1 Response to Referee #1

Thank you very much for your informative and detailed comments. It's obvious that we will have to work on two aspects of the paper during the revision process: firstly, We have to motivate the need for this method better and secondly we have to present the method in a more accessible way.

As for the motivation:

The main motivation to process the raw data by a statistical model is to improve the signal-to-noise ratio. The raw lightning data contains a lot of noise due to the high variability of processes generating lightning. In general this is not only true for lightning, but also for most other atmospheric variables, e.g. precipitation, wind speed and direction, etc. The GAM applied in our study aims at filtering effects/signals associated with altitude, day of the year and space and to separate these from the noise.

We extended the introduction of the manuscript with respect to this point in order to enhance the motivation of the method.

Furthermore, we have to present the method in a more accessible way for readers unfamiliar with advanced regression methods. We are aware that the mathematics behind the methods is complex and maybe even daunting for readers with no or little statistical background. However, we are willing to work on that issue, i.e., presenting the method in a more accessible way, in order to encourage more scientists from the lightning community to work with GAMs.

General point 1:

I have read this paper dealing with the lightning climatology in Austria. While the paper is well written, clear and at a high level of English, I am not sure why a model is needed to describe the lightning climatologies, when the raw data is already available to the authors. The authors state that for risk assessment or when climatology is used as a benchmark weather forecast this model will be valuable. But why do we need a model when we have the actual real lightning climatology. If we need to know what the probability is of lightning hitting location A, we can calculate this from the raw data.

Answer:

This point is already partly addressed by the motivation given above. Regarding the estimation of the probability of lightning at location A we have to admit that this is also possible by averaging the observed lightning days over all years in the data base and maybe also to include some neighboring locations in order to smooth the estimate. However, things are getting more complex when this has to be applied to several locations simultaneously. This is especially true for regions with complex terrain, where a smoothing is desired not only over space and time but also over the altitude. Estimating climatological values by averaging easily leads to an arbitrary selection of smoothness. Thus our aim is to present a neat method that helps researchers to produce valuable climatologies from their raw lightning data.

General point 2:

In addition, the model is developed using the lightning data itself, and then tries to predict the same lightning data. So the model input and output are not independent of each other. A correct model should use parameters A, B and C to predict D. not A, B and D to predict D. Furthermore, the model should be developed for a specific period, i.e. 1992 to 2000 (for example) and then tested on the year 2001 to see if the model can reproduce the lightning of 2001. In fact, it would be interesting and valuable to compare the model output (2001) with the real data (2001). How well correlated are the lightning estimates by the model for 2001 (based on a model constructed with input data from 1992-2000) with the real lightning from 2001. That is a legitimate test of the model.

Answer:

This is absolutely right. We applied this procedure by cross-validation. 6 years of data are available. The parameters of the model are estimated based on 5 years of the data and validated on the remaining year. This is done 6 times with every single year serving as validation period once. All scores presented in the study are based on this procedure which is state-of-the-art in statistics.

In the new version of the manuscript we describe this procedure in more detail in an extra verification paragraph which is part of the method section.

General point 3:

Finally, if the model is a physical model, then it should be applicable to other regions of Austria. How well does this model predict lightning in other regions of Austria (or Europe)? If it is only good for Carinthia, then why bother? Just use the real observed data for risk maps.

Answer:

It is not a physical model, but a statistical one. The presented method can still be applied to other regions.

Specific comments:

Page 2

Line 2: of Austria vary – Corrected.

Line 24: data are – Corrected.

line 30: period are –Corrected.

Page 3

line 1: What is the detection efficiency of the intra-cloud flashes relative to the CG flashes?

Answer: Because it is impossible to determine the detection efficiency of intra-cloud flashes without a locally installed VHF network (e.g. LMA), there was no attempt made in Schulz (2016) to characterize the detection efficiency of intra-cloud flashes.

This information is added to the data section.

line 6: what about the detection efficiency in %?

Answer: Schulz (2016) show that the flash detection efficiency is greater than 96% (100%) if one of the return strokes in a flash had a peak current greater than 2kA (10kA).

This information is added to the data section.

line 13: data are – Corrected.

line 24: Is it a correct assumption to assume a random process?
Answer: Yes, it is a general assumption made for statistical modelling.

Page 4: This reminds me of the KISS principle.....(Keep it Simple, Stupid)
Answer: Yes, we tried to keep it simple. However, I agree that the reader might feel overwhelmed. Nevertheless, eq. (3) is important as it introduces the smoothing parameter lambda which is tuned by cross-validation.

Page 5

line 27: The raw data also shows that the main lightning season is from June until end of August. So what is so great about this model? Why do we need it? To tell us something we already know from the raw data? I dont understand the logic behind the model. What can it tell us that we dont already know.
Answer: Cf. general point 1.

Page 8
line 21: But if we HAVE the climatology, why do we need this tool?
Answer: Cf. general point 1.

line 24: I do not understand why this is a useful tool when the raw data give a better estimate of the climatology.
Answer: Cf. general point 1.

line 26: smooth estimates can also be obtained by averaging the raw data temporally and spatially.
Answer: That is true. This procedure would refer to a k-nearest-neighbor estimation. However, one would have to find the optimal width for the smoothing windows in time and space. The analogy between finding the smoothing parameter lambda in a GAM framework and finding the width for a smoothing window is illustrated by Hastie et al. (2009, chapter 6.2).

line 27: Why not simply use the real raw data? I do not understand why a model is needed.
Answer: Cf. general point 1.

Page 12

Figure 2 caption: cells with – Corrected.

Page 13

Figure 3 (which is now Figure 4): What are units of y-axis in upper plots?
Answer: No units. The values are on the scale of the additive predictor, i.e. the right hand side of eq. (1).

Page 14

Figure 5: How does this differ from the raw data climatology. Maybe show one next to the other.
Answer: Thanks. That's a good idea. The new Figure 3 shows such a climatology as comparison.

Page 15

Figure 7 (which is now Figure 8): What are the numbers in the key of the figure on right. 2.1? 7?.
Answer: The legend shows expected number of flashes accumulated over the lightning season.

2 Response to Referee #2

Thank you very much for your informative comments. They will clearly help improving the quality of the manuscript.

General Comments:

"This paper is basically comprehensible, well structured and written in good English. Moreover, the general idea of the paper is interesting and the given approach is straight forward and certainly viable."

Answer: Thanks for acknowledging this. We agree that the approach is straightforward for someone with experience in GAMs. However, we feel that this is not necessarily the case for all readers of this manuscript at the intersection of lightning science, climatology, and applied statistics. Hence one objective of the manuscript is to bridge some of the gaps and make GAMs more accessible to researchers in the field of lightning science. Both your comments and those of Referee 1 show that we haven't fully accomplished this goal and hence we are grateful for your suggestions for improvements.

"Since I got the impression that a major asset is the modeling for a complex terrain, I would like to know what is the benefit of adding an altitude effect to the statistical model, whereas the lon/lat part seems to be the most influential effect? Moreover, I am not sure whether spatial function and altitude function are really distinct. Isn't it just sufficient to take the location into account because it implicitly contains the altitude?"

