universität innsbruck

Research Platform
Empirical and Experimental Economics

working paper

eeecon
[triple:e:con]

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

Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis

Working Papers in Economics and Statistics

2017-23

University of Innsbruck
http://uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

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

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

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

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

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

Manuel Gebetsberger
University of Innsbruck

Jakob W. Messner
University of Innsbruck

Georg J. Mayr
University of Innsbruck

Achim Zeileis
University of Innsbruck

Abstract

Non-homogeneous regression models are widely used to statistically post-process numerical ensemble weather prediction models. Such regression models are capable of forecasting full probability distributions and correct 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 non-homogeneous regression, where results show discrepancies for the classical Gaussian assumption. The heavy-tailed logistic and Student-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.

Keywords: ensemble post-processing, maximum likelihood, CRPS minimization, probabilistic forecasting, distributional regression models.

1. Introduction

Non-homogeneous regression is a popular regression-based technique to statistically correct an ensemble of numerical weather prediction models (NWP, Leith 1974). 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, Thorpe, and Brunet 2015) so that the raw forecasts are often biased and uncalibrated.

In statistical post-processing, various approaches have been developed to correct such ensembles (Roulston and Smith 2003; Raftery, Gneiting, Balabdaoui, and Polakowski 2005; Gneiting, Raftery, Westveld, and Goldman 2005; Wilks 2009) but none of them has appeared as a best single post-processing strategy (Wilks and Hamill 2007). However, non-homogenous 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; Thorarinsdottir and Gneiting 2010; Messner, Mayr, Wilks, and Zeileis 2014a; Messner, Mayr, Zeileis, and Wilks 2014b; Scheuerer 2014; Hemri, Haiden, and Pappenberger 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 Sec. 22.3) Gneiting *et al.* (2005) established NGR parameter estimation by minimizing the continuous ranked probability score (CRPS, Hersbach 2000). Post-processing studies for meteorological applications have used this estimation approach frequently since then (Raftery *et al.* 2005; Vrugt, Clark, Diks, Duan, and Robinson 2006; Hagedorn, Hamill, and Whitaker 2008; Scheuerer 2014; Scheuerer and Büermann 2014; Mohammadi, Rahmani, and Azadi 2015; Feldmann, Scheuerer, and Thorarinsdottir 2015; Scheuerer and Hamill 2015; Scheuerer and Möller 2015; Taillardat, Mestre, Zamo, and Naveau 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; Grimit, Gneiting, Berrocal, and Johnson 2006). Because of this higher sensitivity to outliers Gneiting *et al.* (2005) found LS minimization to lead to overdispersive forecasts.

Figure 1 (left graphic, "Gaussian") illustrates this overdispersion exemplarily for 2m air temperature forecasts, where NGR is employed at an Alpine site for +24h forecasts (see Sec. 33.1 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. Although the CRPS approach indicates a better calibration, 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 mis-specification of the NGR in terms of its distributional tail. The right plot in Figure 1 shows the PIT histograms of a non-homogeneous regression model with a heavier-tail Student-t 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 define when and why results from LS and CRPS minimization will differ. 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-t probability distributions. In

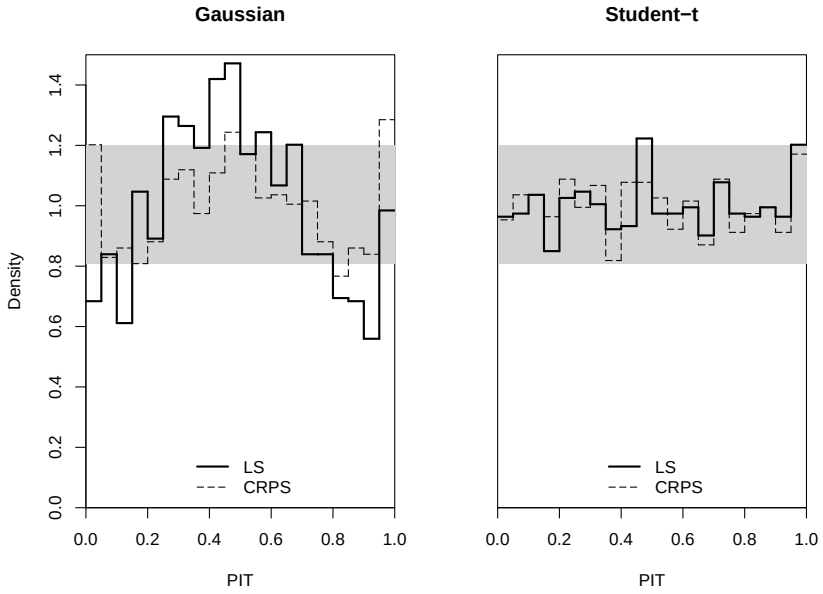

Figure 1: PIT histogram for temperature forecasts at an Alpine site at lead time +24h, shown for the Gaussian (left) and Student-t (right) models, estimated with LS (solid) or CRPS (dashed) minimization. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

particular, the Student-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 of the statistical models. Sections 3 and 4 present and discuss results for probabilistic temperature post-processing and synthetic simulations, respectively. Finally, Section 5 gives a conclusion.

2. Methods

This section briefly describes the distributions, along with the corresponding statistical models which 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.

