Probabilistic Forecasting of Thunderstorms in the Eastern Alps

THORSTEN SIMON, PETER FABSIĆ, GEORG J. MAYR, NIKOLAUS UMLAUF, AND ACHIM ZEILEIS

University of Innsbruck, Innsbruck, Austria

(Manuscript received 4 December 2017, in final form 31 May 2018)

ABSTRACT

A probabilistic forecasting method to predict thunderstorms in the European eastern Alps is developed. A statistical model links lightning occurrence from the ground-based Austrian Lightning Detection and Information System (ALDIS) detection network to a large set of direct and derived variables from a numerical weather prediction (NWP) system. The NWP system is the high-resolution run (HRES) of the European Centre for Medium-Range Weather Forecasts (ECMWF) with a grid spacing of 16 km. The statistical model is a generalized additive model (GAM) framework, which is estimated by Markov chain Monte Carlo (MCMC) simulation. Gradient boosting with stability selection serves as a tool for selecting a stable set of potentially nonlinear terms. Three grids from 64×64 to 16×16 km² and five forecast horizons from 5 days to 1 day ahead are investigated to predict thunderstorms during afternoons (1200–1800 UTC). Frequently selected covariates for the nonlinear terms are variants of convective precipitation, convective potential available energy, relative humidity, and temperature in the midlayers of the troposphere, among others. All models, even for a lead time of 5 days, outperform a forecast based on climatology in an out-of-sample comparison. An example case illustrates that coarse spatial patterns are already successfully forecast 5 days ahead.

1. Introduction

Predicting thunderstorms in complex terrain is a challenging task since one of the main tools, numerical weather prediction (NWP) systems, cannot fully resolve convective processes or circulations and exchange processes over complex topography. Thus, NWP output is statistically postprocessed to enhance its value for thunderstorm forecasts. Logistic regression (Schmeits et al. 2008; Gijben et al. 2017) or artificial neural networks (Collins and Tissot 2015) are often employed for predicting whether thunderstorms will occur.

However, two difficulties are present. First, the response variable, the probability of thunderstorms (equivalently the occurrence of lightning), might nonlinearly depend on

individual covariates from the NWP. Second, an abundance of potential covariates provided by NWP systems could be included in the statistical model. Thus, a statistical framework capable of both handling nonlinear relationships between the response and covariates and of objectively selecting the important covariates is needed.

Nonlinearities can be captured either by transformations of covariates (e.g., power or log transformation) or by nonlinear regression models, [e.g., a generalized additive model (GAM); Hastie and Tibshirani 1990; Wood 2017]. GAMs can be formulated in a Bayesian framework (Brezger and Lang 2006), which allows us to estimate GAMs using Markov chain Monte Carlo (MCMC) simulations. This approach is particularly attractive for inference of complex GAMs (Umlauf et al. 2018). GAMs have been used for postprocessing NWP output to capture complex spatiotemporal characteristics of temperature (Dabernig et al. 2017) and precipitation (Stauffer et al. 2017).

Selection is classically performed by testing all possible subsets of potential covariates (Miller 2002). This procedure becomes computationally intractable for large numbers of covariates, as in our case. Thus, nonexhaustive methods, such as stepwise selection, are more common (Miller 2002). In recent years, regularization methods

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-17-0366.s1>.

Corresponding author: Thorsten Simon, thorsten.simon@uibk.ac.at

DOI: 10.1175/MWR-D-17-0366.1

© 2018 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

have also become popular for variable selection in the field of postprocessing NWP output: for example, the least absolute shrinkage and selection operator (LASSO; [Wahl 2015](#)) and boosting ([Messner et al. 2017](#)).

Gradient boosting was first established in the field of machine learning ([Freund and Schapire 1995](#)) and generalized later by [Bühlmann and Hothorn \(2007\)](#) for regression models such as GAMs. A broad overview of algorithms for this technique can be found in [Mayr et al. \(2012\)](#). However, selecting the right-sized subset of covariates ([Meinshausen and Bühlmann 2010](#))—that is, avoiding selecting some noise variables ([Hofner et al. 2015](#))—remains challenging. A solution to this issue is combining gradient boosting as a method of regularization with stability selection ([Hofner et al. 2015](#)).

The aim of this study is to develop a probabilistic forecasting method for the occurrence of thunderstorms in the eastern Alps and their surroundings. To achieve this objective, we propose a novel combination of the statistical methods introduced above. A GAM serves as framework to account for potentially nonlinear relationships between response and covariates. Within this framework, an objective variable selection scheme (i.e., gradient boosting with stability selection) is performed to select a stable subset of the available covariates. In a final step, the GAM, comprising the selected terms, is estimated using MCMC sampling. This allows us to draw inferential conclusions, such as credible intervals of effects, predictions, or out-of-sample scores.

