

## DISTRIBUTIONAL REGRESSION FORESTS FOR PROBABILISTIC PRECIPITATION FORECASTING IN COMPLEX TERRAIN

BY LISA SCHLOSSER\*, TORSTEN HOTHORN<sup>†,1</sup>, RETO STAUFFER\* AND  
ACHIM ZEILEIS\*

*Universität Innsbruck\* and Universität Zürich<sup>†</sup>*

To obtain a probabilistic model for a dependent variable based on some set of explanatory variables, a distributional approach is often adopted where the parameters of the distribution are linked to regressors. In many classical models this only captures the location of the distribution but over the last decade there has been increasing interest in distributional regression approaches modeling all parameters including location, scale and shape. Notably, so-called nonhomogeneous Gaussian regression (NGR) models both mean and variance of a Gaussian response and is particularly popular in weather forecasting. Moreover, generalized additive models for location, scale and shape (GAMLSS) provide a framework where each distribution parameter is modeled separately capturing smooth linear or nonlinear effects. However, when variable selection is required and/or there are nonsmooth dependencies or interactions (especially unknown or of high-order), it is challenging to establish a good GAMLSS. A natural alternative in these situations would be the application of regression trees or random forests but, so far, no general distributional framework is available for these. Therefore, a framework for distributional regression trees and forests is proposed that blends regression trees and random forests with classical distributions from the GAMLSS framework as well as their censored or truncated counterparts. To illustrate these novel approaches in practice, they are employed to obtain probabilistic precipitation forecasts at numerous sites in a mountainous region (Tyrol, Austria) based on a large number of numerical weather prediction quantities. It is shown that the novel distributional regression forests automatically select variables and interactions, performing on par or often even better than GAMLSS specified either through prior meteorological knowledge or a computationally more demanding boosting approach.

**1. Introduction.** In regression analysis a wide range of models has been developed to describe the relationship between a response variable and a set of covariates. The classical model is the linear model (LM) where the conditional mean of the response is modeled through a linear function of the covariates (see the left panel of Figure 1 for a schematic illustration). Over the last decades this has been extended in various directions including:

---

Received November 2018; revised February 2019.

<sup>1</sup>An extended research stay of Torsten Hothorn in Innsbruck (August 2017 to January 2018) was financially supported by the Swiss National Science Foundation, grant number SNF IZSEZ0 177091.

*Key words and phrases.* Parametric models, regression trees, random forests, recursive partitioning, probabilistic forecasting, GAMLSS.

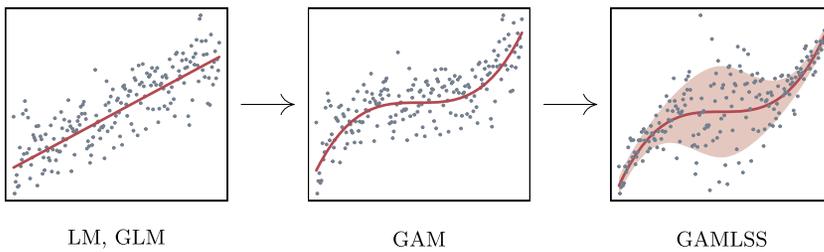


FIG. 1. *Parametric modeling developments. (Generalized) linear models (left), generalized additive models (middle), generalized additive models for location, scale and shape (right).*

- *Generalized linear models* (GLMs, [Nelder and Wedderburn \(1972\)](#)) encompassing an additional nonlinear link function for the conditional mean.
- *Generalized additive models* (GAMs, [Hastie and Tibshirani \(1986\)](#)) allowing for smooth nonlinear effects in the covariates (Figure 1, middle).
- *Generalized additive models for location, scale and shape* (GAMLSS, [Rigby and Stasinopoulos \(2005a\)](#)) adopting a probabilistic modeling approach. In GAMLSS, each parameter of a statistical distribution can depend on an additive predictor of the covariates comprising linear and/or smooth nonlinear terms (Figure 1, right).

Thus, the above-mentioned models provide a broad toolbox for capturing different aspects of the response (mean only vs. full distribution) and different types of dependencies on the covariates (linear vs. nonlinear additive terms).

While in many applications conditional mean regression models have been receiving the most attention, there has been a paradigm shift over the last decade towards distributional regression models. An important reason for this is that in many fields forecasts of the mean are not the only (or not even the main) concern but instead there is an increasing interest in probabilistic forecasts. Quantities of interest typically include exceedance probabilities for certain thresholds of the response or quantiles of the response distribution. Specifically, consider weather forecasting where there is less interest in the mean amount of precipitation on the next day. Instead, the probability of rain vs. no rain is typically more relevant or, in some situations, a prediction interval of expected precipitation (say from the expected 10% to 90% quantiles). Similar considerations apply for other meteorological quantities and hence attention in the weather forecasting literature has been shifting from classical linear deterministic models ([Glahn and Lowry \(1972\)](#)) towards probabilistic models such as the nonhomogeneous Gaussian regression (NGR) of [Gneiting et al. \(2005\)](#). The NGR typically describes the mean of some meteorological response variable through the average of the corresponding quantity from an ensemble of physically-based numerical weather predictions (NWP). Similarly, the variance of the response is captured through the variance of the ensemble of NWPs. Thus, the NGR considers both the mean as well as the uncer-

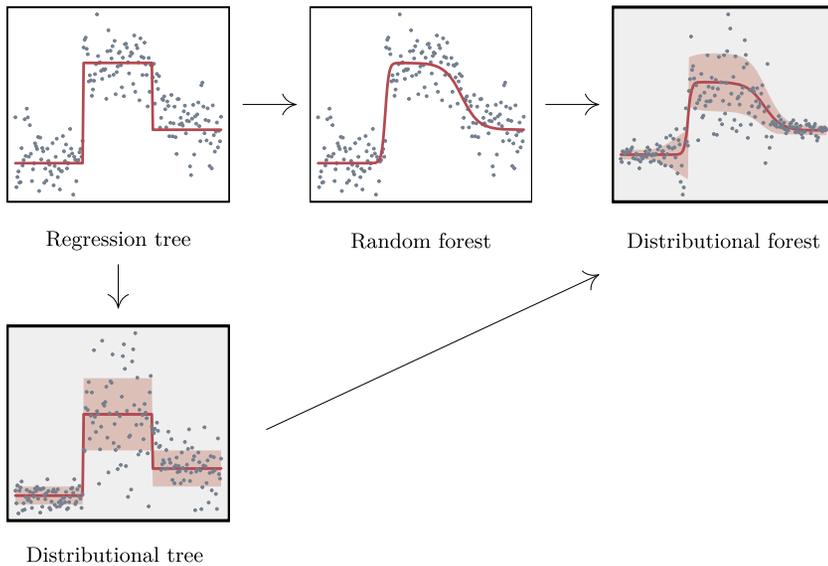


FIG. 2. *Tree and forest developments. Regression tree (top left), distributional tree (bottom left), random forest (top middle) and distributional forest (top right).*

tainty of the ensemble predictions to obtain probabilistic forecasts calibrated to a particular site.

In summary, the models discussed so far provide a broad and powerful toolset for parametric distributional fits depending on a specified set of additive linear or smooth nonlinear terms. A rather different approach to capturing the dependence on covariates are tree-based models.

- *Regression trees* (Breiman et al. (1984)) recursively split the data into more homogeneous subgroups and can thus capture abrupt shifts (Figure 2, top left) and approximate nonlinear functions. Furthermore, trees automatically carry out a forward selection of covariates and their interactions.
- *Random forests* (Breiman (2001)) average the predictions of an ensemble of trees fitted to resampled versions of the learning data. This stabilizes the recursive partitions from individual trees and hence better approximates smooth functions (Figure 2, top middle).

While classical regression trees and random forests only model the mean of the response we propose to follow the ideas from GAMLSS modeling—as outlined in Figure 1—and combine tree-based methods with parametric distributional models, yielding two novel techniques:

- *Distributional regression trees* (for short: *distributional trees*) split the data into more homogeneous groups with respect to a parametric distribution, thus captur-

ing changes in any distribution parameter like location, scale or shape (Figure 2, bottom left).

- *Distributional regression forests* (for short: *distributional forests*) utilize an ensemble of distributional trees for obtaining stabilized and smoothed parametric predictions (Figure 2, top right).

In the following, particular focus is given to distributional forests as a method for obtaining probabilistic forecasts by leveraging the strengths of random forests: the ability to capture both smooth and abruptly changing functions along with simultaneous selection of variables and possibly complex interactions. Thus, these properties make the method particularly appealing in case of many covariates with unknown effects and interactions where it would be challenging to specify a distributional regression model like GAMLSS. However, distributional forests should not be considered as a replacement of GAMLSS but rather as a complementing technique for flexible distributional regression—much like GAMs and random forests are complements for conditional mean regression.