Answer: It is true, the altitude is a function of longitude and latitude. In general the presented method would be capable to model the influence of the altitude within the spatial effect implicitly. However, the shape of the altitude in the region of interest is very complex. Thus, a spatial effect with a large degree of freedom would be required in order to account for the complex altitude shape. As we know the shape of the altitude we can pass it to the GAM as an isolated effect. The altitude effect contains only information associated with the altitude while the remaining effects are captured by the lon/lat term.

This aspect is now further explained the results section.

"Finally, in terms of verification, it is not clear what kind of scores were calculated or used and what their results are."

Answer: The log-likelihood is applied, also called logarithmic score in the literature on proper scoring rules (see Gneiting 2007). We tried to avoid showing the results of the scores in detail, which would mean showing a longish table with proposed values of the smoothing parameter and associated scores from which the best is selected. Instead we wanted to put more emphasis on the results.

We added a paragraph in the methods section to discuss the verification score.

A table summarizing the verification scores would look like the table in this rejoinder.

Specific Comments:

Title

"I am afraid that the title 'Spatio-temporal smoothing of lightning climatologies' is misleading, because spatio-temporal smoothing implies some kind of grid-wise and time-wise moving average or filter, while the main idea of your study is to decompose the signal into a seasonal, spatial and also altitude effect by a statistical model. Reading the paper, I would have entitled it something like 'Statistical modeling of lightning climatologies for complex terrains' or 'Spatio-temporal smoothing of lightning climatologies for complex terrains'..."

Answer: Thanks for pointing this out and for suggesting the alternatives. We will change the title to "Spatio-temporal modeling of lightning climatologies for complex terrain".

λ	Q. 2.5%	Median	Q. 97.5%	d.o.f.
–	762455	765275	768364	0.00
5e+09	754052	756881	759698	1.07
1e+09	751785	754598	757377	1.27
5e+08	749880	752880	755764	1.45
1e+08	746750	749558	752554	2.11
5e+07	**746356**	**749251**	**752243**	**2.51**
1e+07	746754	749571	752341	3.77
1e+06	748266	751320	754269	6.73
100000	753277	756236	759019	11.78
10000	764341	767496	770352	19.38
0	786475	789802	792789	29.00

Table 1: 6-fold cross-validated negative log-likelihood for different smoothness parameters of the temporal effect. The dash in the λ-column indicates that no temporal effect was included into the model. Median, 2.5% quantile and 97.5% quantile was generated by bootstrapping 1000 times.

Introduction

"Reading your introduction, I got the impression that thunderstorms/lightning tends to occur at regions with moderate or lower altitude (page 2, lines 4-8). But your figures 3 and 4, top-left implying a positive and linear relationship between altitude and occurrence/intensity. Why doesn't the GAM fits a function with maxima for lower/moderate altitudes?"

Answer: The observed maxima at moderate or lower altitude are not the general case, but special cases associated with local effects. Thus it is not visible in the altitude effect. E.g., the maximum in the Gurktal Alps cannot be explained by its altitude, but the maximum has to be a consequence of local attributes of the terrain in that region.

We added a sentence in the results section where the effect is introduced.

Data

"Page 3, line 1: Reading this, with little experience on this scientific field, I would like to know the distinction between lightning, flash and stroke?"

Answer: Lightning is defined as a transient, high-current (typically tens of kiloamperes) electric discharge in the air whose length is measured in kilometers. The lightning discharge in its entirety is usually termed a 'lightning flash' or just a 'flash'. Each flash typically contains several 'strokes' which is the basic element of a lightning discharge.

We added a sentence in the data section to clarify the nomenclature.

"Maybe, it would be interesting to show a figure with the spatial climatologies of the number of flashes in Carinthia for the raw data."

Answer: Such a climatology is now part of the manuscript (the new Figure 3).

Methods

"page 4, line 6: As mentioned before, are altitude and horizontal space (lon/lat) really distinct. Thus, eq. (1) probably would have the form: $g(\theta) = \beta_0 + f_1(doy) + f_2(lon, lat, logalt)$"

Answer: As pointed out above one could just use $f_2(lon, lat)$ because logalt is a function of lon/lat but that would necessitate a very complex lon/lat term (using many degrees of freedom). The suggestion $f_2(lon, lat, logalt)$ could be interpreted as a spatially varying logalt effect. This is, in principle, also possible but is also more challenging to estimate. The additive decomposition $f_2(lon, lat) + f_3(logalt)$ is "the usual" trick of using an additive decomposition of the effect which leads to relative parsimonious effects $f_2 and f_3$. Of course, there may be even better parametrizations but this seems to work well and is (relatively) easy to interpret for practitioners.

Results

"page 5, line 20: How does the 1000 day-wise block-bootstrapping work?"

Answer: With day-wise block-bootstrapping we mean the following: We draw 738 dates of all available days with repetition and pick all the data observed on these days spatially. If we would relax the day-wise structure we would draw 7309152 samples with repetition from all available data points.
An explaination is now given in the verification paragraph of the methods section.

"page 6, line 4: Is there any explanation for the maximum in the Gurktal Alps, although this region is quite low elevated?"

Answer: We haven't found an explanation yet. However, in a follow-up study we will set the focus on analysis of single events and associated synoptical situations. Hopefully, this study will provide more insights.

"page 6, line 11: Is there any explanation for the flatter shape of the altitude effect function?"

Answer: We haven't found a sound and strong explanation for that shape.

Discussion

"In my opinion the part where the authors explain that cross-validation with day-wise blocks is much smoother and subsequently recommend to explore dependence structure of the data first would be more suitable for the method section."

Answer: Yes, we agree that this could be part of the methods section which would probably be the more natural section for readers with experience in flexible regression modeling (with GAMs). However, we deferred it to the discussion in order to make the methods section more accessible for readers not so familiar with GAMs. Hence we felt it would be easier for that audience if the the cross-validation is explained along the concrete example rather than abstract formulae. To better accomodate readers with experience in GAMs we have now added a forward reference in the methods section with only some short comments.

Conclusion
"Page 8, line 30-32: As far as I understand, in section 4.2 the higher spatial variability of the intensity model is explained due to local constructions, that trigger the number of flashes without affecting the occurrence. However, in the conclusion part one get the impression that higher spatial variability of the intensity model is distinct from local maxima in the vicinity of radio towers. Thus, I would suggest a sentence like: 'In particular the spatial effect of the intensity model varies more strongly than the corresponding effect of the occurrence model, because local intensity maxima are triggered in vicinity of radio towers. Moreover other new features were exhibited like...'"

Answer: We adopted this sentence.

Spatio-temporal ~~smoothing~~ modelling of lightning climatologies for complex terrain

Thorsten Simon[1,2], Nikolaus Umlauf[2], Achim Zeileis[2], Georg J. Mayr[1], Wolfgang Schulz[3], and Gerhard Diendorfer[3]

[1]Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Austria
[2]Department of Statistics, University of Innsbruck, Austria
[3]OVE-ALDIS, Vienna, Austria

Correspondence to: T. Simon (thorsten.simon@uibk.ac.at)

Abstract. This study develops methods for estimating lightning climatologies on the $d^{-1}km^{-2}$ scale for regions with complex terrain and applies them to summertime observations (2010 – 2015) of the lightning location system ALDIS in the Austrian state of Carinthia in the Eastern Alps.