The region we focus on is the eastern Alps in Europe and their surroundings ([Fig. 1](#)). The region is exposed to severe thunderstorms and lightning during summer ([Schulz et al. 2005](#); [Simon et al. 2017](#)). Furthermore, the eastern Alps are characterized by a complex terrain. Elevation within the study domain extends from sea level up to 3798 m MSL. The atmospheric processes leading to the strong convective events and the occurrence of thunderstorms in this region cover the gamut from small to large scales. Interactions of orography, solar heating, and winds influence the lightning activity ([Bertram and Mayr 2004](#); [Houze 2014](#)). On the other hand, large-scale circulations (e.g., the North Atlantic Oscillation) might influence the lightning patterns in Europe ([Piper and Kunz 2017](#)). Studies investigating the climatological patterns of lightning activity in the region of interest and its surroundings found maxima along the northern and southern rims of the Alps ([Schulz et al. 2005](#); [Feudale et al. 2013](#); [Wapler 2013](#)).

This manuscript is structured as follows. The region of interest, the lightning detection data, and the covariates are presented in [section 2](#). The statistical methods—GAMs, gradient boosting with stability selection, and MCMC—are introduced in [section 3](#). In [section 4](#), the

FIG. 1. Topography of the region of interest (m MSL). Thick white lines indicate the $64 \times 64 \text{ km}^2$ spatial grid, and thin lines indicate the $16 \times 16 \text{ km}^2$ grid. The circle and the triangle show the location of ZRH and VIE, respectively.

results of the selection scheme are presented in detail for one model, the predictive performance of the models across different spatial and temporal scales is analyzed, and a forecast case is visualized. Discussions and conclusions are given in [section 5](#).

2. Data

In the following, the response variable based on lightning detection data and the covariates from the European Centre for Medium-Range Weather Forecasts (ECMWF) high-resolution run (HRES) are introduced. The study covers the times when most thunderstorms occur ([Bertram and Mayr 2004](#); [Schulz et al. 2005](#); [Simon et al. 2017](#)): the afternoons (1200–1800 UTC) of the convective season May–August of the years 2010–15, during which the horizontal mesh of the ECMWF high-resolution run remained unchanged at 16 km. The region is horizontally divided into multiples of the ECMWF grid to study the dependence of the forecast performance on spatial resolution. The three meshes are 64×64 , 32×32 , and $16 \times 16 \text{ km}^2$. The grids with the coarsest and finest spatial resolutions are highlighted in [Fig. 1](#) by white lines.

a. Thunderstorms based on lightning detection data

A thunderstorm is taken to have occurred in a grid cell when at least one lightning stroke—both IC and CG flashes are considered—was registered by the ground-based Austrian Lightning Detection and Information System (ALDIS) network ([Schulz et al. 2005](#)) between 1200 and 1800 UTC.

The resulting sample sizes of the datasets for the three resolutions (64×64 , 32×32 , and $16 \times 16 \text{ km}^2$) are 51 660, 221 400, and 885 600 with the unconditional probability of lightning activity of 30.3%, 19.7%, and

12.5%, respectively. Two-thirds of the data (2010–13) are used to train the statistical model (section 3), with one-third (2014–15) for out-of-sample verification.

The ALDIS detection network is integrated into the European Cooperation for Lightning Detection (EUCLID; Schulz et al. 2016), which has a detection efficiency greater than 96%, given a flash peak current greater than 2 kA and a median location accuracy of 157 m for Austria.

b. Covariates from the ECMWF high-resolution run

The covariates for predicting the occurrence of lightning activity are derived from the ECMWF-HRES initialized at 0000 UTC. The horizontal mesh of $16 \times 16 \text{ km}^2$ remained unchanged during the study period of 2010–15. A list of variables selected for this study is given in Table 1.

The variables are prepared for the lead times 12/15/18 h (day 1), 36/39/42 h (day 2), 60/63/66 h (day 3), 84/87/90 h (day 4), and 108/111/114 h (day 5), which are used to build the base for five different models with respect to each resolution.

Additional covariates are derived from the variables in Table 1. The mean, maximum, and minimum of the values at 1200, 1500, and 1800 UTC of a specific variable are denoted by the name of the variable and the suffixes “.mean,” “.max,” and “.min,” respectively. Differences between different times are marked by suffixes with four digits (e.g., t700_1812 is the temperature difference at 700 hPa between 1800 and 1200 UTC). The first and the last two digits of the suffix correspond to different times. Finally, anomalies computed by subtracting the mean values are marked by suffixes with two digits (e.g., t700_12 for the temperature anomaly at 700 hPa at 1200 UTC). This procedure leads to a total of 126 potential covariates derived from the NWP model.

