest.

## Data availability

We use ERA-5 global total daily precipitation data at resolution  $0.25^\circ$ . Hourly values were downloaded from the ECMWF MARS server (a valid ECMWF account required): <https://apps.ecmwf.int/data-catalogues/era5/?type=fc&class=ea&stream=oper&expver=1> Forecast steps 6 to 17, variable: Total precipitation (228.128). The MARS / EMOSLIB interpolation library has been used.

## 3.7 Appendix

### 3.7.1 Difference between the regional and the local fittings

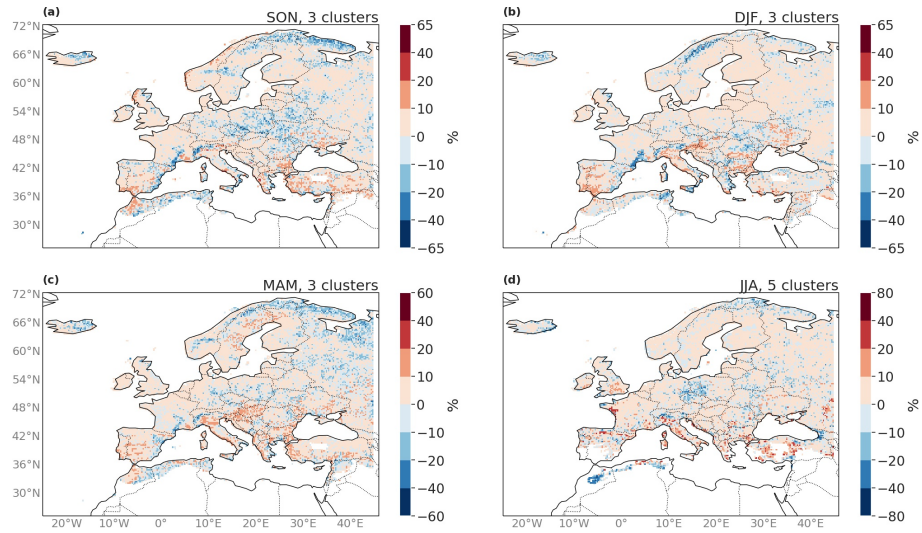


FIGURE 3.7: *Relative difference between the 10-year return levels computed with the regional fitting and the local fitting.*

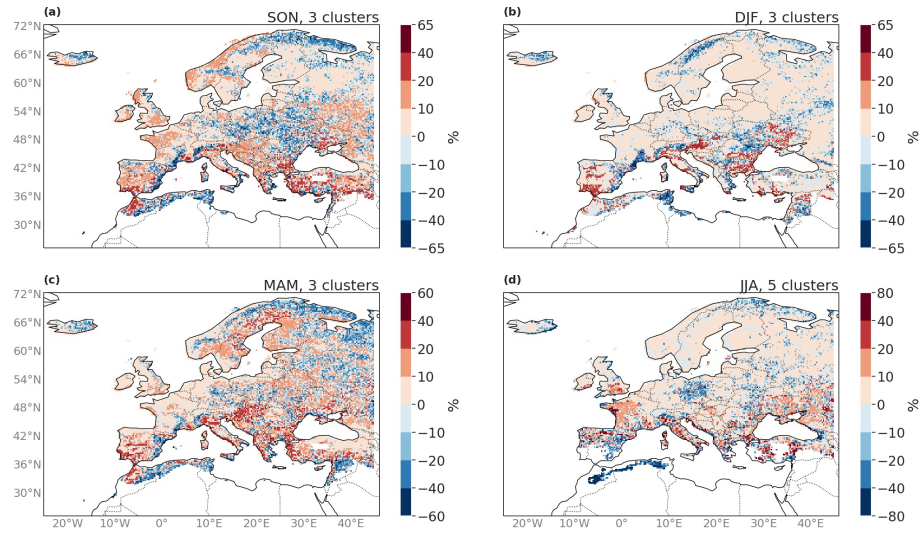


FIGURE 3.8: *Relative difference between the 100-year return levels computed with the regional fitting and the local fitting.*

### 3.7.2 Return levels in Switzerland

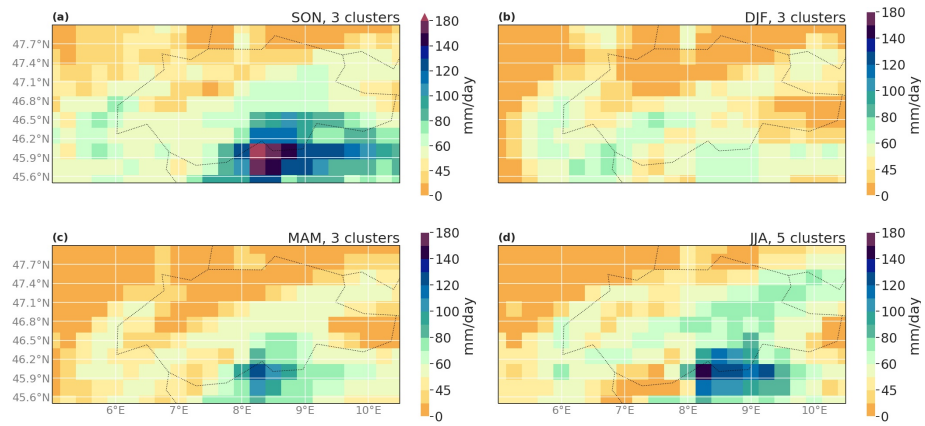


FIGURE 3.9: 10-year return levels in Switzerland computed with the regional fitting, for a comparison with the one provided by MeteoSwiss (2019).

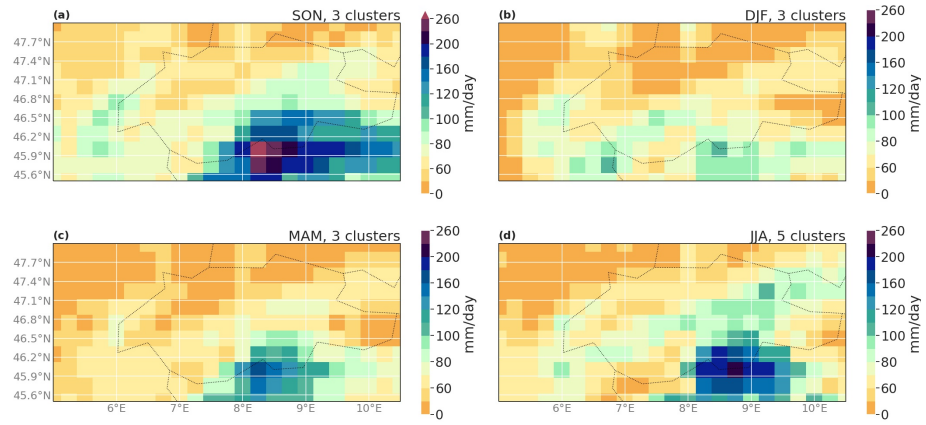


FIGURE 3.10: 50-year return levels in Switzerland computed with the regional fitting, for a comparison with the one provided by MeteoSwiss (2019).

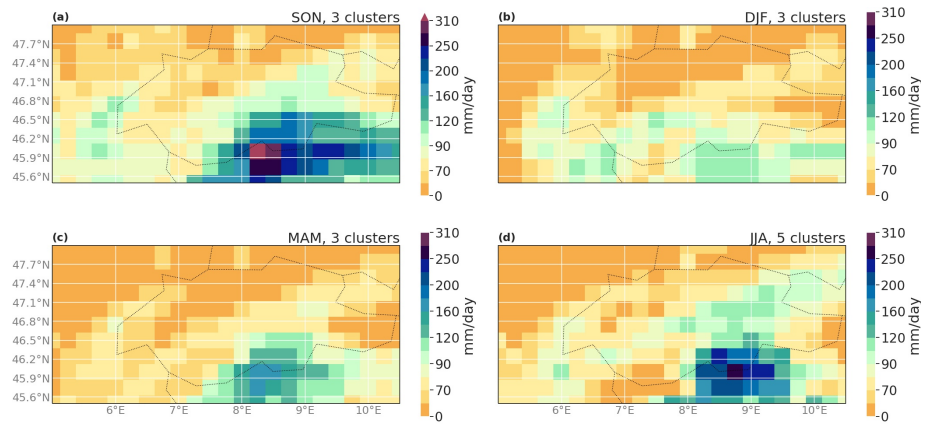


FIGURE 3.11: *100-year return levels in Switzerland computed with the regional fitting, for a comparison with the one provided by MeteoSwiss (2019).*

CHAPTER

4

A VERIFICATION OF EXTREME  
PRECIPITATION OCCURRENCE IN S2S  
FORECASTS OVER EUROPE

## Abstract

Heavy precipitation can lead to floods and landslides, hazards causing large damage and casualties. Some of these impacts can be mitigated if good forecasts and warnings are available. Of particular interest is seasonal to subseasonal (S2S) prediction timescale. It is receiving more and more attention in the research community because of its importance for many sectors. However, very few forecast skill assessments of precipitation extremes in S2S forecast data have been conducted. The goal of this chapter is to provide a methodology to assess the skill of extreme precipitation forecasts on S2S time scales. I apply this methodology to verify extreme precipitation events over Europe in the S2S forecast model from the European Centre for Medium-Range Weather Forecasts. The verification is conducted against ERA-5 reanalysis precipitation. The extreme events are defined as daily precipitation accumulations exceeding the seasonal 95<sup>th</sup> percentile. In addition to the classical Brier score, I use a binary loss function to assess extremes. I analyse daily events locally and spatially aggregated, as well as counts of extremes in a 7-day window. Results consistently show a higher skill in winter compared to summer. The regions showing the highest skill are Norway, Portugal and the south of the Alps. The Mediterranean region has a relatively good skill in winter. The skill is increasing when aggregating the extremes spatially or temporally. The verification methodology can be adapted and applied to other variables, e.g. temperature extremes, river discharge, etc.



## 4.1 Introduction

Extreme precipitation is one of the most relevant weather-related hazard, in terms of human lives, economic impact and number of disasters (see e.g. the impact of storms and flood quantified in WMO, 2021). The mitigation of weather-related hazards depends on the ability to forecast them. It is therefore crucial to improve the predictability of precipitation extremes for a better preparedness (Merz et al., 2020).

Subseasonal-to-seasonal (S2S) forecasting refers to forecasting on timescales from two weeks to a season. S2S prediction has a large range of applications (White et al., 2017, 2021), including the humanitarian sector, public health, energy, water management and agriculture. Forecast skill for this time scale is key to manage natural hazards (Merz et al., 2020). S2S predictions aim to fill the gap between weather forecasts and monthly or seasonal outlooks (White et al., 2017). Providing skilful predictions on subseasonal or monthly timescales is challenging (Hudson et al., 2011). Unlike short-range forecasts and seasonal outlooks that are already operational for many years, the S2S timescale was until recently a “predictability desert” (Vitart et al., 2012). The scientific community working with S2S forecasts has recently expanded (Mariotti et al., 2018; Merryfield et al., 2020; Domeisen et al., 2022). Many research organisations contribute actively to improving S2S forecast skill, as evidenced for example by the *Challenge to improve Sub-seasonal to Seasonal Predictions using Artificial Intelligence* (S2S-challenge, 2021).

Precipitation is a complex variable to predict and S2S forecasts of precipitation extremes have limited skill compared to other types of hazards (see e.g. case study Domeisen et al., 2022; Tian et al., 2017; Endris et al., 2021). The analysis of the S2S precipitation forecast skill could allow to identify regions, seasons and weather regimes with good or bad performance of the forecast. With this information, the user of the datasets can know where and when the forecast model is useful or if it would require improvement (with for example post processing Specq and Batté, 2020). This information is also useful to identify potential sources of prediction and windows of opportunity (i.e. intermittent time periods with higher skill, Mariotti et al., 2020). Most of the existing research on S2S prediction of precipitation extremes is focusing on North America (Zhang et al., 2021a; DeFlorio et al., 2019), Africa (de Andrade et al., 2021; Olaniyan et al., 2018) and Asia (Yan et al., 2021; Li et al., 2019). However little is known about the skill of S2S extreme precipitation prediction over Europe (Monhart et al., 2018; Domeisen et al., 2022). This chapter aims to fill this gap.

A simple analysis of the impact of the North Atlantic Oscillation (NAO, Kenyon and Hegerl, 2010) on the forecast skill is also proposed. The NAO characterizes the atmospheric pressure fluctuations over the North Atlantic ocean through the difference of atmospheric pressure at sea level (SLP) between the Icelandic Low and the Azores High (Hurrell et al., 2003). NAO presents temporal persistence which is a source of forecast skill, therefore the influence of NAO on the prediction skill is investigated. Moreover, NAO can have a large influence on extreme precipitation occurrence, by modulating the moisture transport (Tabari and Willems, 2018). I therefore analyse the influence of the NAO phases on the forecast skill.

An ensemble forecast is a forecast consisting of several equally probable members, i.e. runs of

the numerical model with slightly different initial conditions (WCS, 2021). Metrics to evaluate the bias and the accuracy of precipitation ensemble forecast include the mean absolute error, the probability integral transform, the interquantile range, the continuous ranked probability score (CRPS) (Hersbach, 2000; Gneiting et al., 2007; Crochemore et al., 2016; Monhart et al., 2018), the Brier score (Brier, 1950), and the mean square skill score (Specq and Batté, 2020). These metrics capture the mean behavior of precipitation: most of them are not directly suitable for a verification focused on extremes. The CRPS can be adapted for a focus on extremes, with the threshold-weighted CRPS (Allen et al., 2021, 2022). Another option is the relative operating characteristic (ROC): it can be used to measure the ability of the forecast to identify above normal precipitation events (Domeisen et al., 2022; Monhart et al., 2018). The quantile probability of detection and quantile false alarm ratio are two complementary metrics to assess binary extreme events (Zhang et al., 2021a), however these metrics are design for deterministic forecasts (only one forecast run). In this study, I transform precipitation extremes into binary “events” and assess the extreme events with the classical Brier score (Brier, 1950). In addition, I propose a simple extension of the binary loss score as introduced by Legrand et al. (2021) to ensemble forecasts. This metric considers only the case of the occurrence of an extreme event in the forecast or in the observation or in both but not the non-events (see section 4.2.3.2).

Hindcasts are used to assess the forecast skill. Hindcasts are forecasts run for past dates over sufficiently long time periods (20 years typically) to assess the quality of the forecast and to identify and correct model biases (Manrique-Suñen et al., 2020). The goal of this chapter is to quantify S2S forecast skill for extreme precipitation events over Europe with the hindcast data from the European Centre for Medium-Range Weather Forecasts (ECMWF, Vitart, 2020), one of the most used and most skilful S2S modeling systems (de Andrade et al., 2019; Li et al., 2019; Stan et al., 2022; Domeisen et al., 2022).

This chapter is structured as follows. Section 4.2 contains a description of the forecast and verification data and the methods, including the Brier score (Brier, 1950) and the BLF (Legrand et al., 2021). I present the results of the analysis in Section 4.3. I discuss these results, draw conclusions and give an outlook in Section 4.4.

## 4.2 Data and Methods

### 4.2.1 Data

The ensemble forecast data corresponds to ECMWF S2S precipitation hindcast data (Vitart, 2020; ECMWF, 2022a, cycle 47r2). It is composed of 11 ensemble members, initialized twice a week, run for a leadtime of 46 days. I focus on Europe, in the spatial box  $[30^{\circ}\text{N} ; 72^{\circ}\text{N}] \times [-15^{\circ}\text{E} ; 49.5^{\circ}\text{E}]$ . The hindcast period is 20 years with 2080 initialisations between 2001-01-04 and 2020-12-30 (twice a week, on Monday and Thursday). The data were downloaded at the model spectral resolution “O320” (ECMWF, 2022b,c). For the analysis, the data was interpolated to a  $0.5^{\circ} \times 0.5^{\circ}$  grid using a first-order conservative remapping (Jones, 1999; CDO, 2018). Note that the surface represented by one grid point depends on the latitude: the longitudinal extent of a grid point is always  $0.5^{\circ}$ , but the width in km depends on the latitude (it is about 55km at  $30^{\circ}\text{N}$

and about 18km at 72°N). The latitudinal extent of a grid point is constant (0.5°, about 55km). ERA-5 precipitation (Hersbach et al., 2019) is here the verification dataset. The choice of a reanalysis dataset is motivated by the regular grid and temporal availability; it also avoids the uncertainties due to the inherent spatial sparsity of weather station networks (Hofstra et al., 2009; Rivoire et al., 2021b). Daily precipitation are extracted over the same time period, from 2001-01-04 to 2020-12-30 plus 46 leadtime days i.e. 2021-02-14, with a spatial resolution of 0.5°×0.5°. In the rest of the chapter, for the sake of simplicity, “observation” refers to ERA-5.

### 4.2.2 Definition of extreme events

I define precipitation extremes as daily binary exceedances over a threshold, the 95<sup>th</sup> seasonal dry percentile (i.e. over all days in March-April-May –MAM–, June-July-August –JJA–, September-October-November –SON– or December-January-February –DJF–). Figure 4.8 in appendix shows the 95<sup>th</sup> percentile in ERA-5 in Europe, for the period from 2001-01-04 to 2021-02-14. The hindcast percentiles are computed seasonally and leadtime-dependent: for a given leadtime day and a given season, the 95<sup>th</sup> percentile is computed on daily precipitation of all the ensemble members pooled together. Figure 4.1 shows the bias of this percentile compared to the ERA-5 observations for four different leadtimes. In this figure and all the following ones, only values at grid points where the 95<sup>th</sup> percentile in the observations is greater than 5mm per day are shown. The sign and the magnitude of the forecast bias are season and leadtime dependent. For a leadtime of one day, the forecast generally underestimates the 95<sup>th</sup> percentile. For leadtimes between 2 and 46 days, some regions have a positive bias (central Europe in spring and summer) and some have a negative bias (the Alps in summer, autumn and winter, Norway in spring, autumn and winter) (Figure 4.1). Generally over Europe, the bias depends on the leadtime and on the season.

### 4.2.3 Metrics

I define extreme events as threshold exceedances and I use the Brier score and the BLF to assess the forecast skill. Except for the weather regime analysis, I compute the Brier score and the BLF for the extended winter (NDJFMA, i.e. November-April) and extended summer (MJJASO, i.e. March-October). When defining the extremes (see previous section), the seasonal dependence was based on usual 3-month long seasons, because of the strong dependence of the intensity distribution on the 3-month season (see figure 4.8). The choice of extended seasons for the skill analysis is a trade-off between having enough extreme events for a robust analysis and observing the impact of the season on the performance of the forecast. As consequence the probability of the extreme events is no longer be exactly 0.05 if non-stationarities within the SON and MAM seasons exist.

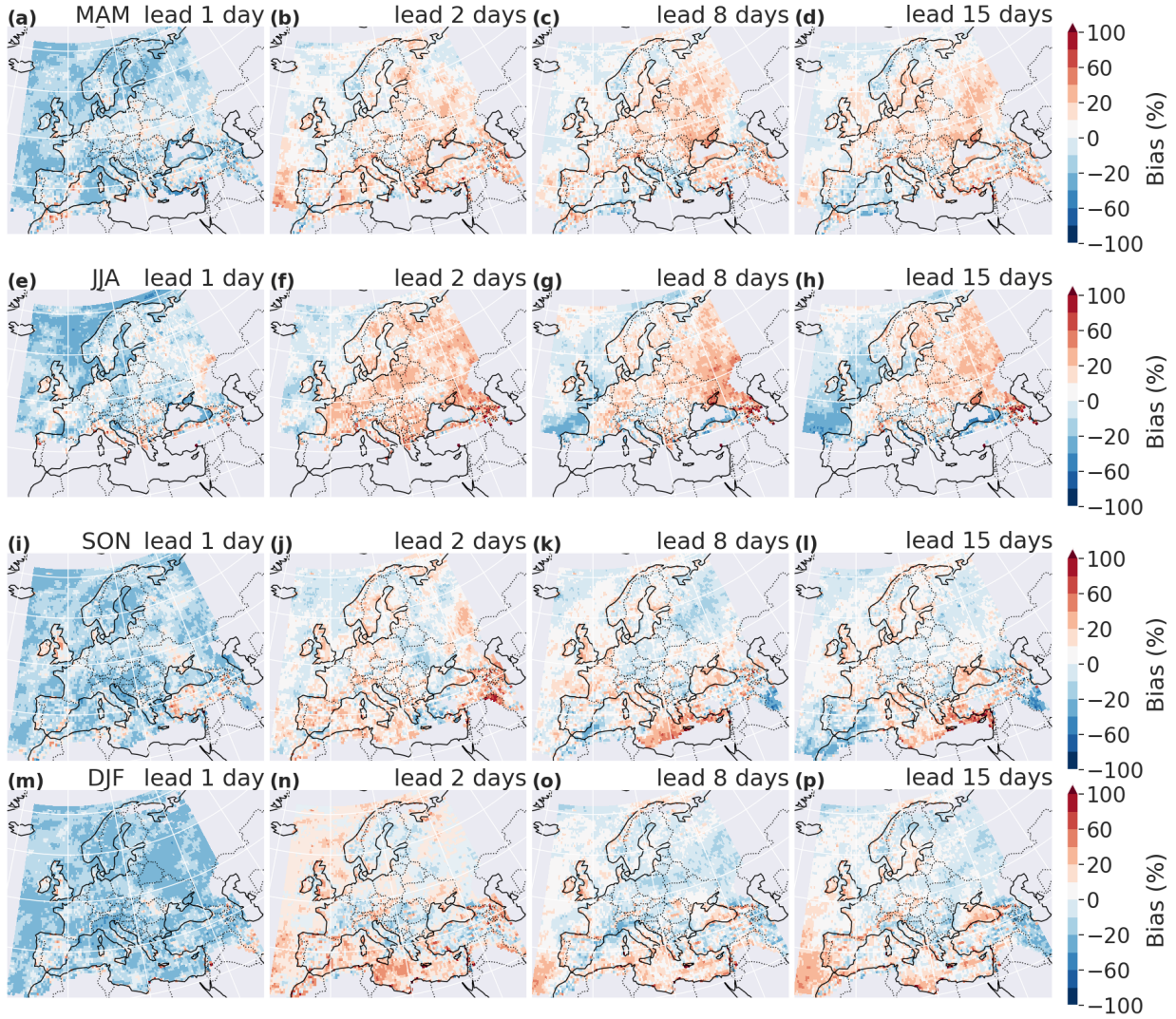


FIGURE 4.1: *Bias of the forecast compared to the observation for the 95<sup>th</sup> percentile, for spring (MAM, (a)-(d)), summer (JJA, (e)-(h)), autumn (SON, (i)-(l)) and winter (DJF, (m)-(p)) for the leadtime day 1 (first column), day 2 (second column), day 8 (third column) and day 15 (last column). Only grid points with a 95<sup>th</sup> percentile greater than 5mm per day are displayed.*

#### 4.2.3.1 Brier Score

The Brier score  $B$  is defined as the mean square difference between forecast probability and binary observation (Brier, 1950):

$$B = \frac{1}{n_D} \sum_{i=1}^{n_D} (f_i - O_i)^2,$$

where  $n_D$  is the total number of days (about 1040 per leadtime, i.e. the number of initialisations in the given extended season);  $O_i$  the binary observation of extreme for day  $i$  ( $O_i=1$  if the daily precipitation exceeds the 95<sup>th</sup> and  $O_i = 0$  otherwise);  $f_i$  is the forecast probability of extreme occurrence for day  $i$ , i.e. the mean of the ensemble members:  $f_i = \frac{1}{M} \sum_{m=1}^M F_{(i,m)}$ , with  $M$  the number of ensemble members (here  $M = 11$ ) and  $F_{(i,m)}$  the binary forecast for a given ensemble member  $m$  for day  $i$ .

$B$  is negatively oriented (the lower the better). The climatological Brier score  $B_{\text{clim}}$  is used as a reference value:

$$B_{\text{clim}} = \frac{1}{n_D} \sum_{i=1}^{n_D} (p - O_i)^2,$$

where  $p$  is the climatological probability of having an extreme. Note that the value of this probability is not exactly 0.05, as two of the 3-month seasons are splitted to form the extended seasons.  $p$  is therefore computed empirically.

The forecast is skilful if its Brier score lower than the climatological Brier score. These scores can be compared using the Brier Skill Score ( $BSS$ ):

$$BSS = 1 - \frac{B_{\text{hind}}}{B_{\text{clim}}}.$$

The  $BSS$  varies between  $]-\infty; 1]$  and is positively oriented (the closer to one, the better). For a given leadtime day, a forecast has skill if  $BSS > 0$ . From here on, the expression “the last skilful day” refers to the largest leadtime day with skill.

The  $BSS$  has limitations because of the unbalanced categories in our case. The extreme events dataset is composed of 95% zeros and 5% ones. A large part of the forecast and observation datasets are matching because of the large presence of “0s” (daily precipitation lower than the 95<sup>th</sup>) in both datasets. To address this issue a binary loss index focusing on extremes (“1s”) is also used.

#### 4.2.3.2 Binary loss function

I define a binary loss index inspired by the binary loss score defined in Legrand et al. (2021). This binary loss index is the ratio between the empirical probability of having an extreme event in either the observation dataset or the forecast dataset, and the empirical probability of having an extreme event in the observations or the forecast (including having an event in both datasets). This index was initially designed for deterministic forecasts. It is here adapted for ensemble forecasts with a dependence on the ensemble member (the interpretation is based on the median

of the members, see the definition of a skilful score below). This adapted index is later on called the binary loss function (*BLF*) and is defined by:

$$BLF_m = \frac{\frac{N_{(m,1)}}{n_D}}{\frac{N_{(m,2)}}{n_D}} \quad \text{i.e.} \quad BLF_m = \frac{N_{(m,1)}}{N_{(m,2)}}$$

where  $m$  is one forecast ensemble member,  $n_D$  is the total number of days,  $N_{(m,1)}$  is the number of days where an extreme event is either observed or predicted by the member  $m$ , i.e.  $N_{(m,1)} = \#\{j \mid F_{(j,m)} \neq O_j\}$  and  $N_{(m,2)}$  is the number of days where an extreme event is observed or predicted by the forecast member  $m$  (including having an event in both datasets), i.e.  $N_{(m,2)} = \#\{j \mid (F_{(j,m)} = 1 \text{ or } O_j = 1)\}$ . In other words,  $N_{(m,1)}$  is the number of false positives and false negatives and  $N_{(m,2)}$  is the number of true positives, false positives and false negatives.

This way, the focus is only on measuring the matching between “1s” (day with extreme event) in the datasets. The case “neither the forecast nor the observation experience an extreme event” is not taken into account in the metric.  $1-BLF$  can be understood as the critical success index (Schaefer, 1990; Legrand et al., 2021). To measure the leadtime dependence of the skill,  $BLF_{(m)}$  is computed for each leadtime day.

$BLF_{(m)}$  varies between  $[0; 1]$  and is negatively oriented (the closer to zero, the better). Note that if the forecast  $F$  and the observation  $O$  are independent (i.e. a forecast with no skill) and if  $\mathbb{P}[F = 1] = \mathbb{P}[O = 1]$ , the  $BLF = \frac{2-2\alpha}{2-\alpha}$ , where  $\alpha = \mathbb{P}[F = 1] = \mathbb{P}[O = 1]$  ( $\alpha = 0.05$  for daily exceedances in a given season). In our case,  $\mathbb{P}[F = 1]$  is not exactly equal to  $\mathbb{P}[O = 1]$  because the index is computed on extended seasons and not on 3-month seasons.

The climatological value (i.e. the no-skill value) of  $BLF$ , referred to as  $BLF_{clim}$ , is also used as reference value. The confidence interval on  $BLF_{clim}$  is computed to determine if the forecast is skilful, i.e. if  $BLF$  is significantly lower than  $BLF_{clim}$ . The confidence intervals are obtained with a bootstrap procedure. For a given bootstrap step, a random time series is formed by drawing values in the observation time series. The  $BLF$  is computed with this random time series as forecast. For a given leadtime day, a forecast is said significantly skilful if the more than 50% of the members are lower than the 5% percentile of the confidence interval on  $BLF_{clim}$ . The expression “last skilful day” introduced for the Brier score is also for the BLF, with the same definition (largest leadtime day with skill, see figure 4.2 for an example).

