



## Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood

**Gebetsberger, Manuel; Messner, Jakob; Mayr, Georg J.; Zeileis, Achim**

*Published in:*  
Monthly Weather Review

*Link to article, DOI:*  
[10.1175/MWR-D-17-0364.1](https://doi.org/10.1175/MWR-D-17-0364.1)

*Publication date:*  
2018

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Gebetsberger, M., Messner, J. W., Mayr, G. J., & Zeileis, A. (2018). Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood. *Monthly Weather Review*, 146(12), 4323-4338. DOI: 10.1175/MWR-D-17-0364.1

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Estimation Methods for Nonhomogeneous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood

MANUEL GEBETSBERGER<sup>a</sup>

*Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, and Division for Biomedical Physics, Medical University of Innsbruck, Innsbruck, Austria*

JAKOB W. MESSNER

*Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark, and Department of Statistics, University of Innsbruck, Innsbruck, Austria*

GEORG J. MAYR

*Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria*

ACHIM ZEILEIS

*Department of Statistics, University of Innsbruck, Innsbruck, Austria*


(Manuscript received 1 December 2017, in final form 27 August 2018)

## ABSTRACT

Nonhomogeneous regression models are widely used to statistically postprocess numerical ensemble weather prediction models. Such regression models are capable of forecasting full probability distributions and correcting for ensemble errors in the mean and variance. To estimate the corresponding regression coefficients, minimization of the continuous ranked probability score (CRPS) has widely been used in meteorological post-processing studies and has often been found to yield more calibrated forecasts compared to maximum likelihood estimation. From a theoretical perspective, both estimators are consistent and should lead to similar results, provided the correct distribution assumption about empirical data. Differences between the estimated values indicate a wrong specification of the regression model. This study compares the two estimators for probabilistic temperature forecasting with nonhomogeneous regression, where results show discrepancies for the classical Gaussian assumption. The heavy-tailed logistic and Student's  $t$  distributions can improve forecast performance in terms of sharpness and calibration, and lead to only minor differences between the estimators employed. Finally, a simulation study confirms the importance of appropriate distribution assumptions and shows that for a correctly specified model the maximum likelihood estimator is slightly more efficient than the CRPS estimator.

## 1. Introduction

Nonhomogeneous regression is a popular regression-based technique to statistically correct an ensemble of numerical weather prediction models (NWP; Leith 1974).

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

<sup>a</sup> Current affiliation: Division for Biomedical Physics, Medical University of Innsbruck, Innsbruck, Austria.

*Corresponding author:* Manuel Gebetsberger, manuel.gebetsberger@gmail.com

Such corrections are often necessary since current NWP models cannot consider all error sources (Lorenz 1963; Hamill and Colucci 1998; Mullen and Buizza 2002; Bauer et al. 2015) so that the raw forecasts are often biased and uncalibrated.

In statistical postprocessing, various approaches have been developed to correct such ensembles (Roulston and Smith 2003; Raftery et al. 2005; Gneiting et al. 2005; Wilks 2009) but none of them has appeared as a best single postprocessing strategy (Wilks and Hamill 2007).



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

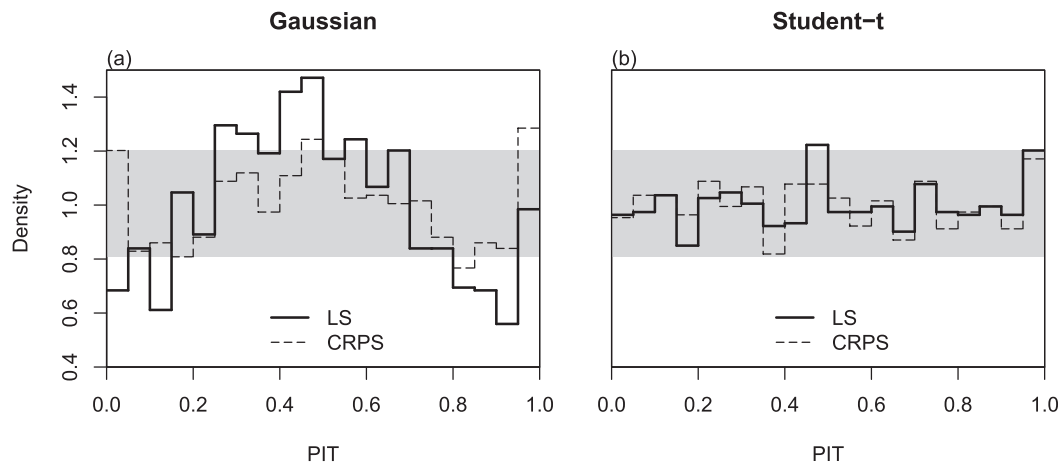


FIG. 1. PIT histogram for temperature forecasts at an Alpine site at +24-h lead time, shown for the (a) Gaussian and (b) Student's  $t$  models, estimated with LS (solid) or CRPS (dashed) minimization. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

However, nonhomogeneous Gaussian regression (NGR) is one of the most widely used techniques (Gneiting et al. 2005) and addresses ensemble errors in terms of regression coefficients, which are estimated on past ensemble forecasts and the corresponding observations. NGR has also been extended from temperature to other meteorological quantities by assuming appropriate forecast distributions (Gneiting et al. 2005; Thorarindottir and Gneiting 2010; Messner et al. 2014a,b; Scheuerer 2014; Hemri et al. 2016).

In the field of statistics, regression coefficients and distribution parameters have traditionally mostly been estimated with maximum likelihood estimation (Aldrich 1997; Stigler 2007). Although the maximum likelihood estimator has certain optimal properties (Huber 1967; Casella and Berger 2002; Winkelmann and Boes 2006, details in section 2c), Gneiting et al. (2005) established NGR parameter estimation by minimizing the continuous ranked probability score (CRPS; Hersbach 2000). Postprocessing studies for meteorological applications have used this estimation approach frequently since then (Raftery et al. 2005; Vrugt et al. 2006; Hagedorn et al. 2008; Scheuerer 2014; Scheuerer and Büermann 2014; Mohammadi et al. 2015; Feldmann et al. 2015; Scheuerer and Hamill 2015; Scheuerer and Möller 2015; Taillardat et al. 2016; Möller and Groß 2016) and often found it to yield sharper and better calibrated probabilistic forecasts than with maximum likelihood estimation.

Likelihood maximization is equivalent to minimizing the log score (LS), which is more sensitive to outliers than the CRPS (Selten 1998; Grit et al. 2006). Because of this higher sensitivity to outliers Gneiting et al. (2005) found LS minimization to lead to overdispersive forecasts.

Figure 1a illustrates this overdispersion exemplarily for 2-m air temperature forecasts, where NGR is employed

at an Alpine site for +24-h forecasts (see section 3a for data). Ideally, for perfect calibration the probability integral transform (PIT) should be distributed uniformly. However, both estimation approaches, LS and CRPS minimization, show a hump in the center bins indicating overdispersive forecasts (i.e., the forecast distribution is too wide so that observations fall overproportionally into the central range of the distribution). Although the CRPS approach indicates a better coverage at center bins, further peaks are found at 0.05 and 0.95, which correspond to the tails of the Gaussian forecast distribution.

The differences between CRPS and LS minimization and the W shape of the CRPS model indicate a misspecification of the NGR in terms of its distributional tail. Figure 1b shows the PIT histograms of a nonhomogeneous regression model with a heavier-tail Student's  $t$  distribution instead of a Gaussian forecast distribution. Both estimation approaches show only small differences and much better calibration. This agrees with theoretical considerations that, given an appropriate distribution, LS and CRPS estimator are consistent and estimate very similar regression coefficients (Winkelmann and Boes 2006; Yuen and Stoev 2014).

In this article we set out to investigate when and why results from LS and CRPS minimization will differ for symmetric distribution assumptions. This is performed in terms of temperature forecasting in central Europe and with simulated data using the NGR as the benchmark approach. Further adjustments of this benchmark include the use of heavy-tailed logistic and Student's  $t$  probability distributions. In particular, the Student's  $t$  distribution allows for flexible adjustment of the distribution tails.

Section 2 provides an overview of the distributions employed and the methods for estimation and evaluation

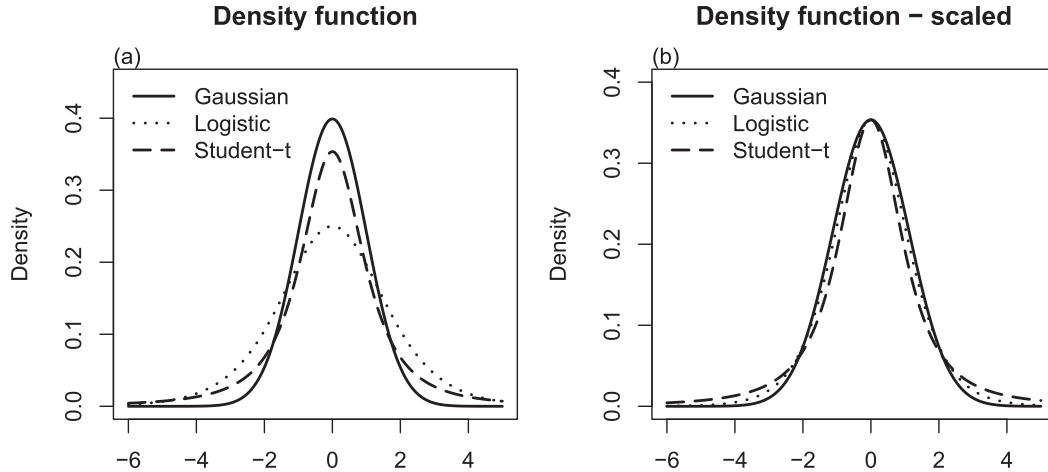


FIG. 2. (a) Probability density functions for a Gaussian (solid), logistic (dotted), and Student’s  $t$  distribution (dashed) with  $\mu = 0$ ,  $\sigma = 1$  for Gaussian and logistic distributions, and the degree of freedom  $\nu = 2$  for the Student’s  $t$  distribution. (b) Scaled density values with respect to the Student’s  $t$  distribution are shown to highlight the tails.

of the statistical models. Sections 3 and 4 present and discuss results for probabilistic temperature postprocessing and synthetic simulations, respectively. Finally, section 5 gives the conclusions.