In weather forecasting, the flexibility of distributional forests is especially appealing in mountainous regions and complex terrain where a wide range of local-scale effects are not yet resolved by the NWP models. Thus, effects with abrupt changes and possibly nonlinear interactions might be required to account for site-specific unresolved features. To illustrate this in practice, precipitation forecasts are obtained with distributional forests at 95 meteorological stations in a mountainous region in the Alps, covering mainly Tyrol, Austria, and adjacent areas (see the map in Figure 8). More specifically, a Gaussian distribution left-censored at zero, is employed to model 24-hour total precipitation so that the zero-censored point mass describes the probability of observing no precipitation on a given day (see Figure 3). Forecasts for July are established based on data from the same month over the years 1985–2012 including 80 covariates derived from a wide range of different NWP quantities. As Figure 3 shows, the station-wise forests yield a full distributional forecast for each day—here for one specific day (July 24) at one station (Axams) over four years (2009–2012)—based on the previous 24 years as learning data. The corresponding observations conform reasonably well with the predictions. In Section 3 we investigate the performance of distributional forests in this forecasting task in more detail. Compared to three alternative zero-censored Gaussian models distributional forests perform at least on par and sometimes clearly better while requiring no meteorological knowledge about the atmospheric processes which drive formation of precipitation for the model specification. The three alternatives are: a standard ensemble model output statistics approach (EMOS, Gneiting et al. (2005)) based on an NGR, a GAMLSS with regressors prespecified based on meteorological expertise (following Stauffer et al. (2017a)) and a boosted GAMLSS (Hofner, Mayr and Schmid (2016)) using nonhomogeneous boosting (Messner, Mayr and Zeileis (2017)) as an alternative technique for variable selection among all 80 available regressors.

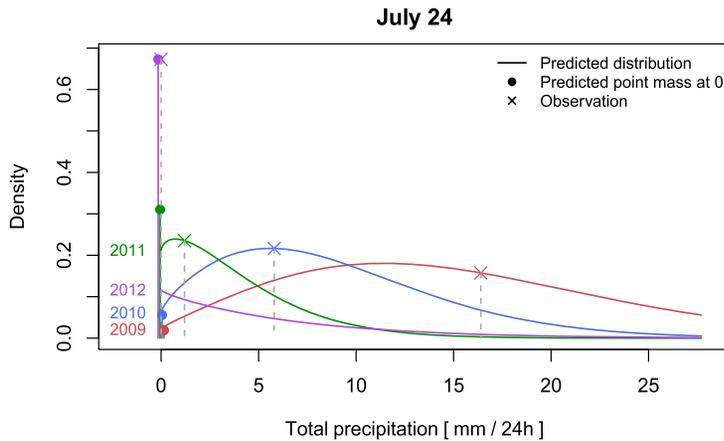


FIG. 3. Total precipitation predictions by a distributional forest at station Axams for July 24 in 2009, 2010, 2011 and 2012 learned on data from 1985–2008. Observations are non-negative and modeled by a Gaussian distribution left-censored at zero. The observations are depicted by crosses and the predicted point mass from the model by filled circles.

**2. Methodology.** To embed the distributional approach from GAMLSS into regression trees and random forests, we proceed in three steps. (1) To fix notation, we briefly review fitting distributions using standard maximum likelihood in Section 2.1. (2) A recursive partitioning strategy based on the corresponding scores (or gradients) is introduced in Section 2.2, leading to distributional trees. (3) Ensembles of distributional trees fitted to randomized subsamples are employed to establish distributional forests in Section 2.3.

The general distributional notation is exemplified in all three steps using the Gaussian distribution left-censored at zero (for short: zero-censored Gaussian). The latter is employed in the empirical case study in Section 3 to model power-transformed daily precipitation amounts.

**2.1. Distributional fit.** A distributional model  $\mathcal{D}(Y, \theta)$  is considered for the response variable  $Y \in \mathcal{Y}$  using the distributional family  $\mathcal{D}$  with  $k$ -dimensional parameter vector  $\theta \in \Theta$  and corresponding log-likelihood function  $\ell(\theta; Y)$ . The GAMLSS framework (Rigby and Stasinopoulos (2005a)) provides a wide range of such distributional families with parameterizations corresponding to location, scale and shape. Furthermore, censoring and/or truncation of these distributions can be incorporated in the usual straightforward way (see e.g., Long (1997, Chapter 7.2)).

To capture both location and scale of the probabilistic precipitation forecasts while accounting for a point mass at zero (i.e., dry days without rain), a zero-censored Gaussian distribution with location parameter  $\mu$  and scale parameter  $\sigma$  is employed. Therefore, the corresponding log-likelihood function with parameter

vector  $\theta = (\mu, \sigma)$  is

$$(2.1) \quad \ell(\mu, \sigma; Y) = \begin{cases} \log \left\{ \frac{1}{\sigma} \cdot \phi \left( \frac{Y - \mu}{\sigma} \right) \right\} & \text{if } Y > 0, \\ \log \left\{ \Phi \left( \frac{-\mu}{\sigma} \right) \right\} & \text{if } Y = 0, \end{cases}$$

where  $\phi$  and  $\Phi$  are the probability density function and the cumulative distribution function of the standard normal distribution  $\mathcal{N}(0, 1)$ . Other distributions  $\mathcal{D}$  and corresponding log-likelihoods  $\ell(\theta; Y)$  could be set up in the same way, for example, for censored shifted gamma distributions (Scheuerer and Hamill (2015)) or zero-censored logistic distributions (Gebetsberger et al. (2017)).

With the specification of the distribution family and its log-likelihood function the task of fitting a distributional model turns into the task of estimating the distribution parameter  $\theta$ . This is commonly done by maximum likelihood (ML) based on the learning sample with observations  $\{y_i\}_{i=1, \dots, n}$  of the response variable  $Y$ . The maximum likelihood estimator (MLE)  $\hat{\theta}$  is given by

$$(2.2) \quad \hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ell(\theta; y_i).$$

Equivalently, this can be defined based on the corresponding first-order conditions

$$(2.3) \quad \sum_{i=1}^n s(\hat{\theta}, y_i) = 0,$$

where  $s(\theta; y_i)$  is the associated score function

$$(2.4) \quad s(\theta; y_i) = \frac{\partial \ell}{\partial \theta}(\theta; y_i).$$

The latter is subsequently employed as a general goodness-of-fit measure to assess how well the distribution with parameters  $\theta$  fits one individual observation  $y_i$ .

*2.2. Distributional tree.* Typically, a single global model  $\mathcal{D}(Y, \theta)$  is not sufficient for reasonably representing the response distribution. Therefore, covariates  $\mathbf{Z} = Z_1, \dots, Z_m \in \mathcal{Z}$  are employed to capture differences in the distribution parameters  $\theta$ . In weather forecasting, these covariates typically include the output from numerical weather prediction systems and/or lagged meteorological observations.

To incorporate the covariates into the distributional model, they are considered as regressors in additive predictors  $g_j(\theta_j) = f_{j,1}(\mathbf{Z}) + f_{j,2}(\mathbf{Z}) + \dots$  in GAMLSS. Link functions  $g_j(\cdot)$  are used for every parameter  $\theta_j$  ( $j = 1, \dots, k$ ) based on smooth terms  $f_{j,k}$  such as nonlinear effects, spatial effects, random coefficients or interaction surfaces (Klein et al. (2015)). However, this requires specifying the additive terms and their functional forms in advance which can be challenging in

practice and potentially require expert knowledge in the application domain, especially if the number of covariates  $m$  is large.

Regression trees generally take a different approach for automatically including covariates in a data-driven way and allowing for abrupt changes, nonlinear and nonadditive effects, and interactions. In the context of distributional models the goal is to partition the covariate space  $\mathcal{Z}$  recursively into disjoint segments so that a homogeneous distributional model for the response  $Y$  can be found for each segment with segment-specific parameters. More specifically, the  $B$  disjoint segments  $\mathcal{B}_b$  ( $b = 1, \dots, B$ ) partition the covariate space

$$(2.5) \quad \mathcal{Z} = \bigcup_{b=1, \dots, B} \mathcal{B}_b,$$

and a local distributional model  $\mathcal{D}(Y, \boldsymbol{\theta}^{(b)})$  (i.e., with segment-specific parameters  $\boldsymbol{\theta}^{(b)}$ ) is fitted to the response  $Y$  in each segment.

To find the segments  $\mathcal{B}_b$  that are (approximately) homogeneous with respect to the distributional model with given parameters, the idea is to use a gradient-based recursive-partitioning approach. In a given subsample of the learning data this fits the model by ML (see equation (2.2)) and then assesses the goodness of fit by assessing the corresponding scores  $s(\hat{\boldsymbol{\theta}}; y_i)$  (see equation (2.4)).