#### 4.2.3.3 Spatio-temporal extension of the metrics and weather-regime dependence

Asking for an exact match of events in the forecast and the observations on the same day and at the same grid point is very strict. Indeed, precipitation is a complex variable to forecast. Moreover, daily precipitation is defined as cumulative precipitation over 24 hours from 00:00 UTC: given a precipitation extreme observed on day  $D$  in the evening, if the forecast accurately catch the accumulated precipitation but with a time delay of few hours, the extreme event would be divided into two non-extreme events on day  $D$  and  $D+1$  in the forecast. A forecast may contain useful information, even if the forecast does not predict the event exactly on the same day or at the same location as in the observation, but in a close neighborhood. The skill scores are

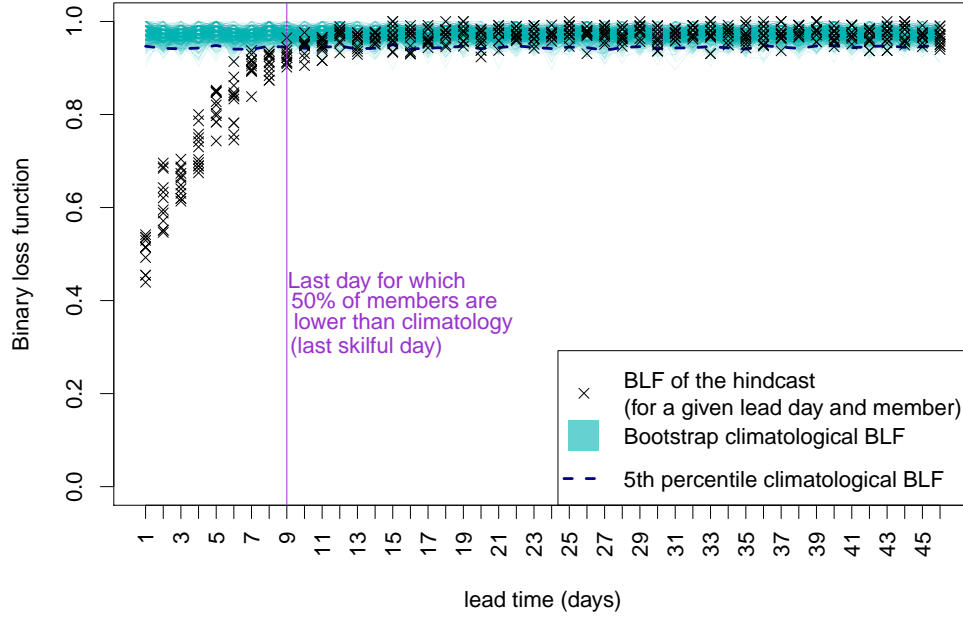


FIGURE 4.2: *Definition of the last skilful day for BLF: example for one grid point and one season.*

therefore additionally computed with temporal and spatial aggregation windows that allow for some flexibility in the exact location or exact timing of the events. The spatial and the temporal aggregations are conducted independently, to analyse the individual impact of each aggregation.

For the temporal aggregation, the number of extreme events in a 7-days window are counted. The forecast skill is measured with the Brier score and *BLF* by defining binary categories  $E_n^t$ , as a function of the number of events  $n$  in the window, for  $n = \{1, \dots, 7\}$ . Either the number  $N^t$  of extreme events in the time window is lower than  $n$  ( $E_n^t = 0$ ) or  $N^t$  is greater or equal to  $n$  ( $E_n^t = 1$ ) :

$$E_n^t = \begin{cases} 1 & \text{if } N^t \geq n \\ 0 & \text{otherwise.} \end{cases}$$

Figure 4.3 provides an example for the definition of  $E_n^t$ . For a given number of events  $n$ , both observation and forecast data are again reduced to binary data ( $E_n^t$ ), therefore the Brier score and *BLF* can be used to quantify the forecast skill. The climatological values are computed in a way to conserve the temporal structure present in the climatology: there is randomness in the bootstrap only when choosing a beginning of time window in the observation. The 6 following days are not random but the 6 following days in the time series of observations. These temporal counts of extremes allow a simple evaluation of the ability of the forecasts to capture the temporal clustering of extreme events

The spatial aggregation is performed by counting extreme precipitation events in neighborhoods. Like for the temporal aggregation, I define two categories, depending on the count of events  $N^s$  in

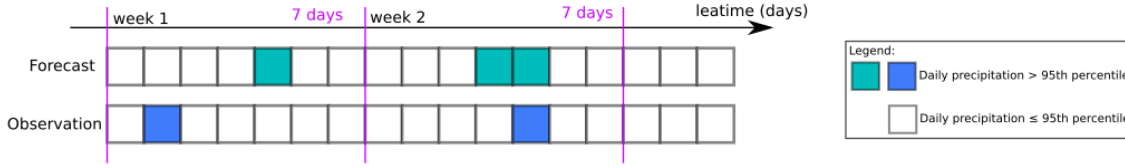


FIGURE 4.3: *Illustration of the weekly aggregation of extremes at one grid point. During week 1, the forecast predicts one extreme and one extreme is observed. For both the forecast and the observation, the number of extreme events in the 7day window is greater or equal to 1:  $E_1^t = 1$  for the two datasets. For both datasets, the number of events in the 7day window is lower than  $n$ , for  $n \geq 2$ :  $E_n^t = 0$  for the two datasets. During week 2, one extreme is observed but the forecast predicts two events. For both datasets, the number of extreme events in the 7day window is greater or equal to 1:  $E_1^t = 1$  for the two dataset. For the observation, the number of events in the 7day window is lower than 2 ( $E_2^t = 0$ ) and this number is greater or equal to 2 for the forecast ( $E_2^t = 1$ ). For the configuration with  $n$  events,  $n \geq 3$ ,  $E_n^t = 0$  for both datasets.*

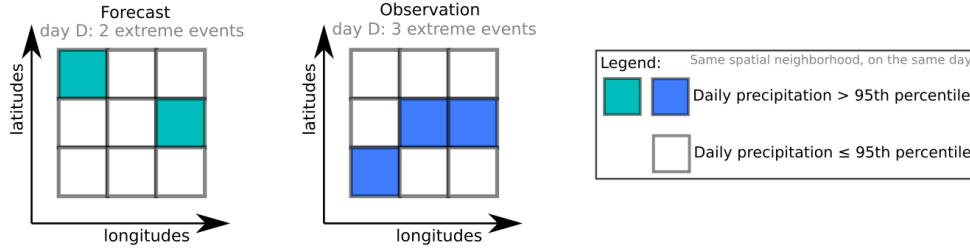


FIGURE 4.4: *Illustration of the spatial aggregation of extremes in one neighborhood. The forecast indicates two extremes in the spatial neighborhood and three events are observed. For both datasets, the number of extreme events in the neighborhood is greater or equal to 1 ( $E_1^s = 1$ ) and greater or equal to 2 ( $E_2^s = 1$ ). For 3 events or more,  $E_3^s = 0$  for the forecast and  $E_3^s = 1$  for the observation. For four events or more,  $E_n^s = 0$  for both datasets for  $n \geq 4$ .*

the spatial neighborhood being greater or lower than a number  $n$  (see figure 4.4 for an example):

$$E_n^s = \begin{cases} 1 & \text{if } N^s \geq n \\ 0 & \text{otherwise.} \end{cases}$$

Precipitation includes some spatial structure, i.e. spatial dependence between points in a neighborhood. When computing the climatology for both scores, the spatial structure is conserved: for one step of the bootstrap only the date is randomly chosen, the spatial neighborhood is the observed neighborhood for that day. I define the neighborhoods as square boxes of about 150km\*150km, i.e. boxes with a latitudinal extent of 1.5°N (3 grid points) and with a longitudinal grid extent that depends on the latitude: from 1.5°E at 30°N (3 gridboxes) to 4.5°E at 70°N (9 gridboxes), see figure 4.11 in appendix for an illustration. Note that both the spatial and temporal neighborhood are non-overlapping, to consider only once a given extreme event.

I additionally investigated the effect of European weather regimes on the forecast skill (as defined



---

in Grams et al., 2017). The forecast skill is computed independently for positive phases and negative phases of the NAO. During positive phases of the NAO, heavy precipitation is reduced in southern and increased in northern Europe (and the opposite configuration in negative phases of the NAO, Haylock and Goodess, 2004; Kenyon and Hegerl, 2010). NAO being a large-scale circulation pattern, I expect the S2S forecasts to have more skill when predicting large-scale patterns than local precipitation extremes.

## 4.3 Results

### 4.3.1 Daily and local comparison

The Brier skill score indicates more skill during the extended winter (skill for up to 12 days, and many regions with a last skilful day greater than 10 days) than during the extended summer (last skilful day below 8 days for most grid points), see figure 4.5. Regions with high skill are Norway, the area South of the Alps and Portugal in the extended winter and the Atlantic coast of France, the South of France, the South of Norway, Central Europe and the South of the Alps in the extended summer. The BLF confirm these patterns, see figure 4.9 in appendix.

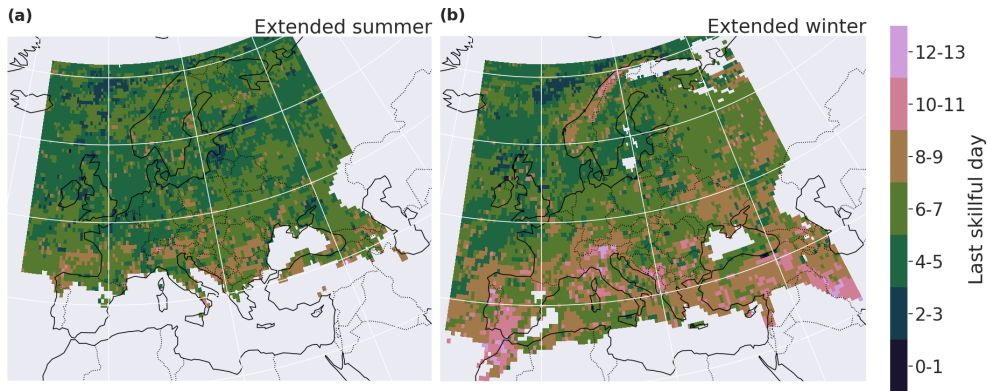


FIGURE 4.5: *Last skilful day for the Brier skill score for local and daily comparison, in extended summer (a) and extended winter (b).*

### 4.3.2 Temporal aggregation

The count of heavy precipitation events over 7 days also exhibits a better score in the extended winter than the extended summer (figure 4.6). For the category “one event or more occurred during the 7 days”, the forecasts at most grid points still have skill during the second week leadtime, i.e. days 8 -14, for both extended seasons. The forecasts for Portugal, Norway, Greece, the Atlantic near France and Portugal even exhibit predictability for the third week leadtime (days 15 -21). The BS is decreasing with increasing number of events per week, however the spatial patterns are constant. The regions showing some skill to capture temporal clustering are Portugal, Norway (especially in winter). The BLF confirms these results, with similar patterns (see figure 4.10 in appendix).

### 4.3.3 Spatial aggregation

When counting extreme events in a spatial neighborhood, the forecast in extended winter also has a longer skilful lead period compared to summer (see figure 4.7). In extended winter for one event or more in the neighborhood, the last skilful leadtime can be up to 13 days in many regions: the western Iberic peninsula, northern Norway, the Atlantic coast of Spain, and the westerly oriented coasts in general. In extended summer, the Atlantic coast of France, Italy, western Europe and

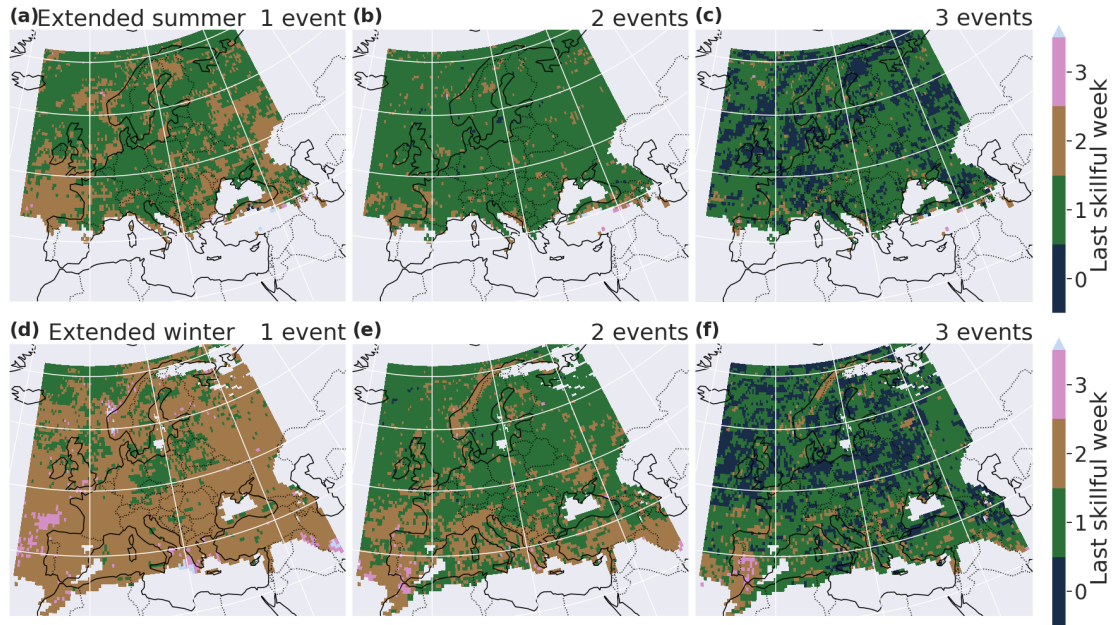


FIGURE 4.6: *Last week of skill for the Brier skill score in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in a 7 days window.*

the coasts of the Iberic peninsula have between 8 and 11 skilful leadtime days. The spatial pattern of the skill remains constant with increase number of events per neighborhood.

Figure 4.12 in appendix shows maps of the last leadtime day with a positive BLF, for different numbers of events aggregated spatially, in extended summer and extended winter. The regions with higher skill are the same for the Brier score and for the BLF. The spatial pattern of the skill also remains constant with increasing number of events per neighborhood.

#### 4.3.4 Dependence on weather regimes

The forecast skill does not seem to have a strong dependence on the NAO phase, although the data was also spatially aggregated to increase robustness (see figures 4.13 and 4.14 in appendix). Forecasts for northern Norway have a higher skill during positive NAO phases than during negative phases, and forecasts for southern Europe have higher skill in negative phases of the NAO. Apart from these regions, no clear patterns appear and the results are more noisy than for the seasonal analysis. In general, in the negative phase of the NAO there is higher skill than in the positive phase and in the positive phase there is a larger latitudinal gradient of skill. However these results are not robust, see figure 4.15 in appendix. The zonal mean of the skill length does not have a stronger gradient in one of the NAO phases when compared to the skill in the extended seasons.

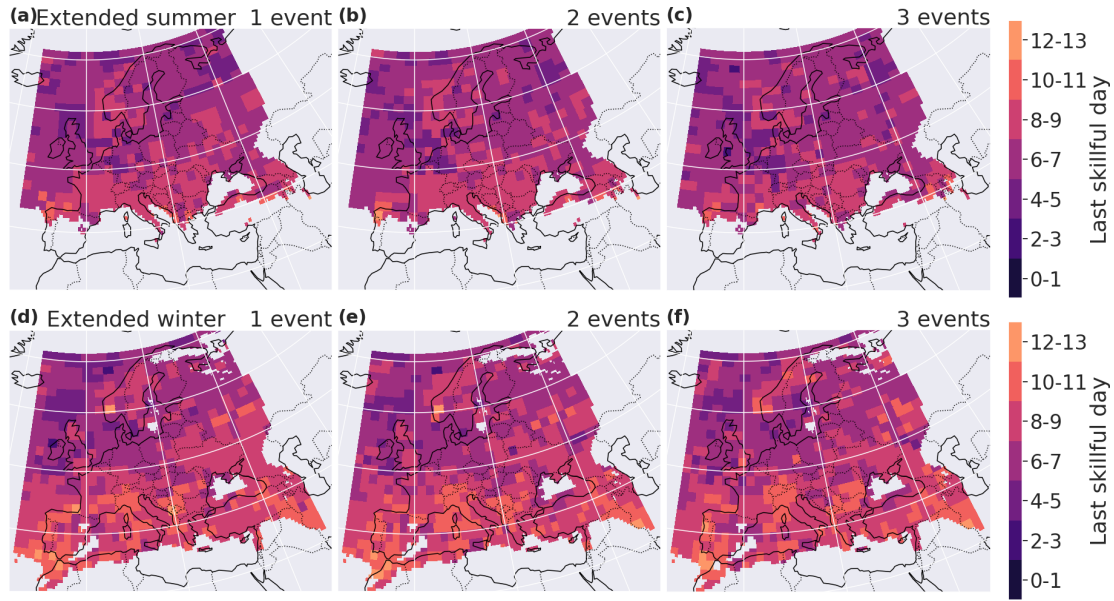


FIGURE 4.7: Last day of skill for the Brier skill score in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in neighborhood of  $150 \times 150 \text{ km}$ .

## 4.4 Discussion and Conclusion

In this chapter, I provide an assessment of the forecast skill of extreme precipitation occurrence over Europe in the ECMWF S2S model. Extremes are defined as exceedances over the seasonal 95<sup>th</sup> percentile. I conduct a verification against ERA-5 precipitation with the Brier score and a binary loss function (*BLF*). The binary loss score is a metric for deterministic forecasts that has the advantage of focusing on extreme event occurrence (hit, false alarm or miss). The index *BLF* proposed here is a simple adaptation of the binary loss score to ensemble forecasts. The *BLF* is qualitatively compared with the Brier score; the skill scores of two metrics agree very well over Europe.

The S2S forecasts have in general a higher skill in predicting extreme precipitation events during winter than during summer. Winter precipitation over Europe is mainly driven by large scale processes, whereas smaller scale, convective events often drive precipitation in summer; the prediction of small scale events is more challenging than large scale ones (Haylock and Goodess, 2004; Kenyon and Hegerl, 2010). This result is in agreement with the existing literature on S2S prediction in other regions (Tian et al., 2017; Kolachian and Saghaian, 2019). Norway, Portugal and the South of the Alps are regions with the larger skill. The orography seems to be a source of skill (like in Norway, the Pyrenees and the South of the Alps): the forecast seems to better capture precipitation events where the complex topography acts as a forcing for precipitation. The Mediterranean region exhibits relatively good skill in winter, along the storm track. Similarly, the coastal regions in general have a higher skill. The water transported from the ocean first rains out next to the coast; a potential explanation is that the model might have more difficulties

to predict where the remaining water in the atmosphere will rain down on continental regions because of an accumulation of uncertainty. Allowing for temporal or spatial flexibility in the evaluation of the forecast extremes confirms the skill patterns, bringing robustness to the analysis. The spatial aggregation conducted here could be adapted for an impact-oriented analysis, by aggregating e.g. over catchments to evaluate the predictability of heavy precipitation, that can potentially result in floods.

A first simple analysis did not highlight a clear link between the forecast skill and the NAO regime. However, this absence of signal should be confirmed with a deeper analysis, by considering some time lag or seasonality for the influence of the teleconnection patterns (Tabari and Willems, 2018) or by aggregating over larger spatio-temporal neighborhoods, to increase the robustness. Other teleconnection patterns could be investigated, such as Scandinavian and East Atlantic patterns, El Niño southern oscillation, the Atlantic multidecadal oscillation (Casanueva et al., 2014) or the state of the stratosphere (Domeisen et al., 2019).

The goal of this chapter is to assess the forecast skill of precipitation extremes for the ECMWF S2S model over Europe. Two independent metrics were used and produced the same spatial features of forecast skill, confirming the robustness of the signal. The BLF used in this dissertation is a simple index. With further research, a probability score for ensemble forecasts could be developed from this index. Moreover, some simulations should be run to conduct a quantitative comparison of the two metrics. By simulating time series of extremes for both the observation and the forecast, one could interpret the difference between the Brier Score and the BLF. An assessment focused on the intensity would also deepen the verification; the precipitation forecast data would then require to be calibrated (Gneiting et al., 2007; Specq and Batté, 2020; Crochemore et al., 2016; Monhart et al., 2018; Huang et al., 2022). The threshold weighted continuous ranked probability score would be an option to measure the intensity forecast skill with a focus on heavy precipitation (see e.g. Pantillon et al., 2018; Allen et al., 2021). Analysing the paradigm of “maximizing the sharpness of the predictive distributions subject to calibration” could also be an extension of this work (Gneiting et al., 2007); the usual evaluation metrics –the probability integral transform histogram, marginal calibration plots, the sharpness diagram– could be applied with a focus on extremes.

As the Brier score can be viewed as a binary version of the more general CRPS, the reader interested in the mathematical properties of such scores could look at Pic et al. (2022). In particular, the limitation of such scores and their weighted versions in terms of scoring extremes events have been pointed by various authors, see e.g. Taillardat et al. (2022). In contrast, the BFL studied by Legrand et al. (2021) offers theoretical mathematical guarantees to judge a forecast with respect to an extreme occurrence.

Note that for practical applications, one needs caution to interpret the skill in an absolute way: a skilful forecast does not mean that the forecast is also useful forecast for practical applications. If the BLF is equal to 0.8 but is outside of climatological confidence interval, the forecast is better than the climatology and therefore skilful. However, it also means that only 20% of the extremes are caught by the forecast. 80% of the time, either the forecast predicted erroneously an extreme (false alarm) or did not predict an extreme that occurred (miss). The definition of

the last skilful day can be adapted depending on the usage of the forecast: the definition can be more conservative than being be, e.g. the last leadtime day for which at least 75% of the extreme events are caught.

Checking if a value of the BLF is significant is a kind of hypothesis test that is repeated for a large number of grid points. One could argue that some regional significance should be investigated. However, when displaying the local significance as “largest leadtime day with skilful forecast”, the results are continuous rather than a strict “yes or no” response. Moreover, the spatial coherence of the results confirms the robustness of the method.

Our method to assess extremes can also be applied to other variables, such as temperature, river discharge, etc. Considering the other side of extremes, evaluating the skill of forecasts to predict droughts would also be of crucial importance. For droughts, the persistence of dry periods matters, rather than the occurrence of precipitation. The method could be adapted accordingly, e.g. adjusting the definition temporal aggregation introduced in this chapter.

## 4.5 Appendix

### 4.5.1 95<sup>th</sup> percentile

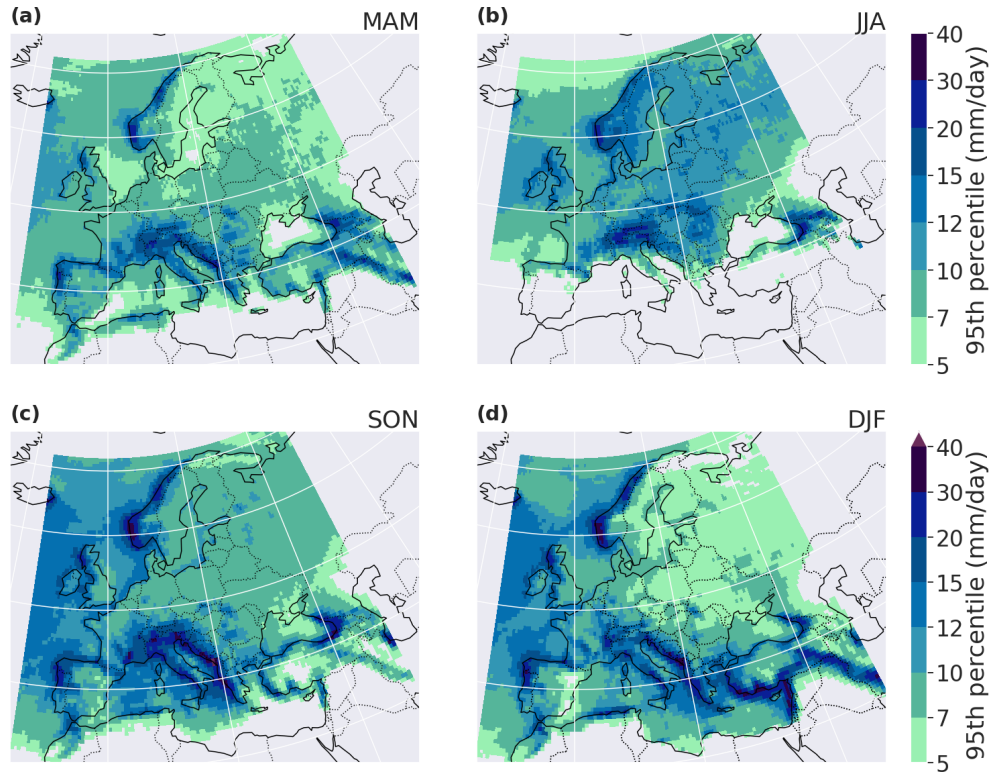


FIGURE 4.8: 95<sup>th</sup> percentile of daily precipitation in ERA-5, 2001-2021.

### 4.5.2 Local and daily comparison of extremes

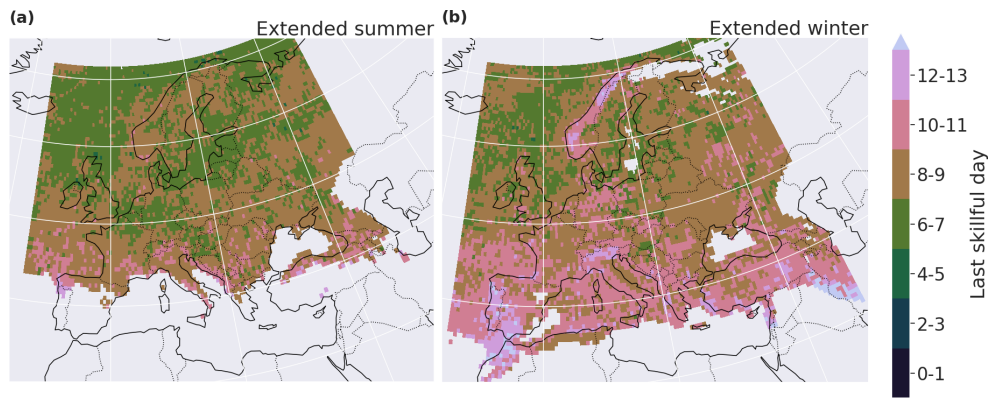


FIGURE 4.9: *Last skilful day for the BLF for local and daily comparison, in extended summer (a) and extended winter (b).*



## 4.5.3 Temporally accumulated extremes

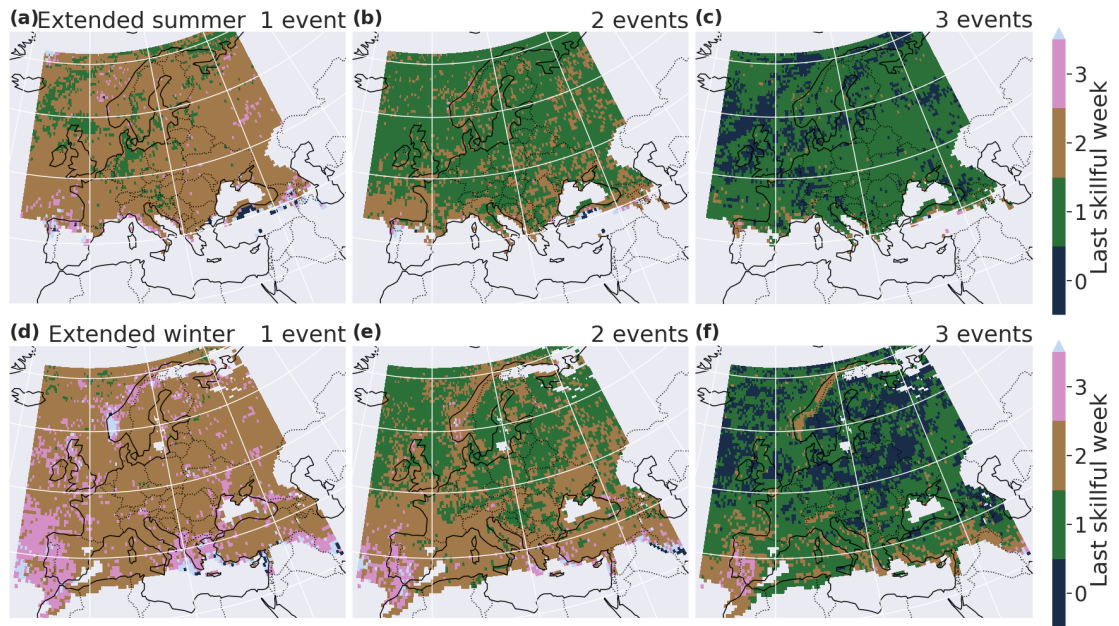


FIGURE 4.10: Last day of skill for the BLF in extended summer (a-c) and extended winter (d-e) for a minimum of 1 (a,d), 2 (b,e) and 3 (c,f) events in a 7 days window.