**2. Methods**

This section briefly describes the distributions, along with the corresponding statistical models that are set up for the real case and simulation studies, and explains the estimation methods and desired estimator properties. Additionally, the comparison setup and verification measures are described.

*a. Distributions used and density functions*

In this article we employ three probability distributions with differences particularly on their tails (Fig. 2a). In the following we overview their key characteristics by their density functions.

The classical NGR approach is based on the Gaussian distribution  $\mathcal{N}(\mu, \sigma)$  with the location parameter  $\mu$  and the scale parameter  $\sigma$ . Its density function  $f_{\mathcal{N}}$  [Eq. (1)] is symmetrical around  $\mu$  (Fig. 2a), and is evaluated at the observed value  $y$  with

$$f_{\mathcal{N}}(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}[(y-\mu)/\sigma]^2}. \tag{1}$$

Similarly, but with a somewhat heavier tail, we use the logistic distribution  $\mathcal{L}(\mu, \sigma)$  with its density function  $f_{\mathcal{L}}$ :

$$f_{\mathcal{L}}(y; \mu, \sigma) = \frac{e^{-(y-\mu)/\sigma}}{\sigma(1 + e^{-(y-\mu)/\sigma})^2}. \tag{2}$$

Note, that the standard deviation of  $\mathcal{L}$  is not equal to the scale parameter  $\sigma$ , as it is the case for  $\mathcal{N}$ , rather than  $\sigma$  times  $\pi/\sqrt{3} \approx 1.8$ .

In addition to  $\mathcal{N}$  and  $\mathcal{L}$ , we make use of the shifted scaled Student’s  $t$  (denoted as “Student- $t$ ” in the following figures for simplicity) distribution  $\mathcal{S}(\mu, \sigma, \nu)$  (Student 1908), which, additionally to the location  $\mu$  and scale  $\sigma$  parameters has a third parameter  $\nu$ , the so-called degree of freedom:

$$f_{\mathcal{S}}(y; \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{\left(\frac{y-\mu}{\sigma}\right)^2}{\nu}\right]^{-\frac{(\nu+1)}{2}}. \tag{3}$$

Herein,  $\Gamma$  denotes the gamma function. The degree of freedom  $\nu$  controls the tails of the Student’s  $t$  distribution with heavier tails for smaller  $\nu$  values. In the limit of  $\nu \rightarrow \infty$  the Student’s  $t$  distribution approaches the Gaussian distribution. Its standard deviation is given by  $\sigma\nu/(\nu-2)$ .

Figure 2 compares the probability density functions of the different distributions where the scaled functions (Fig. 2b) highlight the different tail behaviors. The logistic distribution has clearly heavier tails than the Gaussian distribution and with  $\nu = 2$ , the Student’s  $t$  distribution can accommodate even heavier tails.

*b. Regression models*

As the basis regression model, we apply the NGR approach of Gneiting et al. (2005). The parameters of the assumed distributions are expressed by linear predictors. Each predictor contains covariates, which are typically provided by the NWP ensemble. This leads to regression models of the following form [Eqs. (4)–(6)], where the parameters  $\mu_i$  and  $\sigma_i$  are used for the Gaussian

and logistic assumptions, and  $\mu_i$ ,  $\sigma_i$ , and  $\nu_i$  for our representation of the Student's  $t$  distribution [Eq. (3)]:

$$\mu_i = \beta_0 + \beta_1 \times \overline{\text{ens}_i}, \quad (4)$$

$$\log(\sigma_i) = \gamma_0 + \gamma_1 \times \log(\text{SD}_{\text{ens},i}), \quad (5)$$

$$\log(\nu_i) = \delta_0. \quad (6)$$

The subscript  $i$  labels one observation–forecast pair. Commonly, the ensemble mean value  $\overline{\text{ens}_i}$  is used as covariate for the location parameter  $\mu_i$  [Eq. (4)], and the ensemble standard deviation  $\text{SD}_{\text{ens},i}$  is used for the scale parameter  $\sigma_i$  [Eq. (5)]. The degree of freedom of the Student's  $t$  model is simply modeled by a constant intercept  $\delta_0$  and not dependent on any covariable. Note that the coefficients for  $\sigma_i$  and  $\nu_i$  are estimated on the logarithmic scale in order to ensure the positivity of  $\sigma_i$  and  $\nu_i$ .

The framework defined in Eqs. (4)–(6) is used in later real data and simulation studies (sections 3 and 4). For the real data studies, sine and cosine of the day of the year ( $\text{DOY}_i$ ) are additionally included in the predictor of the location parameter  $\mu_i$ , to better represent seasonal variation of temperature (Dabernig et al. 2017; Messner et al. 2017):

$$\begin{aligned} \mu_i = & \beta_0 + \beta_1 \times \overline{\text{ens}_i} + \beta_2 \times \sin(\text{DOY}_i) \\ & + \beta_3 \times \cos(\text{DOY}_i). \end{aligned} \quad (7)$$

Clearly, the framework of Eqs. (4)–(6) can be extended by including additional covariates and also nonlinear terms (e.g., as in Stauffer et al. 2017). Also, other probability distributions such as the generalized extreme value distribution (Scheuerer 2014) could be used in this framework. Therefore, the models defined in this article as well as in Gneiting et al. (2005), Scheuerer (2014), and Stauffer et al. (2017) can be generally regarded as distributional regression models (Klein et al. 2015), where any probability distribution can be assumed for a response variable where each distribution parameter is linked to explanatory variables.

### c. Estimation methods

Estimation by the use of CRPS and LS belong to the class of M estimation (White 1994), where “M” stands for maximization or minimization. The idea is to find the set of parameters  $\hat{\theta}$  so that a function  $q$  (LS or CRPS in our case) is minimized:

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}} q(y; \theta). \quad (8)$$

More generally,  $\Theta = \mathbb{R}^p$  defines the parameter space with  $p$  being the number of regression coefficients,  $y = (y_1, y_2, \dots, y_N)$  is a vector of observed values, and  $N$

is the number of observations in a training dataset. In our specific regression framework,  $\hat{\theta}$  includes all the estimated regression coefficients ( $\beta$ ,  $\gamma$ ,  $\delta$ ) as defined in Eqs. (4)–(6). Estimators such as LS or CRPS should address the two properties of consistency and asymptotic normality:

$$\hat{\theta} \xrightarrow{p} \theta_0 \quad \text{as } N \rightarrow \infty, \quad (9)$$

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[0, I(\theta_0)^{-1}]. \quad (10)$$

Consistency derives from the law of large numbers (LLN), and normality derives from the central limit theorem (CLT). An estimator is consistent if it approaches the true parameter  $\theta_0$  in probability as the sample size  $N$  increases to infinity [Eq. (9)]. Furthermore, the difference  $\sqrt{N}(\hat{\theta} - \theta_0)$  approaches a Gaussian distribution  $\mathcal{N}$  [Eq. (10)] with the variance  $I(\theta)^{-1}$ . Herein,  $I(\theta)$  defines the Fisher information matrix, and its inverse defines the smallest possible variance achievable for any consistent estimator. Moreover, this variance describes the efficiency of an estimator. (Winkelmann and Boes 2006)

Consistency and asymptotic normality can be mathematically proven for both estimators under certain regularity conditions (Winkelmann and Boes 2006; Yuen and Stoev 2014), where the properties for the CRPS estimator are proven under mild regularity conditions (Yuen and Stoev 2014). Under strong conditions (e.g., where the probability density function is regular,  $\Theta$  is “well behaved” so that an interior solution exists), the LS estimator is also asymptotic efficient among all consistent estimators since it reaches the so-called Cramér–Rao lower bound [chapter 3.3. in Winkelmann and Boes (2006)]. This means that the LS estimator can achieve the smallest variance or has the least uncertainty around the true parameters. Hence, by assuming a correct specification of the regression model, both estimators are supposed to be consistent in finding the “true” parameters, whereas the LS estimator should additionally be more efficient by showing a smaller variance.

The main difference between the scoring rules CRPS and LS is the penalization of individual unlikely events in the tails of the distribution, which is compared in the following. The LS [Eq. (11)] is simply the negative log-likelihood, which is averaged over  $N$  events, where each event  $i$  is evaluated by the negative logarithmic density value  $\log f$ :

$$\text{LS} = \frac{1}{N} \sum_{i=1}^N -\log f(y_i; \mu_i, \sigma_i, \nu_i). \quad (11)$$

This score defines a local score as one single forecast distribution is evaluated only at the observed value  $y_i$  with a logarithmic penalty.