3. Methods

In this section, the framework of GAMs, which allows for modeling potentially nonlinear smooth functions of the covariates, is described. Furthermore, we give an explanation of how variable selection is performed using gradient boosting with stability selection and how inference of the final selected model is made by MCMC sampling.

a. Generalized additive models

The statistical framework to model lightning activity falls in the class of GAMs. A comprehensive introduction to GAMs is given by Wood (2017).

The dichotomous variable of observing “lightning” or “no lightning” activity in a grid cell follows a Bernoulli distribution with the parameter π , which is the probability

TABLE 1. An overview of the base covariates from the ECMWF-HRES forecast. Covariates derived from this base set are discussed in the data section.

Abbreviation	Description
sqrtp	Square root of convective precipitation.
sqrtp.cape	Square root of convective available potential energy.
t500, t700, t2m	Temperature at 500 hPa, 700 hPa, and 2 m.
d2m	Dewpoint temperature at 2 m.
r700	Relative humidity at 700 hPa.
w500, w700	Pressure vertical velocity at 500 and 700 hPa.
ws700	Wind speed at 700 hPa.
wdir700	Wind direction at 700 hPa.
mls	Proxy for midlayer stability: $\text{mls} \propto \text{t500} - \text{t700}$.
tcc	Total cloud cover.
slhf	Surface latent heat flux.
ssr	Surface net solar radiation.
str	Surface net thermal radiation.
e	Evaporation.

of observing lightning activity. The logit function links π to an additive predictor η :

$$\text{logit}(\pi) = \eta = \beta_0 + f_{\text{time}}(\text{doy}) + f_{\text{space}}(\text{lon}, \text{lat}) + f_3(x_3) + \dots + f_p(x_p). \quad (1)$$

The intercept in the additive predictor is β_0 , f_j are potentially nonlinear functions modeled here by P-splines (Wood 2017), and the covariates $\mathbf{x}_j$ are derived from the ECMWF-HRES (Table 1).

Two further additive terms account for seasonal and spatial variations: f_{time} depends on the day of the year (doy), and f_{space} depends on longitude (lon) and latitude (lat). Thus, the total number of potential terms of the GAM is $p = 128$.

The first three components of Eq. (1)—intercept, temporal, and spatial effect—are employed to build a baseline model that describes the climatological probability of lightning:

$$\text{logit}(\pi_{\text{baseline}}) = \beta_0 + f_{\text{time}}(\text{doy}) + f_{\text{space}}(\text{lon}, \text{lat}). \quad (2)$$

The response variable follows a Bernoulli distribution with the associated log-likelihood function,

$$\ell(\pi) = y \times \log(\pi) + (1 - y) \times \log(1 - \pi), \quad (3)$$

for an individual observation $y \in \{0, 1\}$.

To ensure regularization of the functions f_j and to prevent overfitting, in the frequentist approach, so-called penalty terms are added to the objective log-likelihood function such that the smoothness of each function is controlled by additional smoothness parameters that need to be estimated—for example, by additionally

minimizing the AIC or by restricted maximum likelihood (REML; Wood 2017). In boosting, the smoothness parameters are utilized to initialize each function f_j with the same degrees of freedom to ensure an equal comparison for the selection of base learners in each boosting iteration (Bühlmann and Hothorn 2007). The Bayesian analog of the frequentist penalty terms is shrinkage priors that are assigned to the corresponding regression coefficients of each function f_j . These priors are commonly based on multivariate normal priors (Umlauf et al. 2018). Hence, the regression coefficients and the smoothness parameters can be estimated simultaneously using MCMC sampling.

In this study, we propose a novel combination of methods in order to obtain a final GAM. First, gradient boosting with stability selection serves for selecting a stable subset of terms. Second, the selected model is estimated using MCMC sampling, which allows drawing inferential conclusions about the selected terms.

b. Gradient boosting with stability selection

The selection of informative nonlinear functions f_j is performed by gradient boosting (Mayr et al. 2012) combined with stability selection (Meinshausen and Bühlmann 2010).

