GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

A COMPARISON OF SPACE-TIME MODELING APPROACHES FOR EXTREME PRECIPITATION IN GERMANY BETWEEN 1996 AND 2016

A thesis paper for the degree of Bachelor of Arts (B.A.)

Wintersemester 2022

By

Ekin Gülhan

Supervisor: Dr. Isa Marques
Enrolment number: 21675680
e-mail: ekin.guelhan(at)stud.uni-goettingen.de
Address: Hannoversche Straße 134, 37077 Göttingen
Date: 2023-02-13

Contents

Some Remarks

Dear Dr. Marques,

this version of the thesis paper is meant to lay out the general argument of my thesis paper and decisions I have made along the way until now, in the hope, that I can get some feedback in terms of disagreement or agreement. With some tasks, I have been successful in implementing them already, with others, I have not. For those ideas I have not finished implementing in R, I may just refer to the concept of their implementation. Also, I do not reference my explanations and statements as of now, which I do not believe to be relevant for our meeting on thursday just yet. Much of my work was focused on the theory behind the regression models, on how to design a reproducible analysis project in R, using best practices of coding and data science. Some of this is already reflected in the way, that the github repository is structured and particularily in the structure of the code scripts. I especially want to point out that I managed to maximize the mobility of the project by using a combination of the packages "here", "renv" and GitHub obviously, such that anyone who installs both packages should be able to execute the scripts and the late markdown file without the usual trouble of setting work directories and caring about package versions, while scripts themselves are trimmed for readibility by rigorous application of the tidyverse logic, which I believe to reflect the philosophy of R of function and data flow based coding well. If you actually want to have a look at the scripts, there are only a few preliminary steps you need to take (this might be redundant, depending on whether you are familiar with "renv" and "here" or not):

1. Open the r project file first
2. Install and require "here" and "renv"
3. Execute "renv::restore()" to load the packages from the renv lock file into the project library in the right versions.

Thanks for reading and see you on thursday.

Abstract

Introduction

The need for a scientific approach to flood risk management and the purpose of the thesis paper

When as a result of continuous extreme precipitation Central Europe was struck by massive floods in 2013, among other affected areas, Germany suffered loss of life and property. During this time of crisis, the public observed German institutions of flood risk management take effect and soon politicians began to publicly announce and evaluate the state of crisis and countermeasures. As they did, critical voices arose also and assertions were made, whereafter the responsibility for the losses at least partially lie with political failure of then contemporary institutions and officials. Research conducted in the aftermath evaluating the preparation for and response to the floods of 2013 indeed indicate the negative effect of political structures on the performance of the flood risk management of 2013, including the relevance of the dissemination of flood related information through government agencies, which the municipalities' officials found difficult to interpret and apply. Where questions related to how to improve local access to crucial information of the time and place of extreme precipitation events (EPE) naturally warrant answers thus, a scientific approach to their answers is relevant all the more.

The design of this thesis project: Science, Data Science and Reproducibility

Against this background, the present thesis paper has set out to determine, which of two spatio-temporal classifier models perform better in the task of predicting extreme precipitation events in Germany, while observations of precipitation height between 1996 and 2012 have been used in training and evaluating the classifier models and observations between 2013 and 2016 have been fed to the trained models for model validation and model comparison respectively. As the author strives for a data scientific career post graduation, a focus has been set on not only applying scientific, but data scientific best practices in particular, which rendered reproducibility of the analyses a major concern. For this reason, the thesis paper has been produced within a rigorously

encapsulated and mobile project directory, to which access is available via a GitHub repository. The R package "here" has been used to allow for such mobility in the first place, as it enables the substitution of absolute file paths with relative file paths, such that a path is no longer defined as the absolute and hierarchical location of the file on the computer, but as the position of the file relative to the topmost folder, where the R project file resides. In this sense, the R project file is the center of the practical implementation of the thesis. Using relative paths then, different parts of the data manipulation, modelling and analysis have been isolated in scripts, that can be executed on their own and independent of the markdown file, from which this thesis is knitted. As is the case elsewhere, computational functionality in those scripts relies on R packages provided by members of the R community. Reproducing the analytical steps however necessitates not only that the packages are installed, but that they are installed in the form of a specific version. To minimize the effort needed for such package version control, the R package "renv" has been utilized, which creates a project package library secondary to the system library. A so called "renv.lock"-file is included in the project repository, which is a reference to all the packages (and their specific versions) required for the reproduction of the thesis and its' analyses. Using the renv package, anyone with access to the thesis project repository can easily install the package in the required versions via the "restore()" function. Finally, since the author considers the tidyverse as the current best practice of data manipulation and programming of data streams in R, packages associated with it and in particular tidyverse pipes have been used rigourously.