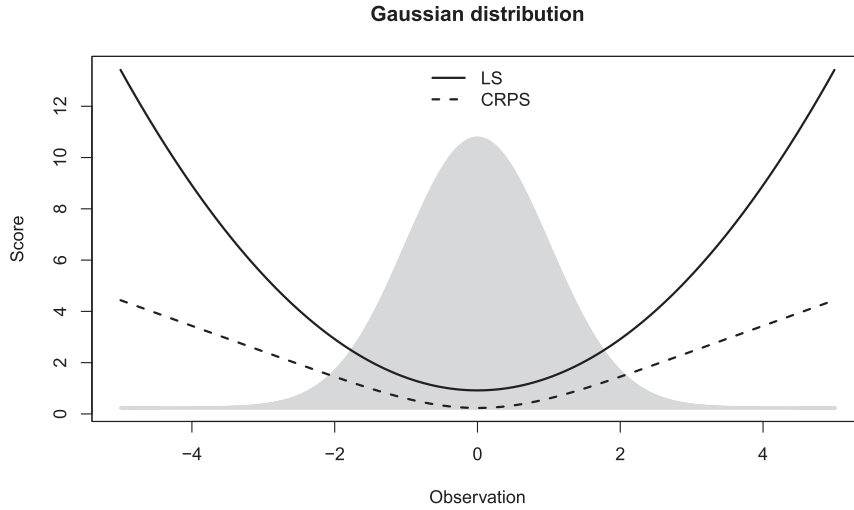


FIG. 3. Continuous ranked probability score (CRPS, dashed) and log score (LS, solid), evaluated at different (theoretical) observed values for an assumed Gaussian distribution with  $\mu = 0$ ,  $\sigma = 1$ , with probability density values sketched as gray area.

In contrast, the continuous ranked probability score for one single event defines a squared error measure, which takes the full forecast distribution into account:

$$CRPS = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{\infty} [F_i(x; \mu_i, \sigma_i, \nu_i) - H_i(x - y_i)]^2 dx. \tag{12}$$

For each observation  $y_i$ ,  $F_i$  denotes the forecasted cumulative distribution function and  $H_i(x - y_i)$  the Heaviside function, which is 0 if  $x < y_i$  and 1 otherwise. Integration over all differences between  $F_i$  and  $H_i$  in  $x$  evaluates the full forecast distribution. Similar to the LS, the CRPS itself defines the average over  $N$  events [Eq. (12)].

The differences between LS and CRPS can be found particularly in the tails of an assumed distribution, as illustrated by the Gaussian example in Fig. 3. If a single observation is located on the distribution tails (above and below  $\pm 2$ ), then larger differences between the scores can be found. The LS penalizes events on these tails more strongly than the CRPS.

*d. Verification*

Different verification approaches are needed for the real data and the simulation study. Regarding the real data, the two estimation approaches are compared in terms of their sharpness and calibration. Sharpness will be evaluated as the average width of the 90% prediction intervals (PIW), defined as the average range between the 0.05 and 0.95 quantile of the forecast distributions. This interval can also be used to assess calibration where 90% of the events should be observed within the 90% prediction intervals [prediction interval coverage (PIC)]

to have perfect calibration. Additionally, calibration is investigated with PIT histograms (Gneiting et al. 2007), which evaluate the forecasted cumulative distribution functions equivalently to the rank histogram (Anderson 1996; Talagrand et al. 1997; Hamill and Colucci 1998). Herein, the CDF values at the observed temperature events can be summarized in a histogram, which should display a uniform distribution of the PIT values. The desired uniformity derives from the statistical forecast consistency (calibration) that is fulfilled if all realizations are statistically indistinguishable from a sample that is drawn from the same predictive distribution. However, since one PIT histogram is obtained for each station, lead time, and statistical model, the differences between the histograms will also be quantified by the reliability index (RI) that computes absolute differences from uniformity for each PIT histogram:

$$RI = \sum_{k=1}^K \left| \kappa_k - \frac{1}{K} \right|. \tag{13}$$

Herein,  $\kappa_k$  defines the relative number of observations in each bin  $k$ , and  $K$  defines the number of used bins.

Furthermore, the overall performance measures for temperature forecasts will be shown in terms of LS and CRPS as defined by Eqs. (11) and (12).

To investigate the characteristics of the two estimators on real temperature data, we perform a tenfold cross validation (CV) to mimic operational conditions. In an operational situation, multiple years of a consecutive time series would be used to estimate a fixed set of regression coefficients that are applied on independent future data where the forecast performance can be assessed. The CV approach is used for scientific research

purposes to obtain stable regression coefficients, and a sufficiently large amount of test data is not used for regression (e.g., as in Hamill et al. 2004; Wilks and Hamill 2007; Hagedorn et al. 2008; Wilks 2009; Messner et al. 2017; Gebetsberger et al. 2017; Rasp and Lerch 2018). Therefore, the data are randomly partitioned into 10 different subsamples and forecasts for each subsample are derived from models trained on the remaining 9 subsamples. More specifically, each trained model is applied on raw ensemble data of the remaining test subsample. This leads to out-of-sample forecasts not used for training, which are verified with PIW, PIC, RI, LS, and CRPS. This approach is repeated for each lead time and station, and estimation method (LS or CRPS).

The CV approach used for temperature data assumes temporal independence and stationarity in the forecast error time series. Since separate CVs are performed for each lead time, temporal independence is a valid assumption and with no major changes in the NWP model and no major changes in the climate over the data period, the ensemble characteristics are not expected to change much either.

However, it has to be mentioned that the original NGR approach uses a rolling training period, where a certain training window (e.g., 30 days for temperature) is used to train the statistical model. This allows us to rapidly update the regression coefficients from day to day to capture seasonality. In the CV approach used for this study, this seasonality is captured by a seasonal effect as explained in section 2b, and must not be updated for each day.

In the simulation study we mainly compare the estimated regression coefficients with their known true values to investigate how well the different estimation approaches estimate the true coefficients. Additionally, calibration is assessed by PIT histograms.

### 3. Probabilistic temperature forecasting

With this real data application it should be investigated if the differences between CRPS and LS minimization, as shown in the introductory example, imply an inappropriate distribution assumption for temperature data. This idea is addressed by the use of heavy-tailed distributions to determine the estimator characteristic on real data and to improve temperature forecasts. For simplicity, statistical models (Gaussian, logistic, and Student's  $t$ ) where CRPS or LS minimization is employed, will be referred to as CRPS or LS models, respectively.

#### a. Temperature data

Temperature records are used from 11 locations over central Europe (Fig. 4) for 3-hourly lead times from +6 to +96 h in the time period between 2011 and 2016.

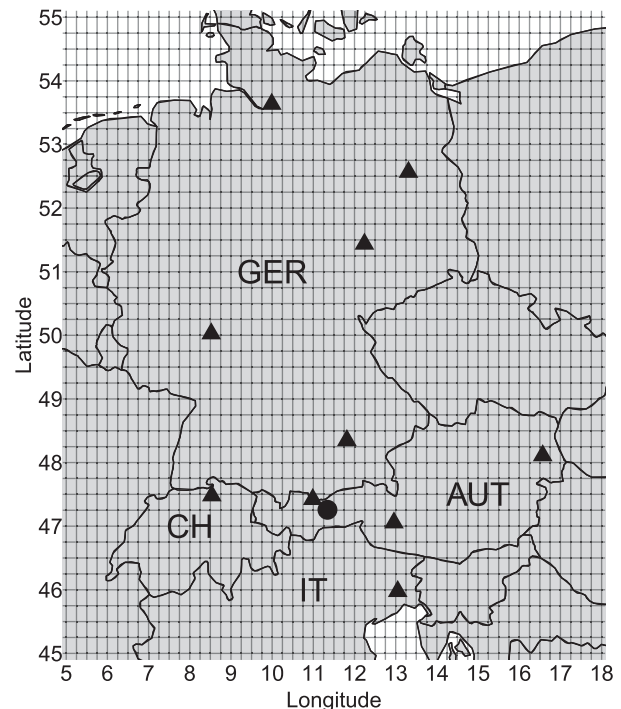


FIG. 4. Study area with the sites in Austria (AUT), Italy (IT), Switzerland (CH), and Germany (GER): the filled circle represents the Alpine site, which is used for the case study. The gray grid illustrates the underlying horizontal grid of the 50-member ECMWF ensemble forecasts.

The corresponding ensemble forecasts of 2-m air temperature are based on the 0000 UTC initialization from the European Centre for Medium-Range Weather Forecasts (ECMWF), of which forecasted mean values and standard deviations of the 50-member ensemble are used. Overall, this yields 581 076 observation–forecast pairs to be validated, which include 311 different regressions for different lead times and stations, since 2 sites had missing observations during nighttime.