Gradient boosting is an iterative gradient descent algorithm, where the term that fits best to the gradient of the log-likelihood is slightly updated in each iteration. The iteration steps are

- 1) Initially all terms (or base learners) are set equal to zero: $f_j(\mathbf{x}_j) = 0$.
- 2) In each iteration k , the negative gradient of the log-likelihood $-\partial\ell/\partial\eta^k$ is evaluated for every observation, leading to a vector of gradients.
- 3) For each term $f_j(\mathbf{x}_j)$, low-degree-of-freedom splines are fitted to the gradient vector using penalized least squares estimation.
- 4) The coefficients of the best fitting term—with respect to the residual sum of squares—are updated by a proportion ν , here $\nu = 0.1$, leading to an updated predictor,

$$\eta^{k+1} = \eta^k + \nu \times f_j(\mathbf{x}_j). \quad (4)$$

- 5) Steps (2)–(4) are repeated for a predefined number of iterations $k_{\max}$ or until a predefined number of terms q has been selected.

If gradient boosting is applied as stand-alone method, the number of iterations $k_{\max}$ —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 selecting important terms f_j . It is desirable to avoid the selection of

numerous noninformative terms. Stability selection is a convenient resampling method for controlling the number of selected noninformative terms by gradient boosting (Meinshausen and Bühlmann 2010; Hofner et al. 2015).

Rather than applying this boosting approach to all n observations of the training data (2010–13), stability selection is based on drawing a subsample of size $n/2$ from the training data, running the boosting algorithm until q base learners are selected. This procedure is repeated many times. Afterward, the relative selection frequencies per base learner are computed. Eventually, the base learners for which the relative selection frequency exceeds a certain threshold are included in the final model [cf. algorithm in Hofner et al. (2015)].

c. Markov chain Monte Carlo sampling

The final model is of a complex form, as it contains several smooth effects. For such a complex model, determining confidence intervals based on asymptotic assumptions might fail. Because of the vast increase of computational power, MCMC simulations offer an attractive toolbox to provide valid credible intervals.

To be able to apply this technique to a GAM, 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 centered at the last state (Gamerman 1997; Fahrmeir et al. 2013; Umlauf et al. 2018). 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).

The ECMWF-based models, selected by gradient boosting with stability selection, and the climatological baseline models are estimated by MCMC sampling. A total of 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

In this section, the selection procedure is illustrated along one example case for one particular spatial resolution and lead time. Afterward, the predictive performance across the different resolutions and forecasts horizons is analyzed.

The selection procedure and the estimation of the final models is performed on data of 4 years (2010–13), leaving data of 2 years (2014–15) for evaluating the predictive performance of the final models.

a. Model selection

In total, 18 models are fitted: five models with ECMWF covariates and one baseline model containing the climatological probability for each of the three spatial resolutions. The variable selection based on boosting and stability selection is performed for the 15 models with ECMWF covariates.

The results of the stability selection for the model referring to the resolution $64 \times 64 \text{ km}^2$ and the forecast horizon of 1 day are visualized in Fig. 2. The boosting algorithm was run on 100 distinct random subsamples of size $n/2$ from the training data until $q = 12$ terms were selected. The bars in Fig. 2 indicate the relative frequency for a term $f_j(\mathbf{x}_j)$ being selected in the 100 boosting runs.

The results of the stability selection for all models can be found in online supplement A.

For example, the term f_{t700_1812} was selected in each of the 100 runs. On the other side of the scale, f_{ssr} was selected only once. However, all 111 terms that are not listed on the y axis have not been selected at all. Neither the seasonal term f_{time} nor the spatial term f_{space} was selected, which indicates that all variability over the considered time of the year and domain can already be explained by the resulting effects from the ECMWF covariates.

All terms for which the relative frequency exceeds the threshold 90% (dotted line) enter the final model. Thus, in this case, the final model contains nine additive terms. For the given number of potential predictors $p = 128$ and the tuning parameters of the stability selection $q = 12$ and a threshold of 90%, this procedure ensures that the expected number of falsely included terms (noninformative effects) is less than 1 [cf. Eq. (6) in Hofner et al. (2015)].

To provide inference for the effects of the final model, a MCMC sampling is performed. Figure 3 shows the resulting effects for the model with the resolution $64 \times 64 \text{ km}^2$ and the forecast horizon of 1 day. The effects are ordered according to their effect size, which is here defined as the absolute difference of the maximum and minimum value of the effect.

Mean relative humidity at 700 hPa ($r700.\text{mean}$) is the most influential covariate (Fig. 3a). The absolute difference between the maximum and minimum values of the effect is 4.13 on the logit scale. Between the lower bound and 80%, $f_{r700.\text{mean}}$ increases, which means that higher values of relative humidity in the ECMWF correspond to higher probabilities in the prediction of thunderstorms or lightning activity. However, at 80%, the effect reaches a maximum and decreases slightly for higher values.