Generalized additive models (GAMs) are used to model both the probability of occurrence and the intensity of lightning. Additive effects are set up for altitude, day of the year (season) and geographical location (longitude/latitude). The performance of the models is verified by 6-fold cross-validation.

The altitude effect of the occurrence model suggests higher probabilities of lightning for locations on higher elevations. The seasonal effect peaks in mid July. The spatial effect models several local features, but there is a pronounced minimum in the Northwest and a clear maximum in the Eastern part of Carinthia. The estimated effects of the intensity model reveal similar features, though they are not equal. Main difference is that the spatial effect varies more strongly than the analogous effect of the occurrence model.

A major asset of the introduced method is that the resulting climatological information vary smoothly over space and time. Thus, the climatology is capable to serve as a useful tool in quantitative applications, i.e., risk assessment and weather prediction.

Key words: lightning location data, generalized additive model, hurdle model, zero-truncated possion distribution

1 Introduction

Severe weather, associated with thunderstorms and lightning, causes fatalities, injuries and financial losses (Curran et al., 2000). Thus, the private and the insurance sector have a strong interest in reliable climatologies for such events, i.e., for risk assessment or as a benchmark forecast of a warning system. For these quantitative purposes, it is crucial to separate signal and noise. Especially when the target variable, i.e., lightning, on the one hand varies strongly in space and time and on the other hand might be explained by other covariates, i.e., altitude. This holds in particular for regions with complex terrain. To this end it is desirable to identify smooth and potentially nonlinear functional dependencies between lightning and variables associated

with space and time. This study aims at testing how generalized additive models can be applied in order to smooth lightning climatologies over complex terrain.

The main motivation to process the raw data by a statistical model is to improve the signal-to-noise ratio. This is in particular true when processing short datasets. The raw lightning data are noisy due to the high variability of processes generating lightning. This is not only true for lightning, but also for other atmospheric variables, e.g. precipitation, wind speed and direction. The GAM applied in our study filters effects/signals associated with altitude, day of the year and space and separates these from the noise.

A study investigating ALDIS data for the period 1992 to 2001 (Schulz et al., 2005) found that flash densities over the complex topography of Austria ~~are~~ vary between 0.5 and 4 flashes $km^{-2}yr^{-1}$ depending on the terrain. Data from the same lightning location system (LLS) was also analyzed to obtain thunderstorm tracks (Bertram and Mayr, 2004). The key finding is that thunderstorms are often initialized at mountains of moderate altitude and propagate towards flat areas afterwards.

Other studies focus on lightning detected in the vicinity of the Alps: A 6-year analysis of lightning detection data over Germany reveals highest activity in the northern foothills of the Alps and during the summer months, where the number of thunderstorm days goes up to ~~$7.5\ km^{-2}yr^{-1}$~~ $7.5\ yr^{-1}$ (Wapler, 2013). Lightning activity is ~~even higher~~ also high along the southern rim of the Alps. Feudale et al. (2013) found ground flash densities up to 11 flashes $km^{-2}yr^{-1}$ in northeast Italy, which is south of our region of interest, Carinthia.

These lightning climatologies provide averages of days with lightning activity and averages of counts of flashes, respectively, on the scale $km^{-2}yr^{-1}$. Since lightning varies strongly in space and time and only short time series of the order of 10 years are available, this procedure might yield spatial fields with strong fluctuations between neighboring cells (cf. Fig. 3). This issue is not a big drawback when the purpose of the analysis is to get an overall qualitative picture of the lightning activity, but it can become a problem in applications where a quantitative assessment is required, i.e., risk assessment or when the climatology has to serve as a benchmark weather forecast. A method is therefore needed to obtain climatologies smoothed in space and time.

In this study generalized additive models (GAMs, see, Hastie and Tibshirani, 1990; Wood, 2006) are applied to estimate a climatology of the probability of lightning, which could also be seen as a thunderstorm climatology with lightning as proxy (e.g., Gladich et al., 2011; Poelman, 2014; Mona et al., 2016), and a climatology of the expected numbers of flashes which can be interpreted as the intensity of thunderstorms.

GAMs have been used to compile climatologies of extremes (Chavez-Demoulin and Davison, 2005; Yang et al., 2016) and full precipitation distributions (Rust et al., 2013; Stauffer et al., 2016). An extension of GAMs is that not only expected values, but also other parameters of a distribution can be modeled (Stauffer et al., 2016).

The manuscript is structured as follows: The lightning detection data, the region of interest, Carinthia, and the pre-processing of the data ~~is~~ are described in Sect. 2. The methods to estimate the lightning climatologies from the data ~~is~~ are based on generalized additive models (Sect. 3). The nonlinear effects estimated for occurrence and intensity model and exemplary climatologies are presented afterwards (Sect. 4). Some special aspects of interest for end-users are discussed in Sect. 5. The study is summarized and concluded at the end of the manuscript (Sect. 6).

2 Data

Lightning is defined as a transient, high-current (typically tens of kiloamperes) electric discharge in the air whose length is measured in kilometers. The lightning discharge in its entirety is usually termed a *lightning flash* or just a *flash*. Each flash typically contains several *strokes* which are the basic elements of a lightning discharge (Rakov, 2016).

In this study 6 years of data (2010 – 2015) from the ALDIS detection network (Schulz et al., 2005) are included. The data within this period ~~is~~ are processed in real-time by the same lightning location algorithm ensuring stationarity with respect to data processing. The summer months May to August are selected, as this is the dominant lightning season in the Eastern Alps. The original ALDIS data contains single strokes and information which strokes belong to a flash. In order to analyze flashes, solely the very first stroke of a flash is taken into account. Both cloud-to-ground and intra-cloud lightning strikes are considered, as both typically indicate thunderstorms which are of interest in this study.

ALDIS is part of the European cooperation for lightning detection (EUCLID) (Pohjola and Mäkelä, 2013; Schulz et al., 2016), which bundles European efforts in lightning detection. Schulz et al. (2016) present an evaluation of the performance of lightning detection in Europe and the Alps by comparison against direct tower observations with respect to detection efficiency, peak current estimation and location accurancy. The median location error was found to be in the order of $100\ m$. Furthermore, they show that the flash detection efficiency is greater than 96% (100%) if one of the return strokes in a flash had a peak current greater than 2kA (10kA). However, it is impossible to determine the detection efficiency of intra-cloud flashes without a locally installed VHF network. Thus no attempt made in Schulz et al. (2016) to characterize the detection efficiency of intra-cloud flashes.

The region of interest is the state of Carinthia in the south of Austria at the border to Italy and Slovenia. Carinthia extends $180\ km$ in west-east direction and $80\ km$ in south-north direction. The elevation varies between $339\ m$ to $3798\ m$ above mean sea level (a.m.s.l.). For invoking elevation as a covariate into the statistical model (Sect. 3) digital elevation model (DEM) data (Kärnten, 2015) is averaged over $1\ km \times 1\ km$ cells (Fig. 1), which leads to a maximum elevation of $3419\ m$ a.m.s.l. with respect to this resolution. As the distribution of the altitude is highly skewed, the logarithm of the altitude serves as covariate, which is distributed more uniformly in the range from 6 to 8.