To sum up, distributional trees are fitted recursively via:

1. Estimate  $\hat{\boldsymbol{\theta}}$  via maximum likelihood for the observations in the current subsample.
2. Test for associations (or instabilities) of the scores  $s(\hat{\boldsymbol{\theta}}, y_i)$  and  $Z_{l,i}$  for each partitioning variable  $Z_l$  ( $l = 1, \dots, m$ ).
3. Split the sample along the partitioning variable  $Z_l^*$  with the strongest association or instability. Choose the breakpoint with the highest improvement in the log-likelihood or the highest discrepancy.
4. Repeat steps 1–3 recursively in the subsamples until these become too small or there is no significant association/instability (or some other stopping criterion is reached).

Different inference techniques can be used for assessing the association between scores and covariates in step 3. In the following we use the general class of permutation tests introduced by [Hothorn et al. \(2006\)](#) which is also the basis of conditional inference trees (CTree, [Hothorn, Hornik and Zeileis \(2006\)](#)). Alternatively, one could use asymptotic M-fluctuation tests for parameter instability ([Zeileis and Hornik \(2007\)](#)) as in model-based recursive partitioning (MOB, [Zeileis, Hothorn and Hornik \(2008\)](#)). More details are provided in the [Appendix](#).

For obtaining probabilistic predictions from the tree for a (possibly new) set of covariates  $\mathbf{z} = (z_1, \dots, z_m)$ , the observation simply has to be “sent down” the tree and the corresponding segment-specific MLE has to be obtained. Thus, in practice  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  does not have to be recalculated for each new  $\mathbf{z}$  but one can simply

extract the parameter estimates for the corresponding segment which have been computed already while learning the tree. However, to understand this estimator conceptually it is useful to denote it as a weighted MLE where the weights select those observations from the learning sample that fall into the same segment:

$$(2.6) \quad w_i^{\text{tree}}(\mathbf{z}) = \sum_{b=1}^B \mathbf{1}((z_i \in \mathcal{B}_b) \wedge (\mathbf{z} \in \mathcal{B}_b)),$$

where  $\mathbf{1}(\cdot)$  is the indicator function. The predicted distribution for a given  $\mathbf{z}$  is then fully specified by the estimated parameter  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  where

$$(2.7) \quad \hat{\boldsymbol{\theta}}(\mathbf{z}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n w_i^{\text{tree}}(\mathbf{z}) \cdot \ell(\boldsymbol{\theta}; y_i).$$

2.3. *Distributional forest.* While the simple recursive structure of a tree model is easy to visualize and interpret, the abrupt changes are often too rough, instable, and impose steps on the model even if the true underlying effect is smooth. Hence, ensemble methods such as bagging or random forests (Breiman (2001)) are typically applied to smooth the effects, stabilize the model and improve predictive performance.

The idea of random forests is to learn an ensemble of trees, each on a different learning data obtained through resampling (bootstrap or subsampling). In each node only a random subset of the covariates  $\mathbf{Z}$  is considered for splitting to reduce the correlation among the trees and to stabilize the variance of the model. For a simple regression random forest the mean of predictions over all trees is considered. In that way changes in the location of the response across the covariates are detected (e.g., in Breiman and Cutler’s random forests, Breiman (2001)). This idea is now taken one step further by embedding it in a distributional framework based on maximum-likelihood estimation. *Distributional forests* employ an ensemble of  $T$  *distributional trees* which pick up changes in the “direction” of any distribution parameter by considering the full likelihood and corresponding score function for choosing splitting variables and split points.

To obtain probabilistic predictions from a distributional forest, it still needs to be specified how to compute the parameter estimates  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  for a (potentially new) set of covariates  $\mathbf{z}$ . Following Hothorn and Zeileis (2017) we interpret random forests as adaptive local likelihood estimators using the averaged “nearest neighbor weights” (Lin and Jeon (2006)) from the  $T$  trees in the forest

$$(2.8) \quad w_i^{\text{forest}}(\mathbf{z}) = \frac{1}{T} \sum_{t=1}^T \sum_{b=1}^{B^t} \frac{\mathbf{1}((z_i \in \mathcal{B}_b^t) \wedge (\mathbf{z} \in \mathcal{B}_b^t))}{|\mathcal{B}_b^t|},$$

where  $|\mathcal{B}_b^t|$  denotes the number of observations in the  $b$ -th segment of the  $t$ -th tree. Thus, these  $w_i^{\text{forest}}(\mathbf{z}) \in [0, 1]$  whereas  $w_i^{\text{tree}}(\mathbf{z}) \in \{0, 1\}$ . Hence, weights cannot

only be 0 or 1 but change more smoothly, giving high weight to those observations  $i$  from the learning sample that co-occur in the same segment  $\mathcal{B}_b^t$  as the new observation  $\mathbf{z}$  for many of the trees  $t = 1, \dots, T$ . Consequently, the parameter estimates may, in principle, change for every observation and can be obtained by

$$(2.9) \quad \hat{\boldsymbol{\theta}}(\mathbf{z}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n w_i^{\text{forest}}(\mathbf{z}) \cdot \ell(\boldsymbol{\theta}; y_i).$$

In summary, this yields a parametric distributional regression model (through the score-based approach) that can capture both abrupt effects and high-order interactions (through the trees) and smooth effects (through the forest).

Distributional forests share some concepts and algorithmic aspects with other generalizations of Breiman and Cutler's random forests. Nearest neighbor weights are employed for aggregation in survival forests (Hothorn et al. (2004)), quantile regression forests (Meinshausen (2006)), transformation forests (Hothorn and Zeileis (2017)) and generalized random forests for causal inferences (Athey, Tibshirani and Wager (2019)). These procedures aggregate over trees fitted to specific score functions (e.g., log rank scores in survival trees, model residuals in transformation or generalized forests). Distributional forests, in contrast to nonparametric approaches, provide a compromise between model flexibility and interpretability: The parameters of a problem-specific distribution (zero-censored Gaussian for precipitation) have a clear meaning but may depend on external variables in a quite general way.

**3. Probabilistic precipitation forecasting in complex terrain.** Many statistical weather forecasting models leverage the strengths of modern numerical ensemble prediction systems (EPSs; see Bauer, Thorpe and Brunet (2015)). EPSs not only predict the most likely future state of the atmosphere but provide information about the uncertainty for a specific quantity and weather situation. This is done by running the NWP model several times using slightly perturbed initial conditions and model specifications to account for uncertainties in both, the initial atmospheric state and the NWP model (and its parametrizations). One frequently-used method based on distributional regression models is the ensemble model output statistics (EMOS) approach first proposed by Gneiting et al. (2005) to produce high-quality forecasts for specific quantities and sites. In case of precipitation forecasting, EMOS typically uses the ensemble mean of "total precipitation" ( $tp$ ) forecasts as the predictor for the location parameter  $\mu$  and the corresponding ensemble standard deviation for the scale parameter  $\sigma$ , for example, assuming the observations to follow a zero-censored Gaussian distribution. This distributional approach of modeling both parameters allows to correct for possible errors of the NWP ensemble in both, the expectation but also the uncertainty of a specific forecast. Thus, a basic EMOS specification typically models the two distribution parameters by two linear predictors, for example,  $\mu = \beta_0 + \beta_1 \cdot \text{mean}(tp)$  and

$\log(\sigma) = \gamma_0 + \gamma_1 \cdot \log(\text{sd}(tp))$  with regression coefficients  $\beta_0$ ,  $\beta_1$ ,  $\gamma_0$  and  $\gamma_1$  (where the log link assures positivity of the scale parameter, following Gebetsberger et al. (2017)).

While this approach alone is already highly effective in the plains, it typically does not perform as well in complex terrain due to unresolved effects in the NWP system (Bauer, Thorpe and Brunet (2015)). For example, in the Tyrolean Alps—considered in the following case study—the NWP grid cells of  $50 \times 50 \text{ km}^2$  are too coarse to capture single mountains, narrow valleys, etc. Therefore, it is often possible to substantially improve the predictive performance of a basic EMOS by including additional predictor variables, either from local meteorological observations or an NWP model. Unfortunately, it is typically unknown which variables are relevant for improving the predictions. Simply including all available variables may be computationally burdensome and can lead to overfitting but, on the other hand, excluding too many variables may result in a loss of valuable information. Therefore, selecting the relevant variables and interactions among all possible covariates is crucial for improving the statistical forecasting model.