The following case study is based on temperature records at an Alpine site (Fig. 4, filled circle) where the complex topography causes a challenging forecasting situation. Distinct differences between the real and NWP topography (valleys that are not well resolved) lead to a cold bias, which can be seen when comparing observations with corresponding ensemble mean forecasts (Fig. 5a). Furthermore, the ensemble is also underdispersive, which is a common problem of many ensemble systems. This underdispersion can be assessed in a rank histogram (Anderson 1996; Talagrand et al. 1997; Hamill and Colucci 1998), which is shown for the bias-corrected ECMWF ensemble forecast for +24 h in Fig. 5c. Here, too many observations are counted below the lowest and above the highest member value (lowest and highest rank), indicating less forecast uncertainty

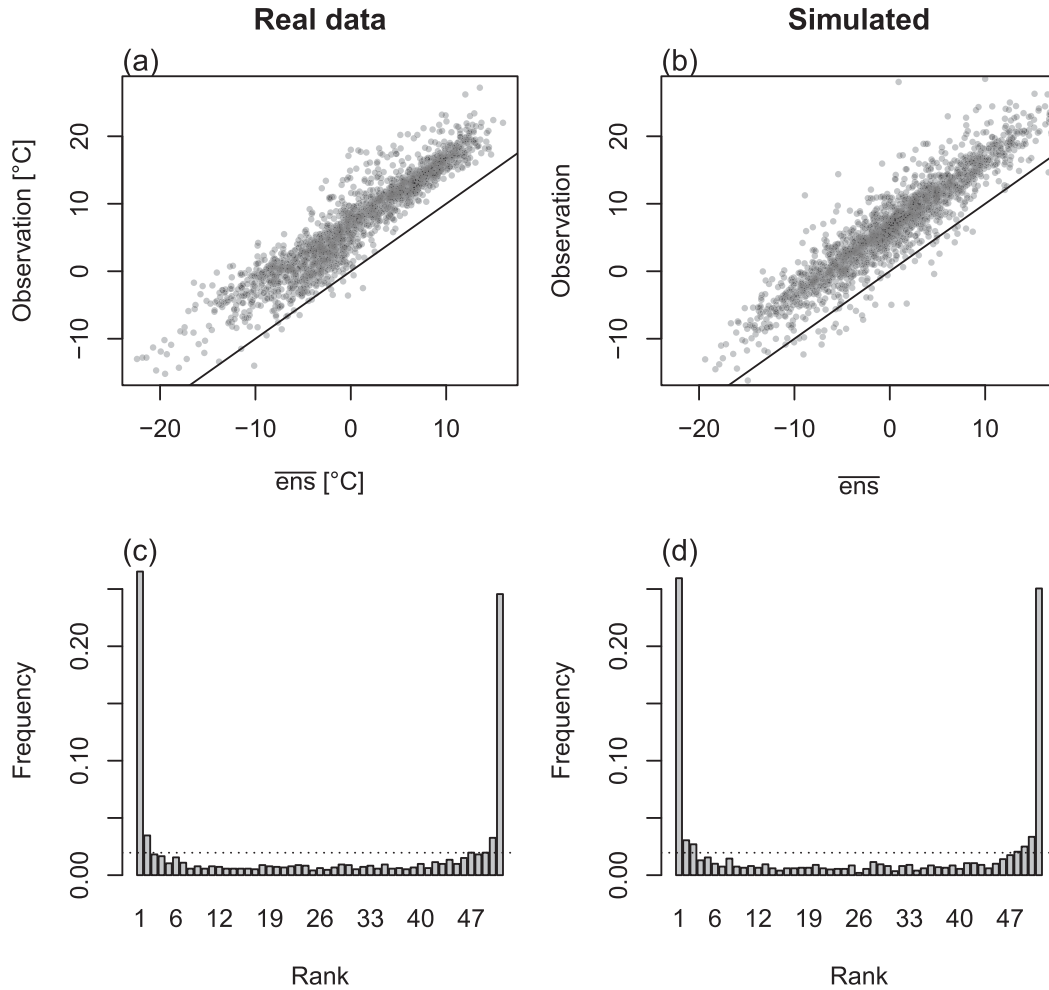


FIG. 5. Error characteristics for real data at the Alpine site for (a),(c) +24-h temperature forecasts and (b),(d) simulated data. (a),(b) Ensemble mean values  $\bar{\text{ens}}$  against observed values, where darker colors indicate a higher scatter density. (c),(d) Rank histograms of the bias-corrected 50-member ensembles, with members randomly drawn from the known Gaussian distribution for the simulated data. Dotted horizontal line indicates perfect calibration.

than needed. Well-calibrated forecasts would result in a display of a uniform distribution. These illustrated ensemble forecasts for +24 h are the basis for later synthetic simulations, using the error characteristics for bias and underdispersion. The empirical values of this dataset have an average ensemble mean value of 0.35 with a standard deviation of 6.91. The corresponding logarithmic standard deviations have an average of  $-0.56$  with a standard deviation of 0.43.

*b. Alpine site case study*

In this subsection we apply the regression framework, as defined in Eqs. (4)–(6), for temperature post-processing at the Alpine site (Fig. 4, filled circle), where individual regressions are performed for each lead time separately.

Figure 6 summarizes RI, PIW, and PIC for the Gaussian and Student’s *t* models, which are estimated with both approaches (CRPS or LS minimization). For the Gaussian models (left panels), there is a clear difference between the LS and CRPS model for certain lead times (e.g., +24 h) where calibration in terms of RI (Fig. 6a) is better for the CRPS model. Additionally, the CRPS model obtains sharper predictions for all lead times, which is shown by a smaller average width of the 90% prediction interval (Fig. 6c). Both estimation approaches show empirical coverages for certain lead times, which do not match the nominal coverage of 90% (Fig. 6e). This empirical coverage should be as close as possible to the nominal coverage of the evaluated prediction interval, where the LS model covers too many observations and the CRPS covers too few in the 90%



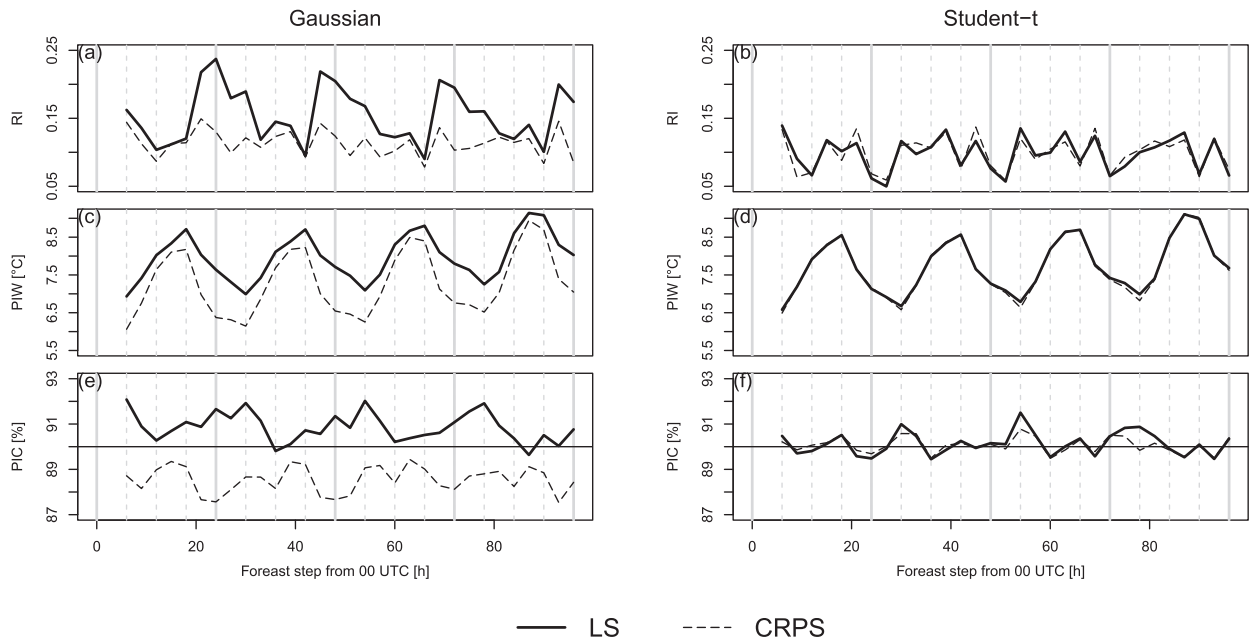


FIG. 6. (a),(b) Reliability index (RI); (c),(d) average width of the 90% prediction interval (PIW); and (e),(f) coverage of the 90% prediction interval (PIC), evaluated for (a),(c),(e) Gaussian and (b),(d),(f) Student's  $t$  models at the Alpine site from lead times +6 to +96 h, estimated with LS (solid) or CRPS (dashed).

interval. The agreement between empirical and nominal coverage is even worse for the 66% interval, particularly for the LS model (not shown).

Therefore, the PIT histograms, which are shown in Fig. 1 for the +24-h example, provide a more complete picture of the calibration. The 95% consistency interval shown as gray area, are derived similar to Bröcker and Smith (2007) and show the expected binwise sampling variations. Thus, as long as the PIT lies within this interval the forecasts can be regarded as calibrated.

Regarding the Gaussian models (Fig. 1a), the smaller sharpness (larger prediction intervals) of the LS model produces a hump-shaped PIT (solid), where too many observations fall in the central bins, and too few fall in the tails (bins close to zero and one). In contrast, the CRPS model (dashed) shows a better calibration especially in central bins, but creates larger peaks on the tails, which results from sharper forecast distributions. Regarding the high standard of PIT histograms showing a uniform distribution, both approaches illustrate shortcomings and differ in the forecasted distribution parameters if the Gaussian distribution is assumed.