Other important effects are associated with the differences of temperature at 700 hPa between 1800 and

FIG. 2. Results of the stability selection procedure for the model for day 1 and $64 \times 64 \text{ km}^2$ resolution. The variable names on the y axis indicate that the associated nonlinear effect was selected. The vertical dotted line highlights the threshold of 90% for the final model.

1200 UTC ($t700_1812$), the square root of convective precipitation (sqrt.cp), and the proxy for mean layer stability (mls.mean ; cf. Table 1). Effects f_{t700_1812} and $f_{\text{mls.mean}}$ decrease both nonlinearly. The effect of the square root of the diagnostic variable convective precipitation is monotonic increasing.

The effect of surface net thermal radiation (Fig. 3e) reveals a very interesting shape. It first increases from -0.17 to 0.89 at a value of $-1.78 \times 10^6 \text{ J m}^{-2}$ and decreases afterward to -1.31 on the logit scale. However, on occasions with high absolute values of longwave heat fluxes (left-hand side of the x scale), the overall model would predict very small probabilities. This is due to a compensation effect between the additive terms. High absolute values of longwave heat fluxes coincide with low values of relative humidity at 700 hPa, for which $f_{r700.\text{mean}}$ is very small. In other words, if surface net thermal radiation would be employed as a single predictor, the increase on the left side of the scale would be more pronounced.

The effect of $d2m.\text{mean}$ is monotonic increasing and spans a range of 2.08. The effects of $w500.\text{min}$, sqrt.cape.mean , and $t2m_1512$ are all less than unity on the logit scale.

This procedure—variable selection by combining gradient boosting and stability selection and fitting the final model by MCMC sampling—was performed for all

FIG. 3. Effects and 95% credible intervals of the model for day 1 and $64 \times 64 km^2$ resolution estimated via MCMC sampling. The effects are displayed on the logit scale. The number in the bottom-right corner of each panel indicates the absolute range of the effect. The x axes are cropped at the 1% and 99% quantiles of the respective covariate to enhance graphical representation.

15 models that build on ECMWF-HRES output. The effects presented for the example above are representative. All of these effects—except $d2m.mean$ and $w500.mean$ —were selected in a majority of models. In addition, the effects of the mean of total cloud cover and the mean of CAPE were selected in more than 50% of the models.

The effects for all models and an overview table can be found in online supplement B.

The selection results can be summarized as follows. Increasing the resolution also increases the number of

selected terms. For models with a longer forecast horizon, the number of selected terms decreases slightly. The median effect size decreases for increasing resolution, as well as for increasing forecast horizons. Effects with large credible intervals are more frequently observed for higher resolutions.

b. Predictive performance

The predictive performance was evaluated on the data from 2014 and 2015 by means of receiver operating

FIG. 4. Predictive performance: ROC diagrams for (left) 64×64 , (middle) 32×32 , and (right) $16 \times 16 \text{ km}^2$ resolution.

characteristics (ROCs) with the associated area under the curve (AUC) and Brier skill score (BSS). The scores are computed globally (i.e., averaged over all grid cells). Additionally, regional variations of the BSS are examined.

The global scores show that the models based on ECMWF covariates are superior to the baseline models (i.e., climatologies). Figure 4 shows the ROC diagrams [displaying true positive rate (TPR) vs false positive rate (FPR); Robin et al. (2011)] for the models with the spatial resolutions 64×64 (left), 32×32 (middle), and $16 \times 16 \text{ km}^2$ (right). The diagram illustrates how well the predictions discriminate between lightning and no lightning. A perfect prediction scheme would include the most top-left point with a TPR of unity and an FPR of zero and, thus, would be associated with an AUC of unity.

The curve for day 1 and $64 \times 64 \text{ km}^2$ shows that the probabilistic forecasts can be transformed to a binary prediction with a TPR greater than 80% and an FPR of less than 20%. For day 5 and $16 \times 16 \text{ km}^2$, TPR and FPR are approximately 70% and 30%, respectively. The AUC summarizes that the ECMWF-based models are superior to the baseline models. The results for the different resolutions are comparable.

The BSS for all models is displayed in Fig. 5. A perfect prediction would result in a BSS of unity, and a BSS of zero would indicate no predictive skill with respect to climatology. The BSS is highest (0.42) for the coarsest resolution ($64 \times 64 \text{ km}^2$) and the shortest forecast horizon (1 day) and smallest (0.11) for the finest resolution ($16 \times 16 \text{ km}^2$) and longest forecast horizon (5 days). The 95% credible intervals of the BSS were obtained using the MCMC samples from the posterior, which incorporate uncertainty of the predicted probabilities.