The lightning data ~~is~~ are transferred to the same $1\ km \times 1\ km$ raster by counting the flashes within one spatial cell and per day. This procedure yields 9904 cells and 738 days for a total of $n = 7309152$ data points, from which 157440 (2.15%) show lightning activity. The amount of cells in which a specific number of flashes was detected decreases rapidly for increasing count numbers. The most extreme data point has 37 flashes per cell and day (Fig. 2). The mean number of detected flashes in the cells *given* lightning activity is 1.75.

Figure 3 shows an example for a climatology based on empirical estimates for July. Here the number of days with lightning is divided by the total number of days for every single grid cell. While some patterns emerge, a large amount of noise is visibily superimposed.

3 Methods

This section introduces the statistical models for estimating the climatologies for lightning occurrence and lightning intensity. The aim of the statistical model is to explain the response, i.e., the probability of occurrence or counts of flashes, by appropriate spatio-temporal covariates, i.e., logarithm of the altitude (`logalt`), day of the year (`doy`) and geographical location (`lon`, `lat`). Since the response might nonlinearly depend on the covariates we choose generalized additive models (GAMs) as a statistical framework, for which a brief introduction is presented in Sect. 3.1.

It is assumed that the number of flashes detected within a cell and day are generated by a random process Y. Realizations of the random process are denoted by $y_i \in \{0, 1, 2, \dots\}$, where $i = 1, 2, \dots, n$ indicates the observation. Two distinct models are set up: first, a model for the probability of the occurrence of lightning $\Pr(Y > 0)$ within a cell and a day; second, a truncated count model to assess the expected number of flashes within a cell *given* lightning activity $\mathrm{E}[Y|Y > 0]$. This procedure refers to a hurdle model (Mullahy, 1986; Zeileis et al., 2008) which has the further benefit to be able to handle the large amount of zero valued data points (97.85%, in our case). The occurrence model and the intensity model are specified in Sect. 3.2 and Sect. 3.3, respectively.

In Sect. 3.4 a short overview over the applied verification techniques is given, i.e., cross-validation, scoring rules and bootstrapping.

3.1 Generalized additive model

The main motivation for using a GAM is the possibility to estimate (potentially) nonlinear relationships between the response and the covariates. In the following, the basic concept of GAMs is introduced for an arbitrary parameter θ of some probability density function $d(\cdot; \theta)$ (PDF). A GAM aiming at modeling a spatio-temporal climatology over complex terrain would have the form,

$$g(\theta) = \beta_0 + f_1(\texttt{logalt}) + f_2(\texttt{doy}) + f_3(\texttt{lon}, \texttt{lat}), \tag{1}$$

where $g(\cdot)$ is a link function that maps the scale of the parameter θ to the real line. The right hand side is called the additive predictor, where β_0 is the intercept term and f_j are unspecified (potentially) nonlinear smooth functions that are modeled using regression splines (Wood, 2006; Fahrmeir et al., 2013). For each f_j a design matrix X_j containing spline basis functions is constructed. Thus the GAM can be written as generalized linear model (GLM),

$$g(\theta) = \beta_0 + \sum_{j=1}^{3} X_j \beta_j. \tag{2}$$

The coefficients $\beta = (\beta_0, \beta_1^\top, \beta_2^\top, \beta_3^\top)$ are estimated by maximizing the penalized log-likelihood,

$$l(\beta) = \sum_{i=1}^{n} \log(d(y_i; g^{-1}(\beta_0 + \sum_{j=1}^{3} X_j \beta_j))) - \frac{1}{2} \sum_{j=1}^{3} \lambda_j \beta_j^\top \mathrm{S}_j \beta_j, \tag{3}$$

where the first term on the right-hand side is the unpenalized log-likelihood. The second term is added to prevent overfitting by penalizing too abrupt jumps of the functional forms. λ_j are the smoothing parameters corresponding to the functions

f_j, respectively. For $\lambda_j = 0$ the log-likelihood is unpenalized with respect to f_j. When $\lambda_j \to \infty$ the fitting procedure will select a linear effect for f_j. The selection of the smoothing parameters λ_j is performed by cross-validation (Sect. 3.4). S_j are prespecified penalty matrices, which depend on the choice of spline basis for the single terms. The reader is referred to Wood (2006) for more details.

Estimation of a GAM for such a large dataset, i.e., 7309152 data points, is feasible, e.g., via function `bam()` (for *big additive models*) implemented in the **mgcv** package (Wood et al., 2015; Wood, 2016) of the statistical software R (R Core Team, 2016).

3.2 Occurrence model

The first component models the probability of lightning $\Pr(Y > 0) = \pi$ to occur within a $1\ km \times 1\ km$ cell and a day. The Bernoulli distribution with the parameter π of the PDF,

$$d_{\mathrm{Be}}(y;\pi) = \pi^y (1-\pi)^{1-y}, \tag{4}$$

will be fitted. Since the data is binomial $y \in \{0,1\}$ indicates no lightning and lightning within the cells, respectively. The model for π takes the form of Eq. 1 with π replacing θ. The complementary log-log $g(\pi) = \log(-\log(1-\pi))$ is implemented as link function.

3.3 Intensity model

The second part is the truncated count component for the expected number of flashes *given* lightning activity. We will refer to this component as the intensity model. It is assumed that the positive counts of flashes within a spatial cell and day follow a zero truncated Poisson distribution with the PDF,

$$d_{\mathrm{ZTP}}(y;\mu) = \frac{d_{\mathrm{Pois}}(y,\mu)}{1 - d_{\mathrm{Pois}}(0,\mu)} \tag{5}$$

where $y \in \{1,2,\dots\}$, and $d_{\mathrm{Pois}}(\cdot;\mu)$ is the PDF of the Poisson distribution with expectation μ. The conditional expectation is $\mathrm{E}[Y|Y>0] = \mu/(1-e^{-\mu})$. The GAM for this component has the form of Eq. 1 with μ replacing θ. The logarithm serves as link function $g(\mu) = \log(\mu)$.

The family for modeling the zero truncated Poisson distribution `ztpoisson()` within a GLM or GAM framework is implemented in the R-package **countreg** (Zeileis and Kleiber, 2015). For more information on and a formal definition of hurdle models the reader is referred to Zeileis et al. (2008).

3.4 Verification

In this section the verification procedures are briefly introduced, namely the cross-validation, the applied scores and the block-bootstrapping.

In order to ensure the verification of the model along independent data, we applied a 6-fold cross-validation (Hastie et al., 2009). 6 years of data are available. The parameters of the model are estimated based on 5 years of the data and validated on the remaining year. This is done 6 times with every single year serving as validation period once.

The log-likelihood is applied as scoring function, which is also called logarithmic score in the literature on proper scoring rules (Gneiting and Raftery, 2007).

To assess confidence intervals of the estimated parameters and effects, *day-wise block-bootstrapping* was performed. With *day-wise block-bootstrapping* we mean the following: We resample the 738 dates of all available days with repetition and pick all the data observed on these days spatially. This procedure is executed 1000 times in order to assess confidence intervals.