In the following, it is illustrated how distributional forests can solve this problem without requiring prior expert knowledge about the meteorological covariates. For fitting the forest only the response distribution and the list of potential predictor variables need to be specified (along with a few algorithmic details) and then the relevant variables, interactions and potentially nonlinear effects are determined automatically in a data-driven way. Here, we employ a zero-censored Gaussian distribution and 80 predictor variables computed from ensemble means and spreads of various NWP outputs. The predictive performance of the forest is compared to three other zero-censored Gaussian models: (a) a basic EMOS, (b) a GAMLSS with prespecified effects and interactions based on meteorological knowledge/experience, and (c) a boosted GAMLSS with automatic selection of smooth additive terms based on all 80 predictor variables.

3.1. *Data.* Learning and validation data consist of observed daily precipitation sums provided by the National Hydrographical Service (BMLFUW (2016)) and numerical weather forecasts from the U.S. National Oceanic and Atmospheric Administration (NOAA). Both observations and forecasts are available for 1985–2012 and the analysis is exemplified using July, the month with the most precipitation in Tyrol.

Observations are obtained for 95 stations all over Tyrol and surroundings, providing 24-hour precipitation sums measured at 0600 UTC and rigorously quality-checked by the National Hydrographical Service. NWP outputs are obtained from the second-generation reforecast data set of the global ensemble forecast system (GEFS, Hamill et al. (2013)). This data set consists of an 11-member ensemble based on a fixed version of the numerical model and a horizontal grid spacing of about  $50 \times 50 \text{ km}^2$  initialized daily at 0000 UTC from December 1984 to present

providing forecasts on a 6-hourly temporal resolution. Each of the 11 ensemble members uses slightly different perturbed initial conditions to predict the situation-specific uncertainty of the atmospheric state.

From the GEFS, 14 basic forecast variables are considered with up to 12 variations each such as mean/maximum/minimum over different aggregation time periods. A detailed overview is provided in Table 1, yielding 80 predictor variables in total.

TABLE 1

*Basic covariates together with the number (#) and the type of variations. Time periods indicate aggregation time periods in hours after NWP model initialization (e.g., 6–30 corresponds to +6 h to +30 h ahead forecasts, 0600 UTC to 0600 UTC of the next day)*

Basic covariates	#	Variations
<i>tp</i> : total precipitation, power transformed (by $1.6^{-1}$ ) <i>cape</i> : convective available potential energy, power transformed (by $1.6^{-1}$ )	12	ensemble mean of sums over 24h, ensemble std. deviation of sums over 24h, ensemble minimum of sums over 24h, ensemble maximum of sums over 24h all for 6–30 ensemble mean of sums over 6h for 6–12, 12–18, 18–24, 24–30 ensemble std. deviation of sums over 6h for 6–12, 12–18, 18–24, 24–30
<i>dswrf</i> : downwards short wave radiation flux (“sunshine”) <i>msl</i> : mean sea level pressure <i>pwat</i> : precipitable water <i>tmax</i> : 2m maximum temperature <i>tcollc</i> : total column-integrated condensate <i>t500</i> : temperature on 500 hPa <i>t700</i> : temperature on 700 hPa <i>t850</i> : temperature on 850 hPa	6	ensemble mean of mean values, ensemble mean of minimum values*, ensemble mean of maximal values, ensemble std. deviation of mean values, ensemble std. deviation of minimum values*, ensemble std. deviation of maximal values, all over 6–30
<i>tdiff500850</i> : temperature difference 500 to 850 hPa <i>tdiff500700</i> : temperature difference 500 to 700 hPa <i>tdiff700850</i> : temperature difference 700 to 850 hPa	3	ensemble mean of difference in mean, ensemble minimum of difference in mean, ensemble maximum of difference in mean all over 6–30
<i>msl_diff</i> : mean sea level pressure difference	1	<i>msl_mean_max</i> – <i>msl_mean_min</i> over 6–30

Note: \*Minimum values of *dswrf* over 24 h are always zero and thus neglected.

To remove large parts of the skewness of precipitation data, a power transformation (Box and Cox (1964)) is often applied, for example, using cubic (Stidd (1973)) or square root (Hutchinson (1998)) transformations. However, the power parameter may vary for different climatic zones or temporal aggregation periods and hence we follow Stauffer et al. (2017b) in their choice of  $1.6^{-1}$  as a suitable power parameter for precipitation in the region of Tyrol. The same power transformation is applied to both the observed precipitation sums and the NWP outputs “total precipitation” (*tp*) and “convective available potential energy” (*cape*).

3.2. *Models and evaluation.* The following zero-censored Gaussian regression models are employed in the empirical case study, see Table 2 for further details:

- *Distributional forest:* All 80 predictor variables are considered for learning a forest of 100 trees. Subsampling is employed for each tree using a third of the predictors in each split of the tree (argument `mtry` in our implementation `distforest`, with more “computational details” provided at the end of the manuscript). Parameters are estimated by adaptive local likelihood based on the forest weights as described in Section 2. The stopping criteria are the minimal number of observations to perform a split (`minsplit` = 50), the minimal number of observations in a segment (`minbucket` = 20) and the significance level

TABLE 2  
*Overview of models with type of covariate dependency and included covariates for each distribution parameter. A \* B indicates an interaction between covariate A and B*

Model	Type	Location ( $\mu$ )	Scale ( $\log(\sigma)$ )
Distributional forest	recursive partitioning	all	all
EMOS	linear	<i>tp_mean</i>	<i>tp_sprd</i>
Prespecified GAMLSS	spline in each	<i>tp_mean</i> , <i>tp_max</i> , <i>tp_mean1218</i> * <i>cape_mean1218</i> , <i>dswrf_mean_mean</i> , <i>tolc_mean_mean</i> , <i>pwat_mean_mean</i> , <i>tdiff500850_mean</i> , <i>mssl_diff</i>	<i>tp_sprd</i> , <i>dswrf_sprd_mean</i> , <i>tp_sprd1218</i> * <i>cape_mean1218</i> , <i>tolc_sprd_mean</i> , <i>tdiff500850_mean</i>
Boosted GAMLSS	spline in each	all	all

for variable selection ( $\alpha = 1$ ). The latter means that no early stopping (or “prepruning”) is applied based on results of the statistical tests.

- *EMOS*: The basic ensemble model output statistics models use the ensemble mean of total precipitation as regressor in the location submodel and the corresponding ensemble standard deviation in the scale submodel. The parameters are estimated by maximum likelihood, using an identity link for the location part and a log link for the scale part (following the advice of Gebetsberger et al. (2017)).
- *Prespecified GAMLSS*: Smooth additive splines are selected for the most relevant predictors based on meteorological expert knowledge following Stauffer et al. (2017a). More specifically, based on the 80 available variables, eight terms are included in the location submodel and five in the scale submodel. Both involve an interaction of *tp* and *cape* in the afternoon (between 1200 UTC and 1800 UTC) to capture the potential for thunderstorms that frequently occur in summer afternoons in the Alps. The model is estimated by maximum penalized likelihood using a backfitting algorithm (Stasinopoulos and Rigby (2007)).
- *Boosted GAMLSS*: Smooth additive splines are selected automatically from all 80 available variables, using noncyclic boosting for parameter estimation (Hofner, Mayr and Schmid (2016), Messner, Mayr and Zeileis (2017)). This updates the predictor terms for the location or scale submodels iteratively by maximizing the log-likelihood only for the variable yielding the biggest improvement. The iteration stops early—before fully maximizing the in-sample likelihood—based on a (computationally intensive) out-of-bag bootstrap estimate of the log-likelihood. The grid considered for the number of boosting iterations ( $m_{\text{stop}}$ ) is: 50, 75, . . . , 975, 1000.

The predictive performance in terms of full probabilistic forecasts is assessed using the continuous ranked probability score (CRPS, Hersbach (2000)). For each of the models this assesses the discrepancy of the predicted distribution function  $F$  from the observation  $y$  by

$$(3.1) \quad \text{CRPS}(y, F) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}(y \leq z))^2 dz$$

where  $\mathbf{1}(\cdot)$  is the indicator function. In the subsequent applications, the mean CRPS is always evaluated out of sample, either using cross-validation or a hold-out data set (2009–2012) that was not used for learning (1985–2008). CRPS is a proper scoring rule (Gneiting and Raftery (2007)) often used within the meteorological community. Lower values indicate better performance.