## 4.5.4 Spatially accumulated extremes

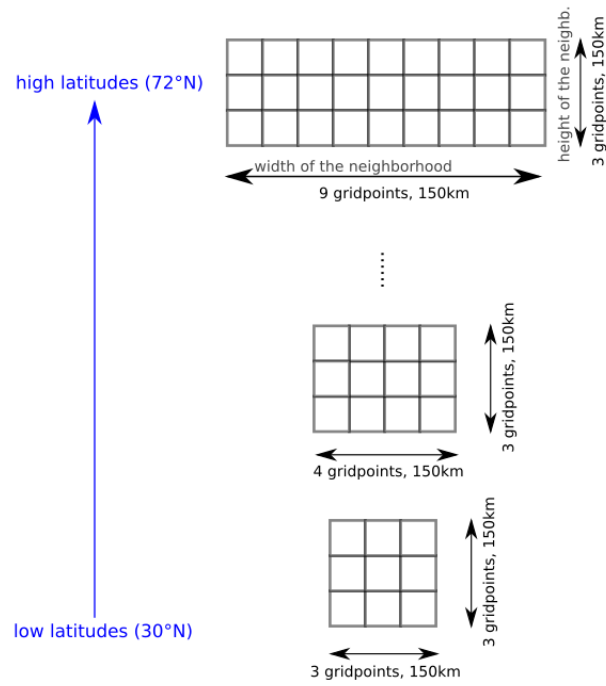


FIGURE 4.11: *Illustration of the width of the spatial neighborhood, in terms of grid points, depending on the latitude for a constant width in kilometers (and for a constant area).*

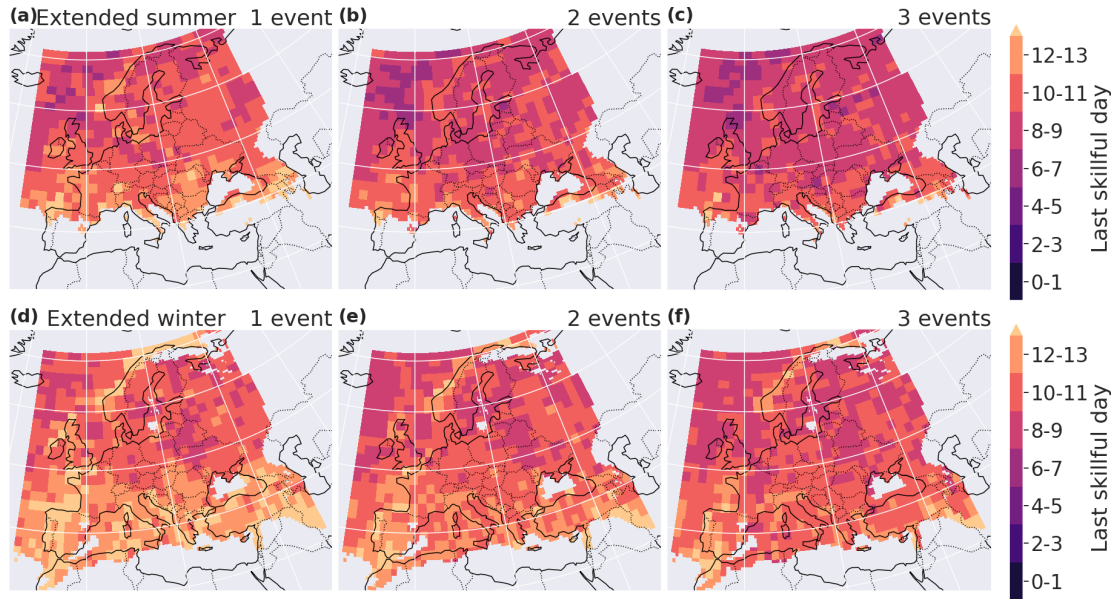


FIGURE 4.12: Last day of skill for the BLF in extended summer (a-c) and extended winter (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in neighborhood of  $150 \times 150 \text{ km}$ .

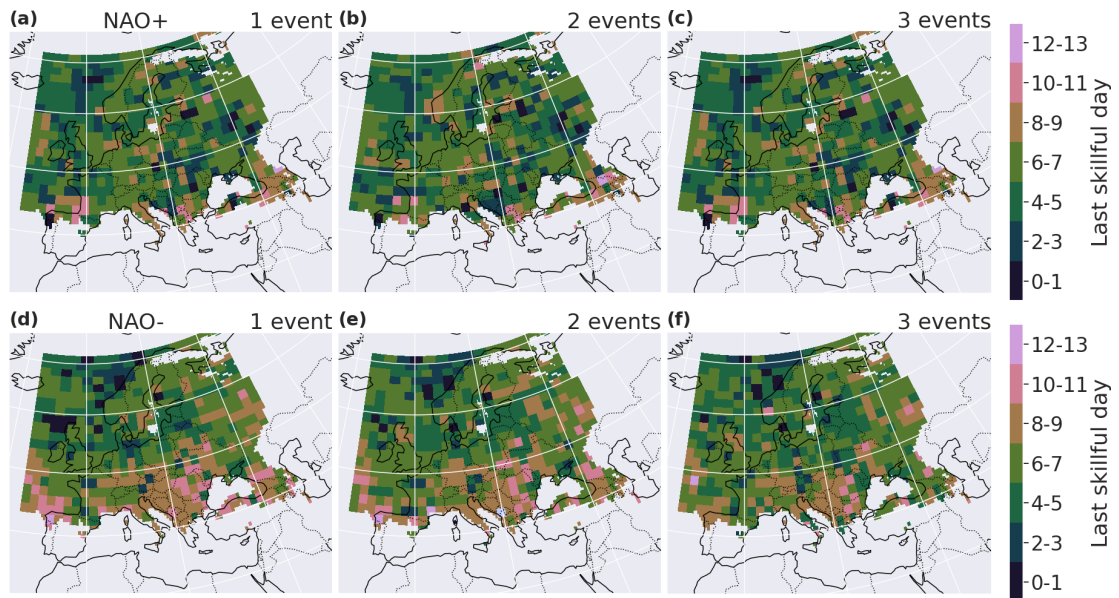


FIGURE 4.13: Last day of skill for the Brier score in positive NAO phase (a-c) and negative NAO phase (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in neighborhood of  $150 \times 150 \text{ km}$ .

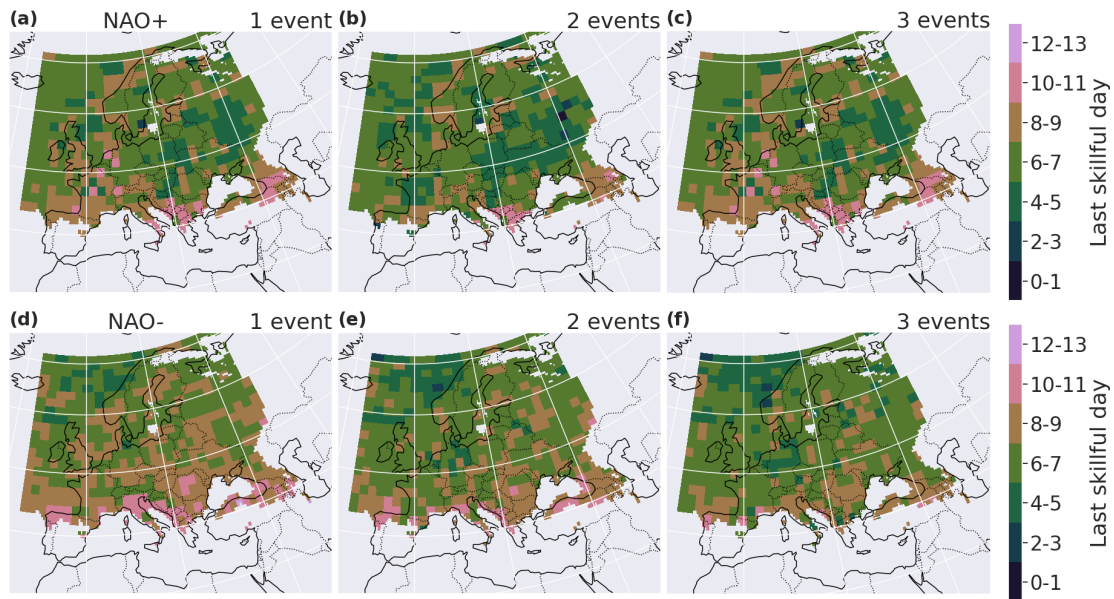


FIGURE 4.14: Last day of skill for the BLF in positive NAO phase (a-c) and negative NAO phase (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in neighborhood of  $150 \times 150 \text{ km}$ .

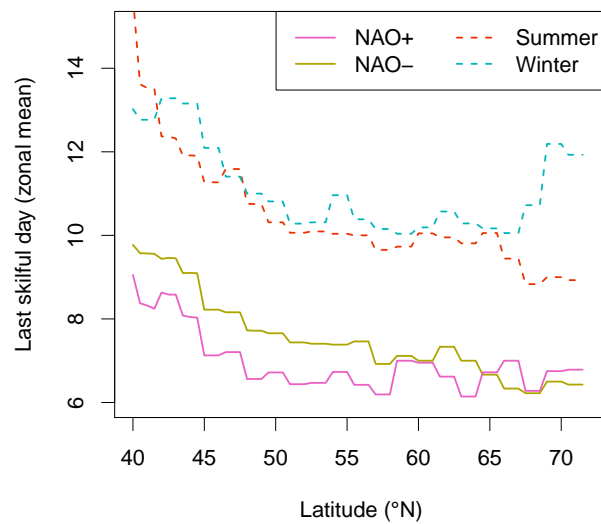


FIGURE 4.15: Zonal mean of the last skillful day for the BLF, during the different NAO phases, and during extended summer and extended winter, for comparison (with spatial aggregation).

## CHAPTER

# 5

# SUMMARY, CONCLUDING REMARKS AND OUTLOOK

## 5.1 Summary and concluding remarks

In this thesis, precipitation was modelled and verified with a focus on extreme precipitation. Extreme value theory was used to model the intensity of precipitation extremes. The research questions formulated in the introduction (Chapter 1) have been answered through the chapters of this thesis.

Chapter 2 contains a validation of precipitation from the ERA-5 reanalysis dataset against two observational datasets, EOBS and CMORPH. EOBS is a European gridded dataset interpolated from station observations. CMORPH provides global precipitation (between  $-60^{\circ}\text{N}$  and  $60^{\circ}\text{N}$ ) derived from a combination of passive micro-wave satellite scans and geostationary satellite infrared data. The strengths and weaknesses of the two observation datasets are documented in the literature. With this information in mind, we assessed both the occurrence and the intensity of precipitation. We used the extended generalized Pareto distribution to model the precipitation intensity. This model can fit the entire range of precipitation distribution using extreme value theory for the lower and upper tails. The precipitation co-occurrence was quantified with the hit rate. The intensities of ERA-5 and the observational datasets were compared using quantile estimates and the Kullback-Leibler divergence. Over Europe, precipitation in EOBS and in ERA-5 are generally in good agreement, particularly in the regions where the EOBS station coverage is dense. The reanalysis dataset shows more discrepancy with the observation in areas where the station coverage is poor. ERA-5 and CMORPH precipitation intensity agree well over the midlatitudes but disagree strongly over the tropics, where ERA-5 underestimates heavy pre-

precipitation. The regular spatial and temporal resolution and the consistency with the large-scale circulation are the strengths of the precipitation data from ERA-5 reanalysis. In general, ERA-5 shows a good agreement with observation-based datasets in the extra-tropics. To my knowledge, a global evaluation of ERA-5 with a specific focus on extremes had never been conducted in the past. The reanalysis dataset provides valuable complementary information to observational data in regions where observational datasets are sparse. The assessment showed that ERA-5 had limited skill for heavy precipitation over the tropics. This limitation has been complemented by Hassler and Lauer (2021), who found a limited performance for low and mean precipitation (but with still a higher skill than other reanalysis datasets). An observational dataset should be preferred over ERA-5 in the tropics. This chapter highlights the usefulness and limitations of ERA-5 precipitation. The analysis presented in this chapter was published in an article (Rivoire et al., 2021b).

The verification of the dataset ERA-5 conducted in Chapter 2 constitutes a basis for the two following chapters.

In Chapter 3, we estimated return levels of extreme precipitation over Europe by fitting the intensity distribution regionally. The procedure consists of two main steps: 1) the identification of homogeneous regions in terms of extreme behavior 2) the fitting of the intensity distribution with a reduced number of parameters over each homogeneous region. We applied the procedure to ERA-5 precipitation over Europe lands. For the first step, the homogeneous regions were identified with the partitioning-around-medoid clustering algorithm. The distance used for the clustering is based on a ratio ( $\omega$ ) of probability-weighted moments that is non-parametric and that characterizes the upper tail of the distribution. The number of clusters was chosen in a way to have a partition not too fragmented in space. The output of the clustering is a physically meaningful partition, e.g. the borders between clusters follow the orography. The ratio  $\omega$ , therefore, captures spatial structures associated with physical features, even without the use of covariates. For the second step, we fitted the intensity distribution within every homogeneous cluster with the extended generalized Pareto distribution. This distribution can be fitted locally with three parameters. However, the region of interest contains 20,000 grid points: a local fitting implies a large number of parameters,  $3 \times 20,000$ . The homogeneous regions were used to reduce this number. Two of the three parameters were regionalized: they were estimated by pooling (and normalizing) the precipitation data of all grid points within a homogeneous region. This way, the number of parameters was divided by approximately 3. The only parameter that is allowed to vary within a homogeneous region brings the flexibility needed to accurately model local precipitation. The return levels for return periods of 10, 50 and 100 years were estimated. The parsimonious fitting we conducted competes well with models with higher complexity. The novelty brought by this chapter is the clustering procedure on extreme precipitation behavior conducted over Europe, as well as the pooling of grid points on large clusters while maintaining the local variability for the estimation of return levels. The identification of regions with similar behavior of heavy precipitation is a central result of this chapter. The analysis presented in this chapter was submitted to the journal *Weather and Climate Extremes* (Rivoire et al., 2021a).

Chapter 4 evaluated extreme events in precipitation forecast. As an illustration, I focused on

precipitating extremes in ECMWF subseasonal forecasts (S2S) over Europe. The verification was conducted against ERA-5 precipitation. The extremes were defined as exceedances over the seasonal 95<sup>th</sup> percentile. This definition has the advantage of providing debiased binary data for extremes. The Brier score and a binary loss function were used to measure the skill of the forecast compared to the observation. The forecast is considered skilful if it has higher skill than a prediction based on the climatology. A forecast can still be useful on the S2S timescale if it can capture an extreme event in a close spatio-temporal neighborhood instead of the exact same day and the same location. Therefore, in addition to a strict local daily verification of extremes, I allowed for spatial and temporal flexibility. For all the configurations, I found a higher forecast skill in winter compared to summer. Norway, Portugal and the south of the Alps are the most skilful regions. The Mediterranean region also has relatively good skill in winter. The spatial and temporal aggregation of the extremes increases the skill over Europe. The assessment conducted in Chapter 4 fills the research gap in S2S forecast skill of precipitation over Europe. The methodology developed in this chapter can be adapted and applied to other variables with large impacts, such as temperature extremes or river discharge. The results presented in Chapter 4 constitute a useful contribution to the knowledge about S2S prediction timescale, a research field in recent expansion.

In this doctoral thesis, several evaluation methods for precipitation dataset were studied. They all have a particular focus on precipitation extremes, analysing both the intensity and the occurrence of heavy precipitation. The developed algorithms for each chapter have been made available on Github (<https://github.com/PauRiv> and [https://github.com/PhilomeneLeGall/RFA\\_regional\\_EGPDk.git](https://github.com/PhilomeneLeGall/RFA_regional_EGPDk.git)) and have been applied to large gridded datasets, with a spatial extent scale of Europe and the whole globe. This thesis lies at the interface between applied statistics and climatological analysis: innovative statistical tools were applied to datasets widely used.

## 5.2 Outlook

The analyses and outcomes of the present thesis opened avenues for further research. In this section, I propose a list of potential research questions, as follow-ups of this doctoral thesis.

- The evaluation of ERA-5 precipitation dataset could be extended further with a focus on other continents than Europe. The methodology developed in Chapter 2 could be used for another comparison of ERA-5 with regional observations over e.g. central America, Africa or Oceania, where the assessment of ERA-5 over large spatial extent (continent-scale) is lacking. The choice of CMORPH for the comparison with ERA-5 was motivated by global availability. However, this satellite dataset is available only for a short time period (2003-2016).
- Even if ERA-5 performs better over the tropics than ERA-interim (the previous reanalysis dataset from ECMWF) and other reanalyses, this region remains a difficulty for this type of dataset (Rivoire et al., 2021b; Hassler and Lauer, 2021). A technical investigation of the reasons for this difficulty would be useful and could help to find ways of potential improvement.
- In Chapter 3, the choice of the number of clusters was guided by visual consideration. Investigating metrics for an automatic decision for the number of clusters would remove the arbitrariness of the choice. One option could be to use geometrical indices to penalize the fragmentation of the clusters (e.g. section 4.3 in Kholodovsky and Liang, 2021).
- The identification of coherent regions in Chapter 3 could be extended by considering both temporal and magnitude considerations for extremes, i.e. both the co-occurrence of extremes and the heavy precipitation behavior. Le Gall et al. (2021) developed a clustering algorithm based on both intensity distribution and temporal dependence and applied it on global climate models. Their method could be applied to ERA-5 precipitation, over Europe or globally.
- In Chapter 4, the simple index derived from the binary loss score still has a lot of room for improvement. With further research, this index could be transformed to a probability score adapted to ensemble forecasts. An assessment of non-binary extremes could also be conducted, working with continuous data in the upper tail of the distribution, with for example a continuous ranked probability score focusing on extreme events (Allen et al., 2021, 2022; Taillardat et al., 2022; Pic et al., 2022). This would require more sophisticated post-processing of the data, e.g. quantile mapping (Crochemore et al., 2016; Monhart et al., 2018) or Bayesian calibration (Specq and Batté, 2020).
- The evaluation of the precipitation extremes could be extended further, with the investigation of the impact the type of precipitation, e.g. with the research question: among convective, cyclonic and orographic precipitation, what is the type of precipitation with the best skill? Moreover, all the available S2S forecast models (in addition to the ECMWF



one) could be evaluated over Europe (Vitart and Robertson, 2018). The spatial aggregation of precipitation over regions where impacts are relevant would bring some depth to the analysis, aggregating for example over catchments (Kopp et al., 2021).

- Finally, the method developed for the verification of extremes could be applied to other variables with available S2S forecasts, such as temperature or streamflow. The S2S hind-cast database of ECMWF (Vitart, 2020) provides a large range of variables. One could investigate the prediction skills of variables having a potential impact on vegetation, such as 2-metre temperature (daily average, 6hourly minimum and 6hourly maximum), total precipitation (6hourly accumulation), surface temperature, soil moisture top 20 cm and 100cm, soil temperature top 20 and 100 cm and 2-metre dewpoint temperature, snow depth water equivalent (daily averages).



# BIBLIOGRAPHY

- Akaike, H.: Factor analysis and AIC, in: Selected papers of Hirotugu Akaike, pp. 371–386, Springer, 1987.
- Alfieri, L. and Thielen, J.: A European precipitation index for extreme rain-storm and flash flood early warning, *Meteorological Applications*, 22, 3–13, <https://doi.org/10.1002/met.1328>, 2015.
- Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F.: Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts, *Quarterly Journal of the Royal Meteorological Society*, 147, 1403–1418, <https://doi.org/10.1002/qj.3983>, 2021.
- Allen, S., Ginsbourger, D., and Ziegel, J.: Evaluating forecasts for high-impact events using transformed kernel scores, pp. 1–34, URL <http://arxiv.org/abs/2202.12732>, 2022.
- Allstadt, A. J., Vavrus, S. J., Heglund, P. J., Pidgeon, A. M., Thogmartin, W. E., and Radeloff, V. C.: Spring plant phenology and false springs in the conterminous US during the 21st century, *Environmental Research Letters*, 10, 104 008, <https://doi.org/10.1088/1748-9326/10/10/104008>, 2015.
- Amjad, M., Yilmaz, M. T., Yucel, I., and Yilmaz, K. K.: Performance evaluation of satellite- and model-based precipitation products over varying climate and complex topography, *Journal of Hydrology*, 584, 124 707, <https://doi.org/10.1016/j.jhydrol.2020.124707>, 2020.
- AMS: Precipitation, Glossary of meteorology, <https://web.archive.org/web/20081009142439/http://amsglossary.allenpress.com/glossary/search?id=precipitation1>, accessed: 2022-05-24, 2003.
- Anderson, T. W. and Darling, D. A.: Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes, *The annals of mathematical statistics*, pp. 193–212, 1952.

- Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rötter, R. P., Cammarano, D., et al.: Uncertainty in simulating wheat yields under climate change, *Nature climate change*, 3, 827–832, <https://doi.org/10.1038/nclimate1916>, 2013.
- Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T.: Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe, *Weather and climate extremes*, 9, 17–24, 2015.
- Barton, Y., Giannakaki, P., von Waldow, H., Chevalier, C., Pfahl, S., and Martius, O.: Clustering of Regional-Scale Extreme Precipitation Events in Southern Switzerland, *Monthly Weather Review*, 144, 347–369, <https://doi.org/10.1175/MWR-D-15-0205.1>, 2016.
- Batjes, N. H.: ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2), Tech. rep., ISRIC-World Soil Information, 2012.
- Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Present and future Köppen-Geiger climate classification maps at 1-km resolution, *Scientific data*, 5, 1–12, 2018.
- Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., Van der Velde, M., and Makowski, D.: Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France, *Nature communications*, 9, 1–10, <https://doi.org/10.1038/s41467-018-04087-x>, 2018.
- Bernard, E., Naveau, P., Vrac, M., and Mestre, O.: Clustering of maxima: Spatial dependencies among heavy rainfall in France, *Journal of Climate*, 26, 7929–7937, 2013.
- Bevacqua, E., De Michele, C., Manning, C., Couasnon, A., Ribeiro, A. F., Ramos, A. M., Vignotto, E., Bastos, A., Blesić, S., Durante, F., Hillier, J., Oliveira, S. C., Pinto, J. G., Ragno, E., Rivoire, P., Saunders, K., van der Wiel, K., Wu, W., Zhang, T., and Zscheischler, J.: Guidelines for Studying Diverse Types of Compound Weather and Climate Events, *Earth’s Future*, 9, 1–23, <https://doi.org/10.1029/2021EF002340>, 2021.
- Bocheva, L. and Pophristov, V.: Seasonal analysis of large-scale heavy precipitation events in Bulgaria, *AIP Conference Proceedings*, 2075, 200 017, <https://doi.org/10.1063/1.5099023>, 2019.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 78, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2), 1950.
- Buizza, R., Milleer, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908, <https://doi.org/10.1002/qj.49712556006>, 1999.

- Buriticá, G., Nicolas, M., Mikosch, T., and Wintenberger, O.: Some variations on the extremal index, arXiv preprint arXiv:2106.05117, 2021.
- C3S, C. C. C. S.: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5>, accessed: 2022-05-24, 2017.
- Carreau, J. and Bengio, Y.: A hybrid Pareto model for asymmetric fat-tailed data: the univariate case, *Extremes*, 12, 53–76, 2009.
- Carreau, J., Naveau, P., and Neppel, L.: Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation, *Water Resources Research*, 53, 4407–4426, 2017.
- Casanueva, A., Rodríguez-Puebla, C., Frías, M. D., and González-Reviriego, N.: Variability of extreme precipitation over Europe and its relationships with teleconnection patterns, *Hydrology and Earth System Sciences*, 18, 709–725, <https://doi.org/10.5194/hess-18-709-2014>, 2014.
- CDO: Regridding with CDO, Running the Remapping (Conservative Method), [https://www.climate-cryosphere.org/wiki/index.php?title=Regridding\\_with\\_CDO#Running\\_the\\_Remapping\\_.28Conservative\\_Method.29](https://www.climate-cryosphere.org/wiki/index.php?title=Regridding_with_CDO#Running_the_Remapping_.28Conservative_Method.29), accessed: 2022-05-19, 2018.
- Chen, Y., Sharma, S., Zhou, X., Yang, K., Li, X., Niu, X., Hu, X., and Khadka, N.: Spatial performance of multiple reanalysis precipitation datasets on the southern slope of central Himalaya, *Atmospheric Research*, 250, 105 365, <https://doi.org/10.1016/j.atmosres.2020.105365>, 2021.
- Climate-Change-Service: <https://climate.copernicus.eu/european-state-of-the-climate>, accessed: 2022-05-24, 2020.
- Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce: NOAA CPC Morphing Technique (CMORPH) Global Precipitation Analyses, URL <https://doi.org/10.5065/D6CZ356W>, accessed: 2022-05-24, 2011.
- Cook, B. I., Mankin, J. S., and Anchukaitis, K. J.: Climate Change and Drought: From Past to Future, *Current Climate Change Reports*, 4, 164–179, <https://doi.org/10.1007/s40641-018-0093-2>, 2018.
- Cooley, D.: Return periods and return levels under climate change, in: *Extremes in a changing climate*, pp. 97–114, Springer, 2013.
- Cooley, D., Nychka, D., and Naveau, P.: Bayesian Spatial Modeling of Extreme Precipitation Return Levels, *Journal of the American Statistical Association*, 102, 824–840, <https://doi.org/10.1198/016214506000000780>, 2007.

- Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., and Jones, P. D.: An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets, *Journal of Geophysical Research: Atmospheres*, 123, 9391–9409, <https://doi.org/10.1029/2017JD028200>, 2018.
- Cortesi, N., Gonzalez-Hidalgo, J. C., Brunetti, M., and de Luis, M.: Spatial variability of precipitation in Spain, *Regional Environmental Change*, 14, 1743–1749, <https://doi.org/10.1007/s10113-012-0402-6>, 2014.
- Crochemore, L., Ramos, M. H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrology and Earth System Sciences*, 20, 3601–3618, <https://doi.org/10.5194/hess-20-3601-2016>, 2016.
- Dai, A., Zhao, T., and Chen, J.: Climate Change and Drought: a Precipitation and Evaporation Perspective, *Current Climate Change Reports*, 4, 301–312, <https://doi.org/10.1007/s40641-018-0101-6>, 2018.
- Dalrymple, T.: Flood-frequency analyses, manual of hydrology: Part 3, Tech. rep., USGPO,, 1960.
- Darwish, M. M., Tye, M. R., Prein, A. F., Fowler, H. J., Blenkinsop, S., Dale, M., and Faulkner, D.: New hourly extreme precipitation regions and regional annual probability estimates for the UK, *International Journal of Climatology*, 41, 582–600, 2021.
- Daryanto, S., Wang, L., and Jacinthe, P.-A.: Global Synthesis of Drought Effects on Maize and Wheat Production, *PloS one*, 11, e0156362, <https://doi.org/10.1371/journal.pone.0156362>, 2016.
- de Andrade, F. M., Coelho, C. A., and Cavalcanti, I. F.: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models, *Climate Dynamics*, 52, 5451–5475, <https://doi.org/10.1007/s00382-018-4457-z>, 2019.
- de Andrade, F. M., Young, M. P., Macleod, D., Hirons, L. C., Woolnough, S. J., and Black, E.: Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability, *Weather and Forecasting*, 36, 265–284, <https://doi.org/10.1175/WAF-D-20-0054.1>, 2021.
- DeFlorio, M. J., Waliser, D. E., Guan, B., Ralph, F. M., and Vitart, F.: Global evaluation of atmospheric river subseasonal prediction skill, *Climate Dynamics*, 52, 3039–3060, <https://doi.org/10.1007/s00382-018-4309-x>, 2019.
- Deidda, R.: A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series, *Hydrology and Earth System Sciences*, 14, 2559–2575, <https://doi.org/10.5194/hess-14-2559-2010>, 2010.
- Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting,

- M.: Insights from Earth system model initial-condition large ensembles and future prospects, *Nature Climate Change*, 10, 277–286, <https://doi.org/10.1038/s41558-020-0731-2>, 2020.
- Desgraupes, B.: Clustering indices, *University of Paris Ouest-Lab Modal’X*, 1, 34, 2013.
- Dikshit, A., Sarkar, R., Pradhan, B., Segoni, S., and Alamri, A. M.: Rainfall induced landslide studies in indian himalayan region: A critical review, *Applied Sciences (Switzerland)*, 10, 1–24, <https://doi.org/10.3390/app10072466>, 2020.
- Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E. P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S. W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling, *Journal of Geophysical Research: Atmospheres*, 125, 1–20, <https://doi.org/10.1029/2019JD030923>, 2019.
- Domeisen, D. I., White, C. J., Afargan-Gerstman, H., Muñoz, Á. G., Janiga, M. A., Vitart, F., Wulf, C. O., Antoine, S., Ardilouze, C., Batté, L., Bloomfield, H. C., Brayshaw, D. J., Camargo, S. J., Charlton-Pérez, A., Collins, D., Cowan, T., del Mar Chaves, M., Ferranti, L., Gómez, R., González, P. L., González Romero, C., Infanti, J. M., Karozis, S., Kim, H., Kolstad, E. W., LaJoie, E., Lledó, L., Magnusson, L., Malguzzi, P., Manrique-Suñén, A., Mastrangelo, D., Materia, S., Medina, H., Palma, L., Pineda, L. E., Sfetsos, A., Son, S.-W., Soret, A., Strazzo, S., and Tian, D.: Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe, *Bulletin of the American Meteorological Society*, pp. 1–76, <https://doi.org/10.1175/bams-d-20-0221.1>, 2022.
- Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., and Zwiers, F. W.: Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets, *Journal of Climate*, 27, 5019–5035, <https://doi.org/10.1175/JCLI-D-13-00405.1>, 2014.
- Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., and Thouvenin, T.: A numerical study of three catastrophic precipitating events over southern France. II: Mesoscale triggering and stationarity factors, *Quarterly Journal of the Royal Meteorological Society*, 134, 131–145, <https://doi.org/10.1002/qj.199>, 2008.
- Dupuis, D.: Exceedances over High Thresholds: A Guide to Threshold Selection, *Extremes*, 1, 251–261, <https://doi.org/10.1023/A:1009914915709>, 1999.
- ECMWF: URL [https://sites.ecmwf.int/era/40-atlas/docs/section\\_B/parameter\\_tp.html#](https://sites.ecmwf.int/era/40-atlas/docs/section_B/parameter_tp.html#), 2006.
- ECMWF: ERA5: data documentation, <https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation>, accessed: 2022-05-25, 2018a.
- ECMWF: Types of Precipitation, <https://confluence.ecmwf.int/display/FUG/Types+of+Precipitation>, accessed: 2022-05-24, 2018b.