4 Results

This section presenting the results of the statistical models is structured as follows: first, the nonlinear effects of the occurrence model are described in Sect. 4.1; second, the effects of the intensity model are presented in Sect. 4.2. Finally, exemplary applications illustrate in Sect. 4.3 how climatological information can be drawn from the models.

4.1 Occurrence model

The estimates of the effect of the occurrence model (Sect. 3.2) are depicted in Fig. 4. The values are on the scale of the additive predictor, i.e. the right hand side of Eq. 1. The additive predictor, takes the value of the intercept term β_0 if the sum of all other effects is equal to zero. Its estimate is $\beta_0 = -3.97$ $(-4.15, -3.80)$ on the complementary log-log scale, which is 1.87% $(1.56\%, 2.21\%)$ in terms of probability of lightning. The numbers in parenthesis are 95% confidence intervals computed from 1000 day-wise block-bootstrapping estimates.

The second term $f_1(\texttt{logalt})$ of the additive predictor models the effect of the logarithm of the altitude. $f_1(\texttt{logalt})$ varies from roughly -0.2 for low altitudes to values greater than 0.5 for altitudes above $2800\ m$ (Fig. 4, top-left). This function takes 7.9 degrees of freedom. Its shape is close to exponential, which suggests that a linear term for the altitude would be sufficient. For altitudes above $2000\ m$, however, the nonlinear term leads to larger values than a linear term $\beta_1 alt$ would do.

The temporal or seasonal effect $f_2(\texttt{doy})$, i.e., the dependence of the target on the day of the year (`doy`), shows a steep increase during May, reaches its maximum in mid July and decreases slowly during August (Fig. 4, top-right). This result indicates that the main lightning season in Carinthia lasts from mid June until end of August. The estimated degrees of freedom are 2.5, which leads to the simple shape of the seasonal effect with one clear maximum.

The spatial effect $f_3(\texttt{lon},\texttt{lat})$, which explains the spatial variations of the linear predictor that cannot be explained by the altitude term $f_1(\texttt{logalt})$, requires 138 degrees of freedom. (Fig. 4, bottom). Most prominent features are the minimum near the northwest and the maximum in the mid to eastern part of Carinthia. In the northwest the highest mountains, the High Tauern, of Carinthia are located. The minimum with values less than -0.3 on the complementary log-log scale suggests that lightning activity is less pronounced in this region. This finding is ~~inline~~ in line with former analyses of the lightning activity in Austria (Troger, 1998; Schulz et al., 2005), which stated that the main alpine crest is an area with a minimum in flash density. The maximum zone with values exceeding 0.3 in the mid to eastern part of Carinthia covers the so-called Gurktal Alps. In comparison with the High Tauern, the Gurktal Alps ~~are on~~ have a lower average elevation and the mountains are not as steep. Such maxima at moderate or low altitude are mostly modeled by the spatial effect, not by the altitude effect.

As the altitude is a function of longitude and latitude, one could ask whether it would be sufficient take only a spatial effect into account that implicitly contains the altitude and skip the explicit altitude effect. In general the presented method would be capable to model the influence of the altitude within the spatial effect implicitly. However, the shape of the altitude in the region of interest is very complex. Thus, a spatial effect with a large degree of freedom would be required in order to account for the complex altitude shape. As we know the shape of the altitude we can pass it to the GAM as an isolated effect. The altitude effect contains only information associated with the altitude while the remaining effects are captured by the spatial term.

4.2 Intensity model

The nonlinear effects of the intensity model (Sect. 3.3) are depicted in Fig. 5. The estimate of the intercept term takes the value $\beta_0 = -0.01$ $(-0.19, 0.14)$ which leads to a expected number of flashes *given* lightning activity of 1.57 $(1.47, 1.68)$ when the sum of all other effects is equal to zero.

The altitude effect $f_1(\texttt{logalt})$ (Fig. 5, top-left), with 5.4 degrees of freedom, reveals a similar functional form as the altitude effect of the occurrence model (Fig. 4, top-left). However, it has a flatter shape for the terrain between $600\ m$–$1200\ m$ and a steeper increase for high altitudes above $2000\ m\ a.m.s.l.$.

The seasonal effect $f_2(\texttt{doy})$ is -0.5 in early May, reaches a maximum of 0.3 in early July and decreases to values around -0.3 until the end of August (Fig. 5, top-right). Thus the amplitude of this effect is not as strong as the seasonal effect of the occurrence model (Fig. 4, top-right) and the location of the maximum is earlier. It has 2.1 degrees of freedom.

The spatial effect $f_3(\texttt{lon},\texttt{lat})$ varies strongly and requires 166 degrees of freedom which is more than the corresponding effect of the occurrence model. However, there are some features common for both effects. For instance, the prominent maximum visible in Fig. 5 (bottom) in the Gurktal Alps appears also in the spatial effect of the count model (Fig. 4, bottom). Common is also the strong minimum in the western part of the domain. The most pronounced new feature is the strong local maximum with values exceeding 0.9 in the south of Carinthia. A $165\ m$ radio tower is installed on the peak of the Dobratsch mountain (location C in Fig. 1), which triggers lightning strokes under suitable conditions, i.e., occurrence of a thunderstorm. Other maxima of this effect could also be attributed to sites of radio towers, which suggests that the number of flashes is more sensitive to local constructions than the probability of lightning.

4.3 Applications

In order to illustrate how climatological information can be drawn from the GAMs, two different kinds of applications are presented. First, maps show spatial climatologies (Fig. 6 and Fig. 7). Here, the occurrence model and/or the intensity model are evaluated for one specific day. Second, the seasonal climatology for selected $1\ km \times 1\ km$ grid cells are discussed (Fig. 8). Here, the models are evaluated with respect to the geographical location of the point of interest and its altitude.

The spatial distribution of climatological probabilities of lightning to occur in a cell for July 20 (close to the seasonal peak) varies from 1.8% to 6.5% (Fig. 6). In the western part of the domain, local valleys and mountain ridges become visible through the altitude effect (Fig. 4, top-left). However, the highest probabilities do not occur over the highest terrain in the northwest,

where the spatial effect counteracts the altitude effect leading to moderate probabilities around 2% to 3%. The spatial effect (Fig. 4, bottom) is responsible for the maximum over the moderate altitude region of the Gurktal Alps. Such a map can also serve as thunderstorm climatology when lightning is taken as a proxy for thunderstorms.

For the same day, July 20, the expected number of flashes is depicted in Fig. 7. This is the product of probabilities of lightning π from the occurrence model and the expected number of flashes *given* lightning activity, which is derived from the intensity model. Values are ranging from 0.028 to 0.166. The lowest values can be found in the northwestern part of Carinthia where the spatial effects of both models reveal a minimum. Next to the maximum in the Gurktal Alps, where also maxima in the spatial effects of both models can be found, a second peak appears at the Dobratsch mountain (location C in Fig. 1) which is due to the local maximum in the spatial effect of the intensity model (Fig. 5, bottom).

