©eeecon

Lightning prediction using model output statistics

**Thorsten Simon, Georg J. Mayr, Nikolaus Umlauf,
Achim Zeileis**

Working Papers in Economics and Statistics

2018-14

University of Innsbruck
https://www.uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform "Empirical and Experimental Economics"
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

Lightning Prediction Using Model Output Statistics

Thorsten Simon
University of Innsbruck

Georg J. Mayr
University of Innsbruck

Nikolaus Umlauf
University of Innsbruck

Achim Zeileis
University of Innsbruck

Abstract

A method to predict lightning by postprocessing numerical weather prediction (NWP) output is developed for the region of the European Eastern Alps. Cloud-to-ground flashes—detected by the ground-based ALDIS network—are counted on the 18×18 km^2 grid of the 51-member NWP ensemble of the European Centre of Medium-Range Weather Forecasts (ECMWF). These counts serve as target quantity in count data regression models for the occurrence and the intensity of lightning events. The probability whether lightning occurs or not is modelled by a binomial distribution. For the intensity a hurdle approach is employed, for which the binomial distribution is combined with a zero-truncated negative binomial to model the counts within a grid cell. In both statistical models the parameters of the distributions are described by additive predictors, which are assembled by potentially nonlinear terms of NWP covariates. Measures of location and spread of approx. 100 direct and derived NWP covariates provide a pool of candidates for the nonlinear terms. A combination of stability selection and gradient boosting selects influential terms. Markov chain Monte Carlo (MCMC) simulation estimates the final model to provide credible inference of effects, scores and predictions. The selection of terms and MCMC simulation are applied for data of the year 2016, and out-of-sample performance is evaluated for 2017. The occurrence model outperforms a reference climatology—based on seven years of data—up to a forecast horizon of 5 days. The intensity model is calibrated and also outperforms climatology for exceedance probabilities, quantiles, and full predictive distributions.

Keywords: lightning detection data, distributional regression, count data model, gradient boosting, MCMC.

1. Introduction

Lightning in Alpine regions is associated with severe events such as convection, thunderstorms, extreme precipitation, high wind gusts, flash floods and debris flows. In order to predict the probability of lightning events (i.e., thunderstorms) numerical weather prediction (NWP) output is often postprocessed by logistic regression (Schmeits *et al.* 2008; Gijben *et al.* 2017; Bates *et al.* 2018) in which lightning detection data serves as proxy for the occurrence of thunderstorms. However, these studies present methods to predict only whether a thunderstorm might take place or not.

The objective of the present work is to extend this approach by modelling the intensity of

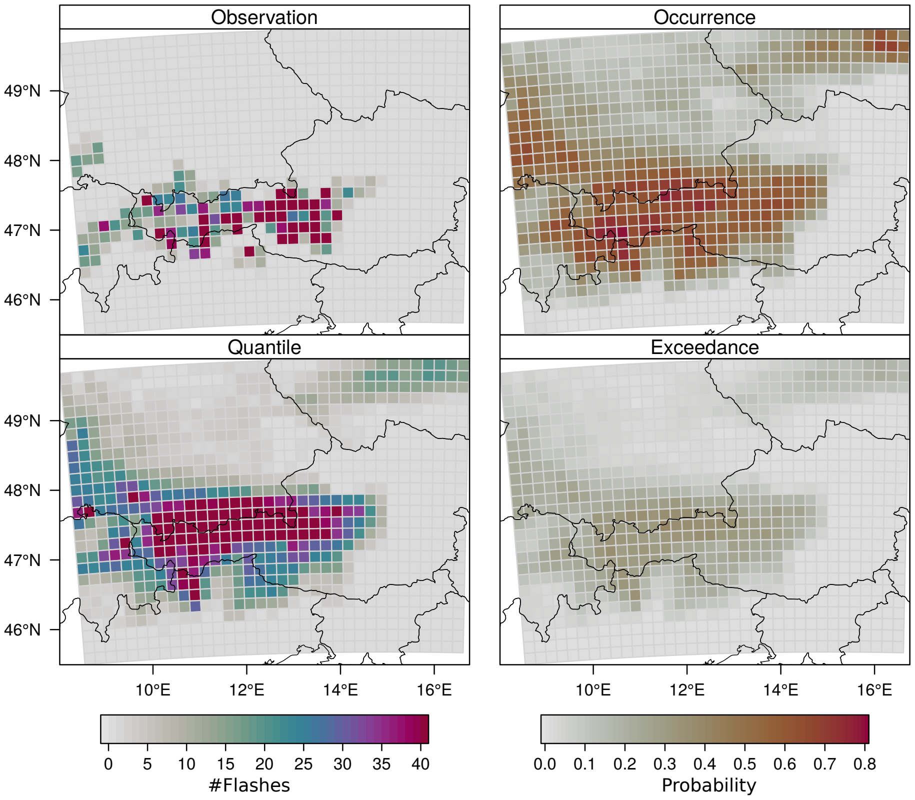

Figure 1: A sample prediction case (2017-07-18) for the lightning count model with a lead time of one day. **Topleft:** Number of observed flashes from 12 to 18 UTC in a $18 \times 18 km^2$ grid cell. **Topright:** Predicted probabilities for the occurrence of lightning events (#flashes > 0). **Bottomleft:** Predicted 90% quantiles. **Bottomright:** Predicted probabilities for exceeding a threshold of 10 flashes in a grid cell.

thunderstorms with a model for lightning counts. Thus a parametric count data model builds the base of this approach. Classically, count data are modelled by a Poisson distribution (Cameron and Trivedi 2013). However, in practical work data are often overdispersed and/or have excess zeros. The issue of overdispersed data can be addressed by applying a negative binomial distribution (Cameron and Trivedi 2013). Excess zeros can be accounted for by splitting the distribution into a binary hurdle and a part for positive counts (Mullahy 1986). The hurdle can be modelled, e.g., by logistic regression and the positive counts by a zero-truncated version of the Poisson or negative binomial distribution.

The combined model predicts a full probability distribution, which allows to derive various quantities such as probabilities for the occurrence of thunderstorms, quantiles, and the exceedance of predefined thresholds. A case study of 18 July 2017 demonstrates the postprocessing model (Fig. 1). The synoptic weak pressure gradient situation allowed local heating to trigger single cell storms. Very high intensities with a reasonable amount of cell exceeding 40 counts were observed along the main Alpine ridge. Yet, a large number of cells remained without a flash.

NWP systems are the most important tool for predicting convection and thunderstorms, although it is challenging to resolve convection directly. For instance, single cell storms have a spatial extent of 2–10 km and a temporal extent of approx. 30 min. Regional NWP systems are run partly with a scale up to 1 km, which is referred to as convection *permitting* or *resolving*. Such models are capable to reproduce bulk heat and water vapour properties (Langhans *et al.* 2012).

In global NWP systems with a coarser resolution convection is simulated by parametric submodels. In modern NWP systems the parameters of such submodels are perturbed stochastically (Buizza *et al.* 1999). By generating ensembles of a NWP one aims at accounting for uncertainties of small scale events such as convection. In this study a set of direct and derived variables from the (global) ECMWF ensemble, are employed as covariates for the statistical model.

Many different output variables of a NWP ensemble system are potential *good* candidates for a lightning prediction, e.g., convective available potential energy (cape) or convective precipitation. However, next to these potential *good* candidates there are variables that could help to improve the prediction even by a small contribution. Moreover, the effect of individual variables might act nonlinearly on the target quantity (lightning counts).

In order to account for nonlinear dependencies we employ additive predictors linked to the parameters of the hurdle model. This statistical framework is often referred to as distributional regression *or* generalized additive models for location, scale and shape (Rigby and Stasinopoulos 2005). The selection of a sparse sufficient set of nonlinear terms from the numerous covariates provided by the NWP ensemble is performed using gradient boosting with stability selection. This concept has been successfully used in several studies (e.g., Simon *et al.* 2018; Thomas *et al.* 2018).