- ECMWF: Climate reanalysis, <https://www.ecmwf.int/en/research/climate-reanalysis>, accessed: 2022-05-24, 2022.
- ECMWF: Changes in ECMWF model, <https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model>, accessed: 2022-05-19, 2022a.
- ECMWF: Re-forecast for medium and extended forecast range, <https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range/re-forecast-medium-and-extended-forecast-range>, accessed: 2022-04-27, 2022b.
- ECMWF: S2S, ECMWF, Reforecasts, Daily averaged, <https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/>, accessed: 2022-04-27, 2022c.
- EEA: National climate change vulnerability and risk assessments in Europe, <https://www.eea.europa.eu/publications/national-climate-change-vulnerability-2018>, accessed: 2022-05-24, 2018.
- Endris, H. S., Hirons, L., Segele, Z. T., Gudoshava, M., Woolnough, S., and Artan, G. A.: Evaluation of the skill of monthly precipitation forecasts from global prediction systems over the greater horn of Africa, *Weather and Forecasting*, 36, 1275–1298, <https://doi.org/10.1175/WAF-D-20-0177.1>, 2021.
- Evin, G., Blanchet, J., Paquet, E., Garavaglia, F., and Penot, D.: A regional model for extreme rainfall based on weather patterns subsampling, *Journal of Hydrology*, 541, 1185–1198, 2016.
- Evin, G., Favre, A.-C., and Hingray, B.: Stochastic generation of multi-site daily precipitation focusing on extreme events, *Hydrology and Earth System Sciences*, 22, 655–672, <https://doi.org/10.5194/hess-22-655-2018>, 2018.
- FAOSTAT: FAO Statistics, Food and Agriculture Organization of the United Nations, Rome, <http://www.fao.org/faostat/en/>, accessed: 2022-05-24, 2020.
- Fawad, M., Ahmad, I., Nadeem, F. A., Yan, T., and Abbas, A.: Estimation of wind speed using regional frequency analysis based on linear-moments, *International Journal of Climatology*, 38, 4431–4444, 2018.
- Forestieri, A., Lo Conti, F., Blenkinsop, S., Cannarozzo, M., Fowler, H. J., and Noto, L. V.: Regional frequency analysis of extreme rainfall in Sicily (Italy), *International Journal of Climatology*, 38, e698–e716, 2018.
- Forkel, M., Thonicke, K., Beer, C., Cramer, W., Bartalev, S., and Schmullius, C.: Extreme fire events are related to previous-year surface moisture conditions in permafrost-underlain larch forests of Siberia, *Environmental Research Letters*, 7, 044 021, <https://doi.org/10.1088/1748-9326/7/4/044021>, 2012.
- Fowler, H. and Kilsby, C.: A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000, *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 23, 1313–1334, 2003.



- Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., Smith, P., van der Velde, M., Vicca, S., Babst, F., Beer, C., Buchmann, N., Canadell, J. G., Ciais, P., Cramer, W., Ibrom, A., Miglietta, F., Poulter, B., Rammig, A., Seneviratne, S. I., Walz, A., Wattenbach, M., Zavala, M. A., and Zscheischler, J.: Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts, *Global Change Biology*, 21, 2861–2880, <https://doi.org/10.1111/gcb.12916>, 2015.
- Friedman, J., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33, <https://doi.org/10.18637/jss.v033.i01>, 2010.
- Fukutome, S., Liniger, M. A., and Sèveges, M.: Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland, *Theoretical and Applied Climatology*, 120, 403–416, <https://doi.org/10.1007/s00704-014-1180-5>, 2015.
- Furnival, G. M. and Wilson, R. W.: Regressions by Leaps and Bounds, *Technometrics*, 16, 499–511, <https://doi.org/10.1080/00401706.1974.10489231>, 1974.
- Gillett, N. P., Cannon, A. J., Malinina, E., Schnorbus, M., Anslow, F., Sun, Q., Kirchmeier-Young, M., Zwiers, F., Seiler, C., Zhang, X., Flato, G., Wan, H., Li, G., and Castellan, A.: Human influence on the 2021 British Columbia floods, *Weather and Climate Extremes*, p. 100441, <https://doi.org/10.1016/j.wace.2022.100441>, 2022.
- Gleixner, S., Demissie, T., and Diro, G. T.: Did ERA5 improve temperature and precipitation reanalysis over East Africa?, *Atmosphere*, 11, 1–19, <https://doi.org/10.3390/atmos11090996>, 2020.
- Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>, 2007.
- Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H.: Balancing Europe’s wind-power output through spatial deployment informed by weather regimes, *Nature Climate Change*, 7, 557–562, <https://doi.org/10.1038/NCLIMATE3338>, 2017.
- Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R.: Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, *Water resources research*, 15, 1049–1054, 1979.
- Grossiord, C., Buckley, T. N., Cernusak, L. A., Novick, K. A., Poulter, B., Siegwolf, R. T. W., Sperry, J. S., and McDowell, N. G.: Plant responses to rising vapor pressure deficit, *New Phytologist*, 226, 1550–1566, <https://doi.org/10.1111/nph.16485>, 2020.
- Gvoždíková, B., Müller, M., and Kašpar, M.: Spatial patterns and time distribution of central European extreme precipitation events between 1961 and 2013, *International Journal of Climatology*, 39, 3282–3297, 2019.

- Halkidi, M., Batistakis, Y., and Vazirgiannis, M.: Clustering validity checking methods: Part II, *ACM Sigmod Record*, 31, 19–27, 2002.
- Hand, D. J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, 77, 103–123, <https://doi.org/10.1007/s10994-009-5119-5>, 2009.
- Hassler, B. and Lauer, A.: Comparison of reanalysis and observational precipitation datasets including ERA5 and WFDE5, *Atmosphere*, 12, <https://doi.org/10.3390/atmos12111462>, 2021.
- Haylock, M. R. and Goodess, C. M.: Interannual variability of European extreme winter rainfall and links with mean large-scale circulation, *International Journal of Climatology*, 24, 759–776, <https://doi.org/10.1002/joc.1033>, 2004.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, *Journal of Geophysical Research Atmospheres*, 113, <https://doi.org/10.1029/2008JD010201>, 2008.
- Hazeleger, W., Wang, X., Severijns, C., Ștefănescu, S., Bintanja, R., Sterl, A., Wyser, K., Semmler, T., Yang, S., Van den Hurk, B., et al.: EC-Earth V2.2: description and validation of a new seamless earth system prediction model, *Climate dynamics*, 39, 2611–2629, <https://doi.org/10.1007/s00382-011-1228-5>, 2012.
- Hénin, R., Liberato, M. L., Ramos, A. M., and Gouveia, C. M.: Assessing the use of satellite-based estimates and high-resolution precipitation datasets for the study of extreme precipitation events over the Iberian Peninsula, *Water (Switzerland)*, 10, <https://doi.org/10.3390/w10111688>, 2018.
- Hennermann, K.: ERA5: data description, <https://confluence.ecmwf.int/pages/viewpage.action?pageId=85402030>, accessed: 2022-05-24, 2020.
- Herrera, S., Kotlarski, S., Soares, P. M., Cardoso, R. M., Jaczewski, A., Gutiérrez, J. M., and Maraun, D.: Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution, *International Journal of Climatology*, 39, 3717–3729, <https://doi.org/10.1002/joc.5878>, 2019.
- Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather and Forecasting*, 15, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.
- Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J. M., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, *ECMWF Newsletter*, 146, 17–24, <https://doi.org/10.21957/vf291hehd7>, 2019.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo,

- G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, pp. 1–51, <https://doi.org/10.1002/qj.3803>, 2020.
- Hofstra, N., Haylock, M., New, M., and Jones, P. D.: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, *Journal of Geophysical Research Atmospheres*, 114, <https://doi.org/10.1029/2009JD011799>, 2009.
- Hofstra, N., New, M., and McSweeney, C.: The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data, *Climate Dynamics*, 35, 841–858, <https://doi.org/10.1007/s00382-009-0698-1>, 2010.
- Hosking, J. R.: The four-parameter kappa distribution, *IBM Journal of Research and Development*, 38, 251–258, 1994.
- Hosking, J. R. M. and Wallis, J. R.: *Regional frequency analysis: an approach based on L-moments*, Cambridge University Press, 2005.
- Hu, G. and Franzke, C. L.: Evaluation of daily precipitation extremes in reanalysis and gridded observation-based data sets over Germany, *Geophysical Research Letters*, 47, e2020GL089624, 2020.
- Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., Liu, Z., Tian, Y., Yan, D., and Chen, X.: A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts, *Journal of Hydrology*, p. 127896, <https://doi.org/10.1016/j.jhydrol.2022.127896>, 2022.
- Hudson, D., Alves, O., Hendon, H. H., and Marshall, A. G.: Bridging the gap between weather and seasonal forecasting: intraseasonal forecasting for Australia, *Quarterly Journal of the Royal Meteorological Society*, 137, 673–689, <https://doi.org/10.1002/qj.769>, 2011.
- Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M.: An overview of the north atlantic oscillation, *Geophysical Monograph Series*, 134, 1–35, <https://doi.org/10.1029/134GM01>, 2003.
- Ibebuchi, C. C.: Patterns of atmospheric circulation in Western Europe linked to heavy rainfall in Germany: preliminary analysis into the 2021 heavy rainfall episode, *Theoretical and Applied Climatology*, 148, 269–283, <https://doi.org/10.1007/s00704-022-03945-5>, 2022.
- Iizumi, T. and Ramankutty, N.: How do weather and climate influence cropping area and intensity?, *Global Food Security*, 4, 46 – 50, <https://doi.org/10.1016/j.gfs.2014.11.003>, 2015.
- IPCC: Summary for Policymakers, p. 3–22, Cambridge University Press, <https://doi.org/10.1017/CBO9781139177245.003>, 2012.

- Isotta, F. A., Frei, C., Weilguni, V., Tadic, M. P., Lassegues, P., Rudolf, B., Pavan, V., Cacciamani, C., Antolini, G., Ratto, S. M., Munari, M., Micheletti, S., Bonati, V., Lussana, C., Ronchi, C., Panettieri, E., Marigo, G., and Vertacnik, G.: The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data, *International Journal of Climatology*, 34, 1657–1675, <https://doi.org/Doi10.1002/Joc.3794>, 2014.
- Jagadish, K. S. V., Kadam, N. N., Xiao, G., Melgar, R. J., Bahuguna, R. N., Quinones, C., Tamilselvan, A., Prasad, P. V. V., and Jagadish, K. S.: Agronomic and Physiological Responses to High Temperature, Drought, and Elevated CO<sub>2</sub> Interactions in Cereals, in: *Advances in Agronomy*, edited by Sparks, D. L., vol. 127 of *Advances in Agronomy*, pp. 111–156, Elsevier Science, Burlington, <https://doi.org/10.1016/B978-0-12-800131-8.00003-0>, 2014.
- Jain, A. K., Murty, M. N., and Flynn, P. J.: Data Clustering: A Review, *ACM Comput. Surv.*, 31, 264–323, <https://doi.org/10.1145/331499.331504>, 1999.
- Jalbert, J., Favre, A.-C., Bélisle, C., and Angers, J.-F.: A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66, 941–962, <https://doi.org/10.1111/rssc.12212>, 2017.
- Jenkinson, A. F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, 81, 158–171, 1955.
- Jentsch, A., Kreyling, J., and Beierkuhnlein, C.: A new generation of climate-change experiments: events, not trends, *Frontiers in Ecology and the Environment*, 5, 365–374, [https://doi.org/10.1890/1540-9295\(2007\)5\[365:ANGOCE\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2007)5[365:ANGOCE]2.0.CO;2), 2007.
- Jiang, Q., Li, W., Fan, Z., He, X., Sun, W., Chen, S., Wen, J., Gao, J., and Wang, J.: Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland, *Journal of Hydrology*, 595, 125 660, <https://doi.org/10.1016/j.jhydrol.2020.125660>, 2021.
- Jones, P. W.: First- and second-order conservative remapping schemes for grids in spherical coordinates, *Monthly Weather Review*, 127, 2204–2210, [https://doi.org/10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2), 1999.
- Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P.: CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution, *Journal of Hydrometeorology*, [https://doi.org/10.1175/1525-7541\(2004\)005<0487:CAMTPG>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2), 2004.
- Kang, S. and Song, J.: Parameter and quantile estimation for the generalized Pareto distribution in peaks over threshold framework, *Journal of the Korean Statistical Society*, 46, 487–501, <https://doi.org/10.1016/j.jkss.2017.02.003>, 2017.

- Kaufman, L. and Rousseeuw, P. J.: Finding groups in data: an introduction to cluster analysis, vol. 344, John Wiley & Sons, 1990.
- Kenyon, J. and Hegerl, G. C.: Influence of modes of climate variability on global precipitation extremes, *Journal of Climate*, 23, 6248–6262, <https://doi.org/10.1175/2010JCLI3617.1>, 2010.
- Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., and Lichtenberger, J.: Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices, *Agricultural and Forest Meteorology*, 260–261, 300–320, <https://doi.org/10.1016/j.agrformet.2018.06.009>, 2018.
- Kholodovsky, V. and Liang, X. Z.: A generalized Spatio-Temporal Threshold Clustering method for identification of extreme event patterns, *Advances in Statistical Climatology, Meteorology and Oceanography*, 7, 35–52, <https://doi.org/10.5194/ascmo-7-35-2021>, 2021.
- Kim, S., Eghdamirad, S., Sharma, A., and Kim, J. H.: Quantification of Uncertainty in Projections of Extreme Daily Precipitation, *Earth and Space Science*, 7, <https://doi.org/10.1029/2019EA001052>, 2020.
- Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O., and Lavrenyuk, A.: Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models, *International Journal of Applied Earth Observation and Geoinformation*, 23, 192–203, <https://doi.org/10.1016/j.jag.2013.01.002>, 2013.
- Kolachian, R. and Saghafian, B.: Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes, *Theoretical and Applied Climatology*, 137, 1479–1493, <https://doi.org/10.1007/s00704-018-2680-5>, 2019.
- Kopp, J., Rivoire, P., Ali, S. M., Barton, Y., and Martius, O.: A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods, *Hydrology and Earth System Sciences*, 25, 5153–5174, <https://doi.org/10.5194/hess-25-5153-2021>, 2021.
- Kotz, M., Levermann, A., and Wenz, L.: The effect of rainfall changes on economic production, *Nature*, 601, 223–227, <https://doi.org/10.1038/s41586-021-04283-8>, 2022.
- Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S.: Cross-validation pitfalls when selecting and assessing regression and classification models, *Journal of Cheminformatics*, 6, 1–15, <https://doi.org/10.1186/1758-2946-6-10>, 2014.
- Kulie, M. S., Bennartz, R., Greenwald, T. J., Chen, Y., and Weng, F.: Uncertainties in microwave properties of frozen precipitation: Implications for remote sensing and data assimilation, *Journal of the Atmospheric Sciences*, 67, 3471–3487, <https://doi.org/10.1175/2010JAS3520.1>, 2010.

- Lamb, R. and Kay, A. L.: Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain, *Water Resources Research*, 40, 1–13, <https://doi.org/10.1029/2003WR002428>, 2004.
- Le Gall, P., Favre, A.-C., Naveau, P., and Prieur, C.: Improved Regional Frequency Analysis of rainfall data, 2021.
- Le Gall, P., Favre, A.-C., Naveau, P., and Tuel, A.: Non-parametric multimodel Regional Frequency Analysis applied to climate change detection and attribution, pp. 1–25, URL <http://arxiv.org/abs/2111.00798>, 2021.
- Légrand, J., Naveau, P., and Oesting, M.: Evaluation of binary classifiers for asymptotically dependent and independent extremes, <https://doi.org/10.48550/ARXIV.2112.13738>, 2021.
- Leng, G., Zhang, X., Huang, M., Asrar, G. R., and Leung, L. R.: The Role of Climate Covariability on Crop Yields in the Conterminous United States, *Scientific Reports*, 6, <https://doi.org/10.1038/srep33160>, 2016.
- Lenggenhager, S. and Martius, O.: Atmospheric blocks modulate the odds of heavy precipitation events in Europe, *Climate Dynamics*, 53, 4155–4171, <https://doi.org/10.1007/s00382-019-04779-0>, 2019b.
- Lenggenhager, S., Croci-Maspoli, M., Brönnimann, S., and Martius, O.: On the dynamical coupling between atmospheric blocks and heavy precipitation events: A discussion of the southern Alpine flood in October 2000, *Quarterly Journal of the Royal Meteorological Society*, 145, 530–545, <https://doi.org/10.1002/qj.3449>, 2019a.
- Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., and Stafford-Smith, M.: A compound event framework for understanding extreme impacts, *Wiley Interdisciplinary Reviews: Climate Change*, 5, 113–128, <https://doi.org/10.1002/wcc.252>, 2014.
- Lesk, C., Rowhani, P., and Ramankutty, N.: Influence of extreme weather disasters on global crop production, *Nature*, 529, 84–87, <https://doi.org/10.1038/nature16467>, 2016.
- Li, K., Yang, X., Liu, Z., Zhang, T., Lu, S., and Liu, Y.: Low yield gap of winter wheat in the North China Plain, *European Journal of Agronomy*, 59, 1–12, <https://doi.org/10.1016/j.eja.2014.04.007>, 2014.
- Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C. Y., and Li, X.: Evaluation and bias correction of S2S precipitation for hydrological extremes, *Journal of Hydrometeorology*, 20, 1887–1906, <https://doi.org/10.1175/JHM-D-19-0042.1>, 2019.
- Liang, W. and Zhang, M.: Summer and winter precipitation in East Asia scale with global warming at different rates, *Communications Earth & Environment*, 2, 1–8, <https://doi.org/10.1038/s43247-021-00219-2>, 2021.

- Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, URL <https://CRAN.R-project.org/doc/Rnews/>, 2002.
- Lobell, D. B.: Changes in diurnal temperature range and national cereal yields, Agricultural and Forest Meteorology, 145, 229–238, <https://doi.org/10.1016/j.agrformet.2007.05.002>, 2007.
- Lobell, D. B. and Asner, G. P.: Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields, Science, 299, 1032, <https://doi.org/10.1126/science.1078475>, 2003.
- Lobell, D. B. and Burke, M. B.: Why are agricultural impacts of climate change so uncertain? The importance of temperature relative to precipitation, Environmental Research Letters, 3, 034007, <https://doi.org/10.1088/1748-9326/3/3/034007>, 2008.
- Lobell, D. B. and Burke, M. B.: On the use of statistical models to predict crop yield responses to climate change, Agricultural and Forest Meteorology, 150, 1443–1452, <https://doi.org/10.1016/j.agrformet.2010.07.008>, 2010.
- Lobell, D. B., Schlenker, W., and Costa-Roberts, J.: Climate Trends and Global Crop Production Since 1980, Science, 333, 616–620, <https://doi.org/10.1126/science.1204531>, 2011.
- Luo, Q.: Temperature thresholds and crop production: a review, Climatic change, 109, 583–598, <https://doi.org/10.1007/s10584-011-0028-6>, 2011.
- Madsen, H., Lawrence, D., Lang, M., Martinkova, M., and Kjeldsen, T. R.: Review of trend analysis and climate change projections of extreme precipitation and floods in Europe, Journal of Hydrology, 519, 3634–3650, <https://doi.org/10.1016/j.jhydrol.2014.11.003>, 2014.
- Mahmood, N., Ahmad, B., Hassan, S., and Bakhsh, K.: Impact of temperature AND precipitation on rice productivity in rice-wheat cropping system of Punjab province, Journal of Animal and Plant Sciences, 22, 993–997, 2012.
- Mahto, S. S. and Mishra, V.: Does ERA-5 Outperform Other Reanalysis Products for Hydrologic Applications in India?, Journal of Geophysical Research: Atmospheres, 124, 9423–9441, <https://doi.org/10.1029/2019JD031155>, 2019.
- Malekinezhad, H. and Zare-Garizi, A.: Regional frequency analysis of daily rainfall extremes using L-moments approach, Atmósfera, 27, 411 – 427, [https://doi.org/10.1016/S0187-6236\(14\)70039-6](https://doi.org/10.1016/S0187-6236(14)70039-6), 2014.
- Manrique-Suñen, A., Gonzalez-Reviriego, N., Torralba, V., Cortesi, N., and Doblas-Reyes, F. J.: Choices in the verification of S2S forecasts and their implications for climate services, Monthly Weather Review, 148, 3995–4008, <https://doi.org/10.1175/MWR-D-20-0067.1>, 2020.
- Maraun, D.: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue, Journal of Climate, 26, 2137–2143, <https://doi.org/10.1175/JCLI-D-12-00821.1>, 2013.

- Mariotti, A., Ruti, P. M., and Rixen, M.: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort, *npj Climate and Atmospheric Science*, 1, 2–5, <https://doi.org/10.1038/s41612-018-0014-z>, 2018.
- Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., Dirmeyer, P. A., Ferranti, L., Johnson, N. C., Jones, J., Kirtman, B. P., Lang, A. L., Molod, A., Newman, M., Robertson, A. W., Schubert, S., Waliser, D. E., and Albers, J.: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond, *Bulletin of the American Meteorological Society*, 101, E608–E625, <https://doi.org/10.1175/BAMS-D-18-0326.1>, 2020.
- Marra, F., Armon, M., Borga, M., and Morin, E.: Orographic effect on extreme precipitation statistics peaks at hourly time scales, *Geophysical Research Letters*, 48, e2020GL091498, 2021.
- Martinkova, M. and Kysely, J.: Overview of observed clausius-clapeyron scaling of extreme precipitation in midlatitudes, *Atmosphere*, 11, 1–16, <https://doi.org/10.3390/ATMOS11080786>, 2020.
- Mason, I.: Dependence of the Critical Success Index on sample climate and threshold probability, *Aust. Meteorol. Mag.*, 37, 75–81, 1989.
- McDowell, N. G., Beerling, D. J., Breshears, D. D., Fisher, R. A., Raffa, K. F., and Stitt, M.: The interdependence of mechanisms underlying climate-driven vegetation mortality, *Trends in Ecology & Evolution*, 26, 523–532, <https://doi.org/10.1016/j.tree.2011.06.003>, 2011.
- McLeod, A., Xu, C., and Lai, Y.: *bestglm: Best Subset GLM and Regression Utilities*, URL <https://CRAN.R-project.org/package=bestglm>, r package version 0.37.3, 2020.
- Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., Ferranti, L., Ilynia, T., Kumar, A., Müller, W. A., Rixen, M., Robertson, A. W., Smith, D. M., Takaya, Y., Tuma, M., Vitart, F., White, C. J., Alvarez, M. S., Ardilouze, C., Attard, H., Baggett, C., Balmaseda, M. A., Beraki, A. F., Bhattacharjee, P. S., Bilbao, R., De Andrade, F. M., DeFlorio, M. J., Díaz, L. B., Ehsan, M. A., Fragkoulidis, G., Grainger, S., Green, B. W., Hell, M. C., Infanti, J. M., Isensee, K., Kataoka, T., Kirtman, B. P., Klingaman, N. P., Lee, J. Y., Mayer, K., McKay, R., Mecking, J. V., Miller, D. E., Neddermann, N., Ng, C. H. J., Ossó, A., Pankatz, K., Peatman, S., Pegion, K., Perlwitz, J., Recalde-Coronel, G. C., Reintges, A., Renkl, C., Solaraju-Murali, B., Spring, A., Stan, C., Sun, Y. Q., Tozer, C. R., Vigaud, N., Woolnough, S., and Yeager, S.: Current and emerging developments in subseasonal to decadal prediction, *Bulletin of the American Meteorological Society*, 101, E869–E896, <https://doi.org/10.1175/BAMS-D-19-0037.1>, 2020.
- Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A.: Impact Forecasting to Support Emergency Management of Natural Hazards, *Reviews of Geophysics*, 58, 1–52, <https://doi.org/10.1029/2020RG000704>, 2020.



- MeteoSwiss: Maps of extreme precipitation, [https://www.meteoswiss.admin.ch/home/climate/swiss-climate-in-detail/extreme-value-analyses/maps-of-extreme-precipitation.html?filters=1-day-sum\\_JJA\\_retlev\\_X100](https://www.meteoswiss.admin.ch/home/climate/swiss-climate-in-detail/extreme-value-analyses/maps-of-extreme-precipitation.html?filters=1-day-sum_JJA_retlev_X100), accessed: 2022-05-24, 2019.
- MeteoSwiss: Spatial Climate Analyses, <https://www.meteoswiss.admin.ch/home/climate/swiss-climate-in-detail/raeumliche-klimaanalysen.html>, accessed: 2022-05-24, 2020.
- Min, S. K., Zhang, X., Zwiers, F. W., and Hegerl, G. C.: Human contribution to more-intense precipitation extremes, *Nature*, 470, 378–381, <https://doi.org/10.1038/nature09763>, 2011.
- Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A.: Skill of Sub-seasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations, *Journal of Geophysical Research: Atmospheres*, 123, 7999–8016, <https://doi.org/10.1029/2017JD027923>, 2018.
- Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *Journal of Geophysical Research: Atmospheres*, 117, 2012.
- Moriondo, M. and Bindi, M.: Impact of climate change on the phenology of typical Mediterranean crops, *Italian Journal of Agrometeorology*, 3, 5–12, 2007.
- Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., and Foley, J. A.: Closing yield gaps through nutrient and water management, *Nature*, 490, 254–257, <https://doi.org/10.1038/nature11420>, 2012.
- Munich Re: Hurricanes, cold waves, tornadoes: Weather disasters in USA dominate natural disaster losses in 2021, <https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2022/natural-disaster-losses-2021.html>, accessed: 2022-05-24, 2022.
- MunichRE: A stormy year: Natural Catastrophes 2017, *Topics Geo*, p. 65, URL <https://www.munichre.com/topics-online/en/climate-change-and-natural-disasters/natural-disasters/topics-geo-2017.html>, 2018.
- Murty, M. N., Jain, A., and Flynn, P.: Data clustering: a review *ACM Comput. Surv*, *ACM Computing Surveys*, 31, 1999.
- NASA: Climate Models, <https://www.climate.gov/maps-data/climate-data-primer/predicting-climate/climate-models>, 2022.
- NASA and WorldBankGroup: Measuring precipitation: on the ground and from space, <https://olc.worldbank.org/sites/default/files/sco/E7B1C4DE-C187-5EDB-3EF2-897802DEA3BF/Nasa/chapter2.html>, 2022.
- Naveau, P., Toreti, A., Smith, I., and Xoplaki, E.: A fast nonparametric spatio-temporal regression scheme for generalized Pareto distributed heavy precipitation, *Water Resources Research*, 50, 4011–4017, 2014.