All forecasts are well calibrated. The skill for all resolutions decreases from short to long forecast horizons due to the decrease in sharpness of the forecasts, which will be discussed further in the case study in the next chapter.

After the evaluation of the global skill achieved by averaging over all grid cells of the forecast, the spatial distribution of BSS (Fig. 6) for the model with the finest resolution of $16 \times 16 \text{ km}^2$ is evaluated for the longest forecast horizon of 5 days. MCMC samples were used to test if BSS values are significantly greater than zero. 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). Considering this correction, a threshold of 3.9% is applied for rejecting a local null hypothesis instead of the nominal test level of 5%. Positive values of the Brier skill score (Fig. 6) mean that predictions from the postprocessing are superior to the climatology. This is given around the Alps and in the northeastern part of the domain. If the BSS in a grid is not significantly positive, the limit of predictive skill is reached. This is the case for regions farther north of the Alps.

c. Case study

One representative example (22 July 2015) is presented in order to highlight the information that can be gained from the introduced models.

The top-left panel of Fig. 7 shows the verifying observation for 22 July 2015 on the resolution $16 \times 16 \text{ km}^2$, where ones (zeros) indicate cells in which lightning was (not) observed. The top-middle panel shows the climatological probability for lightning activity in the cells for the same day compiled by the baseline model.

FIG. 5. BSS for all models with the baseline climatology as reference. The 95% intervals are derived from MCMC samples.

The baseline model reveals areas at the northern and southern rims of the Alps in which lightning activity is relatively likely, with climatological probabilities ranging up to 26.5% on the southern rim. The lowest values around 10% can be found in the northern part of the domain. This pattern is in line with earlier studies (Feudale et al. 2013; Wapler 2013). The mean of the climatological probabilities for this day is 16.2%.

The bottom panels of Fig. 7 illustrate how the predictions made by the GAMs with ECMWF predictors evolve from longer forecast horizons to shorter forecast horizons. The bottom-right panel shows the forecast with the model based on the ECMWF-HRES data with the lead times 108, 111, and 114 h (i.e., 5 days before 22 July). The mean of the predicted probabilities is 28.9% and, thus, clearly above the climatological value. However, probabilities spread homogeneously over the domain with mid-50% of the values lying between 19.0% and 38.7%.

The spatial pattern of the forecast from 3 days before the event (bottom-middle panel; Fig. 7) is already visible. There is a region with low values in the northwest of the domain, which can be distinguished from the rest of the domain with higher values.

The forecast made for the lead times 12–18 h (bottom-left panel) reveals sharp edges between the regions with high and low probabilities. The lower quarter of the predicted probabilities ranges from 0% to 1.6%, and the upper quarter ranges from 59.3% to 83.7%. Thus, the forecast provides sharp information about the spatial pattern of the forthcoming weather event. In comparison with the verifying observation (top

FIG. 6. Spatial predictive performance: the BSS for the model with $16 \times 16 \text{ km}^2$ resolution and a forecast horizon of 5 days. All BSS greater than 0.015 (shown in blue) are significantly positive at the 5% level. Tests are based on MCMC samples.

panel; Fig. 7), the spatial pattern is well reproduced by the prediction.

All out-of-sample predictions can be found in online supplement C. For further comparison forecasts on the coarsest grid with a mesh size of 64 km are collected in online supplement D.

For the same case (22 July 2015), the temporal evolution of predicted probabilities is highlighted for two sample locations: the grid cells associated with the airports of Zurich (ZRH) and Vienna (VIE). Figure 8 shows the probabilities for thunderstorms dependent on the forecast horizon.

Five days before the event, probabilities of 36.1% and 32.8% were predicted for ZRH and VIE, respectively. These values are clearly greater than the corresponding climatological probabilities, 15.8% and 12.5%. When coming closer to the date of interest, the probabilities for ZRH increase, and those for VIE decrease. For a forecast horizon of 3 days, the predicted probability at VIE drops below the climatological one. For day 1, the predicted probability for VIE cannot be distinguished from zero. The value for ZRH on the shortest forecast horizon is 75.6%. On 22 July 2015, lightning was observed in the grid cell containing ZRH, but not VIE.

5. Discussion and conclusions

This section discusses the issue that the temporal period of this study is longer than the stationary lifetime of the employed NWP; what other statistical frameworks could be used for predicting thunderstorms; and how the method can be transferred to other regions of the world.

a. Stationarity of NWP

The temporal period (2010–15) used for training and testing is longer than the stationary lifetime of the

FIG. 7. Spatial example. (top) Observed lightning activity for the afternoon (1200–1800 UTC) on 22 Jul 2015 and baseline climatology for the probability of thunderstorms for the same time interval. (bottom) Probabilistic forecasts for thunderstorms compiled 1, 3, and 5 days before 22 Jul 2015.