The final model resulting from the selection procedure is still of complex form. Different approaches for estimating the model terms are proposed, i.e., penalized maximum likelihood (Wood 2017), gradient boosting (Mayr *et al.* 2012) or Markov chain Monte Carlo (MCMC) simulations based on a Bayesian formulation of the problem (Brezger and Lang 2006). In this study we follow the Bayesian approach which ensures stable estimation and valid credible intervals for the regression coefficients of such a complex model as the present count data distribution (Klein *et al.* 2015). The MCMC samples allow drawing inferential conclusions about the effects and the predictive performance.

The manuscript is structured as follows. The lightning detection data and the NWP covariates are described first (Sect. 2). Afterwards the statistical method—count data model, the selection procedure and MCMC simulations—are introduced (Sect. 3). The selected terms and the out-of-sample performance is presented in the Section 4. The Section 5 discusses the relation of this study to previous studies and concludes the manuscript.

2. Data

This section describes the lightning detection data (Sect. 2.1) and the numerical weather prediction ensemble data (Sect. 2.2). The data is collected for the region of the European Eastern Alps (Fig. 2) which is exposed to thunderstorms and severe lightning events during summer (Schulz *et al.* 2005; Simon *et al.* 2017).

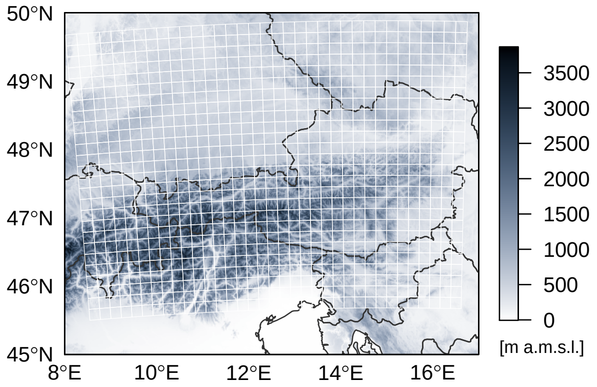

Figure 2: Topography of the European Eastern Alps (from SRTM, Farr *et al.* 2007). Lightning flashes are counted in white grid cells of 18×18 km^2.

2.1. Lightning Detection Data

The proxy for thunderstorm occurrence and intensity is derived for lighting data detected by the ALDIS network (Schulz *et al.* 2005). The lighting data is available for the period 2010–2017, from which the summer month, May–August, are selected. The raw data is aggregated on the 18×18 km^2 grid. One count refers to one cloud-to-ground flash which might contain several strokes.

Table 1: Unconditional and conditional (given positive counts) probabilities [%] of lightning counts.

	0	1	2	3	4	5	6	7	8	9	>9
P	88.05	2.90	1.36	0.92	0.68	0.54	0.44	0.38	0.33	0.29	4.11
cond. P		24.27	11.37	7.68	5.65	4.48	3.71	3.21	2.78	2.44	34.41

The lightning counts aggregated on this scale reveals a large amount of zeros (88.05%). Roughly a quarter (24.27%) of the cells with positive counts contain only a single flash, while approximately a third (34.41%) of the cells contain 10 or more flashes (Tab. 1). The sample mean and the sample variance of the data is 1.8 and 136.3, respectively—and 15.05 and 941.06 only for positive counts, i.e., cells in which lightning occurred. Thus, the data are heavily skewed with the variance much larger than the mean, which is called *overdispersion* in count data literature (Cameron and Trivedi 2013).

For the given aggregation scale the region is described by 910 grid cells. The season from May–August consists of 123 day, which leads to a sample size of 910×123=111930 for each year.

2.2. Numerical Weather Predictions

Covariates are derived from the ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF ENS). Since March 2016 the ECMWF ENS comes with a native resolution of approximately 18 km. The summer of 2016 and 2017 serve as training and evaluation period, respectively. Moreover, 5 forecast horizons are considered, where day 1 refers to lead times of 12–18 h of the ensemble initialized at 00 UTC. Analogously, day 2, day 3, day 4 and day 5 refer to lead times of 36–42 h, 60–66 h, 84–90 h and 108–114 h, respectively. The variables are interpolated bilinearly to the same grid as the lightning data (Fig. 2).

Additional variables are derived by computing vertical differences—i.e., a proxy for mid layer stability, the layer thickness between 700 hPa and 500 hPa and the difference of vertical wind for the same two pressure levels—and by taking the square root of convective precipitation and convective available potential energy (cape). A full list of direct and derived variables is given in Table 2.

Table 2: An overview of the base covariates from the ECMWF-EPS forecast. The asterisk ($\star$) indicates accumulated variables. Covariates derived from this base set are discussed in the data section.

Abbreviation	Description
d2m	Dew point temperature at 2 meters.
e*	Evaporation.
layth	Layer thickness: $(\texttt{z500} - \texttt{z700})/9.81 m/s^2$.
mls	Proxy for mid-layer stability: $\texttt{t500} - \texttt{t700} + 13K$, where $13K$ mimics a humid adiabatic profile between 700 hPa and 500 hPa.
r	Relative humidity at 700 hPa and 500 hPa.
slhf*	Surface latent heat flux.
sqrt_cape	Square root of convective available potential energy.
sqrt_cp*	Square root of convective precipitation.
ssr*	Surface net solar radiation.
str*	Surface net thermal radiation.
t700, t500	Temperature at 700 hPa and 500 hPa.
t2m	Temperature at 2 meters.
tcc	Total cloud cover.
u700, u500 v700, v500	Components of horizontal wind at 700 hPa and 500 hPa.
vgw	Vertical gradient of vertical wind: $\texttt{w500} - \texttt{w700}$.
w700, w500	Pressure vertical velocity at 500 hPa and 700 hPa.
z700, z500	Geopotential at 500 hPa and 700 hPa.

For all variables, except for the accumulated fields, the mean over the afternoon, the difference between the values for 18 UTC and 12 UTC and anomalies of the three afternoon values from the mean are computed.

Finally, two statistics are computed over the ensemble space, namely the median and the interquartile range (igr) as measures for location and spread, respectively, of the covariates over the ensemble.

Hereafter the notation of the covariates is as follows. For accumulated fields the name of the variable as listed in Table 2 and the applied statistic over the ensemble is separated by a dot. For all other variables the computation applied over the time dimension (mean, difference or anomaly) is placed in the middle separated by dots.

3. Methods

This section introduces the statistical framework of a count data model with additive predictors (Sect. 3.1), the selection of nonlinear terms using gradient boosting and stability selection (Sect. 3.2), and Markov chain Monte Carlo simulation used for inference (Sect. 3.3).

3.1. Count Data Regression

To account for the large amount of excess zeros and the strong overdispersion present in the lightning counts $y \in \{0, 1, 2, \dots\}$ a hurdle model (Mullahy 1986) is employed. The hurdle model consists of two parts: One part explicitly models the probability of the occurrence of lightning events, i.e., at least one lightning flash is observed with a grid cell. The second part models the number of flashes given a lightning event takes place.

Hereafter, the two parts of the hurdle model are denoted as *binary hurdle part* and *truncated count part*. A logit binomial model for the probability π of lightning (non-zero) events constitutes the binary hurdle part. The positive counts are modelled using a zero-truncated negative binomial distribution, which handles overdispersion and is determined by two parameters for location $\mu > 0$ and dispersion $\theta > 0$. The zero-truncated negative binomial builds on the negative binomial with the probability mass at zero redistributed towards positive counts (cf. Appendix A).

The hurdle model has the density,

$$f(y \mid \pi, \mu, \theta) = \begin{cases} 1 - \pi & y = 0 \\ \pi \cdot f_{\text{ZTNB}}(y \mid \mu, \theta) & y \in \{1, 2, \dots\}, \end{cases} \tag{1}$$

where f_{ZTNB} is the density of the zero-truncated negative binomial.