Next to the spatial information one can extract seasonal climatologies for different locations (Fig. 8). These are computed exemplary for five sites (Table 1). The left panel of Fig. 8 shows the climatologies of lightning probability. Differences between the annual cycles of the probabilities are due to the altitude effect and the spatial effect of the occurrence model (Fig. 4). The highest probabilities between 4% and 5% are modeled in July for location B (dashed line), which is located at the southwestern border of Carinthia in vicinity of a local maximum of the spatial effect (Fig. 4, bottom). This climatology exhibits a strong seasonality, as probabilities fall below the 1% level. Though located at a similar altitude, the climatology of location A (solid line) reveals maximum values less than 2%. This difference is due to the spatial effect, which exhibits a clear minimum in northwestern Carinthia. The climatology of location D (dashed dotted line) in the lower plains in the eastern part of Carinthia show moderate chance of lightning with values around 3% during the peak of the season.

The climatologies of the expected number of flashes are depicted in the right panel of Fig. 8. The order of location has changed. In particular the highest number of flashed are expected for location C which is the Dobratsch mountain. This is caused by the strong local maximum in the spatial effect of the intensity model (Fig. 5, bottom). ~~Accumulating the probabilities over the season~~ The legend shows expected number of flashes accumulated over the lightning season, which leads to values between 2.1 for location A and 7.6 for location C. These values are in good agreement with the analysis by Schulz et al. (2005, Fig 5. therein).

5 Discussion

This section addresses two points helpful for end-users. The first one is on how to choose the cross-validation score in order to avoid overfitting of the seasonal effect (speaking technically the selection of its smoothing parameter λ). The second point is a discussion on how the introduced model (Eq. 1) can be extended towards a weather prediction tool, i.e., for warning purposes.

For illustration of the first point a subset of the large dataset is selected. We pick all data points in a 5×5 neighborhood around the location E. Thus, only $6\ years \times 123\ days \times 25\ cells = 18450\ data\ points$ remain. The probability of lightning π is the target variable. Furthermore, altitude and spatial effects are omitted for simplicity for such a small region (cf. the smooth spatial effect of the occurrence model in Fig. 4, bottom). Thus the GAM has the form,

$$g(\pi) = \beta_0 + f(\texttt{doy}). \tag{6}$$

The model is fitted twice: first, with the selection of the smoothing parameter λ by six-fold cross-validation where the observations made on a single day are kept together in a block; second, λ is determined by six-fold cross-validation without the *day-wise* blocks. In both cases the maximal number of degrees of freedom is set to 30.

Fig. 9 shows the estimates of the two models. The estimate resulting from the cross-validation with *day-wise* blocks is much smoother (1.9 degrees of freedom) than the estimate resulting from the cross-validation without daily blocks (29.9 degrees of freedom). Thus the latter estimate takes roughly the maximal degree of freedom and is obviously overfitted.

The reason for the distinct estimates lies in the dependence structure of the data. For one cell the probability to detect lightning on one day *given* lightning was detected on the previous day is 6.7%. Spatial dependence is much stronger. Provided that lightning occurs in one cell, the probability of lightning to occur in the adjacent cell is 41%. This strong spatial dependence comes with a physical meaning. First, the preconditions for thunderstorms and lightning to take place vary much stronger from day-to-day than in the course of a single day. Second, thunderstorm systems, i.e., multi-cell thunderstorms or super-cell thunderstorms, cover a large area or even travel over a larger area (Markowski and Richardson, 2011).

For this reason we recommend to explore the dependence structure of the data first and to define the cross-validation score according to this dependence structure.

Finally, we discuss how the introduced model (Eq. 1) can be extended in order to serve as a weather prediction tool. It is possible to add further predictors from a numerical weather prediction system to the right hand side of Eq. 1. In the case of lightning and thunderstorms suitable predictors could be convective inhibition energy (CIN), convective available potential energy (CAPE), vertical shear of horizontal winds or large scale circulation patterns (e.g., Bertram and Mayr, 2004; Chaudhuri and Middey, 2012). Within the GAM framework nonlinear effects and interactions of these predictors can be modeled. Another major benefit of this procedure is that the climatology is nested within the additive predictor. Thus the performance of the prediction tool would be at least on the quality level of the climatology, but would not fall below.

6 Conclusions

This study presented how generalized linear models (GAMs) (e.g., Hastie and Tibshirani, 1990; Wood, 2006) provide a useful tool for building a lightning climatology or a climatology for the occurrence of thunderstorms. The main concept is to decompose the signal into different effects: an altitude effect, a seasonal effect and a spatial effect. The most beneficial aspect of this method is that smooth estimates for these effects are obtained, which makes the resulting climatology a valuable tool for quantitative purposes, e.g., risk assessment or benchmarking in weather prediction.

Moreover, a hurdle approach was employed which is also capable to handle the large amount of zeros in the data. Thus two aspects of lightning are captured by the models: the probability of lightning to occur and the number of flashes detected within a grid cell. The effects of the two models are similar though not equal. In particular the spatial effect of the intensity model varies more strongly than the corresponding effect of the occurrence model. ~~Moreover, new features were exhibited like local maxima~~ For instance, local intensity maxima are triggered in vicinity of radio towers.

In sum, the occurrence model and the count model took roughly 150 and 180 degrees of freedom, respectively. This is a relatively small number compared to the degrees of freedom required by other methods~~, i.e. , counting~~ . Counting and averaging flashes with respect to a resolution of $km^{-2}yr^{-1}$ would lead to 9904 degrees of freedom in the introduced case without capturing the seasonal cycle. Thus the GAM approach leads to a smooth, nonlinear and sparse quantification of the climatologies.

Author contributions. Thorsten Simon, Georg J. Mayr and Achim Zeileis defined the scientific scope of this study. Thorsten Simon performed the statistical modeling, evaluation of the results and wrote the paper. Georg J. Mayr supported on the meteorological analysis. Nikolaus Umlauf and Achim Zeileis contributed to the development of the statistical methods. Wolfgang Schulz and Gerhard Diendorfer were in charge of data quality and advised on the lightning related references. All authors discussed the results and commented on the manuscript.

Acknowledgements. We acknowledge the funding of this work by the Austrian Research Promotion Agency (FFG) project *LightningPredict* (Grant No. 846620). ~~Furthermore, we thank G. Diendorfer and W. Schulz from ALDIS for data support, discussion on data characteristics and measurement network. This work was supported by the Austrian Ministry of Science BMWF as part of the UniInfrastrukturprogramm of the Focal Point Scientific Computing at the~~ The computational results presented have been achieved using the HPC infrastructure LEO of the University of Innsbruck.

References

Bertram, I. and Mayr, G. J.: Lightning in the Eastern Alps 1993–1999, part I: Thunderstorm tracks, Natural Hazards and Earth System Science, 4, 501–511, 2004.

Chaudhuri, S. and Middey, A.: A composite stability index for dichotomous forecast of thunderstorms, Theoretical and Applied Climatology, 110, 457–469, 2012.

Chavez-Demoulin, V. and Davison, A. C.: Generalized additive modelling of sample extremes, Journal of the Royal Statistical Society: Series C (Applied Statistics), 54, 207–222, 2005.

Curran, E. B., Holle, R. L., and López, R. E.: Lightning casualties and damages in the United States from 1959 to 1994, Journal of Climate, 13, 3448–3464, 2000.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B.: Regression: Models, methods and applications, Springer Science & Business Media, 2013.