- Naveau, P., Huser, R., Ribereau, P., and Hannart, A.: Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection, *Water Resources Research*, 52, 2753–2769, <https://doi.org/10.1002/2015WR018552>, 2016.
- Nogueira, M.: Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: Process-based analysis of systematic and random differences, *Journal of Hydrology*, 583, 124 632, <https://doi.org/10.1016/j.jhydrol.2020.124632>, 2020.
- Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., Scott, R. L., Wang, L., and Phillips, R. P.: The increasing importance of atmospheric demand for ecosystem water and carbon fluxes, *Nature Climate Change*, 6, 1023–1027, <https://doi.org/10.1038/nclimate3114>, 2016.
- Olaniyan, E., Adefisan, E. A., Oni, F., Afiesimama, E., Balogun, A. A., and Lawal, K. A.: Evaluation of the ECMWF sub-seasonal to seasonal precipitation forecasts during the peak of West Africa Monsoon in Nigeria, *Frontiers in Environmental Science*, 6, 1–15, <https://doi.org/10.3389/fenvs.2018.00004>, 2018.
- Oppenheimer, M., Campos, M., Warren, R., Birkmann, J., Luber, G., O'Neill, B., Takahashi, K., Brklacich, M., Semenov, S., Licker, R., et al.: Emergent risks and key vulnerabilities, in: *Climate Change 2014 Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects*, pp. 1039–1100, Cambridge University Press, <https://doi.org/10.1017/CBO9781107415379.024>, 2015.
- Ouarda, T., St-Hilaire, A., and Bobée, B.: Synthèse des développements récents en analyse régionale des extrêmes hydrologiques, *Revue des sciences de l'eau/Journal of Water Science*, 21, 219–232, 2008.
- Pan, S., Yang, J., Tian, H., Shi, H., Chang, J., Ciais, P., Francois, L., Frieler, K., Fu, B., Hickler, T., Ito, A., Nishina, K., Ostberg, S., Reyer, C. P., Schaphoff, S., Steinkamp, J., and Zhao, F.: Responses of terrestrial carbon fluxes to temperature and precipitation: carbon extreme versus climate extreme, *Journal of Geophysical Research: Biogeosciences*, p. e2019JG005252, <https://doi.org/10.1029/2019JG005252>, 2020.
- Pansera, W. A., Gomes, B. M., Boas, A., and Mello, E. L.: Clustering rainfall stations aiming regional frequency analysis, *Journal of Food, Agriculture & Environment*, 11, 877–885, 2013.
- Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U.: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble, *Quarterly Journal of the Royal Meteorological Society*, 144, 1864–1881, <https://doi.org/10.1002/qj.3380>, 2018.
- Panziera, L., Gabella, M., Germann, U., and Martins, O.: A 12-year radar-based climatology of daily and sub-daily extreme precipitation over the Swiss Alps, *International Journal of Climatology*, 38, 3749–3769, <https://doi.org/10.1002/joc.5528>, 2018.
- Papastathopoulos, I. and Tawn, J. A.: Extended generalised Pareto models for tail estimation, *Journal of Statistical Planning and Inference*, 143, 131–143, 2013.

- Pendergrass, A. G. and Hartmann, D. L.: Changes in the distribution of rain frequency and intensity in response to global warming, *Journal of Climate*, 27, 8372–8383, <https://doi.org/10.1175/JCLI-D-14-00183.1>, 2014.
- Pfahl, S. and Wernli, H.: Quantifying the Relevance of Cyclones for Precipitation Extremes, *Journal of Climate*, 25, 6770–6780, <https://doi.org/10.1175/JCLI-D-11-00705.1>, 2012.
- Pic, R., Dombry, C., Naveau, P., and Taillardat, M.: Mathematical Properties of Continuous Ranked Probability Score Forecasting, URL <http://arxiv.org/abs/2205.04360>, 2022.
- Pickands III, J. et al.: Statistical inference using extreme order statistics, *Annals of statistics*, 3, 119–131, 1975.
- Porter, J. R. and Gawith, M.: Temperatures and the growth and development of wheat: a review, *European Journal of Agronomy*, 10, 23–36, [https://doi.org/10.1016/S1161-0301\(98\)00047-1](https://doi.org/10.1016/S1161-0301(98)00047-1), 1999.
- Poschlod, B.: Using high-resolution regional climate models to estimate return levels of daily extreme precipitation over Bavaria, *Natural Hazards and Earth System Sciences Discussions*, pp. 1–32, 2021.
- Poschlod, B., Ludwig, R., and Sillmann, J.: Ten-year return levels of sub-daily extreme precipitation over Europe, *Earth System Science Data*, 13, 983–1003, 2021.
- Prahl, B. F., Boettle, M., Costa, L., Kropp, J. P., and Rybski, D.: Data Descriptor: Damage and protection cost curves for coastal floods within the 600 largest European cities, *Scientific Data*, 5, 1–18, <https://doi.org/10.1038/sdata.2018.34>, 2018.
- Prein, A. F. and Gobiet, A.: Impacts of uncertainties in European gridded precipitation observations on regional climate analysis, *International Journal of Climatology*, 37, 305–327, <https://doi.org/10.1002/joc.4706>, 2017.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>, 2019.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>, 2020.
- Rawson, H. M., Begg, J. E., and Woodward, R. G.: The Effect of Atmospheric Humidity on Photosynthesis, Transpiration and Water Use Efficiency of Leaves of Several Plant Species, *Planta*, 134, 5–10, <https://doi.org/10.1007/BF00390086>, 1977.
- Reder, A. and Rianna, G.: Exploring ERA5 reanalysis potentialities for supporting landslide investigations: a test case from Campania Region (Southern Italy), *Landslides*, <https://doi.org/10.1007/s10346-020-01610-4>, 2021.

- Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J.: Clustering rules: a comparison of partitioning and hierarchical clustering algorithms, *Journal of Mathematical Modelling and Algorithms*, 5, 475–504, 2006.
- Rhodes, R. I., Shaffrey, L. C., and Gray, S. L.: Can reanalyses represent extreme precipitation over England and Wales?, *Quarterly Journal of the Royal Meteorological Society*, 141, 1114–1120, <https://doi.org/10.1002/qj.2418>, 2015.
- Ribeiro, A. F. S., Russo, A., Gouveia, C. M., Páscoa, P., and Zscheischler, J.: Risk of crop failure due to compound dry and hot extremes estimated with nested copulas, *Biogeosciences*, 17, 4815–4830, <https://doi.org/10.5194/bg-17-4815-2020>, 2020.
- Rivoire, P., Le Gall, P., Favre, A.-C., Naveau, P., and Martius, O.: High return level estimates of daily ERA-5 precipitation in Europe estimated using regionalised extreme value distributions, URL <http://arxiv.org/abs/2112.02182>, 2021a.
- Rivoire, P., Martius, O., and Naveau, P.: A Comparison of Moderate and Extreme ERA-5 Daily Precipitation With Two Observational Data Sets, *Earth and Space Science*, 8, <https://doi.org/10.1029/2020EA001633>, 2021b.
- Rosenzweig, C., Tubiello, F. N., Goldberg, R., Mills, E., and Bloomfield, J.: Increased crop damage in the US from excess precipitation under climate change, *Global Environmental Change*, 12, 197–202, [https://doi.org/10.1016/S0959-3780\(02\)00008-0](https://doi.org/10.1016/S0959-3780(02)00008-0), 2002.
- Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., et al.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, *Proceedings of the National Academy of Sciences*, 111, 3268–3273, <https://doi.org/10.1073/pnas.1222463110>, 2014.
- Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, 20, 53–65, 1987.
- Ruane, A. C., Goldberg, R., and Chryssanthacopoulos, J.: Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation, *Agricultural and Forest Meteorology*, 200, 233–248, <https://doi.org/10.1016/j.agrformet.2014.09.016>, 2015.
- Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, *Nature Communications*, 10, 2553, <https://doi.org/10.1038/s41467-019-10105-3>, 2019.
- S2S-challenge: Challenge to improve Sub-seasonal to Seasonal Predictions using Artificial Intelligence, <https://s2s-ai-challenge.github.io/>, accessed: 2022-03-05, 2021.

- Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, *Global Ecology and Biogeography*, 19, 607–620, <https://doi.org/10.1111/j.1466-8238.2010.00551.x>, 2010.
- Saf, B.: Regional flood frequency analysis using L-moments for the West Mediterranean region of Turkey, *Water Resources Management*, 23, 531–551, 2009.
- Sang, H. and Gelfand, A. E.: Hierarchical modeling for extreme values observed over space and time, *Environmental and ecological statistics*, 16, 407–426, 2009.
- Schaefer, J. T.: The Critical Success Index as an Indicator of Warning Skill, *Weather and Forecasting*, pp. 570–575, 1990.
- Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Pugh, T. A. M., Rolinski, S., Schaphoff, S., Schmid, E., Wang, X., Schlenker, W., and Frieler, K.: Consistent negative response of US crops to high temperatures in observations and crop models, *Nature communications*, 8, 13 931, <https://doi.org/10.1038/ncomms13931>, 2017.
- Scherrer, S. C., Fischer, E. M., Posselt, R., Liniger, M. A., Croci-Maspoli, M., and Knutti, R.: Emerging trends in heavy precipitation and hot temperature extremes in Switzerland, *Journal of Geophysical Research*, 121, 2626–2637, <https://doi.org/10.1002/2015JD024634>, 2016.
- Schmidli, J., Schmutz, C., Frei, C., Wanner, H., and Schar, C.: Mesoscale precipitation variability in the region of the European Alps during the 20th century, *International J. Climatol.*, 22, 1049–1074, 2002.
- Scholz, F. W. and Stephens, M. A.: K-sample Anderson–Darling tests, *Journal of the American Statistical Association*, 82, 918–924, 1987.
- Schubert, E. and Rousseeuw, P. J.: Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms, *Information Systems*, p. 101804, 2021.
- Seyfert, F.: Phänologie, vol. 255 of *Die neue Brehm-Bücherei*, VerlagsKG Wolf, Magdeburg, nachdr., 2. unveränd. Aufl. edn., 1960.
- Shah, N. and Paulsen, G.: Interaction of drought and high temperature on photosynthesis and grain-filling of wheat, *Plant and Soil*, 257, 219–226, <https://doi.org/10.1023/a:1026237816578>, 2003.
- Sharifi, E., Eitzinger, J., and Dorigo, W.: Performance of the state-of-the-art gridded precipitation products over mountainous terrain: A regional study over Austria, *Remote Sensing*, 11, 1–20, <https://doi.org/10.3390/rs11172018>, 2019.
- Shi, W., Tao, F., and Zhang, Z.: A review on statistical models for identifying climate contributions to crop yields, *Journal of Geographical Sciences*, 23, 567–576, <https://doi.org/10.1007/s11442-013-1029-3>, 2013.

- Singh, A., Phadke, V. S., and Patwardhan, A.: Impact of Drought and Flood on Indian Food Grain Production, in: *Challenges and Opportunities in Agrometeorology*, pp. 421–433, Springer Berlin Heidelberg, [https://doi.org/10.1007/978-3-642-19360-6\\_32](https://doi.org/10.1007/978-3-642-19360-6_32), 2011.
- Sippel, S., Zscheischler, J., and Reichstein, M.: Ecosystem impacts of climate extremes crucially depend on the timing, *Proceedings of the National Academy of Sciences*, 113, 5768–5770, <https://doi.org/10.1073/pnas.1605667113>, 2016.
- Specq, D. and Batté, L.: Improving subseasonal precipitation forecasts through a statistical–dynamical approach : application to the southwest tropical Pacific, *Climate Dynamics*, 55, 1913–1927, <https://doi.org/10.1007/s00382-020-05355-7>, 2020.
- St-Hilaire, A., Ouarda, T., Lachance, M., Bobée, B., Barbet, M., and Bruneau, P.: La régionalisation des précipitations: une revue bibliographique des développements récents, *Revue des sciences de l’eau/Journal of Water Science*, 16, 27–54, 2003.
- Stan, C., Zheng, C., Chang, E. K. M., Domeisen, D. I. V., Garfinkel, C. I., Jenney, A. M., Kim, H., Lim, Y.-K., Lin, H., Robertson, A., Schwartz, C., Vitart, F., Wang, J., and Yadav, P.: Advances in the prediction of MJO-Teleconnections in the S2S forecast systems, *Bulletin of the American Meteorological Society*, -1, 2022.
- Stein, M. L.: Parametric models for distributions when interest is in extremes with an application to daily temperature, *Extremes*, <https://doi.org/10.1007/s10687-020-00378-z>, <https://doi.org/10.1007/s10687-020-00378-z>, 2020.
- Stocker, B. D., Zscheischler, J., Keenan, T. F., Prentice, I. C., Seneviratne, S. I., and Peñuelas, J.: Drought impacts on terrestrial primary production underestimated by satellite monitoring, *Nature Geoscience*, 12, 264–270, <https://doi.org/10.1038/s41561-019-0318-6>, 2019.
- Stocker, T., Qin, D., Plattner, G.-K., Alexander, L., Allen, S., Bindoff, N., Bréon, F.-M., Church, J., Cubasch, U., Emori, S., Forster, P., Friedlingstein, P., Gillett, N., Gregory, J., Hartmann, D., Jansen, E., Kirtman, B., Knutti, R., Krishna Kumar, K., Lemke, P., Marotzke, J., Masson-Delmotte, V., Meehl, G., Mokhov, I., Piao, S., Ramaswamy, V., Randall, D., Rhein, M., Rojas, M., Sabine, C., Shindell, D., Talley, L., Vaughan, D., and Xie, S.-P.: Technical Summary, book section TS, p. 33–115, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, <https://doi.org/10.1017/CBO9781107415324.005>, 2013.
- Sugar, C. A. and James, G. M.: Finding the number of clusters in a dataset: An information-theoretic approach, *Journal of the American Statistical Association*, 98, 750–763, 2003.
- Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L.: A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons, *Reviews of Geophysics*, 56, 79–107, <https://doi.org/10.1002/2017RG000574>, 2018.
- Tabari, H. and Willems, P.: Lagged influence of Atlantic and Pacific climate patterns on European extreme precipitation, *Scientific Reports*, 8, 1–10, <https://doi.org/10.1038/s41598-018-24069-9>, 2018.

- Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R.: Extreme events evaluation using CRPS distributions, URL <http://arxiv.org/abs/1905.04022>, 2022.
- Tapiador, F. J., Turk, F. J., Petersen, W., Hou, A. Y., García-Ortega, E., Machado, L. A., Angelis, C. F., Salio, P., Kidd, C., Huffman, G. J., and de Castro, M.: Global precipitation measurement: Methods, datasets and applications, *Atmospheric Research*, 104-105, 70–97, <https://doi.org/10.1016/j.atmosres.2011.10.021>, 2012.
- Tarek, M., Brissette, F. P., and Arsenault, R.: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America, *Hydrology and Earth System Sciences*, 24, 2527–2544, <https://doi.org/10.5194/hess-24-2527-2020>, 2020.
- Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G.: Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount, *Environmetrics*, 31, e2582, <https://doi.org/https://doi.org/10.1002/env.2582>, 2020.
- Tian, D., Wood, E., and Yuan, X.: CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States, *Hydrology and Earth System Sciences*, 21, 1477–1490, <https://doi.org/10.5194/hess-21-1477-2017>, 2017.
- Tian, Y., Peters-Lidard, C. D., Choudhury, B. J., and Garcia, M.: Multitemporal analysis of TRMM-based satellite precipitation products for land data assimilation applications, *Journal of Hydrometeorology*, 8, 1165–1183, <https://doi.org/10.1175/2007JHM859.1>, 2007.
- Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, 58, 267–288, 1996.
- Timmermans, B., Wehner, M., Cooley, D., O’Brien, T., and Krishnan, H.: An evaluation of the consistency of extremes in gridded precipitation data sets, *Climate Dynamics*, 52, 6651–6670, <https://doi.org/10.1007/s00382-018-4537-0>, 2019.
- Touma, D., Stevenson, S., Swain, D. L., Singh, D., Kalashnikov, D. A., and Huang, X.: Climate change increases risk of extreme rainfall following wildfire in the western United States, *Science Advances*, 8, 1–12, <https://doi.org/10.1126/sciadv.abm0320>, 2022.
- Tramblay, Y., Neppel, L., Carreau, J., and Najib, K.: Non-stationary frequency analysis of heavy rainfall events in southern France, *Hydrological Sciences Journal*, 58, 280–294, <https://doi.org/10.1080/02626667.2012.754988>, 2013.
- Tschumi, E. and Zscheischler, J.: Countrywide climate features during recorded climate-related disasters, *Climatic Change*, 158, 593–609, <https://doi.org/10.1007/s10584-019-02556-w>, 2020.
- Umbricht, A., Fukutome, S., Liniger, M. A. and Frei, C., and Appenzeller, C.: Seasonal variation of daily extreme precipitation in Switzerland, Report, MeteoSwiss, 2013.
- Van der Wiel, K., Stoop, L., van Zuijlen, B., Blackport, R., van den Broek, M., and Selten, F.: Meteorological conditions leading to extreme low variable renewable energy production and

- extreme high energy shortfall, *Renewable and Sustainable Energy Reviews*, 111, 261 – 275, <https://doi.org/10.1016/j.rser.2019.04.065>, 2019a.
- Van der Wiel, K., Wanders, N., Selten, F., and Bierkens, M.: Added value of large ensemble simulations for assessing extreme river discharge in a 2 °C warmer world, *Geophysical Research Letters*, 46, 2093–2102, <https://doi.org/10.1029/2019GL081967>, 2019b.
- Van der Wiel, K., Selten, F. M., Bintanja, R., Blackport, R., and Screen, J. A.: Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions, *Environmental Research Letters*, 15, 034 050, <https://doi.org/10.1088/1748-9326/ab7668>, 2020.
- Van Rossum, G. and Drake Jr, F. L.: *Python reference manual*, Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- Vautard, R., Yiou, P., van Oldenborgh, G.-J., Lenderink, G., Thao, S., Ribes, A., Planton, S., Dubuisson, B., and Soubeyroux, J.-M.: Extreme Fall 2014 Precipitation in the Cévennes Mountains, *Bulletin of the American Meteorological Society*, 96, S56–S60, <https://doi.org/10.1175/BAMS-D-15-00088.1>, 2015.
- Vitart, F.: List of parameters of S2S reforecast data from ECMWF, <https://confluence.ecmwf.int/display/S2S/Parameters>, accessed: 2022-05-24, 2020.
- Vitart, F. and Robertson, A. W.: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events, *npj Climate and Atmospheric Science*, 1, 1–7, <https://doi.org/10.1038/s41612-018-0013-0>, 2018.
- Vitart, F., Robertson, A. W., and Anderson, D. L.: Subseasonal to Seasonal Prediction Project: Bridging the Gap between Weather and Climate, *WMO Bulletin*, 61, 2012.
- Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N., and Frieler, K.: The effects of climate extremes on global agricultural yields, *Environmental Research Letters*, 14, 054 010, <https://doi.org/10.1088/1748-9326/ab154b>, 2019.
- Vogel, J., Rivoire, P., Deidda, C., Rahimi, L., Sauter, C. A., Tschumi, E., Van Der Wiel, K., Zhang, T., and Zscheischler, J.: Identifying meteorological drivers of extreme impacts: An application to simulated crop yields, *Earth System Dynamics*, 12, 151–172, <https://doi.org/10.5194/esd-12-151-2021>, 2021.
- Wang, C., Graham, R. M., Wang, K., Gerland, S., and Granskog, M. A.: Comparison of ERA5 and ERA-Interim near surface air temperature and precipitation over Arctic sea ice: Effects on sea ice thermodynamics and evolution, *The Cryosphere Discussions*, pp. 1–28, <https://doi.org/10.5194/tc-2018-245>, 2018.
- WCS: The Difference Between Deterministic and Ensemble Forecasts, <https://www.worldclimateservice.com/2021/10/12/difference-between-deterministic-and-ensemble-forecasts/>, 2021.



- Westra, S., Alexander, L. V., and Zwiers, F. W.: Global increasing trends in annual maximum daily precipitation, *Journal of Climate*, 26, 3904–3918, <https://doi.org/10.1175/JCLI-D-12-00502.1>, 2013.
- White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E.: Potential applications of subseasonal-to-seasonal (S2S) predictions, *Meteorological Applications*, 24, 315–325, <https://doi.org/10.1002/met.1654>, 2017.
- White, C. J., Domeisen, D. I. V., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw, D. J., Browell, J., Büeler, D., Charlton-perez, A., Christel, I., Coelho, C. A. S., Deflorio, M. J., Monache, D., Giuseppe, F. D., García-solórzano, A. M., Gibson, P. B., Goddard, L., Romero, C. G., Graham, R. J., Graham, R. M., Grams, C. M., Halford, A., Huang, W. T. K., Jensen, K., Kilavi, M., Lawal, K. A., Lee, W., Macleod, D., Manrique-suñén, A., Martins, E. S. P. R., Carolyn, J., Merryfield, W. J., Muñoz, Á. G., Olaniyan, E., Otieno, G., Oyedepo, A., Palma, L., Pechlivanidis, I. G., Pons, D., Ralph, F. M., Dirceu, S., Jr, R., Remenyi, T. A., Risbey, J. S., Robertson, D. J. C., Andrew, W., Smith, S., Soret, A., Sun, T., Todd, M. C., Tozer, C. R., Jr, F. C. V., Vigo, I., Waliser, D. E., Wetterhall, F., and Wilson, G.: Advances in the application and utility of subseasonal-to-seasonal predictions, *Bulletin of the American Meteorological Society*, pp. 1–57, <https://doi.org/10.1175/bams-d-20-0224.1>, 2021.
- Wilks, D. S.: Statistical Forecasting, vol. 100, International Geophysics, <https://doi.org/10.1016/B978-0-12-385022-5.00007-5>, 2011.
- WMO: Weather-related disasters increase over past 50 years, causing more damage but fewer deaths, <https://public.wmo.int/en/media/press-release/weather-related-disasters-increase-over-past-50-years-causing-more-damage-fewer>, accessed: 2022-04-25, 2021.
- Xie, P., Joyce, R., Wu, S., Yoo, S. H., Yarosh, Y., Sun, F., and Lin, R.: Reprocessed, bias-corrected CMORPH global high-resolution precipitation estimates from 1998, *Journal of Hydrometeorology*, 18, 1617–1641, <https://doi.org/10.1175/JHM-D-16-0168.1>, 2017.
- Xu, X., Frey, S. K., Boluwade, A., Erler, A. R., Khader, O., Lapen, D. R., and Sudicky, E.: Evaluation of variability among different precipitation products in the Northern Great Plains, *Journal of Hydrology: Regional Studies*, 24, 100 608, <https://doi.org/10.1016/j.ejrh.2019.100608>, 2019.
- Yan, Y., Liu, B., Zhu, C., Lu, R., Jiang, N., and Ma, S.: Subseasonal forecast barrier of the North Atlantic oscillation in S2S models during the extreme mei-yu rainfall event in 2020, *Climate Dynamics*, 0123456789, <https://doi.org/10.1007/s00382-021-06076-1>, 2021.

- Yu, Z., Yu, H., Chen, P., Qian, C., and Yue, C.: Verification of tropical cyclone-related satellite precipitation estimates in mainland China, *Journal of Applied Meteorology and Climatology*, 48, 2227–2241, <https://doi.org/10.1175/2009JAMC2143.1>, 2009.
- Yuan, W., Zheng, Y., Piao, S., Ciais, P., Lombardozzi, D., Wang, Y., Ryu, Y., Chen, G., Dong, W., Hu, Z., Jain, A. K., Jiang, C., Kato, E., Li, S., Lienert, S., Liu, S., Nabel, J. E., Qin, Z., Quine, T., Sitch, S., Smith, W. K., Wang, F., Wu, C., Xiao, Z., and Yang, S.: Increased atmospheric vapor pressure deficit reduces global vegetation growth, *Science Advances*, 5, eaax1396, <https://doi.org/10.1126/sciadv.aax1396>, 2019.
- Zhang, L., Kim, T., Yang, T., Hong, Y., and Zhu, Q.: Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous U.S., *Journal of Hydrology*, 603, 127 058, <https://doi.org/10.1016/j.jhydrol.2021.127058>, 2021a.
- Zhang, Q., Xiao, M., Singh, V. P., and Li, J.: Regionalization and spatial changing properties of droughts across the Pearl River basin, China, *Journal of Hydrology*, 472, 355–366, 2012.
- Zhang, S., Tao, F., and Zhang, Z.: Spatial and temporal changes in vapor pressure deficit and their impacts on crop yields in China during 1980–2008, *Journal of Meteorological Research*, 31, 800–808, <https://doi.org/10.1007/s13351-017-6137-z>, 2017.
- Zhang, X., Nie, J., Cheng, C., Xu, C., Xu, X., and Yan, B.: Spatial pattern of the population casualty rate caused by super typhoon Lekima and quantification of the interactive effects of potential impact factors, *BMC public health*, 21, 1260, <https://doi.org/10.1186/s12889-021-11281-y>, 2021b.
- Zheng, B., Chenu, K., Doherty, A., and Chapman, S.: The APSIM-wheat module (7.5 R3008), *Agricultural Production Systems Simulator (APSIM) Initiative*, 2014.
- Zscheischler, J., Mahecha, M. D., Harmeling, S., and Reichstein, M.: Detection and attribution of large spatiotemporal extreme events in Earth observation data, *Ecological Informatics*, 15, 66–73, <https://doi.org/10.1016/j.ecoinf.2013.03.004>, 2013.
- Zscheischler, J., Fatichi, S., Wolf, S., Blanken, P. D., Bohrer, G., Clark, K., Desai, A. R., Hollinger, D., Keenan, T., Novick, K. A., and Seneviratne, S. I.: Short-term favorable weather conditions are an important control of interannual variability in carbon and water fluxes, *Journal of Geophysical Research: Biogeosciences*, 121, 2186–2198, <https://doi.org/10.1002/2016JG003503>, 2016.
- Zscheischler, J., Westra, S., van den Hurk, B., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., and Zhang, X.: Future climate risk from compound events, *Nature Climate Change*, 8, 469–477, <https://doi.org/10.1038/s41558-018-0156-3>, 2018.

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



## APPENDIX

### A

# IDENTIFYING METEOROLOGICAL DRIVERS OF EXTREME IMPACTS: AN APPLICATION TO SIMULATED CROP YIELDS

This chapter contains an article published in 2021 with the title “Identifying meteorological drivers of extreme impacts: an application to simulated crop yields” in the journal *Earth System Dynamics* (Vogel et al., 2021). Johannes Vogel and Pauline Rivoire contributed equally to this paper. This article was written together with Cristina Deidda, Leila Rahimi, Christoph A. Sauter, Elisabeth Tschumi, Karin van der Wiel, Tianyi Zhang, and Jakob Zscheischler. This work emerged from the Training School on Statistical Modelling organized by the European COST Action DAMOCLES.

## Abstract

Compound weather events may lead to extreme impacts that can affect many aspects of society including agriculture. Identifying the underlying mechanisms that cause extreme impacts, such as crop failure, is of crucial importance to improve their understanding and forecasting. In this study we investigate whether key meteorological drivers of extreme impacts can be identified using Least Absolute Shrinkage and Selection Operator (Lasso) in a model environment, a method that allows for automated variable selection and is able to handle collinearity between variables. As an example of an extreme impact, we investigate crop failure using annual wheat yield as simulated by the APSIM crop model driven by 1600 years of daily weather data from a global climate model (EC-Earth) under present-day conditions for the Northern Hemisphere. We then apply Lasso logistic regression to determine which weather conditions during the growing season lead to crop failure. We obtain good model performance in Central Europe and the eastern half of the United States, while crop failure years in regions in Asia and the western half of the United States are less accurately predicted. Model performance correlates strongly with annual mean and variability of crop yields, that is, model performance is highest in regions with relatively large annual crop yield mean and variability. Overall, for nearly all grid points the inclusion of temperature, precipitation and vapour pressure deficit is key to predict crop failure. In addition, meteorological predictors during all seasons are required for a good prediction. These results illustrate the omnipresence of compounding effects of both meteorological drivers and different periods of the growing season for creating crop failure events. Especially vapour pressure deficit and climate extreme indicators such as diurnal temperature range and the number of frost days are selected by the statistical model as relevant predictors for crop failure at most grid points, underlining their overarching relevance. We conclude that the Lasso regression model is a useful tool to automatically detect compound drivers of extreme impacts, and could be applied to other weather impacts such as wildfires or floods. As the detected relationships are of purely correlative nature, more detailed analyses are required to establish the causal structure between drivers and impacts.