Within the log-likelihood derived from the density (Eq. 1) one term solely depending on π, i.e., the *binary hurdle part*, and one term depending on μ and θ, i.e., the *truncated count part*, can be identified (Appendix A). As a consequence the two parts of the hurdle model can be handled independently for estimation, term selection, and prediction.

For the *binary hurdle part* the probability π for non-zero events is conditioned on (NWP) covariates by an additive predictor,

$$\text{logit}(\pi) = \underbrace{\beta_0 + f_1(\texttt{doy}) + f_2(\texttt{lon}, \texttt{lat})}_{\text{baseline climatology}} + f_3(\texttt{x}_3) + \cdots + f_p(\texttt{x}_p). \tag{2}$$

where the logit function maps the probability π to the real line. Within the right hand side of Eq. 2 $f_\star$ are potentially nonlinear functions modelled by P-splines (Wood 2017). $f_1(\texttt{doy})$ accounts for an annual cycle, where the day of the year doy serves as covariate. $f_2(\texttt{lon}, \texttt{lat})$ is a spatial effect depending on geographical location, i.e., longitude lon and latitude lat. The covariates $\texttt{x}_3$, …, $\texttt{x}_p$ are the direct and derived parameters from the ECMWF ensemble (Sect. 2.2).

Not all functions f_1, …, f_p are included in the final model, but the relevant terms are selected using gradient boosting combined with stability selection (Sect. 3.2). The resulting final model is estimated using Markov chain Monte Carlo simulation (Sect. 3.3).

For the *truncated count part* the parameters μ and θ are conditioned on covariates by additive predictors analogously to the right hand side of Eq. 2. To ensure positive values for μ and θ, the logarithm serves as link function. The two additive predictors for $\log(\mu)$ and $\log(\theta)$ can encompass different nonlinear terms, which will be selected using gradient boosting combined with stability selection (Sect. 3.2).

3.2. Stability Selection with Gradient Boosting

The selection of the most important nonlinear terms within the predictors associated with the parameters π, μ and θ is performed using gradient boosting combined with stability selection. Gradient boosting is an iterative gradient descent algorithm, where the term which fits best to the gradient of the log-likelihood is slightly updated in each iteration. The estimates converge to the maximum likelihood estimates, when the number of iterations approaches infinity.

The selection of terms for $\text{logit}(\pi)$ (binary hurdle part), and for $\log(\mu)$ and $\log(\theta)$ (truncated count part) is performed separately. Hence the binary hurdle part is determined by exactly one parameter (π), the additive predictor for $\text{logit}(\pi)$ is updated in each iteration. Within the truncated count part, which is determined by two parameters (μ and θ), either the additive predictor of $\log(\mu)$ *or* $\log(\theta)$ is updated in each iteration, depending on which update contributes larger to the log-likelihood. This updating scheme, called *noncyclic* in the boosting literature (Mayr *et al.* 2012), is presented in Appendix B.

If gradient boosting is applied as stand-alone method the number of iterations—and thus the degree of regularization—can be determined by means of information criteria or cross-validation. Here the main purpose of gradient boosting is to select important terms f_j. It is desirable to avoid the selection of numerous non-informative terms. Stability selection is a convenient resampling method for controlling the number of selected non-informative terms by gradient boosting (Meinshausen and Bühlmann 2010; Hofner *et al.* 2015).

Rather than applying the boosting algorithm to all observations, stability selection is based on drawing a subsample half the size of the training data, running the boosting algorithm until a predefined number of terms is selected. This procedure is repeated many times. Afterwards the relative selection frequencies per nonlinear term are computed. Finally the terms for which the relative selection frequency exceeds a certain threshold are included in the final model (cf. algorithm in Hofner *et al.* 2015).

3.3. Markov Chain Monte Carlo Simulation

The final model is of a complex form as it contains several nonlinear terms. For such a complex model determining confidence intervals based on asymptotic assumptions might fail. Markov chain Monte Carlo (MCMC) simulations offer an attractive toolbox to provide valid credible intervals.

To be able to apply this technique to models with additive predictors, the posterior distribution has to be formulated (Brezger and Lang 2006). MCMC samples of the posterior distribution can be efficiently generated by approximating a full-conditional distribution using a second order Taylor series expansion of the log-posterior centred at the last state (Gamerman 1997; Fahrmeir *et al.* 2013; Umlauf *et al.* 2017). Moreover, in most situations the structure of the sampling scheme reduces to an iteratively weighted least squares (IWLS) updating step for which highly efficient algorithms are available (Lang *et al.* 2014).

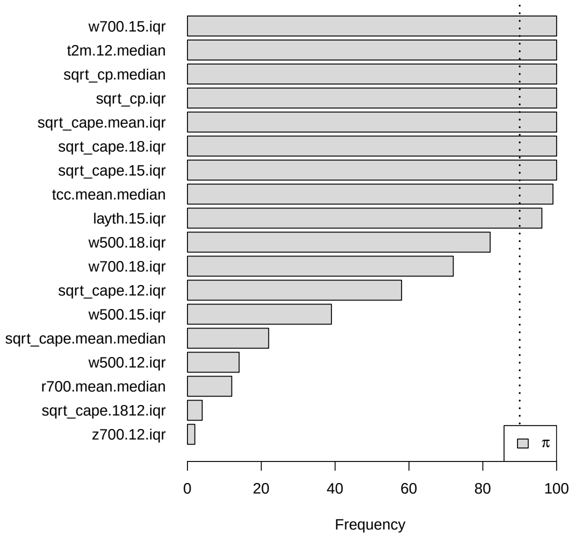

Figure 3: Results of the stability selection procedure for the binary hurdle part of the hurdle model for day 1. The variable names on the y axis serve as placeholder for the associated nonlinear effect. The vertical dotted line marks the threshold of 90% above which terms are added to the final model.

The ECMWF based models, selected by gradient boosting with stability selection, and the climatological baseline models are estimated by MCMC sampling. 1000 independent realizations of the regression coefficients are drawn from the Markov chains, which enables inference of the effects, predictions, and out-of-sample scores.

4. Results

This section is structured as follows. Firstly, we present the results of the selection procedure of nonlinear terms for the binary hurdle part, and the truncated count part. Secondly, we evaluate the performance of the binary hurdle part as an isolated model for the occurrence of lightning events, and the hurdle model (Eq. 1) as a model for the intensity of lightning events.

4.1. Model Selection

Binary Hurdle Part

The selection of nonlinear terms for the binary hurdle part, i.e., the additive predictor for π, for a lead time of one day is visualized in Fig. 3. The gradient boosting algorithm is applied on 100 distinct random subsamples each half the size of the whole training data until 12 terms are selected. The bars in Fig. 3 indicate the relative frequencies for the terms being selected in the 100 boosting runs.

Nine terms are selected in this case, of which five can be associated either with convective precipitation (cp) or convective available potential energy (cape). Neither the seasonal term f_1 nor the spatial term f_2 are selected, which indicates that temporal and spatial variability is well explained by effects depending on the ECMWF ensemble covariates.

The selected effects build the (reduced) additive predictor in Eq. 2. 1000 samples of the coefficients for this final model are drawn using MCMC simulation (Sect. 3.3). The mean effects and associated credible intervals (Fig. 4), are computed from these samples. All effects show a smooth and most a monotonic behaviour. The effect of the median of the square root of convective precipitation (sqrt_cp.median) is close to linearity (Fig. 4d). The effect of variables measuring the spread of the ensemble, i.e., the interquartile range (iqr), have in common that they first increase steeply and flatten after some point (Fig. 4b, c, e, g, h, i).