Framing the problem of spatio-temporally predicting EPE as a binary classification problem

As a preliminary to modelling the prediction of EPE spatio-temporally, firstly, a conceptual grounding of such EPE is due. Since this is no priority in this analysis however, such grounding will be attended to as necessary only. And indeed, it is not outright obvious when a precipitation event (PE) is to be regarded as extreme. A plethora of approaches have their common origin as early as making a choice of what constitutes a precipitation event. Generally, researchers define a precipitation event based on precipitation height, duration or intensity or based on a specific combination of the three. Here, a precipitation event is defined by the daily precipitation height only, as it is believed to be a potent predictor of the environmental impact and damage to infrastructure in itself. Among the possible temporal resolutions, the choice fall on days and against subdaily resolutions (hourly precipitation) or more aggregated resolutions (monthly or annual precipitation), for reasons persistent with the previous argument: Daily precipitation totals are likely better predictors of environmental impact of precipitation than other resolutions, which ultimately is the primary concern for local risk managers. These two choices for essentially basing the concept of precipitation events on daily total precipitation height have already cut away much of the variety of approaching the conceptual grounding thus. At this point, where a daily PE is characterized by the daily precipitation height associated, to also complete the definition of EPE, a threshold (or cut-off) value needs to be decided upon for a specific standard of thruth for classification within this thesis.

A specific percentile of the empirical distribution of precipitation could be used as such threshold value and this approach is indeed being applied in the literature in some cases. Obvious related shortcomings however are the lack of scientific research that indicates, that any specific quantile of the empirical distribution performs well in classifications, that reliably predict risks of environmental hazard. Alternatively, there is a rich whole area of statistics to draw concepts for the cut-off value from, called extreme value theory. This theory has a more differentiated view of the analysis of the values at the precipitation height distributions tail. Attempts to tie it into this thesis however would have increased the difficulty unnecessarily, however. Hence the choice for the threshold value fell on a easily comprehensible, yet powerful concept: Hereafter, a PE is classified

as extreme, if the daily total precipitation height exceeds the average monthly total precipitation height for the month, in which the daily precipitation height was observed. In other words: The standard of truth for the classification of a PE as an EPE is, that a PE is an EPE, if the daily precipitation height exceeds the typical total precipitation of the month it belongs to.

Developing the modelling GAMM with a logit link function as a solution to the classification problem

Continuing the previous argument, that the mitigation or management of the risk of floods caused by EPE fundamentally requires a solid model for the spatio-temporal prediction of prediction of EPE and that the problem of such a prediction is to be conceptualized modelling a spatio-temporal classifier model, here, another argument will be developed, whereafter Generalized Additive Mixed Models with a binomial/logit link function are a suitable solution to the respective classification problem.

And the first step to this is awareness for the desired properties of the classifier. The first such property is that the predictor of the model includes spatial and temporal predictor variables, whithout whom, the occurrence of a EPE can not be depicted for a specific place within German boundaries at a specific time. Secondly, the model has to be capable of handling a binary response variable such as EPE. Thirdly, as the classifier models the dependency of the binary response EPE on the different covariates, including location in time and place, it is also required to account for spatial, temporal and spatio-temporal dependencies, as there evidence strong enough to safely assume their existence. Fourthly, as the various effects on the occurence of an EPE are non-linear presumably, while training the classifier model it should be able to pick up non-linear patterns.

One way of showing how a GAMM with a logit link function meets these conceptual requirements is to built a step by step from a linear regression model forward and show, through which model transformations specific shortcomings of the respective model with regard to the three requirements above can be mitigated.

Classical Linear Regression Model -> Logistic Regression Model -> Generalized Linear Mixed Model with binomial link function -> Generalized Additive Mixed Model with binomial Link Function

Session Info