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: 2023-02-20

Contents

1 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 (you may need to execute this command several times).

Thanks for reading and see you on thursday.

2 Abstract

3 Introduction

3.1 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.

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

Against this background, the present thesis paper is set out for determining, 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. 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. Finally for readibility, within mathematical expressions used in this thesis, vectors are denoted through bold-font while matrices are denoted through bold-font plus a description of the dimension as subscript.

4 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 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 fell 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.

5 A classifier based on a GAMM with a logit-link for the spatio-temporal classification of EPE

The following section will argue, that on a conceptual level, the problem of modelling the spatio-temporal classification of PE as EPE can be approached reasonably by building a generalized additive mixed model with a logit link function, which is then superimposed by a filtering probability threshold layer. The threshold layer superimposing the built GAMM is thereby ultimately responsible for separating the EPE from the PE based on the criteria, whether the predicted probability succeeds the threshold layer or not. In this context, *reasonably* refers to the fact, that by building a classifier based on a GAMM, all of the essential criteria for the model can be met. The criteria in turn follow from the general objective of this thesis, to built a classifier model that facilitates flood risk management, as follows: Firstly, the classifier model has to be capable of translating an input of the spatial and temporal location of the precipitation event in question into a classification decision. In terms of the model, this relates to the presence of both the spatial and temporal location of the event as covariates. Modelling the classification spatio-temporally does also neccessitate the model to identify spatial, temporal and spatio-temporal dependencies within the data, since neglecting those dependencies would result in structurally classification decisions / prediction errors. Next to the space-time character of the model, more generally, it should also be possible to train the classifier to identify non-linear relations between the classification decision and the covariates. Research indicates in fact, that the effect of variables such as altitude of spatial locations and mean-temperature, which are generally considered strong predictor variables, is non-linear. A classifier model finally involves a binary response, where in this case, a predicted value of 1 translates into "PE in class of EPE", whereas a predicted value of zero translates into "PE not in class of EPE". It comes therefore naturally, that the respective model must be able to handle a binary response variable conceptually.

One way to show how a classifier based on a GAMM with a logit link meets all these criteria is to start with the arguably most powerful, prevelant and comprehensible regression model, the classical linear regression model. Although the classical linear regression model only meets one requirement of modelling the dependency of precipitation events on spatial and temporal covariates, it can be shown, that following successive interventions within the model, step by step more of the requirements to the classifier model are met, such that the classifier based on the GAMM with logit-link ultimately is developed conceptually as a model to meet all criteria.

5.1 Classical Linear Regression and its shortcomings

Classical linear regression involves modelling the model response as the conditional expected value of a variable and this expected value as being dependent on one or more covariates in a linear fashion, plus error term. The random error terms are assumed to be identically and independently Gaussian distributed. The covariates form a linear predictor, such that based on the modelled dependence of the response on both the random errors and the linear predictor, the response is conceptualized as a random variable as well, whose distribution, beyond adopting the i.i.d. features, is additionally influenced by the linear predictor. Importantly, as the error terms conceptually account for random, unsystemic differences between the response and the linear predictor as well as measuring errors they are assumed to be independent not only of themselves, but also of the linear predictor. Those characteristics collectively render the linear regression fundamentally inadequate to the present purpose: For once is the response as mentioned assumed to be Gaussian distributed, whereas for a binary response such as a classification decision derived from a probability threshold layer, a binomial distribution need be assumed. In addition to this reason for the inaptitude for modelling a binary response like the classification of PE as EPE, a second reason stems from the inconsistency of strong evidence for the existence of spatial, temporal and spatio-temporal dependencies between precipitation heights and within the conditional dependency of the precipitation height of the covariates on the one side and the assumed independence among the responses and among the error terms on the other side: If it is (safely) assumed, that the responses (the classification decisions) are non-independent - autocorrelation - then this dependencies have to either be captured through the linear predictor or the error terms. As the linear predictor however is inapt for modelling any other than linear effects of the covariates on the responses, these independencies would conceptually be assumed to be reflected in the error terms. The dependencies would therefore reflect in the error terms, particularly as autocorrelation of the error terms, which contradicts the assumption of i.i.d. error terms. Hence, as linear regression is by construct inapt to meet the requirements of modelling a binary response and modelling spatio-temporal dependencies, and as this inaptitude finally also extends to modelling non-linear dependencies, linear regression overall is ill-suited for modelling the classification of PE as EPE, as this thesis is set out for.