Truncated Count Part

The count data part of the hurdle model takes only grid cells with values greater than zero. Thus the sample size of the training data decreases from 111930 to 14099. On this subset of the data the stability selection with gradient boosting is applied in order to find the most relevant effects for the parameters μ and θ of the zero-truncated negative binomial. The gradient boosting was run 100 times, each time until 8 terms were selected. The result of this procedure is shown in Fig. 5 for the truncated count part with a forecast horizon of one day. Three terms are selected for the parameter μ, which is the expectation of the underlying negative binomial distribution, and none for the dispersion parameter θ. Thus, only a intercept β_0 is estimated within the final model of $\log(\theta)$.

The estimated effects from the MCMC simulation are presented in Fig. 6 on the log scale. The effect with the largest range is the median (over the ensemble) of the mean (over the afternoon) of the square root of cape, which increases monotonically but nonlinearly. The spread (iqr over the ensemble) of the 18 UTC anomaly of the vertical velocity at 500 hPa (w500) is associated with a nearly linear effect, higher spread leads to a larger μ. The median of the 12 UTC anomaly of total cloud cover (tcc) reveal a nearly linear effect with a negative slope. The estimated value for θ is 0.199 (0.179, 0.220) which reflects the strong overdispersion of the data.

4.2. Performance

Occurrence of Lightning Events

For the evaluation of the predictive performance for the occurrence of lightning events only the probability π is considered. The models with ECMWF ensemble covariates have been estimated on data from 2016 and the data from 2017 is used for an out-of-sample assessment of the performance of the models. The predictions are compared against a climatology, which accounts for seasonal and spatial variations by nonlinear terms (Eq. 2) and is estimated with data from 2010–2016. First we present the global scores—averaged over all grid cells—and afterwards the spatial distribution of skill is analysed.

The Brier score (BS) and *area under curve* (AUC) derived from the receiver operating characteristics (ROC) are applied as verification measures. Both scores and their associated skill scores reveal that the postprocessed ECMWF predictions outperform the climatologies up to

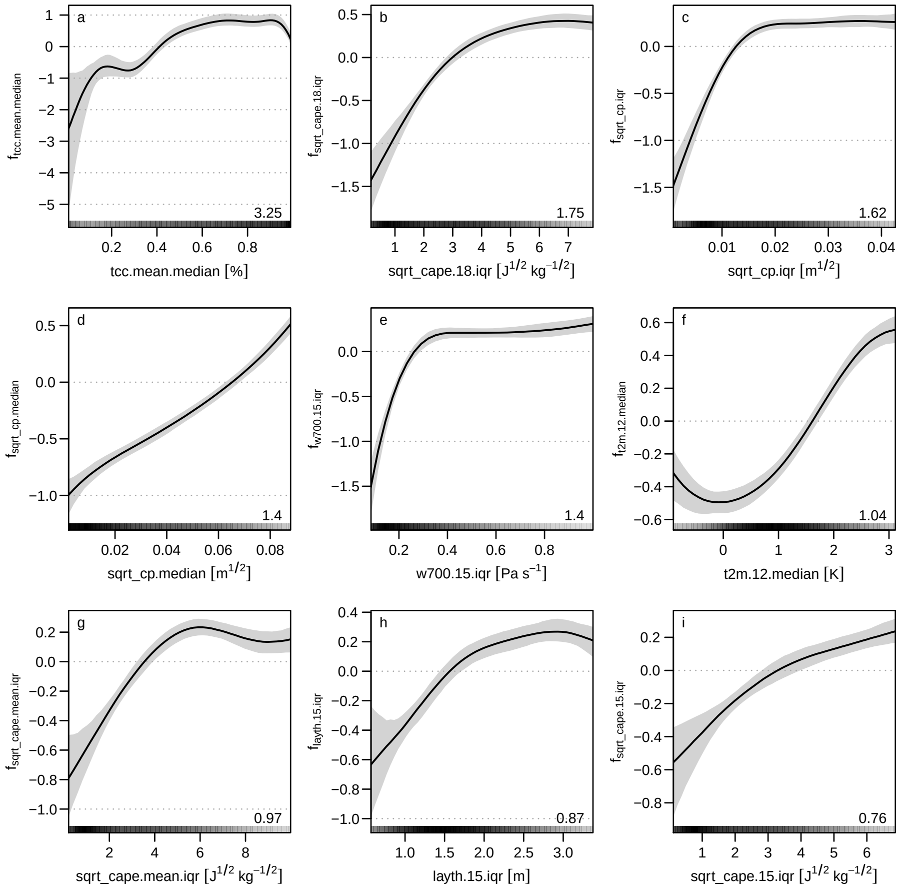

Figure 4: Effects and 95% credible intervals of the occurrence model for day 1 fitted using MCMC simulation. The effects are displayed on the logit scale. The number in the bottom right corner of each panel gives the absolute range of the effect. The shading at the bottom of each panel indicates the density distribution of the corresponding covariate. The x axes are cropped at the 1% and 99% percentile of the respective covariate to enhance graphical representation.

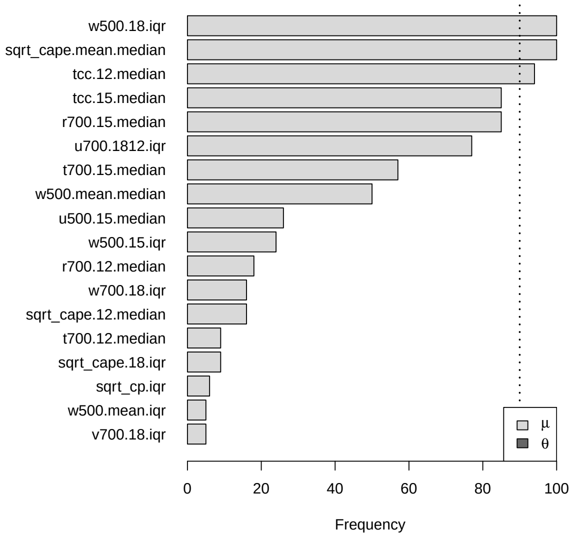

Figure 5: As Fig. 3 but for the truncated count part of the hurdle model for day 1. The grey value indicates whether the term is assigned to the predictor of μ or θ (Note: In this case no terms are selected for the predictor of θ).

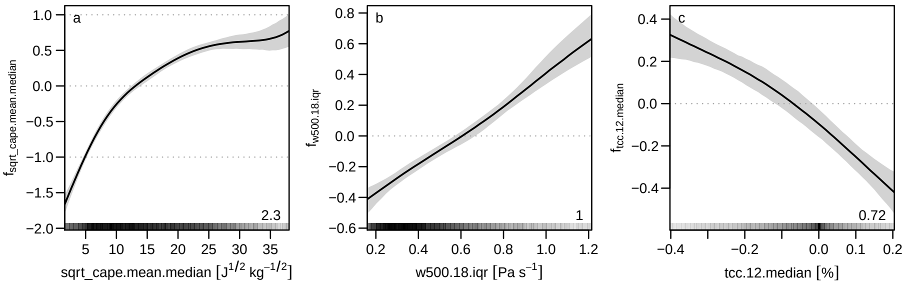

Figure 6: As Fig. 4 but for the intensity model for day 1. All effects are assigned to the predictor of μ and are displayed on the log scale.

Table 3: Out-of-sample performance of the occurrence model. 95% credible intervals based on MCMC samples are given in parentheses.