However, if the Student's  $t$  model is applied, both approaches yield almost the same results. Similar values can be verified for calibration (RI) and sharpness (PIW), as illustrated in Figs. 6b, 6d, and 6f. Regarding the overall calibration in terms of PIT, the example for +24 h yields almost uniform histograms for the Student's  $t$  models for both minimization approaches (Fig. 1b). The corresponding

estimated degree of freedom  $\nu$  is shown for all lead times in Fig. 7. A daily pattern can be identified, with the highest values during daytime (e.g., at +15 h) and the lowest values during nighttime (e.g., at +24 h). Small values of  $\nu$  imply that heavier distribution tails are estimated, whereas high values for  $\nu$  ( $\nu \approx 100$ ) create only a slightly heavier tail than a Gaussian distribution would have. Additionally, there is a slight indication that  $\nu$  increases with lead time after accounting for the diurnal behavior.

### c. Overall performance

The previously shown case study for the Alpine site is now extended to other locations in our study area, again with individual regressions for each lead time. The mean scores over all the individual LS and CRPS emphasize the benefit of the heavy-tailed models for which score values are smallest. Not surprisingly, CRPS models perform better in CRPS evaluation and LS models in LS evaluation (Figs. 8a,c).

Figures 8b and 8d summarize differences in LS and CRPS values between each regression model and the Gaussian benchmark model, where negative values report a better performance than the benchmark model. LS models refer to the Gaussian LS model and CRPS models refer to the Gaussian CRPS models. Absolute differences are chosen rather than relative changes as skill scores cannot be computed for the LS (Gneiting et al. 2005). The variability in the score difference is smaller for CRPS models than for LS models evaluated

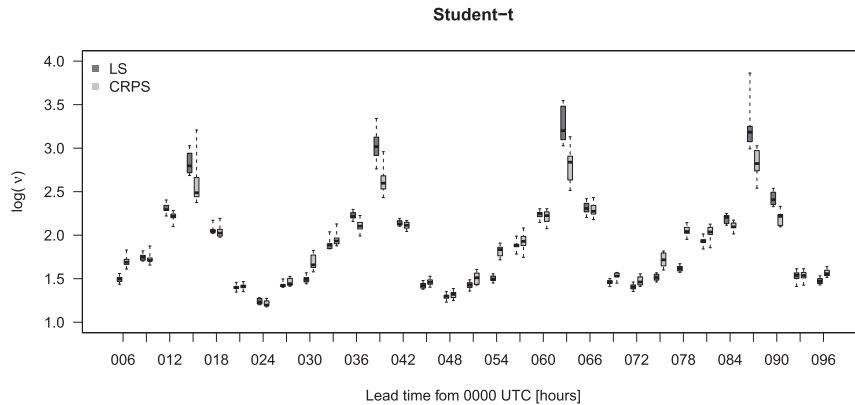


FIG. 7. Estimated degree of freedom  $\nu$  (y axis) for the Student's  $t$  models at the Alpine site for the respective lead time (x axis) using LS (dark gray) and CRPS (gray) estimation. Note that  $\nu$  is illustrated on the log scale. Each boxplot contains 10 estimated values obtained from the tenfold cross validation.

on the CRPS (Fig. 8b), and smaller for LS models than CRPS models evaluated on the LS (Fig. 8d).

However, Figs. 8b and 8d illustrate a clear benefit for all individual regressions if heavy-tailed distribution models (logistic, Student's  $t$ ) are applied, indicated by a negative difference. In terms of CRPS evaluation (Fig. 8b), the logistic models can improve the Gaussian benchmarks in 59% and 76% of all locations and lead times, when estimated with CRPS and LS, respectively. Even smaller CRPS values are obtained in 80% and 86% of the Student's  $t$  models.

A similar picture is visible for LS evaluation (Fig. 8d). A total of 84% and 82% of the evaluated logistic models show smaller LS values for CRPS or LS minimization, respectively. Student's  $t$  models obtain smaller LS values than the Gaussian benchmark for 93% (CRPS minimization) and 97% (LS minimization) of all regressions.

On average CRPS and LS, the Student's  $t$  models perform best (Figs. 8a,c). However results also imply that the logistic models already improve the benchmark sufficiently, and the further improvement of the Student's  $t$  models is small. Hence, there are situations with real data where the logistic models might be good enough and where the tail flexibility of the Student's  $t$  model is not necessary.

An example of the good calibration of logistic models is shown in Fig. 9b, which consists of predictions for all stations at lead time +18 h. Similarly to the Alpine site as shown in Fig. 1a, the Gaussian models in Fig. 9a illustrate an overdispersive W shape over all locations. The PIT histogram of the CRPS model is more pronounced on the tails (dashed), and the PIT histogram for the LS model is more pronounced in the middle (solid). Contrary to this, the heavy tail of the logistic distribution leads to nearly perfect and similar calibration for both approaches (middle). Additionally, the heavy tail created by the

Student's  $t$  models (Fig. 9c) seems to be too heavy for this particular lead time where too few events occur on the tails (right). In this case Student's  $t$  models can clearly improve calibration compared to the Gaussian models, but the tails are not captured appropriately and suggest to assume a logistic distribution. Moreover, the assumption of a constant  $\nu$  in Eq. (6) might be too simple, and a seasonal variation of  $\nu$  as in Eq. (7) would be more reasonable.

To give an impression about the estimated degree of freedom  $\nu$ , Table 1 summarizes the estimated values on the log scale for the entire study area and lead times. For each station and lead time, 10 estimated values of  $\nu$  are obtained, which is why averages for each station and lead time are analyzed. In 75% (third quartile) of these averages, the values are below 2.67 and 3.05 for LS and CRPS estimation, respectively, which correspond to  $\nu = 14.4$  and 21.1, respectively. As the Student's  $t$  distribution approaches the Gaussian distribution with increasing  $\nu$ , values of  $\nu \approx 20$  still produce heavier tails compared to the Gaussian tail. For values larger than 200, the Student's  $t$  distribution is already very close to a Gaussian distribution.

#### 4. Simulation study

In the following simulation study, "ensemble" and "observation" data with similar error characteristics as those at the Alpine site are generated. These data are generated such that the true distribution parameters and regression coefficients are known and can directly be compared with estimated values. Furthermore they are used to evaluate which minimization approach is more efficient and to confirm findings from the real data application.

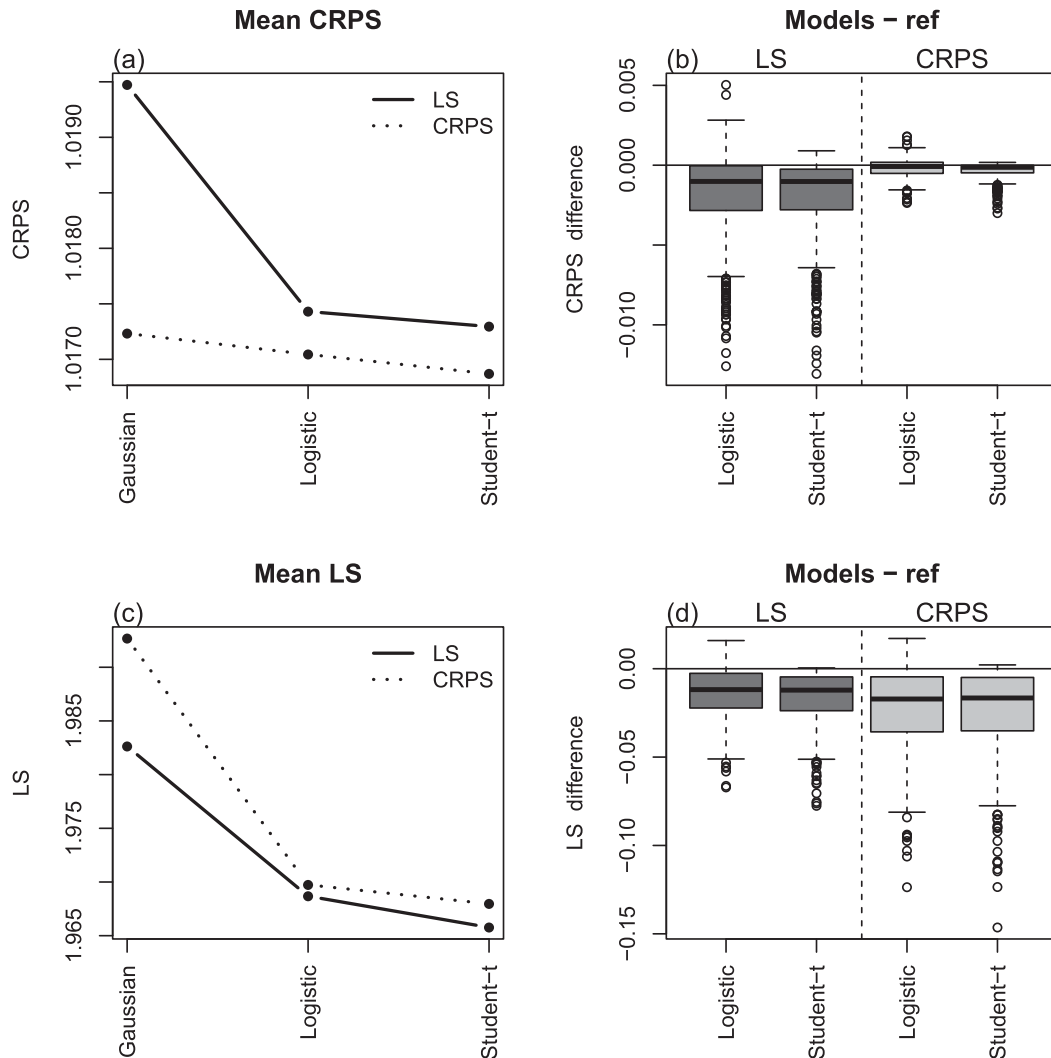


FIG. 8. (a),(c) Mean and (b),(d) differences of averaged scores for each station and lead time for LS-minimized (solid line and dark gray) and CRPS-minimized (dotted line and light gray) models, evaluated with the (a),(b) CRPS and (c),(d) LS. References are the Gaussian models for LS or CRPS minimization, respectively. Each boxplot contains results for 311 individual regression fits for each lead time and station.