ECMWF high-resolution run (i.e., parameterizations affecting convection and thunderstorms change over this period). These changes could lead to biased estimates of the regression coefficients. However, the benefit of this long training period is to reduce the variance of the estimated coefficients. This is particularly beneficial, as the target quantity is rare events. Moreover, the good out-of-sample performance of the predictions supports the choice of a long training period.

b. Other postprocessing methods

Within the wide field of machine learning, there are more methods that capture nonlinearities and are capable of term selection (e.g., artificial neural networks or random forests). Collins and Tissot (2015) applied artificial neural networks to postprocess NWP output for a thunderstorm prediction, but only considered lead times up to 15 h. Furthermore, the method was not designed to draw inferential conclusions. For the proposed GAM, a sound inferential framework is applied. The MCMC samples from the posterior proved to be helpful for examining the nonlinear effects (Fig. 3), assessing the predictive skill (e.g., Fig. 6), and interpreting the predictions (Fig. 8).

c. Transfer of method

The presented framework can be easily transferred to other regions of the world. The key to this

transferability is the objective selection scheme based on a combination of gradient boosting and stability selection. Given local observations, the method is capable to select the nonlinear terms with the highest potential to increase predictive performance.

FIG. 8. Temporal example for the grids in which the airports ZRH and VIE are located (Fig. 1). The predicted probabilities are connected with solid (ZRH) and dashed (VIE) lines. The 95% intervals are based on MCMC samples. Climatologies are added in light colors.

d. Summary

This study explores generalized additive models (GAMs) and gradient boosting with stability selection as a tool for predicting thunderstorms by making use of numerical weather prediction (NWP) output. The eastern Alps in Europe serve as the study region. Observations of lightning strokes provide a proxy for the occurrence of thunderstorms. GAMs capture the potential nonlinear relationship between the covariates and the response, while boosting with stability selection offers an objective way to select a stable subset of covariates and to control the number of falsely selected terms. Inferential conclusions on effects, scores, and predictions can be derived from MCMC samples.

The resulting predictions are skillful to the longest evaluated forecast horizon of 5 days and the finest spatial resolution of $16 \times 16 \text{ km}^2$. 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).

6. Computational details

The statistical modeling has been carried out using the software environment R (R Core Team 2017). The add-on package bamls (Umlauf et al. 2018) offers a flexible toolbox for complex regression models such as GAMs. 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()`.

Acknowledgments. We acknowledge the funding of this work by the Austrian Research Promotion Agency (FFG) project LightningPredict (Grant 846620). The computational results presented have been achieved using the HPC infrastructure LEO of the University of Innsbruck. Furthermore, we are grateful to the editor and two anonymous reviewers for their valuable comments.