	Brier score	Brier skill score
Clim.	0.106 (0.106, 0.106)	
Day 1	0.079 (0.079, 0.080)	0.26 (0.25, 0.26)
Day 2	0.084 (0.083, 0.085)	0.21 (0.20, 0.22)
Day 3	0.089 (0.089, 0.089)	0.16 (0.16, 0.17)
Day 4	0.089 (0.089, 0.090)	0.16 (0.15, 0.16)
Day 5	0.093 (0.092, 0.094)	0.12 (0.11, 0.13)
	Area under curve	Area under curve skill score
Clim.	0.622 (0.620, 0.624)	
Day 1	0.893 (0.892, 0.894)	0.72 (0.71, 0.72)
Day 2	0.872 (0.871, 0.873)	0.66 (0.66, 0.66)
Day 3	0.853 (0.852, 0.854)	0.61 (0.61, 0.62)
Day 4	0.845 (0.843, 0.847)	0.59 (0.58, 0.60)
Day 5	0.815 (0.813, 0.817)	0.51 (0.51, 0.52)

a forecast horizon of 5 days (Tab. 3). Inference is based on the samples from the MCMC simulations.

Further, the Brier skill score (BSS) is investigated over space for a lead time of 5 days (Fig. 7). A 7-year climatology encompassing a spatial and seasonal effect (Eq. 2) serves as reference forecast. Highest skill can be found in the southern half of the same domain as well as in the north eastern region. Inference based on MCMC samples reveals significant positive skill along the main Alpine ridge. In order to account for multiple testing, due to testing each individual cell, we apply the correction for minimizing the *false discovery rate* (Benjamini and Hochberg 1995) which is robust to spatial dependence within the field of the test (Wilks 2016).

Intensity of Lightning Events

The evaluation of the predictive performance with respect to the intensity of lightning events takes the hurdle model (Eq. 1) into account. We investigate the global performance of the forecasts, firstly, by averaging scores over all grid cells, secondly, by visualizing rootograms for a graphical portrayal of calibration, and, thirdly, by looking at the spatial distribution of skill scores.

For every day a probability mass is predicted for every possible outcome $y \in \{0, 1, 2, \dots\}$, which are evaluated (Tab. 4) with the *ranked probability score* (RPS, Epstein 1969) and log-likelihood of the hurdle negative binomial distribution, i.e., the combination of the logit binomial and the zero truncated negative binomial. The predictions are compared against a reference climatology in which each parameter—π, μ and θ—is modelled by a seasonal effect and a spatial effect. The models based on the ECMWF covariates outperform the climatology up to a forecast horizon of 5 days.

Marginal calibration of the predicted distributions is assessed by the use of rootograms (Fig. 8). Rootograms compare the observed frequencies for every possible outcome $\{0, 1, 2, \dots\}$

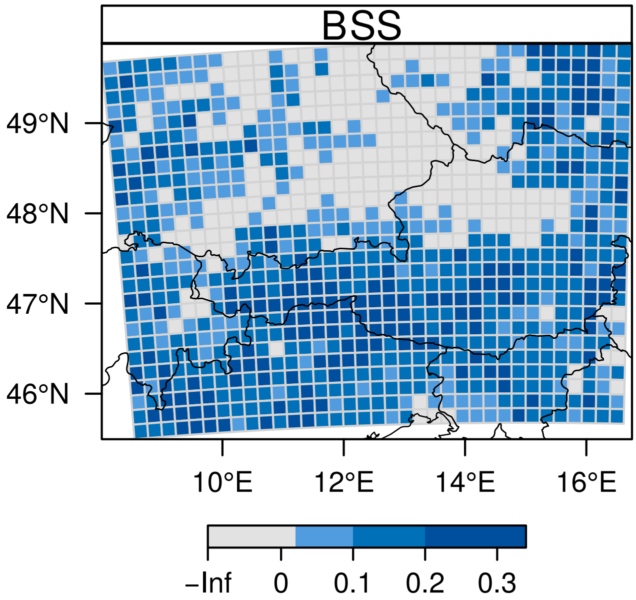

Figure 7: Spatial distribution of Brier skill scores for a lead time of 5 days evaluating the occurrence of lightning events (#flashes>0). Blueish colours indicate significantly positive values.

Table 4: Out-of-sample performance of the intensity model. 95% credible intervals based on MCMC samples are given in parentheses.

	Ranked probability score	Ranked probability skill score
Clim.	1.58 (1.58, 1.58)	
Day 1	1.36 (1.36, 1.37)	0.137 (0.134, 0.139)
Day 2	1.41 (1.40, 1.42)	0.108 (0.102, 0.112)
Day 3	1.46 (1.46, 1.47)	0.074 (0.068, 0.079)
Day 4	1.47 (1.46, 1.47)	0.072 (0.066, 0.076)
Day 5	1.49 (1.49, 1.50)	0.056 (0.052, 0.059)
	Log-likelihood	Log-likelihood skill score
Clim.	−87457 (−87498, −87413)	
Day 1	−73798 (−74204, −73641)	0.156 (0.152, 0.158)
Day 2	−75960 (−76373, −75740)	0.131 (0.127, 0.134)
Day 3	−77698 (−79944, −77374)	0.112 (0.086, 0.115)
Day 4	−78363 (−78808, −78128)	0.104 (0.099, 0.107)
Day 5	−80533 (−81287, −80013)	0.079 (0.071, 0.085)

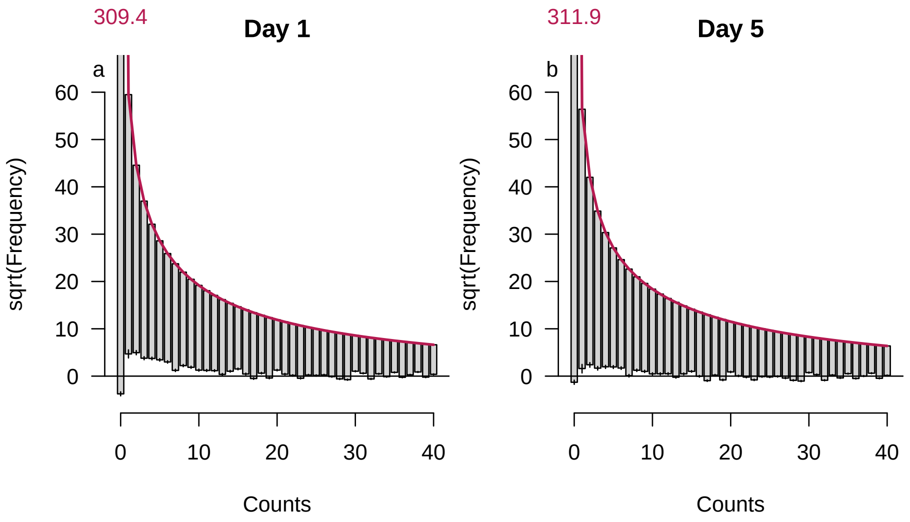

Figure 8: Hanging Rootograms for the intensities models with a forecast horizon of 1 and 5 days. The curve shows the expected frequencies and bars the observed frequencies on the square root scale. The lines at the bottom ends of the bars show the 95% credible intervals from MCMC sampling of the difference between expected and observed frequencies.

with the expected frequencies—the sum of the predicted densities over all samples—on the square root scale (Kleiber and Zeileis 2016). In a *hanging* rootogram bars indicating the square root of the observed frequencies are hanging from a curve showing the square root of expected frequencies.

The rootogram for day 1 reveals that the amount of zero counts is underestimated and that counts in the range from 1 to approx. 10 are overestimated. For higher counts the rootogram reveals good calibration of the model. The rootogram for the model with a forecast horizon of 5 days shows slightly better calibration for counts in the lower range. Although the bottom end of the bar for zero counts is closer to the x-axis, the 95% credible intervals from the MCMC sampling reveal that the model also underestimates the amount of zeros.

Finally, we investigate the spatial distribution of different skill scores for a lead time of 5 days (Fig. 9). From the hurdle model a probability forecast for exceeding 10 flashes per grid cell, a prediction of the 90% quantile, and the full probability distribution as prediction per se are derived. A 7-year climatology encompassing spatial and seasonal effects for the three parameters—π, μ, and θ—of the hurdle model serves as reference forecast. The three spatial distributions of skill reveal the same pattern as the skill score of the occurrence model (Fig. 7), with highest skill in the north eastern corner of the domain and in the southern half which includes the main Alpine ridge.