### a. Simulated dataset

First, a series of  $N = 5000$  simulated ensemble mean values [ $\overline{\text{ens}_i}$ , Eq. (14)] and logarithmic standard deviations [ $\log(\text{SD}_{\text{ens},i})$ , Eq. (15)] were simulated from a Gaussian distribution  $\mathcal{N}$ :

$$\overline{\text{ens}_i} = \mathcal{N}(0.35, 6.91), \quad (14)$$

$$\log(\text{SD}_{\text{ens},i}) = \mathcal{N}(-0.56, 0.43), \quad (15)$$

with the distribution parameters taken from the empirical means and standard deviations of the ECMWF ensemble at the Alpine site (section 3a). Observations are simulated from logistic distributions, which we found in

section 3c to describe temperature data quite well. The location ( $\mu_i^{\text{true}}$ ) and scale ( $\sigma_i^{\text{true}}$ ) parameters of these distributions are modeled as functions of the simulated ensemble statistics  $\overline{\text{ens}_i}$  and  $\text{SD}_{\text{ens},i}$ :

$$\mu_i^{\text{true}} = \beta_0^{\text{true}} + \beta_1^{\text{true}} \times \overline{\text{ens}_i}, \quad (16)$$

$$\log(\sigma_i^{\text{true}}) = \gamma_0^{\text{true}} + \gamma_1^{\text{true}} \times \log(\text{SD}_{\text{ens},i}), \quad (17)$$

where  $(\beta_0^{\text{true}}, \beta_1^{\text{true}}) = (6.5, 1)$  and  $(\gamma_0^{\text{true}}, \gamma_1^{\text{true}}) = (0.9, 1.3)$  are chosen such that the simulated forecasts exhibit a cold bias and underdispersion similar to the real data (Figs. 5b,d).

Thus, a dataset of length 5000 is available with forecasts and corresponding observations that have similar

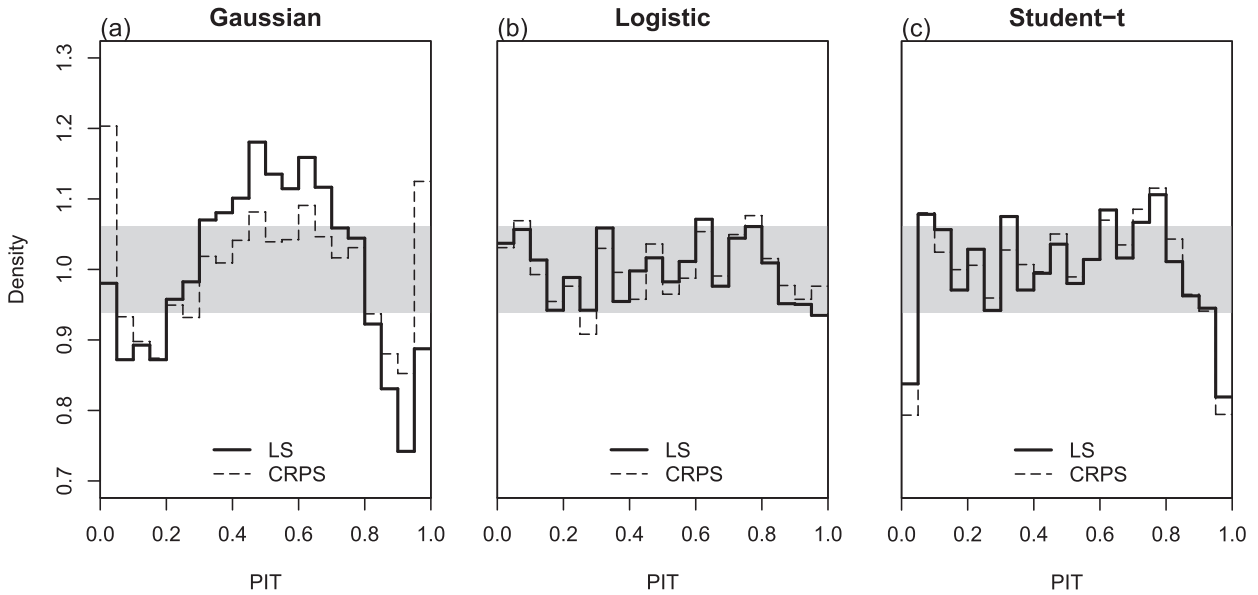


FIG. 9. PIT value for (a) Gaussian, (b) logistic, and (c) Student’s  $t$  models with LS (solid) or CRPS (dashed) minimization. Analysis includes 11 stations for +18-h lead times. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

properties as the real data used in section 3a. However, different to the real data the true coefficients  $\beta_0^{\text{true}}, \beta_1^{\text{true}}, \gamma_0^{\text{true}},$  and  $\gamma_1^{\text{true}}$  are known and can directly be compared to estimated coefficients  $\beta_0, \beta_1, \gamma_0,$  and  $\gamma_1$  from nonhomogeneous regression models of the form of Eqs. (4)–(5).

In the following, we fit models with Gaussian and logistic distribution assumptions and repeat the simulations 1000 times to account for sampling effects.

*b. Simulation results*

Figure 10a compares the two estimation approaches for the Gaussian models. By repeating the simulation 1000 times, both approaches estimate the true coefficients for the location submodel ( $\beta_0, \beta_1$ ) on the median. However, differences occur in the scale submodel ( $\gamma_0, \gamma_1$ ). Although the slope coefficient  $\gamma_1$  expresses the true value on the median, clear differences can be found for the intercept  $\gamma_0$ . Both approaches do not calculate

the true coefficient of 0.9 and estimate a larger value. This is mainly the consequence of the scaling by approximately 1.8 since the standard deviation of the logistic distribution is  $1.8 \times 0.9 = 1.62$ .

Furthermore, this difference is caused by the response data, which are sampled from a logistic distribution that has a heavier tail than the Gaussian distribution. To account for those “extreme” events, both approaches have to estimate a larger intercept and make the “forecast uncertainty” large enough. Furthermore, the LS model produces a larger intercept than the CRPS model, which is caused by the larger penalty of extremes by the logarithm.

However, if the same simulation is performed with logistic models (Fig. 10b), then both approaches estimate the true “errors” (coefficients) on median. By looking on the variance or range of the estimated coefficients, respectively, it can be seen that the LS model is slightly more efficient than the CRPS model. More specifically, the LS model reports a smaller interquartile range than the CRPS model. This finding also agrees with Yuen and Stoev (2014), where CRPS shows a smaller efficiency than LS estimation.

Finally, Fig. 11 shows PIT histograms of the different models for different lengths of the simulated datasets. As expected and similar to the real data case study, the Gaussian “forecasts” humps at central PIT values show the lack of calibration (Fig. 11a). Although this hump is less visible for the CPRS model than for the LS model, the peaks on the tails for the CRPS model are more

TABLE 1. Summary of estimated degree of freedom  $\log(\nu)$  for LS and CRPS estimation over all stations and lead times. (from left to right) Minimum, first quartile (25% quantile), median (50% quantile), mean, third quartile (75% quantile), and maximum. Values are based on  $\log(\nu)$  values that are averaged over the 10 cross-validation blocks of each station and lead time separately.