To assess differences in the improvement of the forests and GAMLSS models over the basic EMOS, a CRPS-based skill score with EMOS as the reference method is computed:

$$(3.2) \quad \text{CRPSS}_{\text{method}} = 1 - \frac{\text{CRPS}_{\text{method}}}{\text{CRPS}_{\text{EMOS}}}.$$

3.3. *Application for one station.* In a first step, we show a detailed comparison of the competing models for one observation site, Axams in Tyrol (in the center of the study area, see Figure 8). As for all other stations, daily precipitation observations and numerical weather predictions are available for the month of July from 1985 through 2012. In Figure 3 in the introduction the probabilistic forecasts from the distributional forest, trained on 1985–2008, for July 24 in 2009–2012 have already been shown as a motivational example. In particular, the figure depicts the forecasted point mass at zero (i.e., the probability of a dry day) along with the forecasted probability density function for the total amount of precipitation. Based on this illustration it can be observed that the four sample forecasts differ considerably in location  $\mu$ , scale  $\sigma$ , and the amount of censoring while conforming quite well with the actual observations from these days. While this is a nice illustrative example we are interested in the overall predictive performance and calibration of the distributional fits. More details of this assessment as well as an application to 14 further meteorological stations is provided in Supplement B (Schlosser et al. (2019b)).

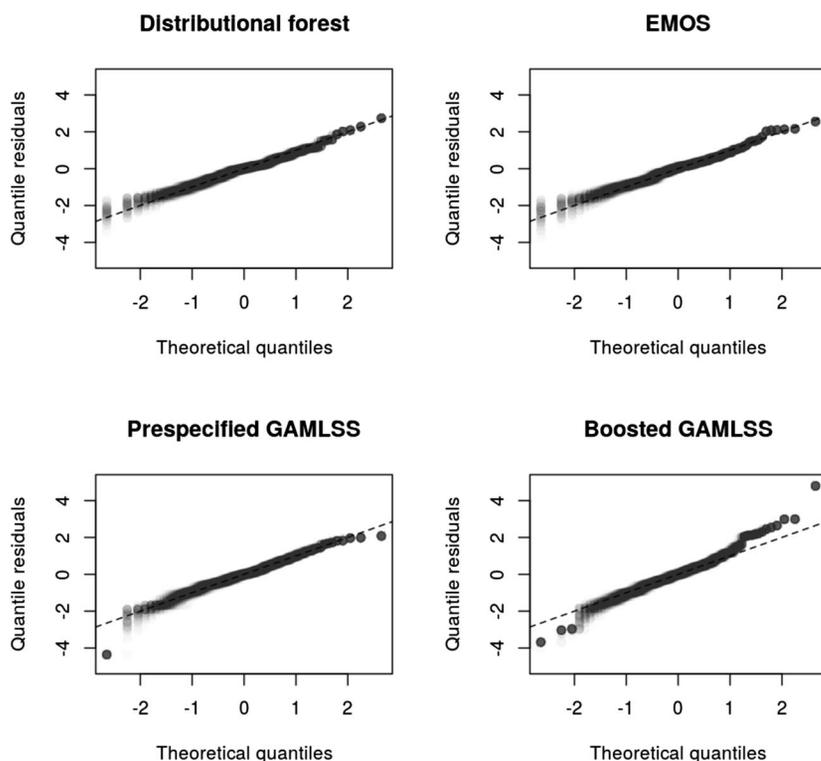


FIG. 4. *Out-of-sample residual QQ plots (2009–2012) for station Axams based on models learned on data from 1985–2008.*

To assess calibration, Figure 4 shows residual QQ plots for out-of-sample predictions (2009–2012) from the different models trained on 1985–2008. Due to the point masses at zero 100 draws from the randomized quantile residuals (Dunn and Smyth (1996)) are plotted in semi-transparent gray. Overall, the randomized quantile residuals conform quite well with the theoretical standard normal quantile (i.e., form a straight line close to the diagonal), indicating that all four models are sufficiently well calibrated. This is also supported by the corresponding probability integral transform (PIT, Gneiting, Balabdaoui and Raftery (2007)) histograms for station Axams in Supplement B (Schlosser et al. (2019b)) which contains a more detailed explanation of residual QQ plots and PIT histograms.

To assess the predictive performance, a full cross-validation is carried out rather than relying on just the one fixed test set for the years 2009–2012. To do so, a 10 times 7-fold cross-validation is carried out where each repetition splits the available 28 years into 7 subsets of 4 randomly-selected (and thus not necessarily consecutive) years. The models are learned on 6 folds (= 24 years) and evaluated on the 7th fold (= 4 years) using the average CRPS across all observations. The resulting 10 CRPS skill scores are displayed by boxplots in Figure 5 using EMOS as the reference model (horizontal line at a CRPSS of 0). Both GAMLSS models and the distributional forest perform distinctly better than the EMOS model. While the two GAMLSS lead to an improvement of around 4 percent, the distributional forest has a slightly higher improvement of around 5.5 percent in median.

Finally, it is of interest how this improvement in predictive performance by the distributional forest is accomplished, that is, which of the 80 covariates are selected in the trees of the forest. As the 100 trees of the forest do not allow to simply assess the variables' role graphically, a common solution for random forests in

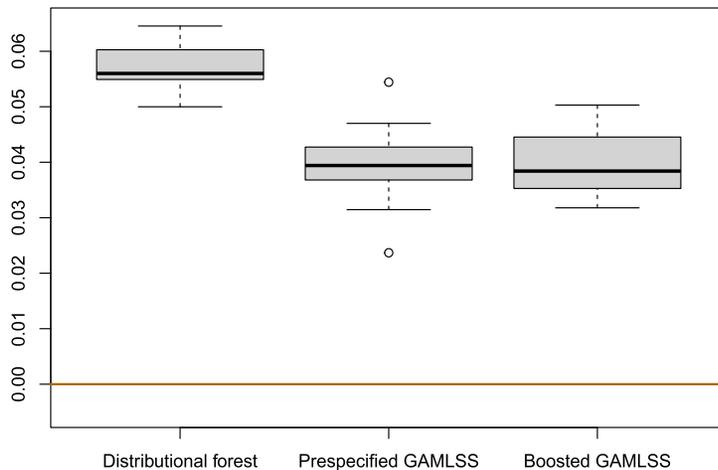


FIG. 5. CRPS skill score from the 10 times 7-fold cross-validation at station Axams (1985–2012). The horizontal orange line pertains to the reference model EMOS.

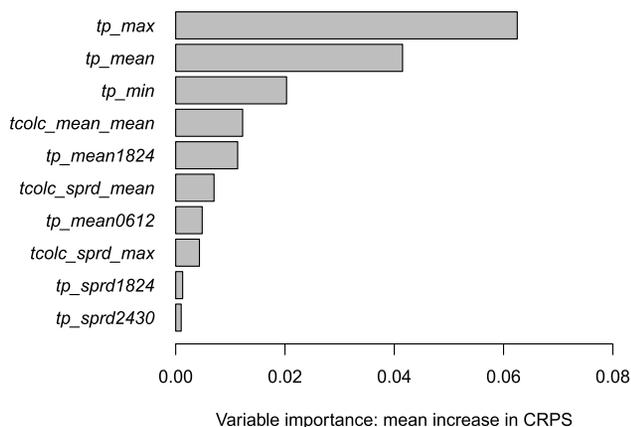


FIG. 6. CRPS-based variable importance for the top 10 covariates in the distributional forest. Based on data for station Axams, learning period 1985–2008 and assessed in 2009–2012.

general is to consider variable importance measures. Here, this is defined as the amount of change in CRPS when the association between one covariate and the response variable is artificially broken through permutation (and thus also breaking the association to the remaining covariates).

Figure 6 shows the 10 covariates with the highest permutation importance (i.e., change in CRPS) for station Axams. As expected the NWP outputs for total precipitation (*tp*) are particularly important along with total column-integrated condensate (*tcolc*). Also, both variables occur in various transformations such as means (either of the full day or certain parts of the afternoon), spreads or minima/maxima. Thus, while the covariates themselves are not surprising, selecting a GAMLSS with a particular combination of all the transformations would be much more challenging.

**3.4. Application for all stations.** After considering only one observational site up to now, a second step evaluates and compares the competing methods on all 95 available stations. As in the previous section, all models are learned on the first 24 years and evaluated by the average CRPS on the last 4 years. More specifically, the CRPS skill score against the EMOS model is computed for the out-of-sample predictions at each station and visualized by parallel coordinates plots with boxplots superimposed in Figure 7. Overall, distributional forests have a slightly higher improvement in CRPSS compared to the two GAMLSS which is best seen by looking at the boxplots and the green line representing the results for station Axams. The underlying parallel coordinates additionally bring out that the prespecified GAMLSS sometimes performs rather differently (sometimes better, sometimes worse) compared to the two data-driven models. Values below zero show that, for some stations, EMOS performs better than the more complex statistical methods.