5. Discussion

This section discusses the relation of the present work with two other studies: Firstly, a work with meteorological background on the prediction of thunderstorms in the Eastern Alps (Simon *et al.* 2018). Secondly, a work from the statistical literature which focuses on gradient

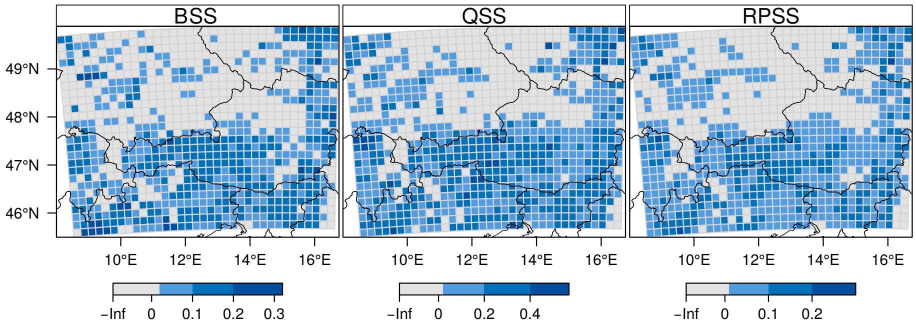

Figure 9: Spatial distribution of skill scores for a lead time of 5 days. Blueish colours indicate significantly positive values. **Left:** Brier skill score for exceeding 10 flashes per grid cell. **Middle:** Quantile skill score for the 90% quantile. **Right:** Ranked probability skill score for the full predictive distribution.

boosting for distributional regression and presents a case study with a count data variable as response (Thomas *et al.* 2018).

Simon *et al.* (2018) use the same methodology—selection using gradient boosting with stability selection and MCMC simulation for estimating the final model—as in this study, but only for the occurrence of thunderstorms and based on the deterministic high resolution ECMWF forecast from 2010–2015. During this time the native resolution of the ECMWF HRES was 16×16 km^2 and thus comparable to the resolution of the target variable in this study. Although the framework was different—longer training period of four year and *only* deterministic NWP forecasts—the resulting out-of-sample scores are comparable: Brier skill score ranges from approx. 0.25 to approx. 0.12 trough out the forecast horizons of 1 to 5 days. The AUC ranges from 0.88 to 0.79. Also, the spatial patterns of the skill match with patterns presented by Simon *et al.* (2018). Hence the evaluation data of that study and the present study do not match, conclusions whether there is a benefit of covariates derived from the ensemble over covariates derived from the deterministic run can not be drawn.

Thomas *et al.* (2018) apply a hurdle model with a zero truncated negative binomial in their study about abundance of wintering sea ducks. The abundance of sea ducks is quantified on a grid which leads to a response quantity with similar properties as the present lightning counts: 75% zeros and overdispersion. They also separate the hurdle model for the selection of terms by gradient boosting with stability selection. However, in Thomas *et al.*'s study also terms for the dispersion parameter θ have been selected, which could also be a consequence of less regularization within the individual boosting runs.

There is one more important difference between the present study and the work by Thomas *et al.* (2018), namely the way in which the final model is estimated. After the selection procedure the final model is fitted by gradient boosting. The optimal amount of regularization—tuning the number of iterations—is found by maximizing the out-of-bootstrap log-likelihood. In the present study the final model is estimated using MCMC simulation. Thus regularization is performed for each individual term by a prior distribution. A major advantage of the Bayesian approach is that inferential conclusions for effects, scores, and predictions can be drawn from the MCMC samples.

To conclude, this study proposes a framework to predict the probability of occurrence and the intensity of lightning events (*or* thunderstorms) in the European Eastern Alps. A hurdle approach—with a binomial hurdle and a zero-truncated negative binomial as count part—is chosen to account for excess zeros and overdispersion in the data. Covariates for nonlinear terms in additive predictors are derived from the ECMWF ensemble prediction system. An objective selection procedure—gradient boosting with stability selection—reduces the set of numerous terms. The final models are estimated using MCMC simulation in order to provide valid credible intervals for effects, predictions, and out-of-sample scores.

Both the occurrence and intensity models outperform a climatology up to a forecast horizon of 5 days. The predictive skill is greater over complex terrain of the Eastern Alps than over regions with fewer orographic features. This pattern can be associated with persistent forcings in regions with complex terrain such as orographic lifting, thermal-induced circulations, and lee effects (Houze 2014).

Computational Details

The statistical modelling has been carried out using the software environment R (R Core Team 2018). The add-on package **bamlss** (Umlauf *et al.* 2017) offers a flexible toolbox for distributional regression models. It allows to perform gradient boosting via the model fitting engine function `boost()`, and to simulate MCMC samples of the posterior distribution with the engine function `GMCMC()`. The **countreg** package (Zeileis *et al.* 2008) provides score functions and the hessian of the zero truncated negative binomial distribution and the high level plotting function `rootogram()`.

Acknowledgements

We acknowledge the funding of this work by the Austrian Research Promotion Agency (FFG) project *LightningPredict* (Contract No. 846620). The computational results presented have been achieved using the HPC infrastructure LEO of the University of Innsbruck.

References

Bates BC, Dowdy AJ, Chandler RE (2018). "Lightning Prediction for Australia Using Multivariate Analyses of Large-Scale Atmospheric Variables." *J. Appl. Meteor. Climatol.*, **57**(3), 525–534. doi:10.1175/JAMC-D-17-0214.1.

Benjamini Y, Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *J. Roy. Stat. Soc. B*, **57**(1), 289–300. URL http://www.jstor.org/stable/2346101.

Brezger A, Lang S (2006). "Generalized Structured Additive Regression Based on Bayesian P-Splines." *Comp. Stat. Data Anal.*, **50**(4), 967–991. doi:10.1016/j.csda.2004.10.011.

Buizza R, Milleer M, Palmer TN (1999). "Stochastic Representation of Model Uncertainties in the ECMWF Ensemble Prediction System." *Quart. J. Roy. Meteor. Soc.*, **125**(560), 2887–2908. doi:10.1002/qj.49712556006.

Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*. Econometric Society Monographs, 2nd edition. Cambridge University Press, Cambridge.

Epstein ES (1969). "A Scoring System for Probability Forecasts of Ranked Categories." *J. Appl. Meteor.*, **8**(6), 985–987. doi:10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). *Regression: Models, Methods and Applications*. Springer, Berlin. doi:10.1007/978-3-642-34333-9.

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J, Werner M, Oskin M, Burbank D, Alsdorf D (2007). "The Shuttle Radar Topography Mission." *Rev. Geophys.*, **45**(2), 1–33. doi:10.1029/2005RG000183.

Gamerman D (1997). "Sampling from the Posterior Distribution in Generalized Linear Mixed Models." *Stat. Comput.*, **7**(1), 57–68. ISSN 0960-3174. doi:10.1023/a:1018509429360.

Gijben M, Dyson LL, Loots MT (2017). "A Statistical Scheme to Forecast the Daily Lightning Threat over Southern Africa Using the Unified Model." *Atmos. Res.*, **194**, 78–88. doi:10.1016/j.atmosres.2017.04.022.

Hofner B, Boccuto L, Göker M (2015). "Controlling False Discoveries in High-Dimensional Situations: Boosting with Stability Selection." *BMC Bioinformatics*, **16**(1). doi:10.1186/s12859-015-0575-3.

Houze RA (2014). *Cloud Dynamics*, volume 104 of *International Geophysics*. Academic Press. doi:10.1016/B978-0-12-374266-7.00003-2.

Kleiber C, Zeileis A (2016). "Visualizing Count Data Regressions Using Rootograms." *Am. Stat.*, **70**(3), 296–303. doi:10.1080/00031305.2016.1173590.