	Min	First quartile	Median	Mean	Third quartile	Max
LS	1.16	1.79	2.19	2.88	2.67	15.83
CRPS	1.21	1.95	2.44	3.52	3.05	14.45

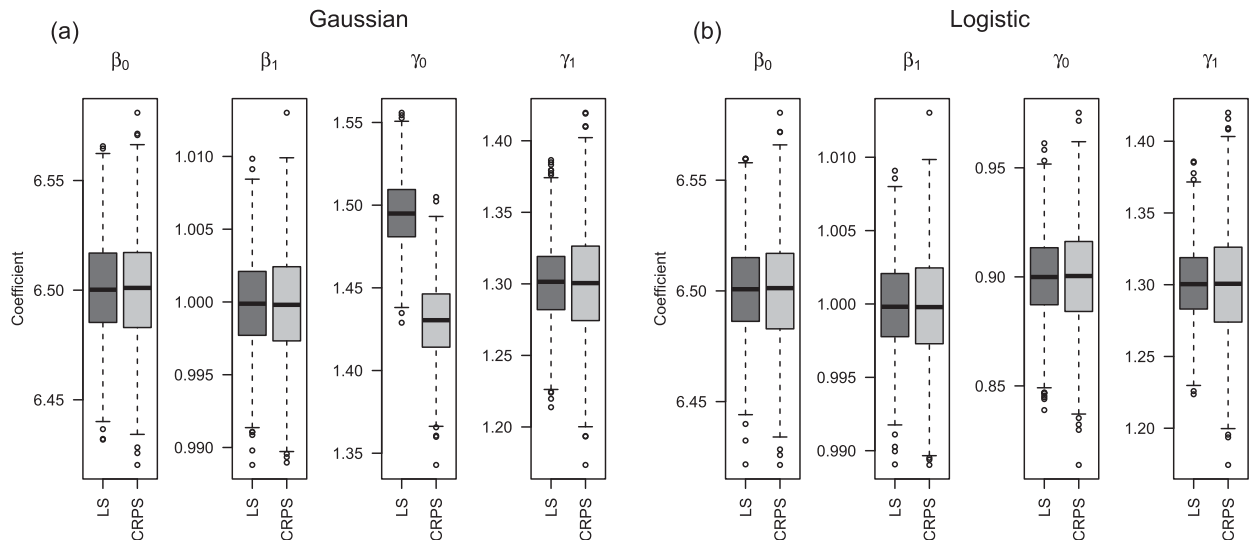


FIG. 10. The estimated regression coefficients ( $\beta_0$ ,  $\beta_1$ ,  $\gamma_0$ ,  $\gamma_1$ ) for the (a) Gaussian models and (b) logistic models, estimated with LS (dark gray) or CRPS (light gray) minimization, respectively. Boxplots are based on the bootstrap procedure of repeating the simulation 1000 times and illustrate the interquartile range (0.25–0.75) in boxes, whiskers for  $\pm 1.5$  times interquartile range, and outliers in solid circles.

pronounced. In contrast, the difference between the estimation approaches becomes smaller if the correct (and known) logistic response distribution is assumed (Fig. 11b). As expected from estimation theory, the differences vanish with increasing sample size for the correct distribution assumption (Figs. 11d,f), and show a well-defined W shape for the wrong assumption (Figs. 11c,e).

This W shape is characteristic in the presented scenario where symmetric heavy-tailed (logistic) data are modeled with the Gaussian assumption. Clearly, this is expected to differ if the response data are drawn from another distribution. As an example, we repeat the simulation with the same setup as described in section 2a, but simulate the observations from a Gaussian instead of the logistic distribution. The characteristic PIT histograms for the respective models are displayed in Fig. 12, which shows an M shape for the logistic models (Fig. 12b). As for the wrong assumption in Figs. 11a,c,e, the two estimation approaches differ most on the tails. This M shape is also visible for the temperature study at +18 h in Fig. 9c for the heavy-tailed Student's  $t$  models, where the PIT histograms for the logistic models show the best calibration and a good agreement between the estimation approaches.

Generally, calibration for symmetric response data in terms of PIT histograms shows the W shape if the assumed distribution tail is too weak, and the M shape if the distribution tail is too heavy. However, the forecast tail shows largest differences between the estimation approaches in both scenarios and agrees with the temperature study of section 3.

To combine results of real and synthetic scenarios, the obtained shapes of the PIT histograms display a useful characterization to identify misspecifications of the distributional assumption. Apart from the presented scenarios, the PIT shapes and differences between the estimation approaches might not be restricted to ensemble postprocessing. For instance, similar calibration results are expected if wrong tails of temperature anomalies are assumed, since anomalies typically have the same distributional properties as the data themselves. Moreover, results are relevant for applications other than probabilistic weather forecasting (e.g., climate), where a future increase in extremes would lead to heavier tails.

## 5. Conclusions

Nonhomogeneous regression is a commonly used postprocessing strategy to statistically correct NWP ensemble forecasts. This approach predicts the outcome of weather quantities of interest with full parametric forecast distributions. To estimate distribution parameters or regression coefficients, scoring rules have to be optimized. Log-score (LS) minimization has a long tradition in statistical modeling, whereas CRPS minimization has become popular in meteorological studies. Although both approaches should theoretically obtain similar results, differences are often found in practical studies. In this article we set out to explain potential differences and use these findings to improve probabilistic temperature forecasts. A comparison of both estimation approaches is performed on air temperature

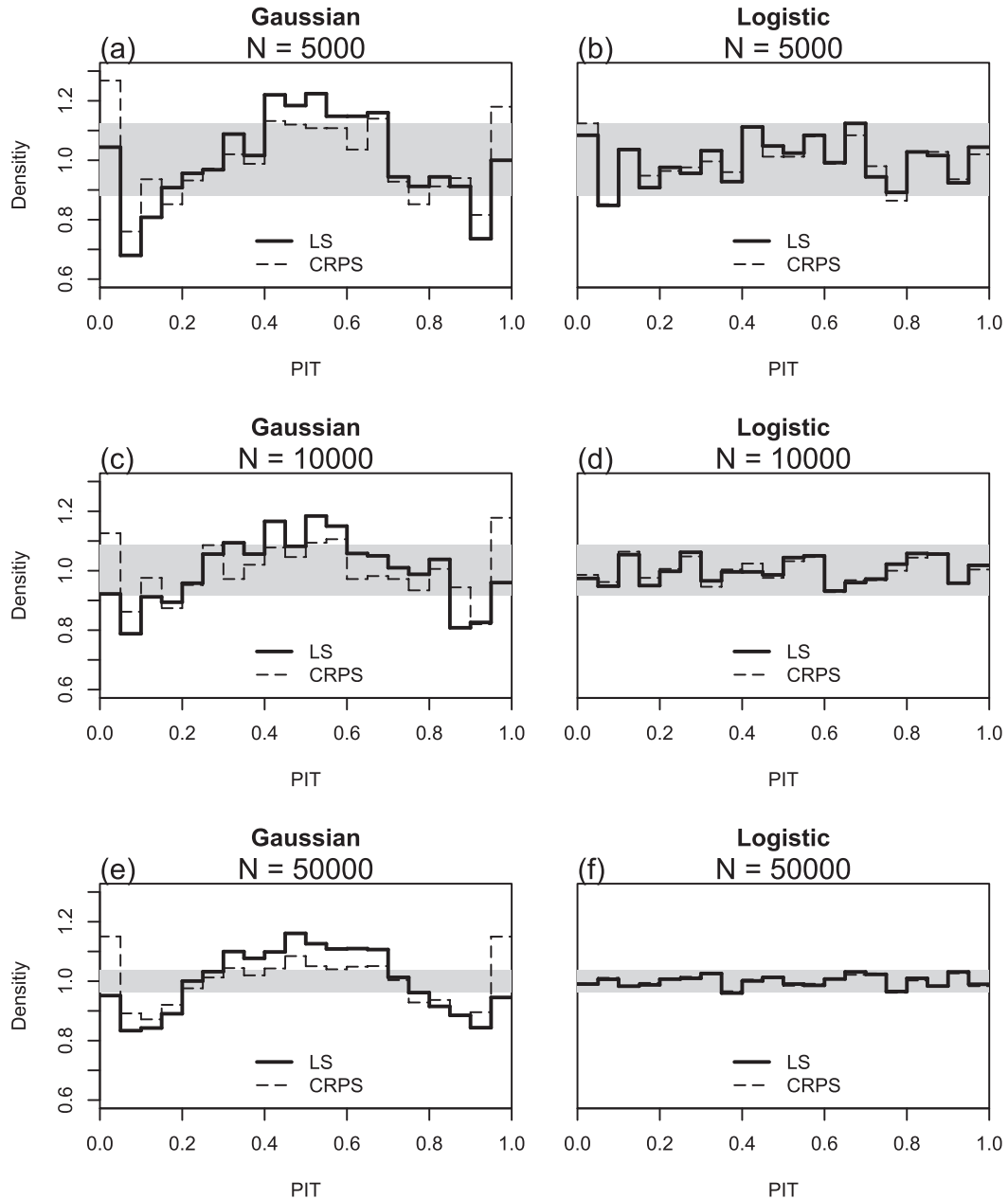


FIG. 11. Calibration in terms of PIT values for one simulation with  $N =$  (a),(b) 5000; (c),(d) 10 000; and (e),(f) 50 000 data using (a),(c),(e) Gaussian or (b),(d),(f) logistic model, estimated with LS (solid) or CRPS (dashed) minimization. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

data from 11 stations in central Europe and in a simulation study.

In principle, LS and CRPS minimization differently penalize “extreme” events or events with larger deviations from the mean forecast, respectively. Consequently, the assumed forecast distribution plays a crucial role to obtain a good forecast performance regarding sharp and calibrated predictions.