Classical Linear Model Equation

$$\begin{aligned}\textbf{Precipitation Height} = \boldsymbol{\beta_0} + \beta_1\textbf{Mean Temperature} + \beta_2\textbf{Altitude} \\ + \beta_3\textbf{Longitude} + \beta_4\textbf{Latitude} + \beta_5\textbf{Julian Date} + \epsilon\end{aligned} \tag{1}$$

5.2 Logistic Regression Models as *Generalized* Linear Models

The classical linear regression model can still at least predict some response from the spatial and temporal location of a PE. With changes of the model structures, it can be raised to a model with a binomially distributed binary response variable such as decisions of classification of PE as EPE. To achieve such a generalization of the linear model, first of all the error term is dropped from the right hand side, such that the response is no longer modeled as dependent on the linear predictor plus an error term, but only on the linear predictor. Instead, the response itself is modeled to consist of a variable plus an error term. In the next step the response is modelled neither a continuous nor binary variable, but as the so called log-odds (logit) of the probability, that a PE belongs to the class of EPE. This step allows for modelling a continuous response as depending on a linear predictor and is one of two pivot points of the generalization. For the final step of the generalization, both sides of the model equation - the response and the linear predictor - are supplied to the inverse the of logit link-function. As a result, the model response is now a probability value, which is modeled as dependent of the inverse of the logit link-function applied to a linear predictor. This generalization yields a probability estimation model, whose predictions or responses respectively are assumed to be binomially distributed. From here, it would only take the small step of superimposing a threshold-layer to raise the logistic regression model with its binomially distributed probability responses to a classifier model with a binomially distributed binary decision of the classification of PE as EPE as response. As a consequence of these steps of generalization, within the linear predictor the coefficients are no longer additively concatenated, which is the key condition for the application of the method of least squares to estimate the coefficients. Instead, the model (coefficients) are estimated via maximizing the model likelihood.

$$\boxed{\textbf{Logistic Regression Model}}$$

$$\mathbf{p(PE\ in\ class\ EPE)} = \mathrm{S}\left(log\left(\frac{p}{1-p}\right)\right) = \mathrm{S}\left(\begin{matrix}\beta_{\mathbf{0}} + \beta_1\textbf{Mean Temperature} + \beta_2\textbf{Altitude}\\ + \beta_3\textbf{Longitude} + \beta_4\textbf{Latitude}\\ + \beta_5\textbf{Julian Date}\end{matrix}\right) \tag{2}$$

where

$$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \text{ are } \textit{estimated} \text{ per } \textit{Maximum Likelihood Estimation},$$

such that

$P(EPE_i|\mathbf{x}_i, \beta) = p_i^{EPE}(1-p_i)^{1-EPE}$ is the likelihood of a single classification decision

and

$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5})}}$ is the predicted probability, that an EPE occurs at spatio-temporal location i

and

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \prod_{i=1}^{n} P(EPE_i|\mathbf{x}_i, \beta) = \prod_{i=1}^{n} p_i^{EPE_i}(1-p_i)^{1-EPE_i}$$

5.3 Generalized *Additive Mixed* Models (GAMM) with a logit link-function and spline-based smooth functions as non-linear special cases of GAMMs

Presently, following the above mentioned steps of generalization and on a conceptual level, the spatio-temporal classification of PE as EPE is arguably already partially modelled: The Logistic

Regression Model has a probability value as its response after the superimposition of a threshold-layer, the classification of the PE as EPE can be modeled as depending on a linear predictor consisting out of covariates such as altitude, mean-temperature, spatial location and temporal location, while the linear predictor is supplied to the inverse of the logit link-function. From here, the remaining two requirements, the model's ability to capture spatial, temporal and spatio-temporal dependencies as well as capture non-linear relations between response and predictor, are introduced in two simultaneous steps: Another generalization, namely the substition of the linear effects through smoothing-terms, as well as the introduciton of random effects. As a consequence, the originally linear predictor is no longer linear and since the smoothing terms are still additively concatenated, ultimately a Generalized Additive Mixed Model with a logit link-function and smoothing terms for effects is built. As this GAMM is able to model the probability of a PE being an EPE as being spatio-temporally dependent on the mentioned covariates, it takes the superimposition of a filtering probability threshold-layer to produce a classifier model that finally meets all the discussed requirements.