Klein N, Kneib T, Lang S (2015). "Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data." *J. Am. Stat. Assoc.*, **110**(509), 405–419. doi:10.1080/01621459.2014.912955.

Lang S, Umlauf N, Wechselberger P, Harttgen K, Kneib T (2014). "Multilevel Structured Additive Regression." *Stat. Comput.*, **24**(2), 223–238. ISSN 0960-3174. doi:10.1007/s11222-012-9366-0.

Langhans W, Schmidli J, Schär C (2012). "Bulk Convergence of Cloud-Resolving Simulations of Moist Convection over Complex Terrain." *J. Atmos. Sci.*, **69**(7), 2207–2228. doi:10.1175/JAS-D-11-0252.1.

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012). "Generalized Additive Models for Location, Scale and Shape for High Dimensional Data—A Flexible Approach based on Boosting." *J. Roy. Stat. Soc. C*, **61**(3), 403–427. doi:10.1111/j.1467-9876.2011.01033.x.

Meinshausen N, Bühlmann P (2010). "Stability Selection." *J. Roy. Stat. Soc. B*, **72**(4), 417–473. doi:10.1111/j.1467-9868.2010.00740.x.

Mullahy J (1986). "Specification and Testing of some Modified Count Data Models." *J. Econometr.*, **33**(3), 341–365. doi:10.1016/0304-4076(86)90002-3.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *J. Roy. Stat. Soc. C*, **54**(3), 507–554. ISSN 1467-9876. doi:10.1111/j.1467-9876.2005.00510.x.

Schmeits MJ, Kok KJ, Vogelezang DHP, van Westrhenen RM (2008). "Probabilistic Forecasts of (Severe) Thunderstorms for the Purpose of Issuing a Weather Alarm in the Netherlands." *Wea. Forecasting*, **23**(6), 1253–1267. doi:10.1175/2008WAF2007102.1.

Schulz W, Cummins K, Diendorfer G, Dorninger M (2005). "Cloud-to-Ground Lightning in Austria: A 10-Year Study Using Data from a Lightning Location System." *J. Geophys. Res.*, **110**(D9). doi:10.1029/2004JD005332.

Simon T, Fabsic P, Mayr GJ, Umlauf N, Zeileis A (2018). "Probabilistic Forecasting of Thunderstorms in the Eastern Alps." *Mon. Wea. Rev.*, **0**(0), 0–0. doi:10.1175/MWR-D-17-0366.1.

Simon T, Umlauf N, Zeileis A, Mayr GJ, Schulz W, Diendorfer G (2017). "Spatio-Temporal Modelling of Lightning Climatologies for Complex Terrain." *Nat. Hazards Earth Syst. Sci.*, **17**(3), 305–314. doi:10.5194/nhess-17-305-2017.

Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B (2018). "Gradient Boosting for Distributional Regression: Faster Tuning and Improved Variable Selection via Noncyclical Updates." *Stat. Comput.*, **28**(3), 673–687. doi:10.1007/s11222-017-9754-6.

Umlauf N, Klein N, Zeileis A (2017). "BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond)." *J. Comput. Graph. Stat.* doi:10.1080/10618600.2017.1407325.

Wilks DS (2016). ""The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It." *Bull. Amer. Meteor. Soc.*, **97**(12), 2263–2273. doi:10.1175/BAMS-D-15-00267.1.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *J. Stat. Softw.*, **27**(1), 1–25. doi:10.18637/jss.v027.i08.

A. The Negative Binomial Hurdle Distribution

In this section we derive the density and the log-likelihood of the hurdle model (Eq. 1). Hurdle models consist of two parts: A *binary hurdle part*—modelling the probability of non-zero events—and a *truncated count part*—modelling the distribution of positive counts.

Hurdle models were introduced by Mullahy (1986). A comprehensive overview of modelling count data is given by Cameron and Trivedi (2013). Zeileis *et al.* (2008) present an implementation of regression models for count data in the software environment R.

In the present study a binomial distribution serves as binary hurdle, and a zero-truncated negative binomial distribution as truncated count part. The binomial distribution has the density,

$$f_{\text{BINOM}}(z \,|\, \pi) = (1-\pi)^{1-z} \cdot \pi^{z}, \quad z \in \{0, 1\}, \tag{3}$$

which is determined by the probability π.

To derive the truncated count part we start with the negative binomial (type 2) distribution (Cameron and Trivedi 2013), with the density,

$$f_{\text{NB}}(z \,|\, \mu, \theta) = \frac{\Gamma(\theta + z)}{\Gamma(\theta) \cdot z!} \cdot \frac{\mu^{z} \cdot \theta^{\theta}}{(\mu + \theta)^{\theta + z}}, \quad z \in \{0, 1, 2, \dots\}, \tag{4}$$

where $\mu > 0$ is the expectation of the distribution, $\mathsf{E}(z) = \mu$, and $\theta > 0$ modifies the variance, $\mathsf{VAR}(z) = \mu + \mu^2/\theta$, in order to account for the overdispersion in the gridded lightning observations.

For truncating the negative binomial the probability mass at zero is redistributed towards positive values leading to the density of the zero-truncated negative binomial,

$$f_{\text{ZTNB}}(y \,|\, \mu, \theta) = \frac{f_{\text{NB}}(y \,|\, \mu, \theta)}{1 - f_{\text{NB}}(0 \,|\, \mu, \theta)}, \quad y \in \{1, 2, \dots\}. \tag{5}$$

The binomial distribution (Eq. 3) and the zero-truncated negative binomial (Eq. 5) are combined to obtain the hurdle model (Eq. 1). From the density of the hurdle model we can derive the log-likelihood function (which serves as objective function during optimization),

$$\ell(\pi, \mu, \theta \,|\, y) = \underbrace{\mathbf{I}_{\{0\}}(y) \cdot \log(1-\pi) + (1 - \mathbf{I}_{\{0\}}(y)) \cdot \log \pi}_{\tilde{\ell}_{\text{BHP}}(\pi \,|\, y)} + \underbrace{(1 - \mathbf{I}_{\{0\}}(y)) \cdot \log(f_{\text{ZTNB}}(y \,|\, \mu, \theta))}_{\tilde{\ell}_{\text{TCP}}(\mu, \theta \,|\, y)}, \tag{6}$$

where $\mathbf{I}_{\{0\}}(y)$ is an indicator function which takes the value one if y equals zero, and zero otherwise. The log-likelihood is a function of the parameters π, μ, and θ. However, it can be separated additivly into a function of π, $\tilde{\ell}_{\text{BHP}}(\pi \,|\, y)$, and a function of μ and θ, $\tilde{\ell}_{\text{TCP}}(\mu, \theta \,|\, y)$. Thus, during optimization the optima for the two functions can be obtained independently from each other.

In particular $\tilde{\ell}_{\text{BHP}}$ and $\tilde{\ell}_{\text{TCP}}$ are equivalent to the log-likelihood of the binomial distribution (Eq. 3) and the zero-truncated negative binomial (Eq. 5), respectively.

B. Noncyclic Gradient Boosting

The steps of the noncyclical algorithm for an arbitrary distribution with the parameters λ_1, λ_2, $\dots \lambda_m$ are as follows:

1. Initially all terms (or *base-learners*) from all predictors $\eta^0(\lambda_\star)$ are set equal to zero, i.e., $f_j(\mathtt{x}_j) = 0$.

2. For each predictor find the term fitting best to the score function:

   (a) Evaluate the negative gradient of the log-likelihood $-\partial\ell/\partial\eta^k(\lambda_\star)$ w.r.t. the current predictor $\eta^k(\lambda_\star)$ for every observation, leading to a vector of gradients.

   (b) Fit low-degree-of-freedom splines for each term $f_j(\mathtt{x}_j)$ to the gradient vector using penalized least squares estimation.