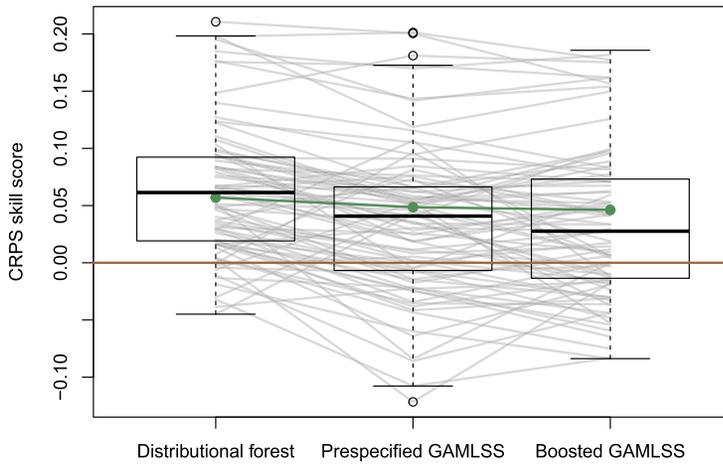


FIG. 7. CRPS skill score for each station (gray lines with boxplots superimposed). Station Axams is highlighted in green and the horizontal orange line pertains to the reference model EMOS. The models are learned on 1985–2008 and validated for 2009–2012.

To assess whether these differences in predictive performance are due to differences in the topography, Figure 8 shows a brief spatial summary of all stations. Each station is illustrated by a symbol that conveys which model performed best in terms of CRPS on the last four years of the data. Additionally, the color of the symbol indicates the CRPS difference between distributional forest and the best-performing other model. Green signals that the distributional forest performs better than the other models whereas red signals that another model performs better. Overall the distributional forest performs on par (gray) or better (green) for the majority of stations. Only for a few stations in the north-east EMOS performs best, and in East Tyrol the prespecified GAMLSS performs particularly well in the validation period (2009–2012). Partially, this can be attributed to random variation as the differences at several stations are mitigated when considering a full cross-validation rather than a single split into learning and validation period (see Supplement B, Schlosser et al. (2019b) and the corresponding discussion in the next section). Further differences are possibly due to East Tyrol lying in a different climate zone, south of the main Alpine Ridge. Hence, long-term climatological characteristics as well as the precipitation patterns in 2009–2012 differ from North Tyrol, conforming particularly well with the additive effects from the prespecified GAMLSS.

**4. Discussion.** Distributional regression modeling is combined with tree-based modeling to obtain a novel and flexible method for probabilistic forecasting. The resulting distributional trees and forests can capture abrupt and nonlinear effects and interactions in a data-driven way. By basing the split point and split variable selection on a full likelihood and corresponding score function, the trees

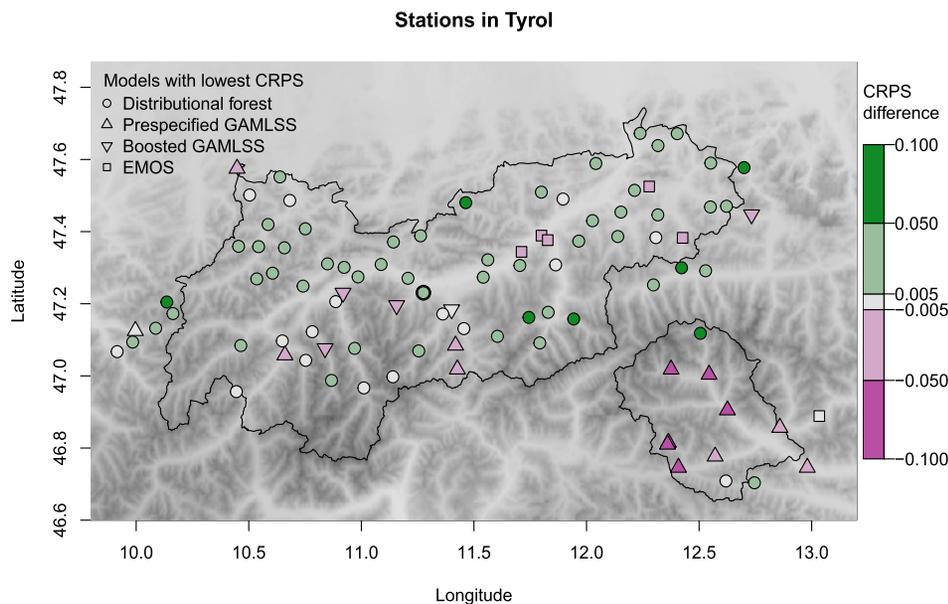


FIG. 8. Map of Tyrol coding the best-performing model for each station (type of symbol) when learned on 1985–2008 and validated for 2009–2012. The color codes whether the distributional forest had higher (green) or lower (red) CRPS compared to the best of the other three models. The gray background shows the local topography (Robinson, Regetz and Guralnick (2014)). Station Axams is highlighted in bold.

and forests can not only pick up changes in the location but also the scale or shape of any distributional family.

Distributional forests are an attractive alternative when prespecifying or boosting all possible effects and interactions in a GAMLSS model is challenging. Distributional forests are rather straightforward to specify requiring only little prior subject matter knowledge and also work well in the presence of many potential covariates. The application to precipitation forecasting in complex terrain illustrates that distributional forests often perform on par or even better than their GAMLSS counterparts. Hence, they form a useful addition to the already available toolbox of probabilistic forecasts for disciplines such as meteorology.

*Variable selection.* Generally, there are many possibilities how to specify the variables that are to be included in a distributional regression model. Especially for a low number of covariates, the GAMLSS approach offers a powerful framework in which penalized estimation of both smooth main effects and corresponding interaction surfaces yields models that often balance good predictive performance with high interpretability (see e.g., Wood, Scheipl and Faraway (2013); Goicoa et al. (2018); Ugarte, Adin and Goicoa (2017b)). However, if the number of covariates is high, including all (or many) main effects and interactions in a GAMLSS

typically becomes challenging both in terms of interpretability and computational complexity/stability (see also Hofner, Mayr and Schmid (2016)).

In the precipitation forecasting application, as presented in Section 3, 80 covariates are considered which corresponds to 3160 potential pairwise interactions (and even more higher-order interactions). Therefore, only main effects are considered for the boosted GAMLSS while the prespecified GAMLSS also includes selected interactions chosen based on meteorological expert knowledge. In contrast, the distributional forest requires no prespecification as covariates and corresponding interactions are selected automatically. Thus, distributional forests are an appealing alternative to (boosted) GAMLSS in weather forecasting tasks as the main concern is typically not so much interpretability but forecasting skill and (semi-)automatic application on a larger domain (see also the discussion in Rasp and Lerch (2018)).

*Distributional specifications for precipitation modeling.* Choosing an adequate distributional family is an important step for establishing a well-fitting model. A zero-censored Gaussian distribution is employed in this manuscript as this has been found to be an appropriate choice for precipitation modeling in earlier literature (e.g., Stauffer et al. (2017b)). To test for robustness against distributional misspecification, two alternative distributional specifications have been considered in Supplement A (Schlosser et al. (2019a)): Using the same evaluations as in Section 3.4, all models are additionally fitted for 15 meteorological stations using a zero-censored logistic distribution in order to account for heavier tails and a two-part Gaussian hurdle model combining a binary model for zero vs. positive precipitation and a separate Gaussian model, truncated at zero, for the positive precipitation observations. Both specifications yield qualitatively similar results as for the zero-censored Gaussian distribution. For some stations the two-part hurdle model leads to small improvements, however at the expense of increased variability across stations (especially for EMOS and the boosted GAMLSS). Overall, the results from this manuscript are quite robust across these distributional specifications, especially for the distributional forests.

Moreover, one could consider a distribution including an additional parameter for capturing skewness (as in Scheuerer and Hamill (2015), Baran and Nemoda (2016)). However, this would go beyond the mean/variance specification of the NGR that is widely used in ensemble post-processing. Therefore, this contribution investigates the effects of using the same distributional family with a novel strategy for specifying dependence on covariates.

*More general distributional specifications.* Beyond the task of modeling precipitation it is of interest how well distributional forests perform in combination with other more general distributional specifications. It has been shown previously in the literature that using a score- or gradient-based selection of splitting variables outperforms a mean-based selection with subsequent flexible distributional modeling: For example, both Athey, Tibshirani and Wager (2019, Figure 2) and

Hothorn and Zeileis (2017, Figure 1) demonstrated (independently) that their respective score-based random forest algorithms outperform the mean-based quantile regression forests of Meinshausen (2006) in a setup where only the variance of a normal response variable changes across the considered covariates. However, if all distribution parameters are closely correlated with the distribution mean the forests with different splitting strategies all perform similarly, provided a sufficiently flexible distribution is employed for the final predictions (see Hothorn and Zeileis (2017), Section 7).