Generally, it turns out that evaluation of CRPS shows better values if CRPS minimization is performed, and evaluation of LS shows better values if LS minimization is employed. However, synthetic simulations and the case studies show that CRPS models can lead to sharper predictions than LS models. This particularly occurs if a wrong distribution with too light tails is assumed. Unfortunately, the increased sharpness of CRPS minimization is obtained

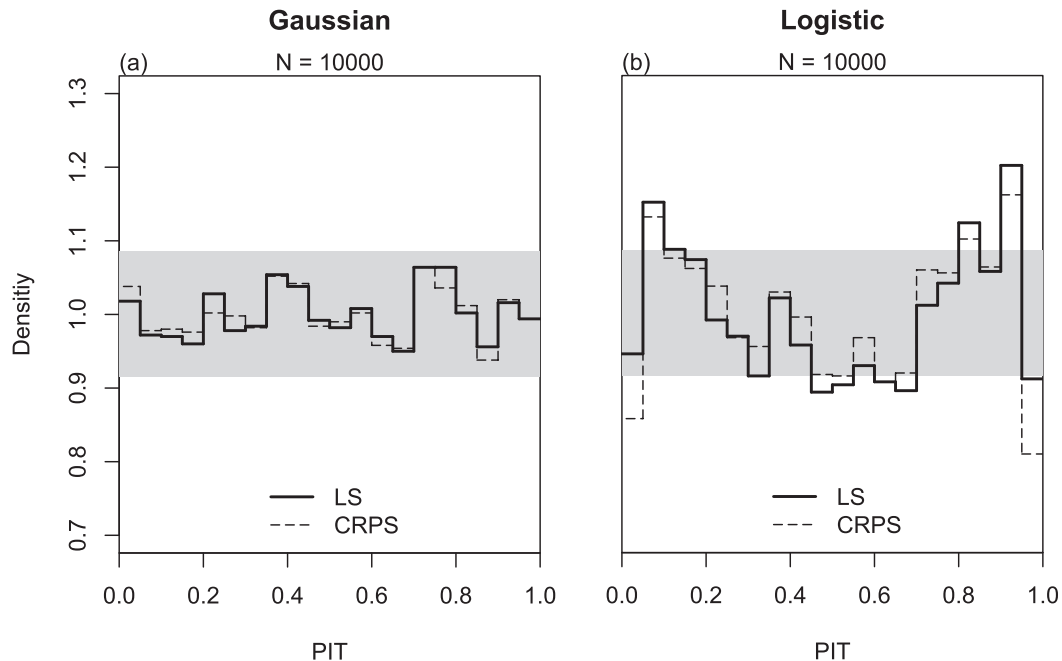


FIG. 12. As in Fig. 11, but for  $N = 10\,000$  observations simulated from Gaussian distributions.

at the expenses of a decreased calibration, where coverages are better at center bins but worse on the tails. CRPS minimization apparently improves coverage, but only at particular prediction intervals. Overall calibration in terms of PIT histograms illustrates that both approaches cannot calibrate appropriately if the wrong distribution is applied, which qualifies the better sharpness of CRPS minimization. Therefore, we cannot conclude that one approach should be applied over the other. In this context, more appropriate distribution assumptions have to be made if PIT calibration highlights problems on the tails, or if differences between the two estimation approaches occur. As a consequence, symmetric or—if needed—asymmetric distributions should be assumed, which better take heavy tails into account if necessary.

To account for a potentially heavier tail, this study introduces and compares the logistic and Student's  $t$  distribution against the classical Gaussian assumption for air temperature. The Gaussian and logistic assumption is found appropriate for air temperature at certain stations and lead times. However, the larger flexibility of the Student's  $t$  distribution to adjust the tail, could clearly improve sharpness with respect to calibration in the overall analysis. This derives from the distribution parameter, which accounts for a possible heavier tail if needed.

If the distributional assumption accounts for the tails, then both approaches lead to very similar results. In this case, the synthetic study highlights that the LS

approach is more efficient in estimating the true regression coefficients.

*Acknowledgments.* We thank the Austrian weather Service (ZAMG) for access to ECMWF EPS and observational data. We also thank the University of Innsbruck, the Faculty of Geo- and Atmospheric Sciences, and the Institute of Atmospheric and Cryospheric Sciences for sharing publication costs. This project was partially funded by doctoral funding of the University of Innsbruck, Vizerektorat für Forschung.

## APPENDIX

### Computational Details

The estimation of regression coefficients is performed in R (R Core Team 2017) using the *crch* package (Messner et al. 2016), which is able to perform minimization of the CRPS or LS. Closed expressions of the CRPS for the Gaussian, logistic, and Student's  $t$  distribution are based on the *scoringRules* package (Jordan et al. 2017).

## REFERENCES

- Aldrich, J., 1997: R. A. Fisher and the making of maximum likelihood 1912-1922. *Stat. Sci.*, **12**, 162-176, <https://doi.org/10.1214/ss/1030037906>.
- Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecast from ensemble model integration.

- J. Climate*, **9**, 1518–1530, [https://doi.org/10.1175/1520-0442\(1996\)009<1518:AMFPAE>2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2).
- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, <https://doi.org/10.1175/WAF993.1>.
- Casella, G., and R. L. Berger, 2002: *Statistical Inference*. 2nd ed. Thomson Learning, 660 pp.
- 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>.
- Feldmann, K., M. Scheuerer, and T. L. Thorarindottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, <https://doi.org/10.1175/MWR-D-14-00210.1>.
- Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Finetuning nonhomogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Wea. Rev.*, **145**, 4693–4708, <https://doi.org/10.1175/MWR-D-16-0388.1>.
- Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.*, **132**, 2925–2942, <https://doi.org/10.1256/qj.05.235>.
- Hagedorn, R., T. Hamill, and J. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, <https://doi.org/10.1175/2007MWR2410.1>.
- Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, [https://doi.org/10.1175/1520-0493\(1998\)126<0711:EOEREP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2).
- , J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, [https://doi.org/10.1175/1520-0493\(2004\)132<1434:ERIMFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2).
- Hemri, S., T. Haiden, and F. Pappenberger, 2016: Discrete post-processing of total cloud cover ensemble forecasts. *Mon. Wea. Rev.*, **144**, 2565–2577, <https://doi.org/10.1175/MWR-D-15-0426.1>.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Huber, P. J., 1967: The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Statistics*, University of California Press, 221–233.
- Jordan, A., F. Krüger, and S. Lerch, 2017: Evaluating probabilistic forecasts with scoringRules. R package, <https://cran.r-project.org/web/packages/scoringRules/vignettes/article.pdf>.
- Klein, N., T. Kneib, S. Lang, and A. Sohn, 2015: Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.*, **9**, 1024–1052, <https://doi.org/10.1214/15-AOAS823>.
- Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, [https://doi.org/10.1175/1520-0493\(1974\)102<0409:TSOMCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2).
- Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, <https://doi.org/10.1175/MWR-D-13-00355.1>.
- , —, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, <https://doi.org/10.1175/MWR-D-13-00271.1>.
- , —, and —, 2016: Heteroscedastic censored and truncated regression with crch. *R J.*, **8** (1), 173–181.
- , —, and —, 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>.
- Mohammadi, S. A., M. Rahmani, and M. Azadi, 2015: Optimization of continuous ranked probability score using PSO. *Decis. Sci. Lett.*, **4**, 373–378, <https://doi.org/10.5267/j.dsl.2015.4.001>.
- Möller, A., and J. Groß, 2016: Probabilistic temperature forecasting based on an ensemble autoregressive modification. *Quart. J. Roy. Meteor. Soc.*, **142**, 1385–1394, <https://doi.org/10.1002/qj.2741>.
- Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size of probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191, [https://doi.org/10.1175/1520-0434\(2002\)017<0173:TIOHRA>2.0.CO;2](https://doi.org/10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2).
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Rasp, S., and S. Lerch, 2018: Neural networks for post-processing ensemble weather forecasts. arXiv:1805.09091 [stat.ML], <https://arxiv.org/abs/1805.09091>.
- R Core Team, 2017: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>.
- Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, <https://doi.org/10.3402/tellusa.v55i1.12082>.
- Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, <https://doi.org/10.1002/qj.2183>.
- , and L. Büermann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. *J. Roy. Stat. Soc. Ser. C Appl. Stat.*, **63**, 405–422, <https://doi.org/10.1111/rssc.12040>.
- , and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- , and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9**, 1328–1349, <https://doi.org/10.1214/15-AOAS843>.
- Selten, R., 1998: Axiomatic characterization of the quadratic scoring rule. *Exp. Econ.*, **1**, 43–62, <https://doi.org/10.1023/A:1009957816843>.



- Stauffer, R., G. J. Mayr, J. W. Messner, N. Umlauf, and A. Zeileis, 2017: Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.*, **37**, 3264–3275, <https://doi.org/10.1002/joc.4913>.
- Stigler, S. M., 2007: The epic story of maximum likelihood. *Stat. Sci.*, **22**, 598–620, <https://doi.org/10.1214/07-STS249>.
- Student, 1908: The probable error of a mean. *Biometrika*, **6**, 1–25, <https://doi.org/10.2307/2331554>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability*, Shinfield Park, Reading, Berkshire, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.
- Thorarindottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc. Ser. A Stat. Soc.*, **173**, 371–388, <https://doi.org/10.1111/j.1467-985X.2009.00616.x>.
- Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson, 2006: Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophys. Res. Lett.*, **33**, 2–7, <https://doi.org/10.1029/2006GL027126>.
- White, H., 1994: The asymptotic distribution of the QMLE and the information matrix equality. *Estimation, Inference and Specification Analysis, Econometric Society Monogr.*, No. 22, Cambridge University Press, 88–129, <https://doi.org/10.1017/CCOL0521252806>.
- Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- Wilks, D., and T. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, <https://doi.org/10.1175/MWR3402.1>.
- Winkelmann, R., and S. Boes, 2006: *Analysis of Microdata*. Springer, 313 pp., <https://doi.org/10.1007/3-540-29607-7>.
- Yuen, R., and S. Stoev, 2014: CRPS M-estimation for max-stable models. *Extremes*, **17**, 387–410, <https://doi.org/10.1007/s10687-014-0185-x>.