   (c) The coefficients of the best fitting term—w.r.t. the residual sum of squares—are updated by a proportion ν, e.g., $\nu = 0.1$, leading to an auxiliary predictor,

$$\tilde{\eta}(\lambda_\star) = \eta^k(\lambda_\star) + \nu \cdot f_j(\mathtt{x}_j). \tag{7}$$

3. Find the auxiliary predictor leading to the largest improvement of the log-likelihood and assign it to its predictor for the next iteration,

$$\eta^{k+1}(\lambda_\star) = \begin{cases} \tilde{\eta}(\lambda_\star) & \text{if } \tilde{\eta}(\lambda_\star) \text{ improves } \ell \text{ best} \\ \eta^k(\lambda_\star) & \text{otherwise.} \end{cases} \tag{8}$$

4. Repeat steps 2 and 3 for a predefined number of iterations $k_{\max}$ or until a predefined number of terms q has been selected.

Affiliation:

Thorsten Simon
Department of Atmospheric and Cryospheric Sciences
University of Innsbruck
Innrain 52f
6020 Innsbruck, Austria
E-mail: Thorsten.Simon@uibk.ac.at

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2018-14 **Thorsten Simon, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Lightning prediction using model output statistics

2018-13 **Martin Geiger, Johann Scharler:** How do consumers interpret the macroeconomic effects of oil price fluctuations? Evidence from U.S. survey data

2018-12 **Martin Geiger, Johann Scharler:** How do people interpret macroeconomic shocks? Evidence from U.S. survey data

2018-11 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Low visibility forecasts for different flight planning horizons using tree-based boosting models

2018-10 **Michael Pfaffermayr:** Trade creation and trade diversion of regional trade agreements revisited: A constrained panel pseudo-maximum likelihood approach

2018-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model

2018-08 **Lisa Schlosser, Torsten Hothorn, Reto Stauffer, Achim Zeileis:** Distributional regression forests for probabilistic precipitation forecasting in complex terrain

2018-07 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Delegated decision making and social competition in the finance industry

2018-06 **Manuel Gebetsberger, Reto Stauffer, Georg J. Mayr, Achim Zeileis:** Skewed logistic distribution for statistical temperature post-processing in mountainous areas

2018-05 **Reto Stauffer, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Hourly probabilistic snow forecasts over complex terrain: A hybrid ensemble postprocessing approach

2018-04 **Utz Weitzel, Christoph Huber, Florian Lindner, Jürgen Huber, Julia Rose, Michael Kirchler:** Bubbles and financial professionals

2018-03 **Carolin Strobl, Julia Kopf, Raphael Hartmann, Achim Zeileis:** Anchor point selection: An approach for anchoring without anchor items

2018-02 **Michael Greinecker, Christopher Kah:** Pairwise stable matching in large economies

2018-01 **Max Breitenlechner, Johann Scharler:** How does monetary policy influence bank lending? Evidence from the market for banks' wholesale funding

2017-27 **Kenneth Harttgen, Stefan Lang, Johannes Seiler:** Selective mortality and undernutrition in low- and middle-income countries

2017-26 **Jun Honda, Roman Inderst:** Nonlinear incentives and advisor bias

2017-25 **Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Probabilistic forecasting of thunderstorms in the Eastern Alps

2017-24 **Florian Lindner:** Choking under pressure of top performers: Evidence from biathlon competitions

2017-23 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood

2017-22 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Forecasting low-visibility procedure states with tree-based statistical methods

2017-21 **Philipp Kneringer, Sebastian J. Dietz, Georg J. Mayr, Achim Zeileis:** Probabilistic nowcasting of low-visibility procedure states at Vienna International Airport during cold season

2017-20 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior

2017-19 **Martin Geiger, Richard Hule:** The role of correlation in two-asset games: Some experimental evidence

2017-18 **Rudolf Kerschbamer, Daniel Neururer, Alexander Gruber:** Do the altruists lie less?

2017-17 **Meike Köhler, Nikolaus Umlauf, Sonja Greven:** Nonlinear association structures in flexible Bayesian additive joint models

2017-16 **Rudolf Kerschbamer, Daniel Muller:** Social preferences and political attitudes: An online experiment on a large heterogeneous sample

2017-15 **Kenneth Harttgen, Stefan Lang, Judith Santer, Johannes Seiler:** Modeling under-5 mortality through multilevel structured additive regression with varying coefficients for Asia and Sub-Saharan Africa

2017-14 **Christoph Eder, Martin Halla:** Economic origins of cultural norms: The case of animal husbandry and bastardy

2017-13 **Thomas Kneib, Nikolaus Umlauf:** A primer on bayesian distributional regression

2017-12 **Susanne Berger, Nathaniel Graham, Achim Zeileis:** Various versatile variances: An object-oriented implementation of clustered covariances in R

2017-11 **Natalia Danzer, Martin Halla, Nicole Schneeweis, Martina Zweimüller:** Parental leave, (in)formal childcare and long-term child outcomes

2017-10 **Daniel Muller, Sander Renes:** Fairness views and political preferences - Evidence from a large online experiment

2017-09 **Andreas Exenberger:** The logic of inequality extraction: An application to Gini and top incomes data

2017-08 **Sibylle Puntscher, Duc Tran Huy, Janette Walde, Ulrike Tappeiner, Gottfried Tappeiner:** The acceptance of a protected area and the benefits of sustainable tourism: In search of the weak link in their relationship

2017-07 **Helena Fornwagner:** Incentives to lose revisited: The NHL and its tournament incentives

2017-06 **Loukas Balafoutas, Simon Czermak, Marc Eulerich, Helena Fornwagner:** Incentives for dishonesty: An experimental study with internal auditors

2017-05 **Nikolaus Umlauf, Nadja Klein, Achim Zeileis:** BAMLSS: Bayesian additive models for location, scale and shape (and beyond)

2017-04 **Martin Halla, Susanne Pech, Martina Zweimüller:** The effect of statutory sick-pay on workers' labor supply and subsequent health

2017-03 **Franz Buscha, Daniel Müller, Lionel Page:** Can a common currency foster a shared social identity across different nations? The case of the Euro.

2017-02 **Daniel Müller:** The anatomy of distributional preferences with group identity

2017-01 **Wolfgang Frimmel, Martin Halla, Jörg Paetzold:** The intergenerational causal effect of tax evasion: Evidence from the commuter tax allowance in Austria

University of Innsbruck

Working Papers in Economics and Statistics

2018-14

Thorsten Simon, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis

Lightning prediction using model output statistics

Abstract
A method to predict lightning by postprocessing numerical weather prediction (NWP) output is developed for the region of the European Eastern Alps. Cloud-to-ground flashes-detected by the ground-based ALDIS network-are counted on the 18x18 km^2 grid of the 51-member NWP ensemble of the European Centre of Medium-Range Weather Forecasts (ECMWF). These counts serve as target quantity in count data regression models for the occurrence and the intensity of lightning events. The probability whether lightning occurs or not is modelled by a binomial distribution. For the intensity a hurdle approach is employed, for which the binomial distribution is combined with a zero-truncated negative binomial to model the counts within a grid cell. In both statistical models the parameters of the distributions are described by additive predictors, which are assembled by potentially nonlinear terms of NWP covariates. Measures of location and spread of approx. 100 direct and derived NWP covariates provide a pool of candidates for the nonlinear terms. A combination of stability selection and gradient boosting selects influential terms. Markov chain Monte Carlo (MCMC) simulation estimates the final model to provide credible inference of effects, scores and predictions. The selection of terms and MCMC simulation are applied for data of the year 2016, and out-of-sample performance is evaluated for 2017. The occurrence model outperforms a reference climatology-based on seven years of data-up to a forecast horizon of 5 days. The intensity model is calibrated and also outperforms climatology for exceedance probabilities, quantiles, and full predictive distributions.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)