## A.1 Introduction

Climate extremes such as droughts, heatwaves, floods and frost events can have substantial impacts on crop health (Shah and Paulsen, 2003; Singh et al., 2011; Lesk et al., 2016; Ben-Ari et al., 2018). However, not all climate extremes lead to an extreme impact, and large impacts can be related to moderate drivers (Zscheischler et al., 2016; Van der Wiel et al., 2019a, 2020; Pan et al., 2020). Whether a large impact occurs does not only depend on a climate hazard but also on the vulnerability of the underlying system (Oppenheimer et al., 2015), which varies strongly for crops during the course of the growing season (Iizumi and Ramankutty, 2015; Ben-Ari et al., 2018). The mechanisms that translate a climate hazard into crop failure are often very complex and associated with lagged effects that are difficult to disentangle (Frank et al., 2015).

While climate extremes may lead to large impacts, extreme climate-related impacts are often the result of multiple contributing factors (Tschumi and Zscheischler, 2020). The concept of compound events has recently been promoted to address climate impacts from an impact-centred perspective. For instance, compound events have been defined as extreme impacts that depend on multiple statistically dependent drivers (Leonard et al., 2014) or, more recently, simply as the combination of multiple drivers that contributes to environmental or societal risk (Zscheischler et al., 2018). Drivers in this context refer to climate and weather processes and phenomena. With respect to yields at the local scale, multiple drivers can compound an impact through a sequence of weather events (temporally compounding); one weather event may also change the vulnerability of the crop to a subsequent weather event (preconditioning); or multiple drivers may interact and impact crops at the same time (multivariate events) (Zscheischler et al., 2020). Understanding the drivers that lead to extreme impacts helps to better predict and mitigate the potential impacts of such events. One way of identifying the relevant drivers of an impact is to perform a bottom-up analysis, that is, start from an impact and identify key drivers through statistical analysis (Zscheischler et al., 2013; Ben-Ari et al., 2018). In this context, linear regression analysis can identify the most relevant drivers of an impact variable and reveal potential interactions between drivers (Forkel et al., 2012; Ben-Ari et al., 2018). More sophisticated approaches such as random forest might yield higher predictive power at the cost of losing explainability (Vogel et al., 2019). When the set of possible predictors is very large, suitable variable selection approaches need to be applied to reduce the number of predictors. In order to be applicable to a large number of locations and a variety of impacts, an automatic approach is desired that only requires a limited amount of expert knowledge and parameter tuning. An example of such an approach is the Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996), or short Lasso regression, which obtains a reduced number of predictors by penalizing the number of variables in the loss function.

The aim of this study is to present a method that can identify drivers of extreme impacts in an automatic manner and that is suitable for many applications. We use crop failure as an example of an extreme impact in a model environment, that is, we use simulated data from a climate and a crop model. End-of-season crop yield is related to climate drivers via highly complex interactions at different temporal scales. Temperature and precipitation are the two

basic climate variables that regulate crop health (Lobell and Asner, 2003; Lobell et al., 2011; Leng et al., 2016). Furthermore, vapour pressure deficit (VPD), the difference of water vapour pressure at saturated condition and its actual value at a given temperature, determines crop photosynthesis and water demand (Rawson et al., 1977; Zhang et al., 2017; Yuan et al., 2019).

Here we use 1600 years of wheat yield data from a global gridded crop model driven by simulated meteorological data under present-day conditions. Based on this large database of yield data we showcase approaches to identify multiple drivers of crop failure in different regions of the world and highlight results for the Lasso regression. Using a model environment to explore new analytical approaches to identify drivers of extreme impacts, we circumvent common limitations associated with observational data, such as a small sample size, measurement uncertainties and data coverage. Among the large amount of information provided by the crop model simulations, the statistical model summarizes the link between crop failure and climate conditions.

This paper is structured as follows. The data and methods used in this study are introduced in section A.2. In this section, the reader can first find a description of the data, including an introduction to the global climate model and the crop model used in this study. We further describe which meteorological variables are considered in the statistical analysis; section A.2 also introduces the Lasso logistic regression to predict years of low yield based on meteorological drivers and the metrics employed to assess the performance of the statistical model. The results of the Lasso regression are shown in section A.3, where the performance and the summary statistics for the variables that have been selected as being critical to predict crop failure events are presented. Finally, we summarize and discuss the Lasso regression’s results in section A.4, and give some perspective to this study in section A.5.

## A.2 Data and Methods

### A.2.1 Climate and crop model simulations

To investigate the influence of natural variability and climatic extreme events, a large ensemble simulation experiment was set up with the EC-Earth global climate model (v2.3, Hazeleger et al., 2012). We use this climate model dataset, consisting of 2000 years of present-day simulated weather, to investigate if we can identify the drivers of extreme low crop yield seasons. Large ensemble modelling is at the forefront of climate science (Deser et al., 2020), due to the computational expenses involved a balance between ensemble size, horizontal resolution and number of climate models has to be found. We have found the climate data used here to be suitable for the present study. A detailed description of these climate simulations is provided in Van der Wiel et al. (2019b), here we provide a short overview of the experimental setup. Present-day was defined as the five year model period in which the simulated global mean surface temperature matched that observed in 2011-2015 (HadCRUT4 data, Morice et al., 2012). Because of a cold bias in EC-Earth, in the model this period is 2035-2039. To create the large ensemble, twenty-five ensemble members were branched off from sixteen long transient climate runs (forced by Representative Concentration Pathway (RCP) 8.5). Each ensemble member was integrated for five years. Differences between ensemble members were forced by choosing different seeds



in the atmospheric stochastic perturbations (Buizza et al., 1999). This resulted in a total of  $16 \times 25 \times 5 = 2000$  years of meteorological data, at T159 horizontal resolution (approximately  $1^\circ$ ).

Biases in the EC-Earth simulations result in unrealistic growing conditions for crops. Therefore, minimum and maximum temperatures and precipitation fields were bias-corrected. The AgMERRA reanalysis (Ruane et al., 2015) was used as ‘truth’. From AgMERRA the years 1981-2010 were used as a training set, while EC-Earth uses the long transient runs (sixteen  $\times$  2005-2034). Daily minimum and maximum temperatures were corrected on a grid point basis, a model bias field was defined as the difference between the model climatology and the AgMERRA climatology. The climatology was defined to be the mean plus the first three annual harmonics. Daily precipitation was corrected towards having the correct number of rainy days and total amount of precipitation. Firstly, for each month the number of rainy days in AgMERRA was computed (threshold 0.1 mm/day), then the same threshold was determined for EC-Earth data, which resulted in the same number of rainy days. All days with simulated precipitation smaller than this threshold were set to 0 mm/day. Lastly, the total amount of precipitation was corrected by means of a multiplicative factor, also on a month-by-month basis. Other meteorological variables were not bias-corrected.

Northern Hemisphere winter wheat yields were simulated using the APSIM-Wheat model (Zheng et al., 2014), which is a process-based model incorporating wheat physiology, water and nitrogen processes under a wide range of growing conditions. It was previously used for field (Li et al., 2014), regional (Asseng et al., 2013) and global scale (Rosenzweig et al., 2014) wheat studies. A grid point-specific sowing date was used based on Sacks et al. (2010). The application of nitrogen was exacted from Mueller et al. (2012). Soil parameters (including pH, soil total nitrogen, organic carbon content, bulk density and soil moisture characteristics curves for each of five 20 cm deep soil layers) were derived from the International Soil Profile dataset (Batjes, 2012). In addition, we also input the grid-specific thermal time accumulation parameters, which were derived from phenology (Sacks et al., 2010) and AgMERRA data. The atmospheric CO<sub>2</sub> concentration was set to 394 ppm. The growing season of winter wheat spans two calendar years (e.g. sowing in November and harvest in June). As such, each climate model integration of five years covers four winter wheat growing seasons, the 2000 years of EC-Earth climate data thus result in 1600 simulated wheat growing seasons. Further details on the settings of the APSIM-Wheat model can be found in Appendix A.5.1. For model validation, the grid-based wheat yield simulations were aggregated to country-level and then validated against the yield statistics during 2011-2015 (FAOSTAT, 2020). Most simulated yields are closely related to observed yields (Fig. A.9), indicating a good model performance.

### A.2.2 Data processing

The APSIM model provided crop data for 995 grid points in the Northern Hemisphere. For our analysis, we chose to discard all grid points for which the annual mean yield is below the 10<sup>th</sup> percentile of annual mean yield across all grid points because many of these grid points were also associated with unrealistically long ( $>365$  days) or short ( $<90$  days) growing seasons or had an

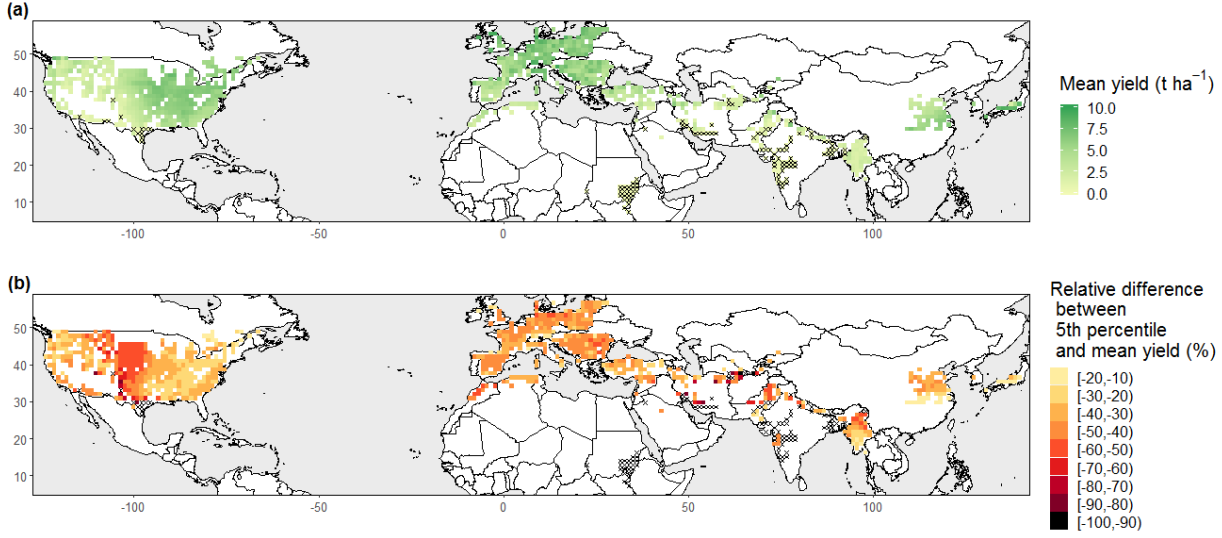


FIGURE A.1: (a) Mean annual yield over the 1600 years (ton/hectare). (b) Relative difference between the 5<sup>th</sup> percentile and the mean annual yield. Grid points discarded for our study are crossed out (specified in the Section A.2.2).

overall average crop yield of 0 kg/ha. 895 grid points remained for the analysis.

At each grid point, a year with yield lower than the 5<sup>th</sup> percentile for this grid point is considered as a year with crop failure, and called “bad year” in the remainder, whereas all other years are referred to as “normal years”. Grid points for which the 5<sup>th</sup> percentile yield was equal to 0 were excluded to avoid the co-occurrence of years without yield in the bad and normal years. This excluded 6 more grid points so that 889 remained for further analysis. Figure A.1a shows the simulated mean annual yield and Fig. A.1b displays the relative difference between the 5<sup>th</sup> percentile and the mean annual yield. These two figures also indicate grid points that were discarded for further analysis. Finally, we discarded individual years with a growing season longer than 365 days, leading to a slightly smaller number of years than 1600 for 82 pixels, i.e. for about 5 % of the grid points.

The data was split into a training and testing dataset by randomly assigning 70 % of the data to the former and 30 % to the latter. For the logistic regression (Section A.2.4) explanatory variables and yield were normalised by rescaling them to a range of  $[-1, 1]$  for each grid point individually.

### A.2.3 Explanatory data analysis

The APSIM model uses six meteorological variables on a daily basis as input (dew point temperature ( $T_d$ ), precipitation (Pr), 10 m wind speed (Wind), incoming shortwave radiation (Rad), maximum temperature ( $T_{max}$ ), and minimum temperature ( $T_{min}$ )). From these variables, we additionally calculated vapour pressure deficit (VPD) as an important variable for plant growth (Rawson et al., 1977; Zhang et al., 2017; Yuan et al., 2019). For a given grid point, the sowing date is the same for the 1600 simulated years, but the harvest dates differ. We therefore define

TABLE A.1: *Meteorological drivers used in the analysis.*

Variable name	Description	Type
$T_{max}$	Maximum temperature	Monthly mean
VPD	Vapour-pressure deficit	Monthly mean
Pr	Precipitation	Monthly mean
dtr	Mean diurnal temperature range in the growing season	Clim. extr. ind.
frs	Number of frost days in the growing season	Clim. extr. ind.
TXx	Maximum temperature in the growing season	Clim. extr. ind.
TNn	Minimum temperature in the growing season	Clim. extr. ind.
Rx5day	Maximum five day precipitation sum in the growing season	Clim. extr. ind.
TX90p	Number of warm days in the growing season with daily maximum temperature above the 90 <sup>th</sup> percentile <sup>a</sup>	Clim. extr. ind.
TN10p	Number of cold days in the growing season with daily minimum temperature below the 10 <sup>th</sup> percentile <sup>a</sup>	Clim. extr. ind.

Note: Percentiles are grid point based, i.e. they are representative for the local climate. Clim. extr. ind. refers to climate extreme indicator.

the growing season for a given grid point as starting on the month containing the sowing date and finishing with the month containing the latest harvest date. Figure A.2 illustrates the temporal evolution of composites of these seven variables over the course of a growing season for normal (blue) and bad years (red) for one grid point in France (1.1° E, 47.7° N, Fig. A.2a). The composites provide some indication about which of the meteorological variables may contribute to crop failure. In addition, the temporal evolution of the two composites reveals during which part of the growing season the different variables are relevant. The various composites suggests that, for this grid point, 30-day Pr, VPD and  $T_{max}$  during the summer (June-August) have a high impact on crop yield (Figs. A.2c, f and h). The other variables appear to be less relevant (Figs. A.2b, d, e and g). Similar composites for grid points in the US (90.0° W, 44.3° N) and in China (118.1° E, 30.8° N) are shown in Figs. A.10 and A.11, respectively.

In addition to the seven meteorological variables, we considered seven climate extreme indicators as potential predictors of crop failure (mean diurnal temperature range, dtr; number of frost days, frs; maximum temperature, TXx; minimum temperature, TNn; maximum five day precipitation sum, Rx5day; number of warm days, TX90p; number of cold days, TN10p; following Vogel et al., 2019) (Table A.1). For both the monthly means of the meteorological variables, as well as for the growing season means/totals of the indicators of climate extremes we calculated the Pearson correlation coefficient between the variables and annual yield (Figs. A.3a and b for the same grid point as in Fig. A.2 and Figs. A.3c and d as average correlation over all grid points). These correlations are computationally and conceptionally very simple and together with Fig. A.2 serve as a first estimation of the importance of the available variables. Some variables, such as wind speed, do not have a discernible influence on yield and thus can be neglected for this study. We use monthly means of  $T_{max}$ , Pr and VPD during the growing season, as well as the seven extreme indicators for further analysis.

#### A.2.4 Lasso regression

The aim of this study is to provide an interpretable statistical model able to predict years with extremely low yields (bad years) with meteorological variables. We use the Least Absolute Shrinkage and Selection Operator (Lasso, Tibshirani, 1996) logistic regression for an automatic selection of meteorological variables that are statistically linked to low yields. The approach is explained below.

For a given grid point, let  $Y \in \{0, 1\}^n$  be the binary yield vector, with  $n$  the number of years. If the year  $i \in \{1, \dots, n\}$  is a bad year, then  $Y_i = 1$ , otherwise  $Y_i = 0$ . Let  $X_1, \dots, X_p \in \mathbb{R}^n$  be the explanatory variables vectors (monthly meteorological variables and climate extreme indicators, rescaled as explained in Section A.2.2). Using a generalized linear model and, more specifically, a logistic regression, we can identify how much of the occurrence of bad yields is explained by which explanatory variable:

$$\mathbb{P}[Y = 1] = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)} \quad (\text{A.1})$$

where  $\beta_0, \beta_1, \dots, \beta_p$  are the regression coefficients.

However, a simple logistic regression presents two challenges here. Firstly, some variables might be highly correlated (e.g. correlation between temperature in May and temperature in June, or the correlation of extreme indices with meteorological variables). This correlation implies a high variability of the coefficients. For instance, if the variables  $X_j$  and  $X_k$  are highly correlated, the information brought by a high absolute value of  $\beta_j$  and a low absolute value of  $\beta_k$  might be the same as the information brought by a low absolute value of  $\beta_j$  and a high absolute value of  $\beta_k$ . Another issue is the large number of potential explanatory variables (up to 43 for some grid points). The relatively straightforward relationship of a generalized linear model (simpler than the crop model equations themselves) allows to reveal which meteorological variables explain bad yields best. However, if the number of *a priori* explanatory variables is very large, the regression becomes rather complex and many coefficients will be close to zero, rendering an interpretation difficult.

Lasso regression tackles both challenges with an automatic variable selection using a regularization by penalizing the number of coefficients different from 0 using the  $\ell_1$  norm on the vector of coefficients (Tibshirani, 1996). Thus, the regression coefficients are obtained by minimizing an objective function consisting of the sum of the usual loss function for logistic regression and a penalty term on the coefficient norm:

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} - \left[ \frac{1}{n} \sum_{i=1}^n y_i (\beta_0 + x_i^T \beta) - \log(1 + e^{\beta_0 + x_i^T \beta}) \right] + \lambda \|\beta\|_1, \quad (\text{A.2})$$

for a fixed  $\lambda > 0$ . The penalty term on the coefficient norms prevents a high variability of these coefficients. Furthermore, the  $\ell_1$  norm implies a variable selection. Coefficients associated to non-relevant explanatory variables are set to 0.

We use the R package `glmnet` (Friedman et al., 2010) to perform the Lasso regression with R

version 3.6 (R Core Team, 2019). Through 10-fold cross-validation in the training dataset, we obtain the optimal  $\lambda_{min}$  and  $\lambda_{1se} = \lambda_{min} + se$  with  $se$  the standard error of the lambda that achieves the minimum loss, and the coefficients  $\beta$ , which are the solution to the optimization in equation (A.2) for  $\lambda = \lambda_{1se}$ . Our preference for  $\lambda = \lambda_{1se}$  is motivated by the balance between number of selected variables and accuracy of the loss function minimization (Friedman et al., 2010; Krstajic et al., 2014). Indeed, less variables are selected with  $\lambda_{1se}$  than with  $\lambda_{min}$ , because  $\lambda_{1se} > \lambda_{min}$  and thus the penalty term on the norm of coefficient is stronger, but the minimization of the equation (A.2) is still sensible, because  $\lambda_{1se}$  lies within the uncertainty range of the optimal  $\lambda$ .

### A.2.5 Other models

To compare the performance of the Lasso regression with other regression methods we also perform the analysis with a Generalized linear model (GLM) and a random forest binary classification.

For the application of the GLM, a pre-selection of the initial variables is required, since the number of predictors is limited. Only the variables with the highest Pearson correlation coefficient ( $\rho > 0.30$ ) were selected as initial predictors from an initial dataset composed by all months of the growing season for each of the three variables ( $T_{max}$ , Pr and VPD) and the seven extreme indicators. Next, the subset of best predictor variables is identified with the leaps algorithm (Furnival and Wilson, 1974). We use the implementation of the R package `bestGLM` (McLeod et al., 2020), using a binomial family with a logit link function. Overall, GLM achieves lower performance (Section A.2.7) compared to the Lasso logistic regression (not shown). The weaknesses of this approach is its sensitivity to outliers and multicollinearity, and overfitting.

Finally, a random forest approach – a common machine learning technique – was also performed using the R package `randomForest` (Breiman, 2001; Liaw and Wiener, 2002) serving as a benchmark for the model performance of the Lasso logistic regression. The random forest binary classification achieves comparable performance (Section A.2.7) but is not superior to the Lasso approach.

### A.2.6 Segregation threshold adjustment

The segregation threshold for assigning a continuous prediction to either a bad or normal year was adjusted grid point-wise to account for the unbalanced dataset with 19-fold higher occurrences of normal years than bad years. Let  $s$  be the local segregation threshold between bad year predicted and good year predicted. In other words, if the probability  $p = \mathbb{P}[Y = 1]$  predicted for a given grid point by the Lasso logistic regression model is greater or equal to  $s$  (resp. lower than  $s$ ), then the year is predicated as a bad year (resp. normal year). We want to choose  $s$  as a good compromise in prediction of normal years and bad years, given that bad years are rare. In other words, we want to find an optimal trade-off between specificity and sensitivity. To this purpose, a cost function  $\mathcal{C} = \mathcal{C}(s)$  is calculated based on the false positive rate  $R_{FP} = R_{FP}(s)$ , the associated cost for a false positive instance  $\mathcal{C}_{FP}$ , the sum of observed normal years  $O_{NY}$ , the false negative rate  $R_{FN} = R_{FN}(s)$ , the associated cost for a false negative instance  $\mathcal{C}_{FN}$  and the sum of observed bad

years  $O_{BY}$  of the training dataset (Hand, 2009). A false positive means that a normal year was observed while a bad year was predicted, and a false negative refers to the observation of a bad year, whereas a normal year was predicted. For a given grid point, FP, FN, TP and TN denote the total number of false positives, false negatives, true positives and true negatives, respectively (Fig. A.4). The value of  $\mathcal{C}(s)$  is given by:

$$\mathcal{C}(s) = R_{FP}(s)\mathcal{C}_{FP}O_{NY} + R_{FN}(s)\mathcal{C}_{FN}O_{BY}, \quad (\text{A.3})$$

where  $R_{FP} = \frac{FP}{FP + TN}$ ,  $R_{FN} = \frac{FN}{FN + TP}$  and  $\mathcal{C}_{FP} = \mathcal{C}_{FN} = 100$ . In this study, the cost associated with false positive  $\mathcal{C}_{FP}$  and false negatives  $\mathcal{C}_{FN}$  are given equal weight.

The optimal segregation threshold  $s^*$  for a given grid point is  $s^* = \operatorname{argmin}_{s \in (0,1)} \mathcal{C}(s)$ . The segregation threshold selected in this study is the mean value of  $s^*$  over all grid points.

### A.2.7 Model performance assessment and sensitivity analysis

Model performance is assessed using the critical success index (CSI). The CSI is frequently used for evaluating the prediction of rare events, as it neglects the number of correct predictions of non-extremes, which dominate the confusion matrix (Mason, 1989). General performance measures such as the misclassification error are biased by the high number of normal years and are therefore not meaningful for the assessment of model performance in unbalanced datasets with underrepresented extreme events. The CSI is defined as

$$\text{CSI} = \frac{TP}{TP + FP + FN}. \quad (\text{A.4})$$

To evaluate the robustness of our model, in addition to the 5<sup>th</sup> percentile threshold we repeated the analysis with thresholds of 2.5 % and 10 %, reaching qualitatively similar performance. In addition to the normalization by rescaling the data to the interval  $[-1, 1]$ , we also performed a z-score transformation, which yielded comparable results. Therefore our choice of normalization is arbitrary to a degree and a z-score transformation can potentially also be applied in the Lasso logistic regression model. Moreover, we applied two more combinations of splitting training and testing dataset, a 60/40 and 80/20 split. With increasing size of the training dataset, the CSI increased slightly, however at the expense of stochastic under-representation of bad yield years in the smaller testing datasets. As a trade-off, we decided for the 70/30 split.

The adjustment of the segregation threshold was carried out with equal weight to false positive and false negative predictions. It can be argued that the latter case – where a normal year is predicted, but crop failure is observed – is more detrimental and should therefore be given a higher weight. Due to the subjectivity in the determination of this weight, an adjustment of the weight term was not applied. However, it should be noted that the attribution of a higher weight of false negative predictions would yield a lower segregation threshold and hence improve the overall CSI.