Feudale, L., Manzato, A., and Micheletti, S.: A cloud-to-ground lightning climatology for north-eastern Italy, Advances in Science and Research, 10, 77–84, 2013.

Gladich, I., Gallai, I., Giaiotti, D., and Stel, F.: On the diurnal cycle of deep moist convection in the southern side of the Alps analysed through cloud-to-ground lightning activity, Atmospheric Research, 100, 371–376, 2011.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 102, 359–378, 2007.

Hastie, T. and Tibshirani, R.: Generalized additive models, vol. 43, CRC Press, 1990.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference and prediction, Springer Series in Statistics Springer, Berlin, 2nd edn., 2009.

Kärnten, L.: Digital terrain model (DTM) Carinthia, CC-BY-3.0: Land Kärnten - data.ktn.gv.at, ALS airborne measurements, 2015.

Markowski, P. and Richardson, Y.: Mesoscale meteorology in midlatitudes, vol. 2, John Wiley & Sons, 2011.

Mona, T., Horváth, Á., and Ács, F.: A thunderstorm cell-lightning activity analysis: The new concept of air mass catchment, Atmospheric Research, 169, 340–344, 2016.

Mullahy, J.: Specification and testing of some modified count data models, Journal of econometrics, 33, 341–365, 1986.

Poelman, D. R.: A 10-year study on the characteristics of thunderstorms in Belgium based on cloud-to-ground lightning data, Monthly Weather Review, 142, 4839–4849, 2014.

Pohjola, H. and Mäkelä, A.: The comparison of GLD360 and EUCLID lightning location systems in Europe, Atmospheric Research, 123, 117–128, 2013.

Rakov, V. A.: Fundamentals of Lightning, Cambridge University Press, doi:10.1017/CBO9781139680370, 2016.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2016.

Rust, H. W., Vrac, M., Sultan, B., and Lengaigne, M.: Mapping weather-type influence on senegal precipitation based on a spatial-temporal statistical model, Journal of Climate, 26, 8189–8209, 2013.

Schulz, W., Cummins, K., Diendorfer, G., and Dorninger, M.: Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system, Journal of Geophysical Research: Atmospheres, 110, 2005.

Schulz, W., Diendorfer, G., Pedeboy, S., and Poelman, D.: The European lightning location system EUCLID – part 1: Performance validation, Natural Hazards and Earth System Science, 3, 5325–5355, 2016.

Stauffer, R., Messner, J. W., Mayr, G. J., Umlauf, N., and Zeileis, A.: Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model, Working papers, Faculty of Economics and Statistics, University of Innsbruck, http://EconPapers.repec.org/RePEc:inn:wpaper:2016-07, 2016.

Troger, W.: Die Einbeziehung des Österreichischen Blitzortungssystems ALDIS in die Meteorologische Analyse von Gewittern, Master's thesis, Institut für Meteorologie und Geophysik der Universität Innsbruck, 1998.

Wapler, K.: High-resolution climatology of lightning characteristics within central Europe, Meteorology and Atmospheric Physics, 122, 175–184, 2013.

Wood, S. N.: Generalized additive models: An introduction with R, CRC Press, 2006.

Wood, S. N.: **mgcv**: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL, http://CRAN.R-project.org/package=mgcv, R package version 1.8-12, 2016.

Wood, S. N., Goude, Y., and Shaw, S.: Generalized additive models for large data sets, Journal of the Royal Statistical Society: Series C (Applied Statistics), 64, 139–155, 2015.

Yang, C., Xu, J., and Li, Y.: Bayesian geoadditive modelling of climate extremes with nonparametric spatially varying temporal effects, International Journal of Climatology, 2016.

Zeileis, A. and Kleiber, C.: **countreg**: Count data regression, http://R-Forge.R-project.org/projects/countreg, R package version 0.1-5, 2015.

Zeileis, A., Kleiber, C., and Jackman, S.: Regression models for count data in R, Journal of Statistical Software, 27, 1–25, 2008.

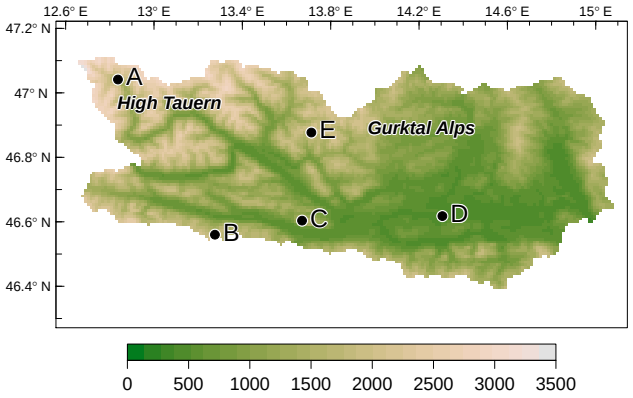

Figure 1. Altitude of Carinthia ($m\ a.m.s.l.$) averaged over $1\ km \times 1\ km$ cells. Attributes of sample locations are listed in Table 1.

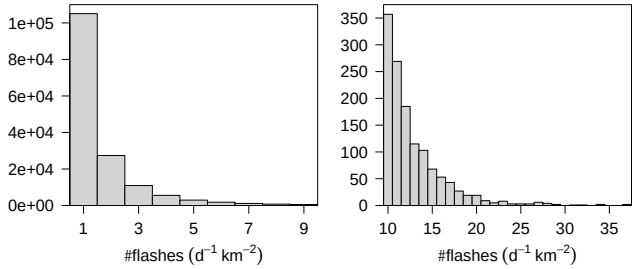

Figure 2. Daily frequency of $1\ km \times 1\ km$ grid cells ~~in~~ with counts of flashes (excluding zeros). The right panel shows a zoom into the tail of the distribution. The percentage of boxes with no flashes detected is 97.85%.

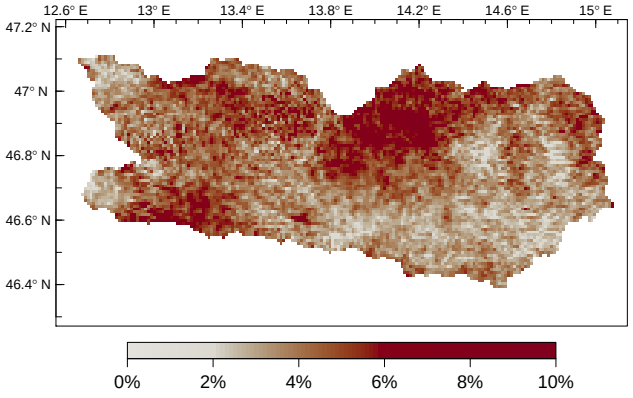

Figure 3. Empirical climatological probability of lightning for a day in July in Carinthia on the $1\ km \times 1\ km$ scale.