REFERENCES

- Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **57B**, 289–300.
- Bertram, I., and G. J. Mayr, 2004: Lightning in the eastern Alps 1993–1999, part I: Thunderstorm tracks. *Nat. Hazards Earth Syst. Sci.*, **4**, 501–511, <https://doi.org/10.5194/nhess-4-501-2004>.
- Brezger, A., and S. Lang, 2006: Generalized structured additive regression based on Bayesian P-splines. *Comput. Stat. Data Anal.*, **50**, 967–991, <https://doi.org/10.1016/j.csda.2004.10.011>.
- Bühlmann, P., and T. Hothorn, 2007: Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.*, **22**, 477–505, <https://doi.org/10.1214/07-STS242>.
- Collins, W., and P. Tisot, 2015: An artificial neural network model to predict thunderstorms within 400 km² South Texas domains. *Meteor. Appl.*, **22**, 650–665, <https://doi.org/10.1002/met.1499>.
- Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quart. J. Roy. Meteor. Soc.*, **143**, 909–916, <https://doi.org/10.1002/qj.2975>.
- Fahrmeir, L., T. Kneib, S. Lang, and B. Marx, 2013: *Regression: Models, Methods and Applications*. Springer, 698 pp.
- Feudale, L., A. Manzato, and S. Micheletti, 2013: A cloud-to-ground lightning climatology for north-eastern Italy. *Adv. Sci. Res.*, **10**, 77–84, <https://doi.org/10.5194/asr-10-77-2013>.
- Freund, Y., and R. E. Schapire, 1995: A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, P. Vitányi, Ed., Lecture Notes in Computer Science Series, Vol. 904, Springer, 23–37, https://doi.org/10.1007/3-540-59119-2_166.
- Gamerman, D., 1997: Sampling from the posterior distribution in generalized linear mixed models. *Stat. Comput.*, **7**, 57–68, <https://doi.org/10.1023/A:1018509429360>.
- Gijben, M., L. L. Dyson, and M. T. Loots, 2017: A statistical scheme to forecast the daily lightning threat over southern Africa using the Unified Model. *Atmos. Res.*, **194**, 78–88, <https://doi.org/10.1016/j.atmosres.2017.04.022>.
- Hastie, T., and R. Tibshirani, 1990: *Generalized Additive Models*. Monographs on Statistics and Applied Probability, Vol. 43, Chapman & Hall/CRC, 352 pp.
- Hofner, B., L. Boccuto, and M. Göker, 2015: Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, **16**, 144, <https://doi.org/10.1186/s12859-015-0575-3>.
- Houze, R. A., 2014: *Cloud Dynamics*. International Geophysics Series, Vol. 104, Academic Press, 496 pp.
- Lang, S., N. Umlauf, P. Wechselberger, K. Harttgen, and T. Kneib, 2014: Multilevel structured additive regression. *Stat. Comput.*, **24**, 223–238, <https://doi.org/10.1007/s11222-012-9366-0>.
- Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid, 2012: Generalized additive models for location, scale and shape for high dimensional data—A flexible approach based on boosting. *J. Roy. Stat. Soc.*, **61C**, 403–427, <https://doi.org/10.1111/j.1467-9876.2011.01033.x>.
- Meinshausen, N., and P. Bühlmann, 2010: Stability selection. *J. Roy. Stat. Soc.*, **72B**, 417–473, <https://doi.org/10.1111/j.1467-9868.2010.00740.x>.
- Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, <https://doi.org/10.1175/MWR-D-16-0088.1>.
- Miller, A., 2002: *Subset Selection in Regression*. 2nd ed. Monographs on Statistics and Applied Probability, Vol. 95, Chapman & Hall/CRC, 258 pp.
- Piper, D., and M. Kunz, 2017: Spatiotemporal variability of lightning activity in Europe and the relation to the North Atlantic Oscillation teleconnection pattern. *Nat. Hazards Earth Syst. Sci.*, **17**, 1319–1336, <https://doi.org/10.5194/nhess-17-1319-2017>.

- R Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, <https://www.R-project.org/>.
- Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, 2011: pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77, <https://doi.org/10.1186/1471-2105-12-77>.
- Schmeits, M. J., K. J. Kok, D. H. P. Vorgelezang, and R. M. van Westrhenen, 2008: Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the Netherlands. *Wea. Forecasting*, **23**, 1253–1267, <https://doi.org/10.1175/2008WAF2007102.1>.
- Schulz, W., K. Cummins, G. Diendorfer, and M. Dorninger, 2005: Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system. *J. Geophys. Res.*, **110**, D09101, <https://doi.org/10.1029/2004JD005332>.
- , G. Diendorfer, S. Pedebay, and D. R. Poelman, 2016: The European lightning location system EUCLID—Part 1: Performance analysis and validation. *Nat. Hazards Earth Syst. Sci.*, **16**, 595–605, <https://doi.org/10.5194/nhess-16-595-2016>.
- Simon, T., N. Umlauf, A. Zeileis, G. J. Mayr, W. Schulz, and G. Diendorfer, 2017: Spatio-temporal modelling of lightning climatologies for complex terrain. *Nat. Hazards Earth Syst. Sci.*, **17**, 305–314, <https://doi.org/10.5194/nhess-17-305-2017>.
- Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, <https://doi.org/10.1175/MWR-D-16-0260.1>.
- Umlauf, N., N. Klein, and A. Zeileis, 2018: BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *J. Comput. Graph. Stat.*, <https://doi.org/10.1080/10618600.2017.1407325>, in press.
- Wahl, S., 2015: Uncertainty in mesoscale numerical weather prediction: Probabilistic forecasting of precipitation. Rheinische Friedrich-Wilhelms-Universität Bonn Rep., 120 pp., <http://hss.ulb.uni-bonn.de/2015/4190/4190.htm>.
- Wapler, K., 2013: High-resolution climatology of lightning characteristics within central Europe. *Meteor. Atmos. Phys.*, **122**, 175–184, <https://doi.org/10.1007/s00703-013-0285-1>.
- Wilks, D. S., 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**, 2263–2273, <https://doi.org/10.1175/BAMS-D-15-00267.1>.
- Wood, S. N., 2017: *Generalized Additive Models: An Introduction with R*. 2nd ed. Texts in Statistical Science Series, Chapman & Hall/CRC, 476 pp.