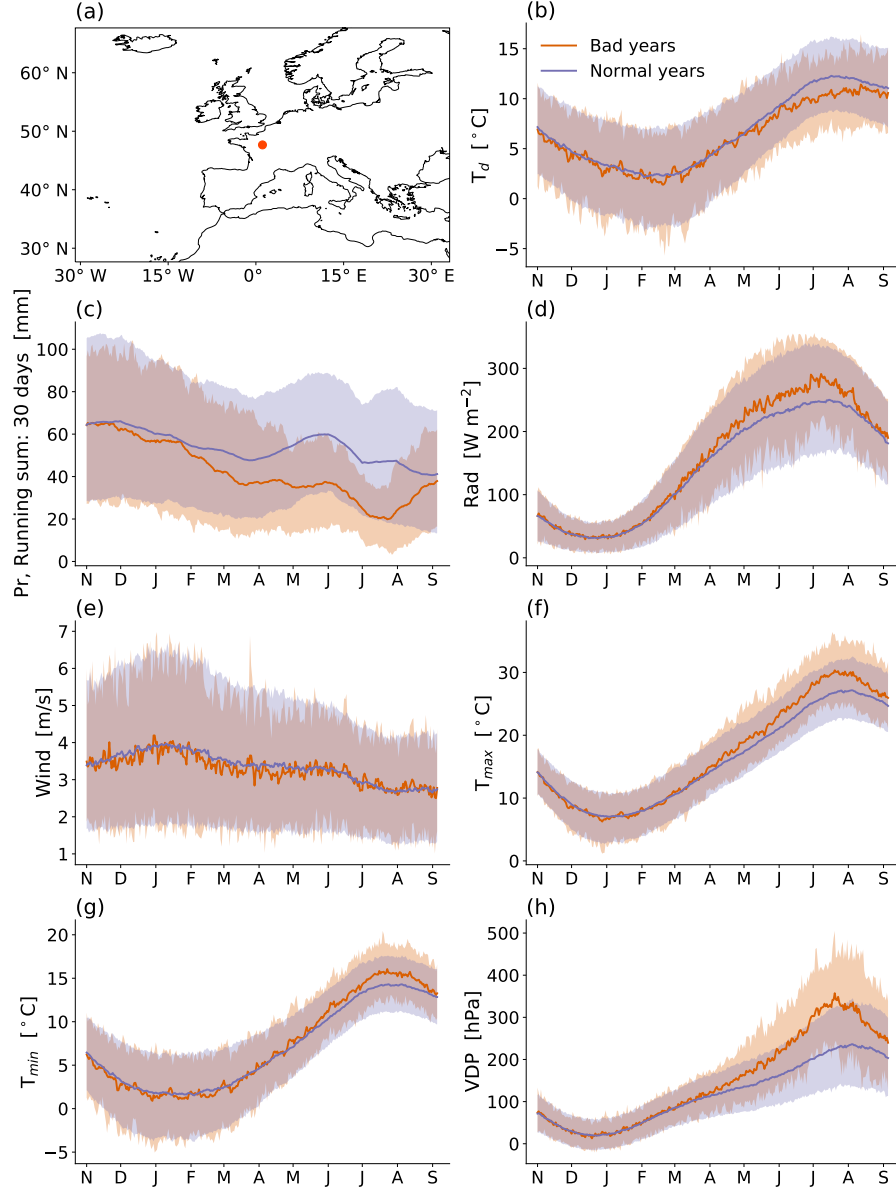


FIGURE A.2: *Daily evolution of meteorological variables used as input for the APSIM model over the course of the year for an exemplary grid point in France ( $1.1^{\circ}$  E,  $47.7^{\circ}$  N, shown as a red dot in (a)). Red lines indicate the composite mean of the bad years (80 seasons), blue lines the composite mean of the normal years (1520 seasons). Shading shows the range between the 10<sup>th</sup> and 90<sup>th</sup> percentile of the respective years. Variables shown are (b) dewpoint temperature, (c) 30-day running sum of precipitation, (d) incoming shortwave radiation, (e) wind speed, (f) maximum temperature, (g) minimum temperature, and (h) vapour pressure deficit (VPD).*

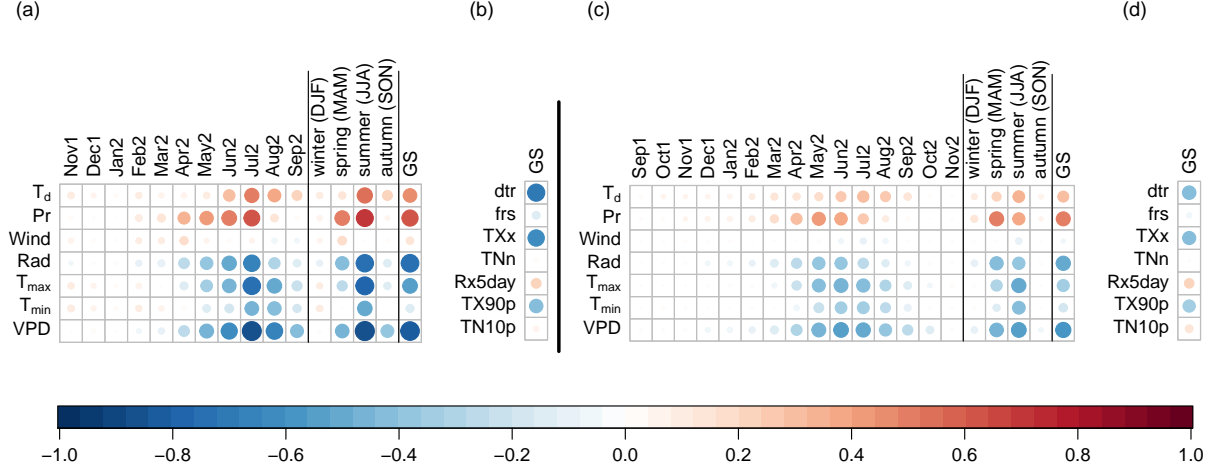


FIGURE A.3: *Linear correlations between potential meteorological predictors and annual yield. (a) Correlation between the monthly, seasonal and growing season (GS) averages of the meteorological variables and annual yield for a grid point in France ( $1.1^{\circ}$  E,  $47.7^{\circ}$  N). (b) Correlation of the climate extreme indicators (Table A.1) and annual yield for the same grid point. (c, d) Average of the same correlations across all Northern Hemisphere grid points. Note that (a) shows the correlation for all months included in the growing season of the grid point in France while (c) shows the average correlation for a given month computed over all grid points containing this month in their growing season.*

		Observed	
		Normal year ( $Y = 0$ )	Bad year ( $Y = 1$ )
Predicted	Normal year ( $Y = 0$ )	TN	FN
	Bad year ( $Y = 1$ )	FP	TP

FIGURE A.4: *Confusion matrix for classification of observed and predicted normal and bad years.*



## A.3 Results

### A.3.1 Overall performance

The Lasso logistic regression model can predict bad years with an average CSI = 0.43 across all grid points. Best performance is obtained in the eastern half of the United States with a maximum of CSI = 0.82 (Fig. A.5), which decreases westwards in the Great Plains and is lowest in the wheat growing regions located close to the Rocky Mountains. Furthermore, especially the most northern and southwestern grid points in North America show a lower performance in general. Also central Europe shows high performances up to CSI = 0.80. A notable regional exception with low performances can be found in the Alps. Many Asian and African growing regions show medium prediction accuracy such as northern China, Myanmar, Turkey and the Maghreb, with exceptions of some regions including Pakistan, southern China and Japan, which show a low performance in general. For 30 grid points, it is not possible to obtain reasonable predictions of bad years with our approach, indicated by a CSI equal to 0. Overall, regions with high prediction accuracy of bad years are often those that also have high mean yields (Fig. A.1). CSI is positively correlated with mean yield with a Pearson's correlation coefficient of  $\rho = 0.46$  (Fig. A.6a), an even stronger correlation is found with yield variability ( $\rho = 0.57$ ) (Fig. A.6b).

### A.3.2 Explanatory variables

Here we summarize the properties of the variables selected by the Lasso logistic regression as relevant explanatory variables, i.e., which are statistically linked to bad years. A median of 11 variables per grid point has been selected as explanatory variables, and for 50 % of grid points the number of selected variables lies between 7 and 14 (Fig. A.7a). The inclusion of extreme indicators provides a useful addition to the monthly predictors, shown by a median number of two selected extreme indicators per grid point (Fig. A.7b). Grid points without extreme indicators are found only in few areas such as eastern Europe, the Alps and southern China. 72 % of all grid points include monthly predictors of VPD, Pr and  $T_{max}$  and almost all grid points (97 %) incorporate VPD (Fig. A.7c). Interestingly, in the Great Plains (USA) in many cases temperature is not included, whereas in most other regions of the USA all meteorological variables are selected to achieve a good prediction. In southern China, temperature is not needed by the models, whereas in the northern areas, usually all meteorological variables are part of the model. In most wheat-growing regions, particularly in northeastern USA, southeastern Europe and Turkey, all four seasons contain relevant predictors for predicting bad years (Fig. A.7d). Generally, the number of seasons included decreases towards the southeastern regions in the USA, whereas in western Europe no clear pattern emerges. In lower latitudes such as southern Asia, growing seasons are generally shorter (Fig. A.12) and consequently often only predictors from one or two seasons are included in the respective models.

At the global scale, VPD in May and June, as well as Pr in April are the predictors which are most often included in the Lasso regression, followed by the climate extreme indicators diurnal temperature range (dtr) and number of frost days (frs) (Fig. A.8a). In nearly all cases the sign of

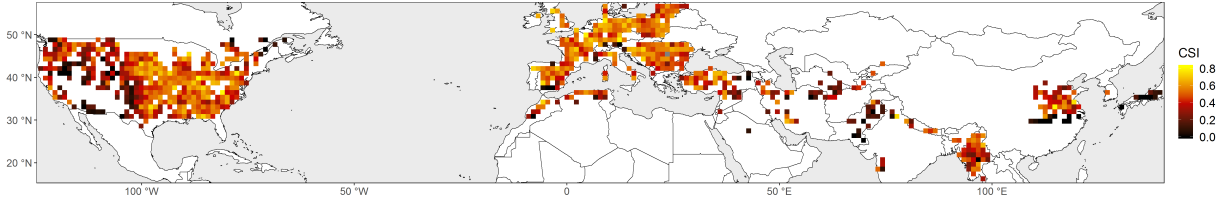


FIGURE A.5: *Critical success index (CSI, equation (A.4)) of the Lasso logistic regression model. (See Section A.2.7 for definition).*

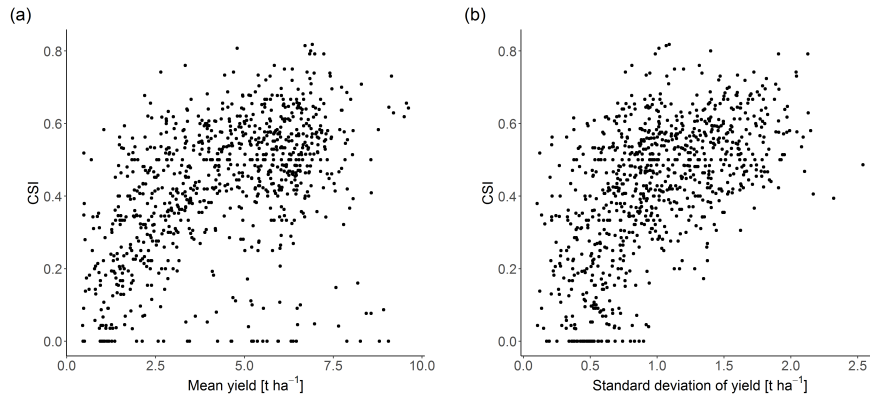


FIGURE A.6: *Correlation between Critical Success Index (CSI) and annual crop yield mean and variability for the 889 pixels included in the Lasso logistic regression model. (a) Scatterplot between CSI and mean annual yield. (b) Scatterplot between CSI and annual yield standard deviation.*

the coefficient is positive for VPD in May and June, and negative for Pr in April. This implicates that higher VPD increases the risk of crop failure, and similar for the other variables. In North America and Europe, in addition to *dtr* and *frs*, VPD and Pr in spring to early summer are the most frequent monthly predictors (Fig. A.8b, c). The growing season for wheat varies with latitude. Consequently, in more northern regions, mostly in Europe and North America, monthly predictors from the months between March and July are included in the Lasso regression, whereas in southern regions such as in Asia and Africa, November to May are usually the most frequent months (Fig. A.8d).

This clear latitudinal shift can be visually identified in North America from February to July, especially for VPD (see GIFs in the Supplementary Material). In central Europe the growing season ends latest, thus VPD in August is usually included in the model. In addition to the most common climate extreme indicators *dtr* and *frs*, *Rx5day* and *TXx* are among the most frequent predictors in Asia and North America, respectively. Overall, *frs* is mostly included in northern grid points, with notable exceptions in western Europe (Fig. A.13a) and mainly with a positive coefficient (higher *frs* leads to more crop failure events), which can likely be attributed to the influence of mild maritime climate in those regions. In contrast, *dtr* is important in most Asian grid points and especially in western Europe and the Maghreb, whereas in the Pannonian

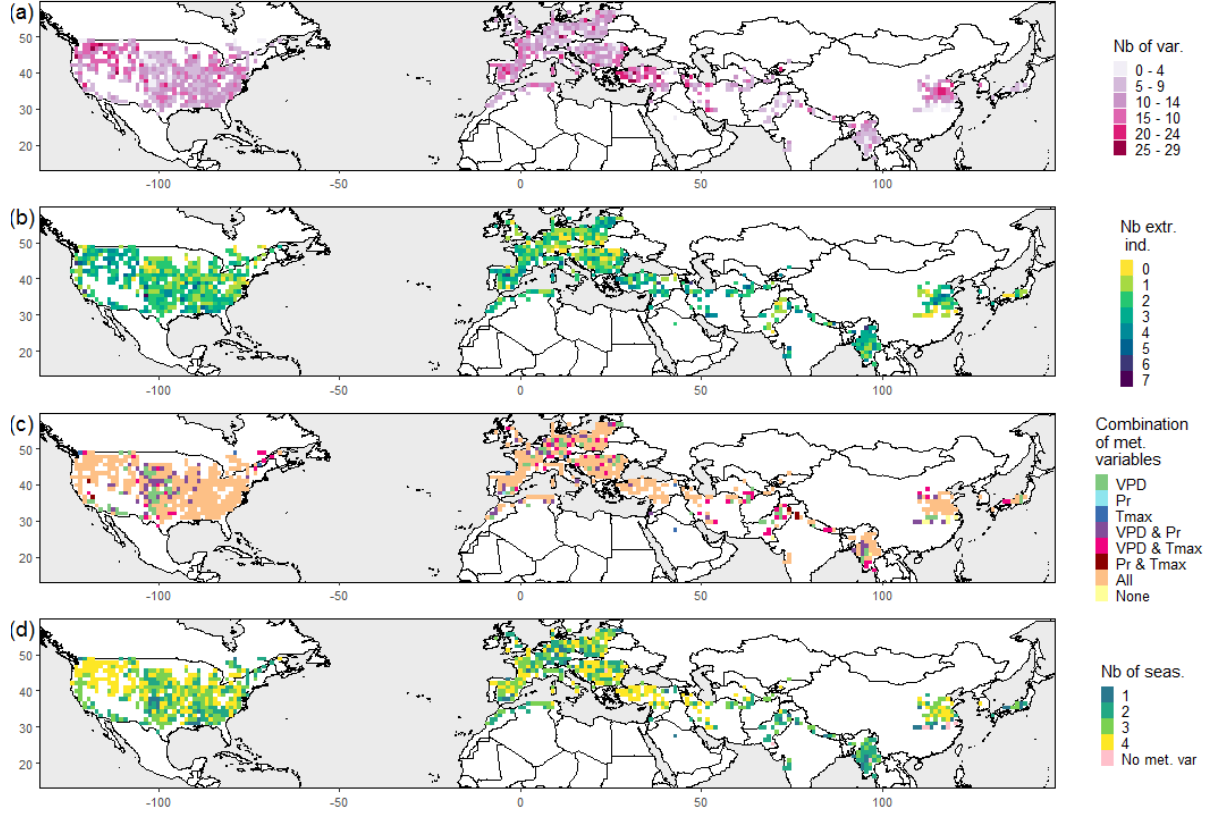


FIGURE A.7: Maps illustrating the selected predictors by the Lasso logistical regression. (a) Total number of selected variables. (b) Number of selected climate extreme indicators. (c) Combination of selected meteorological variables. “None” means that only climate extreme indicators were selected, “All” means that at least one month from each of the three meteorological variables (VPD, Pr,  $T_{max}$ ) is selected. (d) Number of selected seasons (out of the four seasons DJF, MAM, JJA, SON).

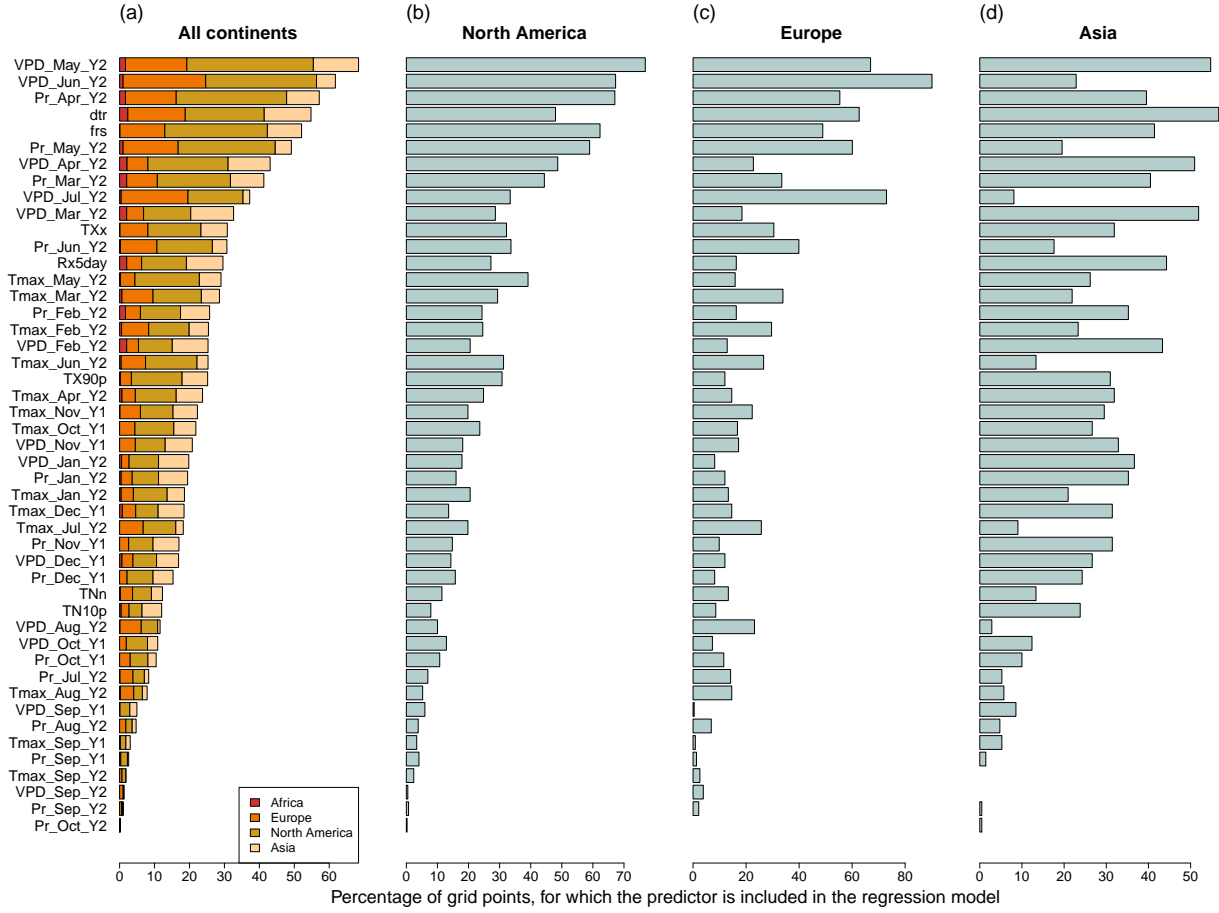


FIGURE A.8: For each possible predictor we show the percentage of grid points for which this predictor has a non-zero coefficient in the Lasso logistic regression. (a) all continents (889 grid points in total), (b) North America (419 grid points), (c) Europe (233 grid points) and (d) Asia (210 grid points). The extension “Y1” means that the respective month belongs to the first calendar year of the growing season, while “Y2” means it belongs to the second calendar year of the growing season.

Basin and Turkey it is a less common predictor (Fig. A.13b). The coefficient associated with *dtr* in the Lasso regression is mainly positive, except in parts of India and Myanmar. Some variability in mean diurnal temperature range might be beneficial for regions close to the equator where the variability in diurnal temperature is usually low. Generally, both low and high *dtr* values can influence wheat yield beneficially depending on the growing region, e.g. a low *dtr* can be advantageous because of a reduced occurrence of frost days, whereas a higher *dtr* might also indicate a favorable effect because of increased solar radiation (Lobell, 2007). *Rx5day* is predominant in the western USA, the western Mediterranean and central Asia (Fig. A.13c), which are all growing regions with comparably low average annual precipitation. *TX90p* is a common variable in low latitudes with a positive coefficient, especially in the southern USA and Myanmar (Fig. A.13d). This indicates that in these regions physiological temperature thresholds are occasionally exceeded, making *TX90p* a crucial variable in these areas.

## A.4 Discussion

### A.4.1 Predicting bad yield years

In this study, we presented a method for identifying drivers of extreme impacts using crop failure as an example. Such approaches are highly sought after to identify compound drivers of large impacts (Zscheischler et al., 2020; Van der Wiel et al., 2020). Our method allows to investigate potential drivers at a global scale using a highly automated scheme based on Lasso regression. The benefits of Lasso regression include its usage for automatic variable selection, its consideration of correlation between explanatory variables, and its performance. Moreover, the statistical model obtained provides a logistic linear relationship between crop failure and selected variables, which is much simpler to interpret than the crop model equations themselves or results obtained with more complex machine learning approaches.

We defined bad years as years where the annual crop yield is below the 5<sup>th</sup> percentile and were able to predict those years by using the Lasso regression with an average CSI of 0.43. This means that on average, the sum of the numbers of false positives and false negatives is slightly higher than the number of true positives (or accurate predictions of bad years). Our model performance is somewhat comparable to results from Vogel et al. (2019), who were able to explain 46 % of variation in spring wheat anomalies using a similar set of predictors based on a random forest algorithm. In our case, more sophisticated machine learning regression models such as random forest did not yield better prediction skill, indicating that performance in the current set-up using monthly predictors for a binary classification of bad years likely cannot be much improved. This is probably also related to the fact that predicting extremes of a continuous variable is challenging because no natural separation between extremes and non-extremes exists. Another challenge arises from the highly asymmetric distribution of observed bad and normal years. Even though in our case the total amount of samples per grid point is relatively large (1600, because we used simulated crop yield data) the number of observed bad years is only 80 and thus can still be considered fairly small.

We analysed the robustness of our results using a) the 10<sup>th</sup> percentile as a threshold to discriminate between bad and normal years and b) a smaller data subset with only 400 entries per grid point (i.e. a quarter of the available data). The spatial patterns of the selected predictors and the CSI using the 10<sup>th</sup> percentile threshold are very similar compared to those of the 5<sup>th</sup> percentile and the average CSI increases slightly from 0.43 to 0.52. Using a sample size of 400 we still obtain an average CSI of 0.33, indicating that performance decreases only slightly with decreasing data size, while the spatial patterns remain consistent (results not shown). Furthermore, the spatial coherence of our results (Fig. A.7) additionally suggests robustness of our analysis. An application of the approach on real data might still be challenging, as observational sample sizes generally are much smaller than even 400 years. In addition, observational data is often not available at such spatial resolution and extent as it is the case for the crop model data used in this study. This will make it difficult to use spatial coherence of the identified drivers as an indicator of model robustness when using observational data. Furthermore, modelling winter wheat yield is particularly challenging due to its long growing season (Vogel et al., 2019).

A limitation to our study design is the pre-selection of potential predictor variables. Here we used monthly mean values and a number of climate extreme indicators. More flexible averaging time periods for the predictors might result in higher prediction accuracy due to better overlap with sensitive periods of the impact variable. For instance, in our crop yield example meteorological predictors need to coincide with the respective phenological development stage because their impact can vary depending on the phenophase. Wheat, for example, is known to require wet conditions in the vegetative phase, however prefers dry conditions during ripening (Seyfert, 1960). Therefore, the application of monthly meteorological predictors might be insufficient for accurate matching of meteorological drivers to the respective phenological phases. We explored the option of automatically generating optimal time periods for the meteorological predictors by maximizing the difference between the composites between normal and bad years. For instance, 30-days cumulative precipitation differs between normal and bad years starting in February and ending in August for a grid point in France (Fig. A.2c), whereas VPD only differs from May to September (Fig. A.2h). Composite plots for a grid point in the US and in China are shown in Figs. A.10 and A.11, respectively. However, deciding when the separation between normal and bad years is large enough to start and end the optimal time periods is challenging and difficult to generalize and thus automate, which was the aim of our method design. Nevertheless, such a well-tuned selection of predictors has the potential to improve the prediction of bad years significantly and should thus be explored in future research.

We find a strong correlation of the yearly mean and standard deviation of annual yield with the Lasso regression performance indicator CSI (Fig. A.6). Low model performance at grid points with low yield variability suggests that the distinction between normal and bad years is challenging at these locations, e.g. in southern China and Japan (Figs. A.1b and A.5). Regions with high annual yield are found primarily in central Europe and the eastern half of the United States, which also represent the regions with highest model performances. In contrast, many regions in Asia generally have lower average yields and lower prediction skill of bad years. This could be related to a calibration bias in the crop model, leading to a better representation of wheat growing processes in regions where wheat reaches higher yields in the real world. A further explanation for this phenomenon could be that the crop model is primarily designed for crop growth at typical environmental conditions, whereas growing regions with conditions at the edge of the ecological niche of wheat might be less well represented.

Our analysis was based on fitting a local model at each location, which is one of the three principal statistical methods used to link crop yield with weather conditions, along with cross section models and panel models, which are global models that adjust for spatial variability using fixed or random effects (Lobell and Burke, 2010; Shi et al., 2013). Collinearity between explanatory variables is a recurrent issue when using these methods (Shi et al., 2013) that we addressed with the Lasso regression. One example is VPD and  $T_{max}$ , that might be highly correlated, but still might individually contribute relevant information because they have a different impact on the plant process, as explained in Kern et al. (2018). Lasso regression did not completely discard one of these two variables, despite their high correlation.

### A.4.2 Important predictors

For most grid points, VPD is the most important monthly predictor of bad years, followed by  $P_r$  and  $T_{max}$  in that order. While their importance in time differs between grid points, depending on the timing of the respective growing season (Sippel et al., 2016), their order changes little across space. In addition, the order of importance of extreme indicators is quite similar in North America, Europe and Asia. One notable distinction is the higher importance of  $Rx5day$  in Asian grid points compared to North America and Europe. The consistent selection of similar predictors across large spatial scales may suggest that the Lasso regression is fairly robust. However, this may also be related to the inevitable simplifications of crop growing processes in the employed crop model. In particular, the same model is applied at all locations likely creating certain homogeneity by default. Kern et al. (2018) conducted a comparable analysis on observed winter wheat crop yield in Hungary. With a linear regression using a step-wise selection of monthly meteorological variables, they found that a positive anomaly in VPD and  $T_{min}$  during May decreases yield. Additionally, April, May, and June appear to be the most relevant months in our global analysis, which is consistent with regional studies (Kern et al., 2018; Kogan et al., 2013; Ribeiro et al., 2020).

Climate extreme indicators are important predictors as the occurrence of an extreme weather event may induce crop failure in a given year. However, in years without such extreme events, crop yields are still governed by the weather during the growing season (Iizumi and Ramankutty, 2015). We found that both climate extreme indicators as well as monthly means of common climate variables have proven to be valuable predictors of years resulting in crop failure. Droughts and heat waves are well known to affect crop yield (Lesk et al., 2016; Jagadish et al., 2014), and temperature and precipitation explain a large fraction of interannual crop yield variability (Lobell and Burke, 2008). In contrast, VPD is often overlooked in statistical analyses of crop yield variability (Zhang et al., 2017). We show that VPD is a key predictor for crop failure. It is known to play a crucial role in plant functioning and is projected to increase as main limiting driver in the face of climate change (Novick et al., 2016; Grossiord et al., 2020). High VPD values can lead to plant mortality via carbon starvation and hydraulic failure (McDowell et al., 2011; Grossiord et al., 2020). However, its covariation with temperature and solar radiation makes it difficult to disentangle their respective effects (Stocker et al., 2019; Grossiord et al., 2020). There are well-defined temperature thresholds for wheat, e.g. a temperature of 31 °C before flowering is considered to evoke sterile grains and thus reduces yield (Porter and Gawith, 1999; Daryanto et al., 2016).  $T_{max}$  is a secondary predictor in our statistical model, which is in line with results based on observed and simulated yields (Schauberger et al., 2017), and can be attributed to the rare exceedance of critical temperature thresholds in the growing season. Crops are particularly vulnerable during key development stages, so extreme events during that time span can lead to large yield reductions, even in case of otherwise favorable weather conditions during the growing season (Porter and Gawith, 1999; Moriondo and Bindi, 2007). The vulnerability of wheat to climatic events depends largely on phenophases and generally wheat possesses a higher sensitivity to temperature and precipitation during its reproductive phase than during its vegetative phase (Porter and Gawith, 1999; Luo, 2011; Daryanto et al., 2016). Future research could investigate

the importance of time of occurrence of extreme indicators (Vogel et al., 2019). For instance, due to climate change false spring events may become more likely in some regions (Moriondo and Bindi, 2007; Allstadt et al., 2015) and thus the timing of frost days could provide a valuable addition to the model.

The frequent inclusion of the extreme indicators *dtr* and *frs* in our regression model highlights that short-term extreme events can potentially have larger impacts than gradual changes over time (Jentsch et al., 2007). The variable *dtr* was also identified as an important predictor by Vogel et al. (2019), whereas *frs* was of minor importance for explaining variation in spring wheat yield. By contrast, *frs* is one of the most predominant predictors in our study, which might be explained by the differing growing season of winter wheat, which is encompassing primarily the cold seasons.

We explored the relevance of interactions between predictors; however, this did not significantly improve model performance. This might hint at the inability of the APSIM crop model to account adequately for such compound effects, which is consistent with Ben-Ari et al. (2018), who linked the crop failure 2016 in France to an extraordinary combination of warm winter temperatures followed by wet spring conditions. The commonly used crop models employed for crop yield forecasts were not able to predict the 2016 yield failure in France (Ben-Ari et al., 2018).

Overall, our results illustrate the omnipresence of compounding meteorological events for crop failure. In nearly all grid points, most seasons and meteorological variables were relevant to predict years with crop failure (Fig. A.7). This suggests that the co-occurrence of certain weather conditions as well as the combination of weather conditions in different seasons are associated with crop failure. With our approach we have identified meteorological conditions that are statistically linked to crop failure. Our results confirm prior findings by Van der Wiel et al. (2020) that such conditions are not necessarily extreme, but can also be moderate. However, for interpretation of the selected variable set one should be aware that the variables in our model are selected based on correlation and thus attributing them as potential physical drivers needs further careful investigation. To identify such causal relationships, more advanced methods from the emerging field of causal inference could be employed (Runge et al., 2019).