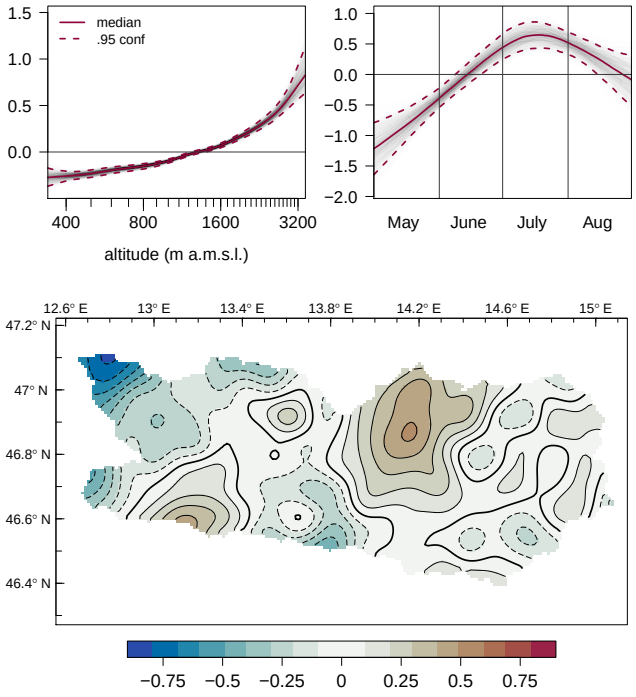

Figure 4. The effects of the occurrence model on the scale of the additive predictor. **Top-Left:** The altitude (`logalt`) effect. Ticks on the x-axis are set in 100 m intervals. The gray lines show 1000 estimates from day-wise block-bootstrapping. The solid red line is the median of the 1000 estimates, the dashed red lines are the 95% confidence intervals. **Top-Right:** The seasonal (`doy`) effect. **Bottom:** The spatial (`lon`, `lat`) effect. The plot shows the median of 1000 estimates from day-wise block-bootstrapping. The difference between two contour lines is 0.1. Dashed contour lines indicate negative values.

id	name	lon (° E)	lat (° N)	alt ($m\ a.m.s.l.$)
A	Heiligenblut	12.84	47.04	1315
B	Nassfeld	13.28	46.56	1525
C	Dobratsch	13.67	46.60	2166
D	Klagenfurt	14.31	46.62	447
E	Rosennock	13.71	46.88	2440

Table 1. Coordinates of the sample locations in Figure 1.

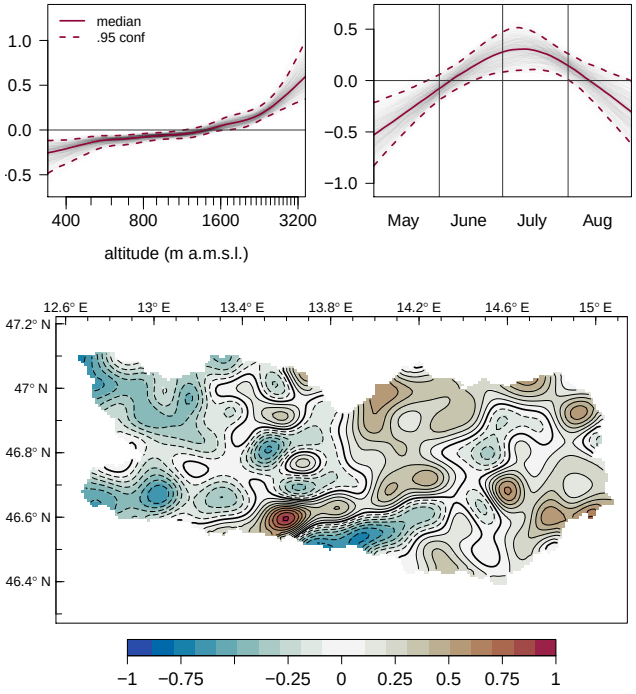

Figure 5. The effects of the intensity model on the scale of the additive predictor. Labeling is analog to Fig. 4. **Top-Left:** The altitude (`logalt`) effect. **Top-Right:** The seasonal (`doy`) effect. **Bottom:** The spatial (`lon`, `lat`) effect.

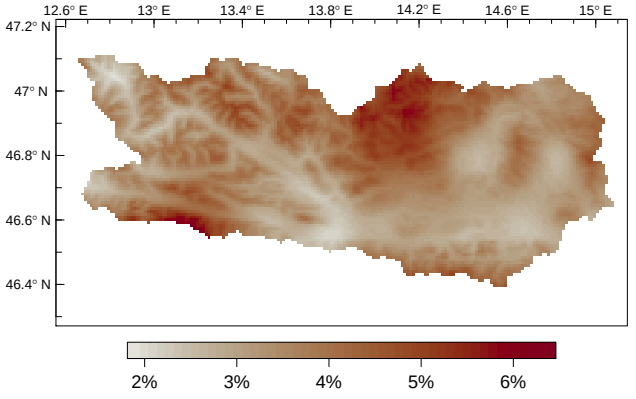

Figure 6. Climatological probability of lightning for July 20 in Carinthia on the $1\ km \times 1\ km$ scale.

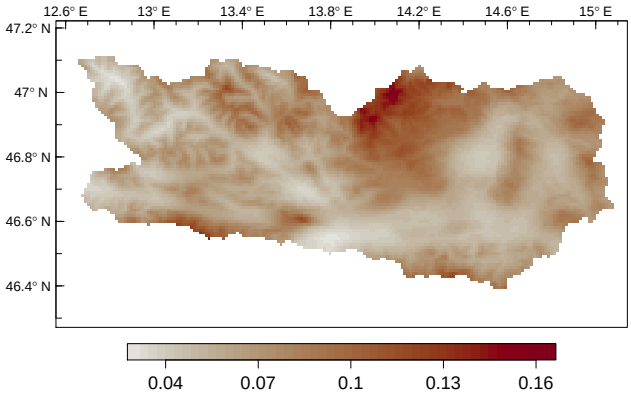

Figure 7. Expected (climatological) number of flashes for July 20 in Carinthia on the $1\ km \times 1\ km$ scale.

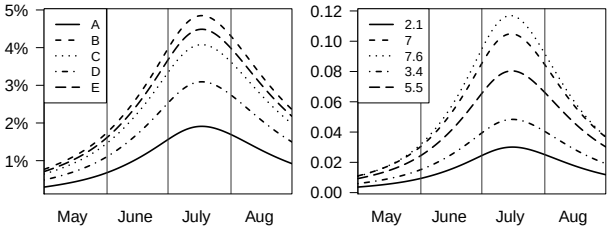

Figure 8. Seasonal climatologies for sample locations, which are highlighted in Figure 1. Left: Occurrence model. Right: Expected number of flashes. The legend shows expected number of flashes accumulated over the lightning season.

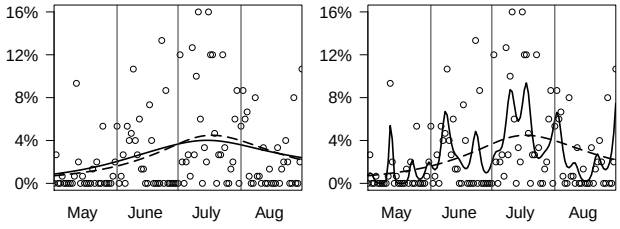

Figure 9. Local fits for the location E. Circles show *empirical* estimates. For comparison the estimate of the full occurrence model is added (dashed line). **Left:** Solid line is the GAM evaluated by cross-validation with *day-wise* blocks. **Right:** Solid line is the GAM evaluated by cross-validation without *day-wise* blocks.