Similarly, the score-based distributional forests introduced in this manuscript proved to be quite robust to the different distributional specifications considered. While all specifications focus on capturing mean-variance effects note that these parameters are never fully orthogonal but can actually become quite closely correlated due to the censoring (or truncation and/or zero-inflation considered in Supplement A (Schlosser et al. (2019a))).

However, exploring extensions to more flexible parametric distributions (e.g., such as the Dagum distribution considered by Klein et al. (2015) in GAMLSS-type models) as well as transformation model specifications (e.g., as in Hothorn and Zeileis (2017)) are of interest for future research.

*Axams vs. other meteorological stations.* Axams was chosen as the meteorological station for the more extensive evaluations in Section 3.3 as it yields fairly typical results and is geographically in the center of the study area and closest to Innsbruck, the capital of Tyrol and the work place of three of the authors. To show that qualitatively similar results are obtained for other meteorological stations, Supplement B (Schlosser et al. (2019b)) carries out the same evaluation for 14 further stations. These cover a wide range of geographical locations/altitudes and a mix of different best-performing models in the single-split setting reported in Section 3.4.

The supplement shows that some of the differences in forecast skill from Figure 8 even out in the cross-validation with distributional forests typically performing at least as well as the best of the other models at most stations. In particular, this also includes three stations in East Tyrol where the prespecified GAMLSS performs best in the single-split setting (learning based on 1985–2008 and validation for 2009–2012).

*Tuning parameters.* Selecting tuning parameters for flexible regression models is important not only in terms of predictive accuracy but also computational complexity. For the application in Section 3 tuning parameters are selected based on advice from the literature as well as our own experiences. As Hastie, Tibshirani and Friedman (2001) and Breiman (2001) recommend to build full-grown trees, early stopping upon nonsignificance is disabled (`alpha = 1`) and low values are used for `minsplit` (= 50) and `minbucket` (= 20), while assuring that

`minsplit` is sufficiently large for reasonably obtaining MLEs of all parameters in each segment of the tree.

Applying the Law of Large Numbers it can be shown that random forests do not overfit as the number of trees increases (Biau and Scornet (2016), Breiman (2001), Hastie, Tibshirani and Friedman (2001)). Therefore, in principle, forests can be built with a very large number of trees (`nntree`) as this cannot deteriorate the predictions. However, “[...] the computational cost for inducing a forest increases linearly with the number of trees, so a good choice results from a trade-off between computational complexity and accuracy” (Biau and Scornet (2016, p. 205)). Following this advice, we decided to build forests consisting of 100 trees.

*Computational difficulties.* As stated by Hofner, Mayr and Schmid (2016) the AIC-based variable selection methods implemented in the R package *gamlss* “[...] can be unstable, especially when it comes to selecting possibly different sets of variables for multiple distribution parameters.” We have noticed computational problems when applying *gamlss* in certain settings within the cross-validation framework as it did not succeed in fitting the model. In these cases the prespecified GAMLSS was not taken into consideration in the comparison of all applied models.

*Computational details.* The proposed methods are implemented in the R package *disttree* (version 0.1.0) based on the *partykit* package (version 1.2.3), both available on R-Forge at (<https://R-Forge.R-project.org/projects/partykit/>). The function `distforest` learns the distributional forests proposed in this manuscript by combining the general `cforest` function from *partykit* with the function `distfit` for fitting distributional models by maximum likelihood. Analogously, *disttree* can learn a single distributional tree by combining `ctree` with `distfit`. All functions can either be used with GAMLSS family objects from the R package *gamlss.dist* (Stasinopoulos and Rigby (2007), version 5.0.6) or with custom lists containing all required information about the distribution family.

In addition to *disttree*, Section 3 employs R package *crch* (Messner, Mayr and Zeileis (2016), version 1.0.1) for the EMOS models, *gamlss* (Stasinopoulos and Rigby (2007), version 5.1.0) for the prespecified GAMLSS and *gamboostLSS* (Hofner, Mayr and Schmid (2016), version 2.0.1) for the boosted GAMLSS.

The fitted distributional forest for July 24 and observation station Axams (including Figure 3) is reproducible using `demo("RainAxams", package = "disttree")`. This also includes fitting the other zero-censored Gaussian models considered in this paper and generating the corresponding QQ plots (Figure 4) and PIT histograms (Schlosser et al. (2019b), Supplement B). Full replication of all results can be obtained with `demo("RainTyrol", package = "disttree")` requiring the companion R package *RainTyrol* (version 0.1.0), also available within the R-Forge project. The results presented

in Supplement A (Schlosser et al. (2019a)) and Supplement B (Schlosser et al. (2019b)) can be reproduced using `demo("RainDistributions", package = "disttree")` and `demo("RainStationwise", package = "disttree")`, respectively.

APPENDIX: TREE ALGORITHM

In the following, the tree algorithm applied in the empirical case study discussed in this paper is explained. For notational simplicity, the testing and splitting procedure is described for the root node, that is, the entire learning sample with observations  $\{y_i\}_{i=1,\dots,n}$ ,  $n \in \mathbb{N}$ . In each child node the corresponding subsample depends on the foregoing split(s).

After fitting a distributional model  $\mathcal{D}(Y, \theta)$  to the learning sample with observations  $\{y_i\}_{i=1,\dots,n}$  as explained in Section 2.1 the resulting estimated parameter  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ ,  $k \in \mathbb{N}$  can be plugged in the score function  $s(\theta, Y)$ . In that way a goodness-of-fit measurement is obtained for each parameter  $\theta_j$  and each observation  $y_i$ . To use this information, statistical tests are employed to detect dependencies between the score values

$$(A.1) \quad s(\hat{\theta}, y) = \begin{pmatrix} s(\hat{\theta}, y_1)_1 & s(\hat{\theta}, y_1)_2 & \dots & s(\hat{\theta}, y_1)_k \\ \vdots & \vdots & \ddots & \vdots \\ s(\hat{\theta}, y_n)_1 & s(\hat{\theta}, y_n)_2 & \dots & s(\hat{\theta}, y_n)_k \end{pmatrix}$$

and each variable  $Z_l \in \{Z_1, \dots, Z_m\}$ . More formally, the following hypotheses are assessed with permutation tests:

$$(A.2) \quad H_0^l : s(\hat{\theta}, Y) \perp Z_l.$$

The permutation tests are based on the multivariate linear statistic

$$(A.3) \quad T_l = \text{vec} \left( \sum_{i=1}^n v_l(Z_{li}) \cdot s(\hat{\theta}, Y_i) \right),$$

where  $s(\hat{\theta}, Y_i) \in \mathbb{R}^{1 \times k}$  and the type of the transformation function  $v_l$  depends on the type of the split variable  $Z_l$ . If  $Z_l$  is numeric then  $v_l$  is simply the identity function  $v_l(Z_{li}) = Z_{li}$  and therefore  $T_l \in \mathbb{R}^k$  as the “vec” operator converts the  $1 \times k$  matrix into a  $k$  column vector. If  $Z_l$  is a categorical variable with  $H$  categories then  $v_l(Z_{li}) = (\mathbf{I}(Z_{li} = 1), \dots, \mathbf{I}(Z_{li} = H))$  such that  $v_l$  is a  $H$ -dimensional unit vector where the element corresponding to the value of  $Z_{li}$  is 1. In this case the statistic  $T_l \in \mathbb{R}^{H \cdot k}$  as the “vec” operator converts the  $H \times k$  matrix into a  $H \cdot k$  column vector by column-wise combination. Observations with missing values are excluded from the sums.

With the conditional expectation  $\mu_l$  and the covariance  $\Sigma_l$  of  $T_l$  as derived by Strasser and Weber (1999) the test statistic can be standardized. The observed multivariate linear statistic  $t_l$  which is either a  $k$ - or  $k \cdot H$ -dimensional vector, depending on the scale of  $Z_l$ , is mapped onto the real line by a univariate test statistic  $c$ .

In the application of this paper a quadratic form is chosen, such that

$$(A.4) \quad c_{\text{quad}}(t_l, \mu_l, \Sigma_l) = (t_l - \mu_l) \Sigma_l^+ (t_l - \mu_l)^\top$$

where  $\Sigma_l^+$  is the Moore–Penrose inverse of  $\Sigma_l$ . Alternatively, the maximum of the absolute values of the standardized linear statistic can be considered ( $c_{\text{max}}$ ).

Strasser and Weber (1999) showed that the asymptotic conditional distribution of the linear statistic  $t_l$  is a multivariate normal with parameters  $\mu_l$  and  $\Sigma_l$ . Hence, the asymptotic conditional distribution of  $c(t_l, \mu_l, \Sigma_l)$  is either normal (for  $c_{\text{max}}$ ) or  $\chi^2$  (for  $c_{\text{quad}}$ ).