6 The Necessity for modelling a ***SPATIO-TEMPORAL*** classification procedure

Within this thesis, I assert, that the clasifier models to be built have to account for spatial, temporal and spatio-temporal dependencies in the precipitation heights data. To support this assertion, this chapter is set out to demonstrate these spatial and temporal dependencies, where the variation of spatial and temporal covariability will be visualized and analyzed accordingly. Tools, that are used to demonstrate these dependencies are

- One spatial and one temporal Semivariogram
- One spatial and one temporal Covariancematrix
- Spatial autocorrelation measured as spatial covariance as a function of latitudinal lag d
- Temporal autocorrelation measured as temporal covariance as a function of temporal lag p

6.1 Spatial Semivariance Equation (for Semivariogram)

$$\gamma_{\text{lon}}(d) = \frac{1}{2N(d)(T-p)} \sum_{i=1}^{N(d)} \sum_{t=p+1}^{T} (rain_{\text{lon},lat_i+d,t} - rain_{\text{lon},lat_i,t-p})^2$$

6.2 Spatial Covariance Equation

$$\hat{C}_{\text{lon}}(lat_i, lat_k) = \frac{1}{T-p} \sum_{t=p+1}^{T} \left(rain_{\text{lon},lat_i,t} - \frac{1}{T_{\text{lat}_\text{i}}} \sum_{t=1}^{T_{\text{lat}_\text{i}}} rain_{\text{lon},lat_i,t} \right) \left(rain_{\text{lon},lat_k,t-p} - \frac{1}{T_{\text{lat}_\text{k}}} \sum_{t=1}^{T_{\text{lat}_\text{k}}} rain_{\text{lon},lat_k,t-p} \right)$$

where lat_i **and** lat_k **are latitudinal degrees, such that** $(\text{lat}_\text{i},\text{lat}_\text{k})$ ϵ [**47.4,54.8**]

and where lon **denotes one of five latitudinal intervals of equal length, as follows:**

- Interval 1: 6.02 to 7.49 degrees longitude[1]
- Interval 2: 7.49 to 8.96 degrees longitude[2]
- Interval 3: 8.96 to 10.43 degrees longitude[3]
- Interval 4: 10.43 to 11.9 degrees longitude[4]
- Interval 5: 11.9 to 13.37 degrees longitude[5]
- Interval 6: 13.37 to 14.95 degrees longitude[6]

6.3 Spatial Latitudinal Autocorrelation Equation

$$\hat{C}_{\text{lon}}(d,p) = \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{\text{lat}_\text{i}} \sum_{\text{lat}_\text{k}} \left(rain_{\text{lon},\text{lat}_\text{i},t} - \frac{1}{T_{\text{lat}_\text{i}}} \sum_{t=1}^{T_{\text{lat}_\text{i}}} rain_{\text{lon},\text{lat}_\text{i},t} \right) \\ \times \left(rain_{\text{lon},\text{lat}_\text{k},t-p} - \frac{1}{T_{\text{lat}_\text{k}}} \sum_{t=1}^{T_{\text{lat}_\text{k}}} rain_{\text{lon},\text{lat}_\text{k},t-p} \right) \delta_{\text{lat}_\text{i}-\text{d},\text{lat}_\text{k}}$$

[1] If the observation belongs to interval 1, then lon = 1.
[2] If the observation belongs to interval 2, then lon = 2.
[3] If the observation belongs to interval 3, then lon = 3.
[4] If the observation belongs to interval 4, then lon = 4.
[5] If the observation belongs to interval 5, then lon = 5.
[6] If the observation belongs to interval 6, then lon = 6.

where

$$\delta_{lat_i-d,lat_k} = \begin{cases} 1, & \text{if } lat_i = lat_k + d \\ 0, & \text{otherwise} \end{cases}$$

7 Session Info