## A.5 Conclusion

In this paper, we presented a robust statistical approach – namely Lasso logistic regression – for predicting crop failure and automatically selecting relevant predictors among a large number of meteorological variables and climate extreme indicators. We illustrated our approach on 1600 years of simulated winter wheat yield for the Northern Hemisphere under present-day climate conditions. Lasso regression can serve as a tool for identifying important variables with automated variable selection, while accounting for collinearity and achieving overall good predictive power. Consistent with earlier knowledge, we find that predicting crop failure requires accounting for a number of different meteorological drivers at different times of the growing season, which is illustrated by the large amount of variables at all seasons included in our statistical model (Fig. A.7). This indicates that compounding effects are ubiquitous across time and meteorological



drivers, and highlights the usefulness of approaches such as Lasso regression to reveal multiple meteorological drivers of crop failure. We identified vapour pressure deficit as one key variable to predict crop failure, which underlines the importance of its consideration in statistical crop yield models, in particular because it is often overlooked in statistical analyses of crop yield variability. Furthermore, climate extreme indicators such as diurnal temperature range and the number of frost days have proven to be valuable additions to the predictive models, highlighting the necessity to address not only monthly mean conditions, but especially also climatic extremes in such models. Overall this study helps to enhance the knowledge required to improve seasonal forecasts and undertake adaptation measures against crop failure. The flexibility of our approach allows an application to other climate impacts that are influenced by a large range of variables varying with seasonality, for instance wildfires or flooding.

**Code and data availability** The code to reproduce the figures is available from GitHub ([https://github.com/jo-vogel/Identify\\_crop\\_yield\\_drivers](https://github.com/jo-vogel/Identify_crop_yield_drivers)). The climate and crop simulations are available from Karin van der Wiel ([wiel@knmi.nl](mailto:wiel@knmi.nl)) and Tianyi Zhang ([zhangty@post.iap.ac.cn](mailto:zhangty@post.iap.ac.cn)) upon request, respectively.

**Video supplement** The Supplementary Material contains GIFs showing monthly binary maps of whether a specific predictor was included to predict crop failure by the Lasso logistic regression. GIFs are provided for a) VPD, b)  $T_{max}$  and c) Pr. The extension “Y1” means that the respective month belongs to the first calendar year of the growing season, while “Y2” means it belongs to the second calendar year of the growing season.

## Appendix

### A.5.1 APSIM-Wheat model settings

Eleven phenological phases are included in the APSIM-Wheat model and the length of each phase is simulated based on thermal time accumulation, which is adjusted for other factors such as vernalisation, photoperiod and nitrogen. To calculate thermal time, crown minimum ( $T_{cmin}$ ) and maximum ( $T_{cmax}$ ) temperatures are first simulated for non-freezing temperatures ( $T_{min}$  and  $T_{max}$ , equations A.5 and A.6) and then used to compute the crown mean temperature ( $T_c$ , equation A.7). Finally, daily thermal time ( $\Delta TT$ ) is calculated based on three cardinal temperatures ( $T_{base}$ ,  $T_{opt}$  and  $T_{ceiling}$ , equation A.8) (Zheng et al., 2014):

$$T_{cmax} = \begin{cases} 2 + T_{max}(0.4 + 0.0018(H_{snow} - 15)^2) & T_{max} < 0 \\ T_{max} & T_{max} \geq 0 \end{cases} \quad (A.5)$$

$$T_{cmin} = \begin{cases} 2 + T_{min}(0.4 + 0.0018(H_{snow} - 15)^2) & T_{min} < 0 \\ T_{min} & T_{min} \geq 0 \end{cases} \quad (A.6)$$

$$T_c = \frac{(T_{cmin} + T_{cmax})}{2} \quad (A.7)$$

$$\Delta TT = \begin{cases} T_c & T_{base} < T_c \leq T_{opt} \\ \frac{T_{opt}}{T_{base}}(T_{ceiling} - T_c) & T_{opt} < T_c \leq T_{ceiling} \\ 0 & T_c \leq T_{base} \text{ or } T_c \geq T_{ceiling} \end{cases} \quad (A.8)$$

where  $H_{snow}$  is set to 0 and  $T_{base}$ ,  $T_{opt}$ , and  $T_{ceiling}$  are set to 0, 26 and 34 °C, respectively. The dry-matter above-ground biomass ( $\Delta Q$ , equation A.12) is calculated as a potential biomass accumulation resulting from radiation interception ( $\Delta Q_r$ ) and soil water deficiency ( $\Delta Q_w$ ) (Zheng et al., 2014). The radiation limited dry-biomass accumulation ( $\Delta Q_r$ , equation A.10) is calculated by the intercepted radiation ( $I$ ), radiation use efficiency ( $RUE$ ), stress factor ( $f_s$ ) and carbon dioxide factor ( $f_c$ ). The stress factor ( $f_s$ ) comprises stresses that crops may encounter during growth and is the minimum value of a temperature factor ( $f_{T,photo}$ ) and a nitrogen factor ( $f_{N,photo}$ ) (equation A.9). The water-limited dry above-ground biomass ( $\Delta Q_w$ , equation A.11) is a function of radiation-limited dry above-ground biomass ( $\Delta Q_r$ ), the ratio between the daily water uptake ( $W_u$ ) and demand ( $W_d$ ):

$$f_s = \min(f_{T,photo}, f_{N,photo}) \quad (A.9)$$

$$\Delta Q_r = I \cdot RUE \cdot f_s \cdot f_c \quad (A.10)$$

$$\Delta Q_w = \Delta Q_r \frac{W_u}{W_d} \quad (A.11)$$

$$\Delta Q = \begin{cases} \Delta Q_r & W_u = W_d \\ \Delta Q_w & W_u < W_d \end{cases} \quad (A.12)$$

### A.5.2 Additional figures

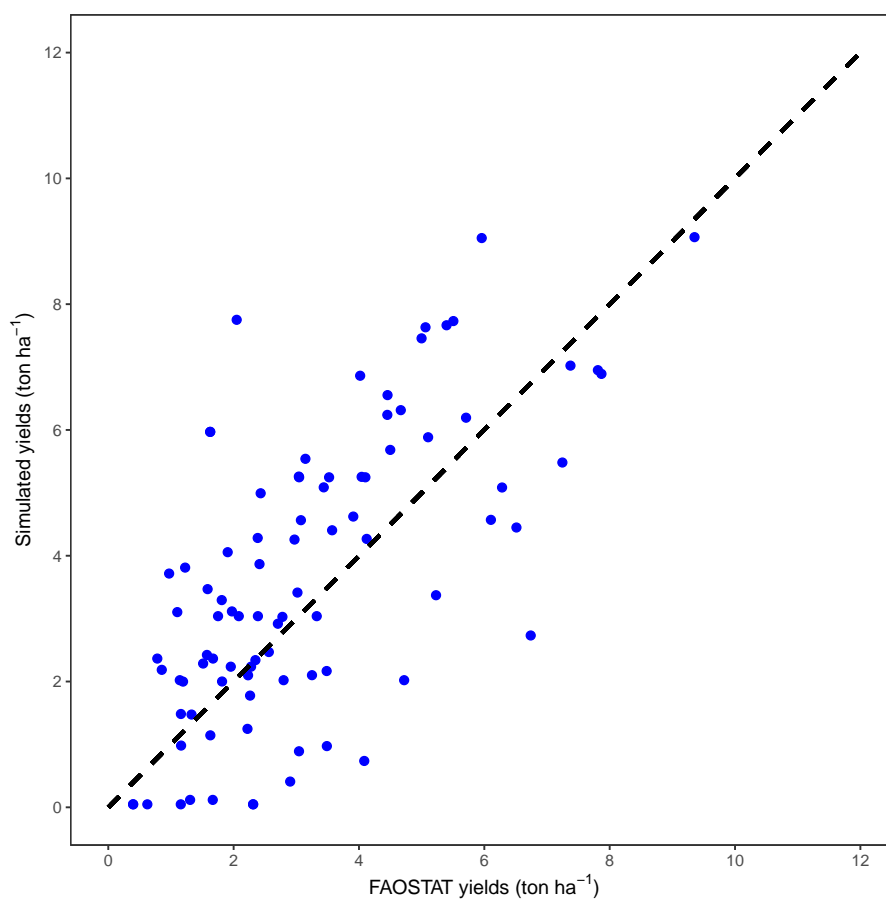


FIGURE A.9: *Comparison between the country-specific simulated yields and yield statistics (FAO-STAT, 2020). The dashed line is the 1:1 line.*

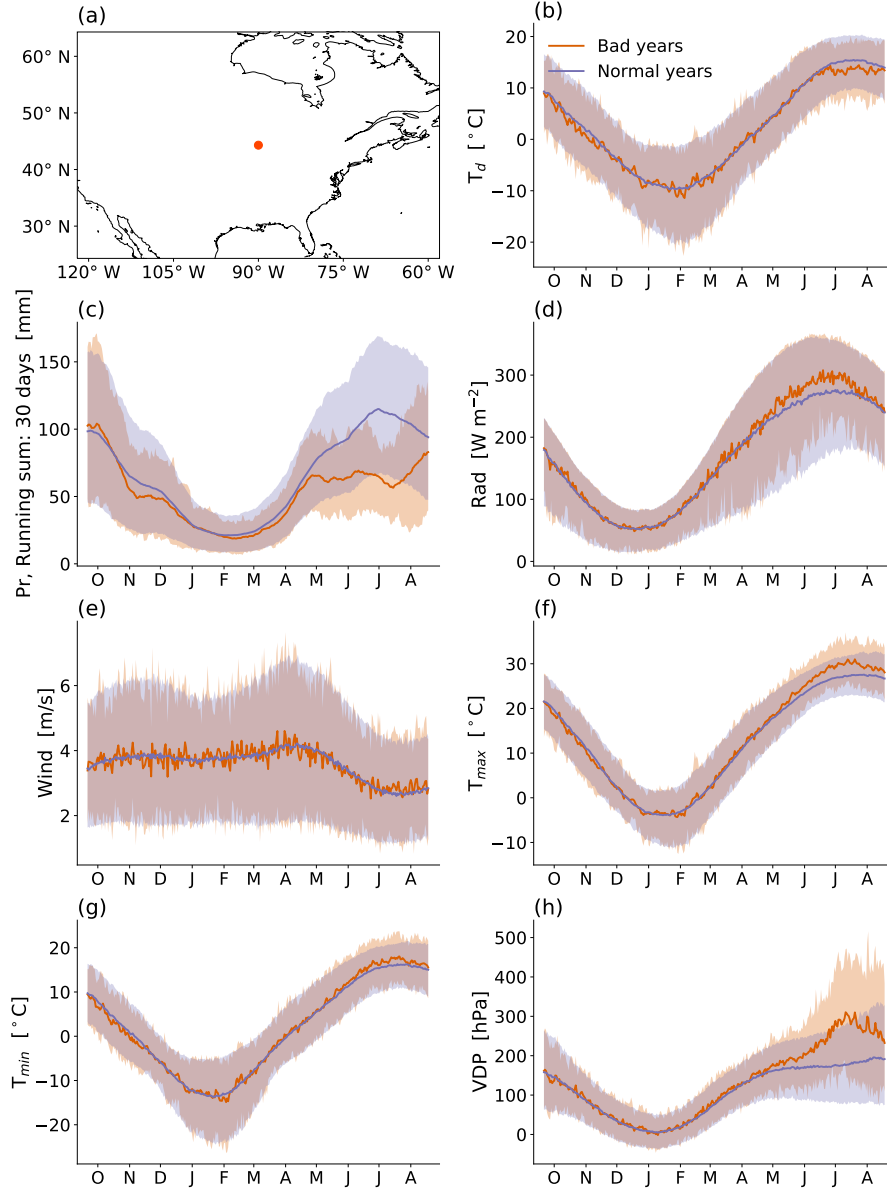


FIGURE A.10: As Figure A.2, but for a grid point in the United States ( $90.0^\circ$  W,  $44.3^\circ$  N).

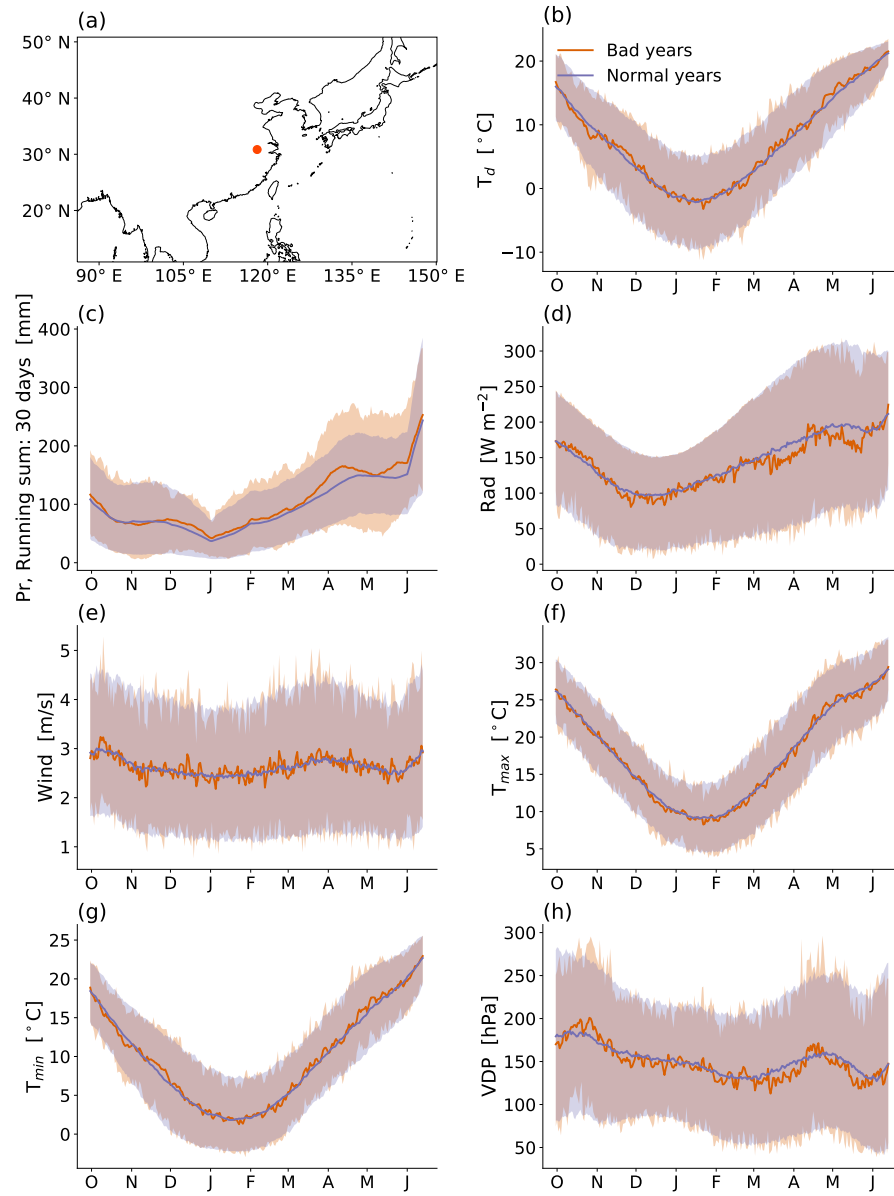


FIGURE A.11: As Figure A.2, but for a grid point in China ( $118.1^{\circ}$  E,  $30.8^{\circ}$  N).

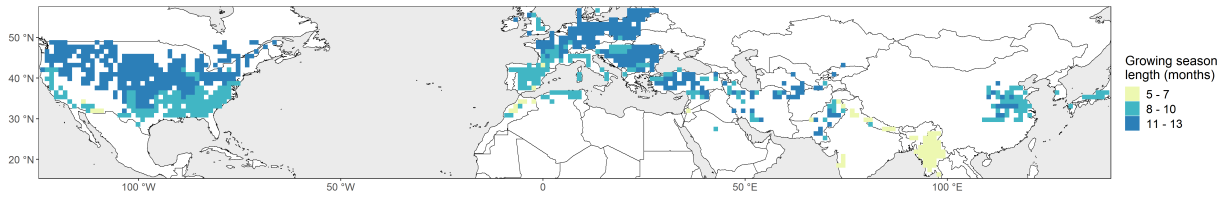


FIGURE A.12: *Number of months in the growing season (number of months between the earliest sowing date and the latest harvest date). The growing season starts at the month containing the sowing date and ends with the month containing the latest harvest date, among the 1600 model years. We discarded years with a harvest date later than 365 days after the sowing date. Some growing seasons are 13 months long because we include both the entire first month and the entire last month.*

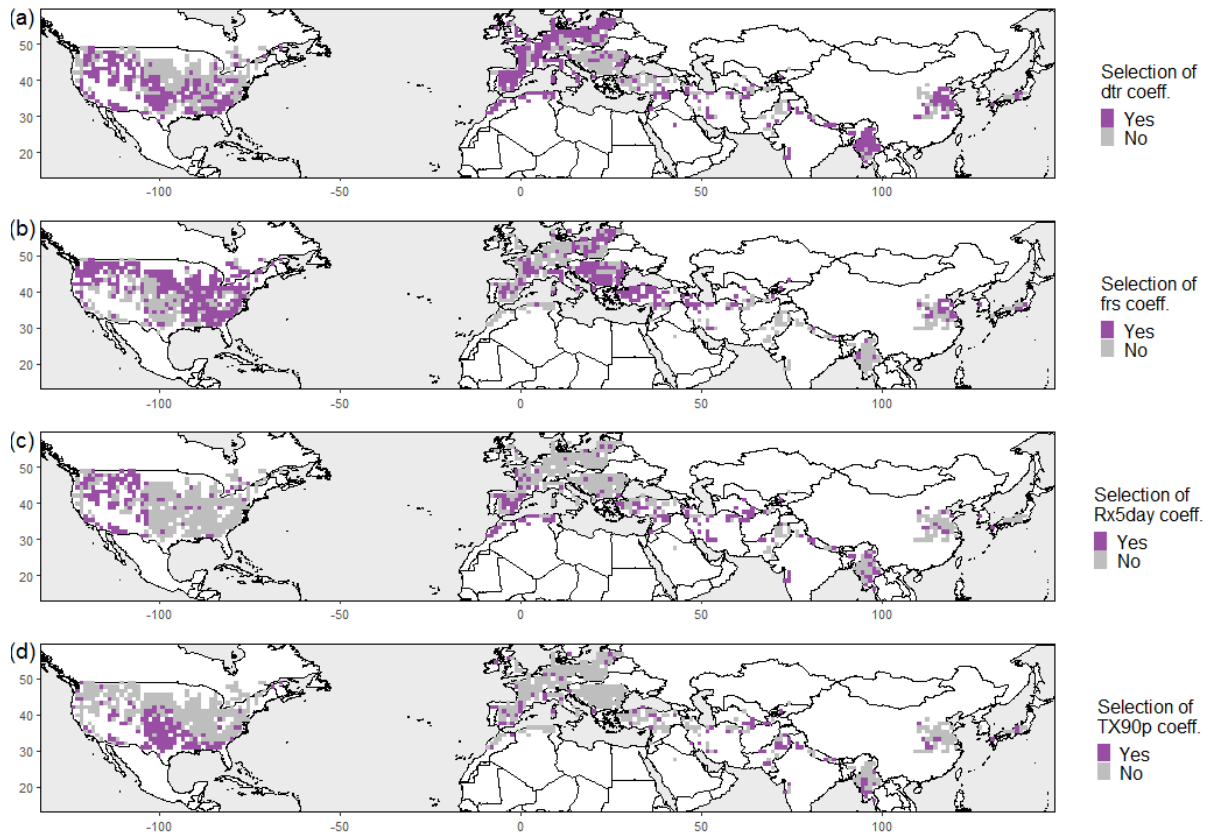


FIGURE A.13: Selected climate extreme indicators (Table A.1) in the Lasso logistic regression model for each location. Diurnal temperature range (*dtr*, a), number of frost days (*frs*, b), *Rx5day* (c) and *TX90p* (d).

**Author contributions** J.Z. and K.v.d.W. conceived the project and supervised the work. J.V. and P.R. conducted most of the data analysis, including the Lasso logistic regression and creation of the key figures. K.v.d.W. performed the climate model simulations with EC-Earth. T.Z. performed the crop model simulations with APSIM. All authors contributed substantially to the data analysis, design of figures and writing of the manuscript.

**Competing interests** The authors declare that they have no competing interests.

**Acknowledgment** This work emerged from the Training School on Statistical Modelling organized by the European COST Action DAMOCLES (CA17109). J.V. acknowledges funding by the DFG research training group “Natural Hazards and Risks in a Changing World” (NatRiskChange GRK 2043). Part of this work was funded by the Swiss National Science Foundation (grant numbers 178751 (P.R.), 179876 (E.T. and J.Z.) and 189908 (J.Z.)). C.A.S. is funded by an EPSRC Doctoral Training Partnership (DTP) Grant (EP/R513349/1). J.Z. acknowledges the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537). K.v.d.W. and T.Z. acknowledge funding for the HiWAVES3 project from the

National Natural Science Foundation of China (41661144006), funding was supplied through JPI Climate and the Belmont Forum (NWO ALWCL.2 016.2 and NSFC 41661144006), T.Z. further acknowledges the National Key Research and Development Project of China (2019YFA0607402).



## APPENDIX

### B

# A NOVEL METHOD TO IDENTIFY SUB-SEASONAL CLUSTERING EPISODES OF EXTREME PRECIPITATION EVENTS AND THEIR CONTRIBUTIONS TO LARGE ACCUMULATION PERIODS

This chapter contains the abstract of an article published in 2021 with the title “A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods” in the journal *Hydrology and Earth System Sciences* (Kopp et al., 2021). This paper was written by Jérôme Kopp, together with Pauline Rivoire (advisor for his master thesis), S. Mubashshir Ali, Yannick Barton and Olivia Martius.

## Abstract

Temporal (serial) clustering of extreme precipitation events on sub-seasonal timescales is a type of compound event. It can cause large precipitation accumulations and lead to floods. We present a novel, count-based procedure to identify episodes of sub-seasonal clustering of extreme precipitation. We introduce two metrics to characterise the prevalence of sub-seasonal clustering episodes and their contribution to large precipitation accumulations. The procedure does not require the investigated variable (here precipitation) to satisfy any specific statistical properties. Applying this procedure to daily precipitation from the ERA5 reanalysis dataset, we identify regions where sub-seasonal clustering occurs frequently and contributes substantially to large precipitation accumulations. The regions are the east and northeast of the Asian continent (northeast of China, North and South Korea, Siberia and east of Mongolia), central Canada and south of California, Afghanistan, Pakistan, the southwest of the Iberian Peninsula, and the north of Argentina and south of Bolivia. Our method is robust with respect to the parameters used to define the extreme events (the percentile threshold and the run length) and the length of the sub-seasonal time window (here 2–4 weeks). This procedure could also be used to identify temporal clustering of other variables (e.g. heat waves) and can be applied on different timescales (sub-seasonal to decadal). The code is available at the listed GitHub repository.

## APPENDIX

### C

# GUIDELINES FOR STUDYING DIVERSE TYPES OF COMPOUND WEATHER AND CLIMATE EVENTS

This chapter contains the abstract of an article published in 2021 with the title “Guidelines for Studying Diverse Types of Compound Weather and Climate Events” in the journal *Earth’s Future* (Bevacqua et al., 2021). This paper was written by Emanuele Bevacqua, together with Carlo De Michele, Colin Manning, Anaïs Couasnon, Andreia F. S. Ribeiro, Alexandre M. Ramos, Edoardo Vignotto, Ana Bastos, Suzana Blesić, Fabrizio Durante, John Hillier, Sérgio C. Oliveira, Joaquim G. Pinto, Elisa Ragno, Pauline Rivoire, Kate Saunders, Karin van der Wiel, Wenyan Wu, Tianyi Zhang and Jakob Zscheischler. The work emerged from a workshop organized by the European COST Action DAMOCLES.

## Abstract

Compound weather and climate events are combinations of climate drivers and/or hazards that contribute to societal or environmental risk. Studying compound events often requires a multidisciplinary approach combining domain knowledge of the underlying processes with, for example, statistical methods and climate model outputs. Recently, to aid the development of research on compound events, four compound event types were introduced, namely (a) preconditioned, (b) multivariate, (c) temporally compounding, and (d) spatially compounding events. However, guidelines on how to study these types of events are still lacking. Here, we consider four case studies, each associated with a specific event type and a research question, to illustrate how the key elements of compound events (e.g., analytical tools and relevant physical effects) can be identified. These case studies show that (a) impacts on crops from hot and dry summers can be exacerbated by preconditioning effects of dry and bright springs. (b) Assessing compound coastal flooding in Perth (Australia) requires considering the dynamics of a non-stationary multivariate process. For instance, future mean sea-level rise will lead to the emergence of concurrent coastal and fluvial extremes, enhancing compound flooding risk. (c) In Portugal, deep- landslides are often caused by temporal clusters of moderate precipitation events. Finally, (d) crop yield failures in France and Germany are strongly correlated, threatening European food security through spatially compounding effects. These analyses allow for identifying general recommendations for studying compound events. Overall, our insights can serve as a blueprint for compound event analysis across disciplines and sectors.

---

## Acknowledgement

---

First of all, I want to express my deep gratitude to my two PhD supervisors. Thank you for both being constantly encouraging and kind, for being so patient and available. Your excellent guidance and your great research ideas have driven my motivation during these four years. You represent complementary sources of inspiration for me. Thank you very much Olivia, for your kindness, a central key for the pleasant working environment. Philippe, mille mercis for your bienveillance, for the nice discussions and the great moments in Aussois.

I would like to thank the university of Bern, the Oeschger Centre for Climate Change Research and the SNF, for making this PhD project possible. Thank you Anne-Catherine for grading this thesis as an external examiner. I would also like to thank the COST action Damocles and my friends from group project 1 of the training school in Como. Thanks should also go to my favorite softwares: R (R Core Team, 2020), L<sup>A</sup>T<sub>E</sub>X and Python (Van Rossum and Drake Jr, 1995).

Merci Philomène, my dear friend and collaboratrice. I am so grateful to have you as a friend and to have been able to share so much of the PhD experience together (et plus encore). Thanks a lot Martin, Hélène and Philomène for proofreading parts of my thesis, your help was extremely precious.

Thank you to my colleagues at GIUB from the 5th and the 4th floor. Many thanks to Muba, this PhD was a nice experience to share with you. Muchas gracias Noemi for cheering me up and for the random super interesting topics of conversation. Thank you Lucas for the nice inspiration for the poem. Thank you Pascal and Simon for the precious input at the beginning of my PhD. Thank you my dear Angie and Saba. Thank you Andrey, Ralf and Eric. Merci Alexandre, for the fruitful conversations.

Merci David, cher coloc en or. Thank you to my friends at the DCB for the good mood, merci Éléonore pour nos folies. Merci Delphine, Candice et Jeanne pour le soutien et la bonne humeur. Merci mes amies françaises, mes racines, Justine, Coralie, Typhanie, Aurélie, Capucine, merci d'avoir été et d'être une belle ressource dans la difficulté mais aussi une éternelle source de légèreté et bonne humeur. Merci Louise, pour ton amitié et les coups de pouce avec Python. Merci à mes camarades de l'AareThéâtre, comme une petite famille d'adoption qui m'a grandement aidée à me sentir chez moi à Berne. Merci aussi à toute la team d'Orsignac.

Un immense merci à ma famille chérie, Papa, Matthieu, Eliot, Catherine, Antoine et Marion. Thank you Victor and Laura for ze presence and ze legereté. Merci Maman pour la force.

And merci du fond du coeur Martin, for encouraging me, empowering me, feeding me, distracting me when needed, for giving me self-confidence when I am hesitant, for being so patient and so kind all the time.



# Declaration

under Art. 28 Para. 2 RSL 05

Last, first name: Rivoire, Pauline

Matriculation number: 18-115-030

Programme: PhD. in Climate Sciences

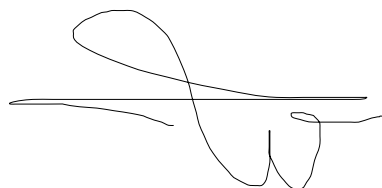
Bachelor ☐ Master ☐ Dissertation ☒

Thesis title: On the Assessment of Precipitation Extremes in Reanalysis and Ensemble Forecast Datasets

Thesis supervisors: Prof. Dr. Olivia Romppainen-Martius,  
Dr. Philippe Naveau

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due acknowledgement has been made in the text. In accordance with academic rules and ethical conduct, I have fully cited and referenced all material and results that are not original to this work. I am well aware of the fact that, on the basis of Article 36 Paragraph 1 Letter o of the University Law of 5 September 1996, the Senate is entitled to deny the title awarded on the basis of this work if proven otherwise.

Bern, Mai 26, 2022

A handwritten signature in black ink, consisting of a large, stylized 'P' followed by a horizontal line and a small 'R'.

Signature