The smaller the  $p$ -value corresponding to the standardized test statistic  $c(t_l, \mu_l, \Sigma_l)$  is the stronger the discrepancy from the assumption of independence between the scores and the split variable  $Z_l$ . After Bonferroni-adjusting the  $p$ -values it has to be assessed whether any of the resulting  $p$ -values are beneath the selected significance level. If so, the partitioning variable  $Z_{l^*}$  with the lowest  $p$ -value is chosen as splitting variable. Otherwise no further split is made in this node as the stopping criterion of no  $p$ -values being below the significance level is fulfilled. This type of early-stopping in building a tree is sometimes also referred to as “prepruning”. For random forests prepruning is often switched off by setting the significance level to 1.

The breakpoint that leads to the highest discrepancy between score functions in the two resulting subgroups is selected as split point. This is measured by the linear statistic

$$(A.5) \quad T_{l^*}^{qr} = \sum_{i \in \mathcal{B}_{qr}} s(\hat{\theta}, Y_i)$$

for  $q \in \{1, 2\}$  where  $\mathcal{B}_{1r}$  and  $\mathcal{B}_{2r}$  are the two new subgroups, without any particular ordering, that are defined by splitting in split point  $r$  of variable  $Z_{l^*}$ . The split point is then chosen as follows:

$$(A.6) \quad r^* = \underset{r}{\operatorname{argmin}} (\min_{q=1,2} (c(t_{l^*}^{qr}, \mu_{l^*}^{qr}, \Sigma_{l^*}^{qr}))).$$

One repeats the testing and splitting procedure in each of the resulting subgroups until some stopping criterion is reached. This criterion can for example be a minimal number of observations in a node or a minimal  $p$ -value for the statistical tests. In that way prepruning is applied in order to find right-sized trees and hence avoid overfitting.

This permutation-test-based tree algorithm is presented in Hothorn, Hornik and Zeileis (2006) as the CTree algorithm. A different framework to build a likelihood-based tree is provided by the MOB algorithm which is based on M-fluctuation tests (Zeileis, Hothorn and Hornik (2008)).

## SUPPLEMENTARY MATERIAL

**Supplement A: Different response distributions** (DOI: [10.1214/19-AOAS1247SUPPA](https://doi.org/10.1214/19-AOAS1247SUPPA); .pdf). To assess the goodness of fit of the Gaussian distribution, left-censored at zero, this supplement employs the same evaluations as in the main manuscript but based on two other distributional assumptions: A logistic distribution, left-censored at zero, is employed to potentially better capture heavy tails—and a two-part hurdle model combining a binary model for zero vs. positive precipitation and a Gaussian model, truncated at zero, for the positive precipitation observations.

**Supplement B: Stationwise evaluation** (DOI: [10.1214/19-AOAS1247SUPPB](https://doi.org/10.1214/19-AOAS1247SUPPB); .pdf). To show that Axams is a fairly typical station and similar insights can be obtained for other stations as well, this supplement presents the same analysis as in Section 3.3 of the main manuscript for 14 further meteorological stations.

## REFERENCES

- ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. [MR3909963](#)
- BARAN, S. and NEMODA, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* **27** 280–292. [MR3521202](#)
- BAUER, P., THORPE, A. and BRUNET, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** (7567) 47–55.
- BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. [MR3493512](#)
- BMLFUW (2016). Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4—Wasserhaushalt. Available at <http://ehyd.gv.at/>. Accessed: 2016–02–29.
- BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. [MR0192611](#)
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. [MR0726392](#)
- DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.
- GEBETSBERGER, M., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Fine-tuning non-homogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Weather Rev.* **145** 4693–4708.
- GLAHN, H. R. and LOWRY, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11** 1203–1211.
- GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. [MR2325275](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GNEITING, T., RAFTERY, A. E., WESTVELD III, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.

- GOICOA, T., ADIN, A., UGARTE, M. D. and HODGES, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch. Environ. Res. Risk Assess.* **32** 749–770.
- HAMILL, T. M., BATES, G. T., WHITAKER, J. S., MURRAY, D. R., FIORINO, M., GALARNEAU JR., T. J., ZHU, Y. and LAPENTA, W. (2013). NOAA’s second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94** 1553–1565.
- HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. [MR0858512](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York. [MR1851606](#)
- HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.
- HOFNER, B., MAYR, A. and SCHMID, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *J. Stat. Softw.* **74** (1) 1–31.
- HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. [MR2291267](#)
- HOTHORN, T. and ZEILEIS, A. (2017). Transformation forests. Available at [arXiv:1701.02110](#).
- HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91.
- HOTHORN, T., HORNIK, K., VAN DE WIEL, M. A. and ZEILEIS, A. (2006). A Lego system for conditional inference. *Amer. Statist.* **60** 257–263. [MR2246759](#)
- HUTCHINSON, M. F. (1998). Interpolation of rainfall data with thin plate smoothing splines—Part II: Analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis* **2** 152–167.
- KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.* **9** 1024–1052. [MR3371346](#)
- LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. [MR2256176](#)
- LONG, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks, CA.
- MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. [MR2274394](#)
- MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2016). Heteroscedastic censored and truncated regression with crch. *The R Journal* **8** (1) 173–181.
- MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Non-homogeneous boosting for predictor selection in ensemble post-processing. *Mon. Weather Rev.* **145** 137–147.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A* **135** 370–384.
- RASP, S. and LERCH, S. (2018). Neural networks for post-processing ensemble weather forecasts. *Mon. Weather Rev.* **146** 3885–3900.
- RIGBY, R. A. and STASINOPOULOS, D. M. (2005a). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. [MR2137253](#)
- ROBINSON, N., REGETZ, J. and GURALNICK, R. P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS J. Photogramm. Remote Sens.* **87** 57–67.
- SCHEUERER, M. and HAMILL, T. M. (2015). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143** 4578–4596.
- SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019a). Different response distributions. Supplement A to “Distributional regression forests for probabilistic precipitation forecasting in complex terrain.” DOI:[10.1214/19-AOAS1247SUPPA](#).

- SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019b). Stationwise evaluation. Supplement B to “Distributional regression forests for probabilistic precipitation forecasting in complex terrain.” DOI:10.1214/19-AOAS1247SUPPB.
- STASINOPOULOS, D. M. and RIGBY, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in *R. J. Stat. Softw.* **23** (7) 1–46.
- STAUFFER, R., UMLAUF, N., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017a). Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Weather Rev.* **45** 955–969.
- STAUFFER, R., MAYR, G. J., MESSNER, J. W., UMLAUF, N. and ZEILEIS, A. (2017b). Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.* **37** 3264–3275.
- STIDD, C. K. (1973). Estimating the precipitation climate. *Water Resour. Res.* **9** 1235–1241.
- STRASSER, H. and WEBER, CH. (1999). The asymptotic theory of permutation statistics. *Math. Methods Statist.* **8** 220–250. Johann Pfanzagl—On the occasion of his 70th birthday. MR1722622
- UGARTE, M. D., ADIN, A. and GOICOA, T. (2017b). One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: Model fitting and model identifiability. *Spat. Stat.* **22** 451–468. MR3732861
- WOOD, S. N., SCHEIPL, F. and FARAWAY, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat. Comput.* **23** 341–360. MR3041440
- ZEILEIS, A. and HORNIK, K. (2007). Generalized  $M$ -fluctuation tests for parameter instability. *Stat. Neerl.* **61** 488–508. MR2351461
- ZEILEIS, A., HOTHORN, T. and HORNIK, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Statist.* **17** 492–514. MR2439970

L. SCHLOSSER  
 R. STAUFFER  
 A. ZEILEIS  
 DEPARTMENT OF STATISTICS  
 FACULTY OF ECONOMICS AND STATISTICS  
 UNIVERSITÄT INNSBRUCK  
 UNIVERSITÄTSSTR. 15  
 6020 INNSBRUCK  
 AUSTRIA  
 E-MAIL: [Lisa.Schlosser@uibk.ac.at](mailto:Lisa.Schlosser@uibk.ac.at)  
[Reto.Stauffer@uibk.ac.at](mailto:Reto.Stauffer@uibk.ac.at)  
[Achim.Zeileis@R-project.org](mailto:Achim.Zeileis@R-project.org)  
 URL: <https://www.uibk.ac.at/statistics/personal/schlosser-lisa/>  
<https://retostauffer.org/>  
<https://eeecon.uibk.ac.at/~zeileis/>

T. HOTHORN  
 INSTITUT FÜR EPIDEMIOLOGIE,  
 BIostatistik und Prävention  
 UNIVERSITÄT ZÜRICH  
 HIRSCHENGRABEN 84  
 CH-8001 ZÜRICH  
 SWITZERLAND  
 E-MAIL: [Torsten.Hothorn@R-project.org](mailto:Torsten.Hothorn@R-project.org)