2.1. Distributions used and density functions

In this article we employ three probability distributions with differences particularly on their tails (Figure 2, left). 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 μ and the scale parameter σ. Its density function $f_{\mathcal{N}}$ (Eq. 1) is symmetrical around

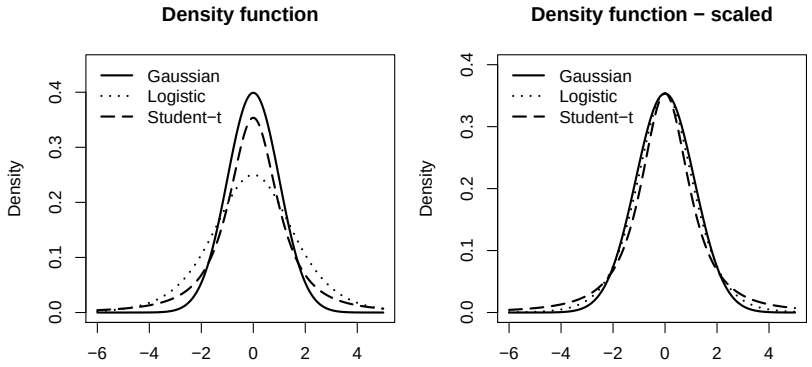

Figure 2: Probability density functions for a Gaussian (solid), logistic (dotted) and Student-t distribution (dashed) with $\mu = 0, \sigma = 1$ for Gaussian and logistic distributions, and the degree of freedom $\nu = 2$ for the Student-t distribution (left figure). Right figure illustrates scaled density values with respect to the Student-t distribution to highlight the tails.

μ (Figure 2, left), and is evaluated at the observed value y with

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

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

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

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

In addition to $\mathcal{N}$ and $\mathcal{L}$, we make use of the shifted scaled Student-t (Student-t in the following) distribution $\mathcal{S}(\mu,\sigma,\nu)$ (Student 1908), which, additionally to the location μ and scale σ parameters has a third parameter ν, 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, Γ denotes the gamma function. The degree of freedom ν controls the tails of the Student-t distribution with heavier tails for smaller ν values. In the limit of $\nu \to \infty$ the Student-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 (right) highlight the different tail behaviors. The logistic distribution has clearly heavier tails than the Gaussian distribution and with $\nu = 2$, the Student-t distribution can accommodate even heavier tails.

2.2. Regression models

As the basis regression model, we apply the non-homogeneous Gaussian regression approach (NGR) 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 (Equations 4–6), where the parameters μ_i, σ_i are used for the Gaussian and logistic assumptions, and μ_i, σ_i, ν_i for our representation of the Student-t distribution (Eq. 3).

$$\mu_i = \beta_0 + \beta_1 \cdot \overline{ens_i} \tag{4}$$

$$\log(\sigma_i) = \gamma_0 + \gamma_1 \cdot \log(SD_{ens,i}) \tag{5}$$

$$\log(\nu_i) = \delta_0 \tag{6}$$

The subscript i labels one observation. Commonly, the ensemble mean value $\overline{ens_i}$ is used as covariate for the location parameter μ_i (Eq. 4), and the ensemble standard deviation $SD_{ens,i}$ for the scale parameter σ_i (Eq. 5). The degree of freedom of the Student-t model is simply modeled by a constant intercept δ_0 and not dependent on any covariable. Note that the coefficients for σ_i and ν_i are estimated on the logarithmic scale in order to ensure the positivity of σ_i, ν_i.

The framework described by Equations 4, 5, and 6 is used in real data and simulation studies. For the real data studies, sine and cosine of the day of the year (DOY_i) are additionally included in the predictor of the location parameter μ_i, to better represent seasonal variation of temperature:

$$\mu_i = \beta_0 + \beta_1 \cdot \overline{ens_i} + \beta_2 \cdot \sin(DOY_i) + \beta_3 \cdot \cos(DOY_i) \tag{7}$$

Clearly, the framework of Equations 4–6 can be extended by including additional covariates and also non-linear terms (e.g., as in Stauffer, Mayr, Messner, Umlauf, and Zeileis 2017). Also, other probability distributions such as the generalized extreme value distribution (Scheuerer 2014) could be used in this framework. As a generalization, the defined models can be viewed as the distributional regression framework of Klein, Kneib, Lang, and Sohn (2015).

2.3. 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:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{N} q(y_i; \theta) \tag{8}$$

More generally, $\Theta = \mathbb{R}^p$ defines the parameter space with p being the number of regression coefficients, y_i an observed value, and N the number of observations in a training data set. In our specific regression framework, $\hat{\theta}$ includes all the estimated regression coefficients (β, γ, δ) as defined in Eq. 4–6. Estimators such as LS or CRPS should address the two properties of consistency and asymptotic normality:

$$\hat{\theta} \xrightarrow{p} \theta_0 \ \ as \ \ N \to \infty \tag{9}$$

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

Consistency derives from the law of large numbers (LLN), and normality from the central limit theory (CLT). An estimator is consistent if it approaches the true parameter θ_0 in probability as the sample size N increases to infinity (Eq. 9). This means that the probability of $|\hat{\theta} - \theta_0|$ to be larger than a certain value ϵ becomes zero. Furthermore, the difference $\hat{\theta} - \theta_0$ approaches a Gaussian distribution $\mathcal{N}$ (Eq. 10) with the variance $I(\theta)^{-1}$. $I(\theta)$ defines the Fisher information matrix and its inverse the smallest possible variance achievable for any consistent estimator (Winkelmann and Boes 2006).

Both properties can be mathematically proven for both estimators under certain regularity conditions (Winkelmann and Boes 2006; Yuen and Stoev 2014). Under strong conditions, the LS estimator is also the most efficient among all consistent estimators. 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.

The main difference between the scoring rules CRPS and LS is the penalization of individual observations, 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 $\log f$.

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

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

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 Figure 3. If a single observation is located on the distribution tails (above and below +/- 2), then larger differences between the scores can be found. The LS penalizes events on these tails more strongly than the CRPS.

2.4. 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 asses 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, Balabdaoui, and Raftery 2007), which should be uniformly distributed. This uniformity

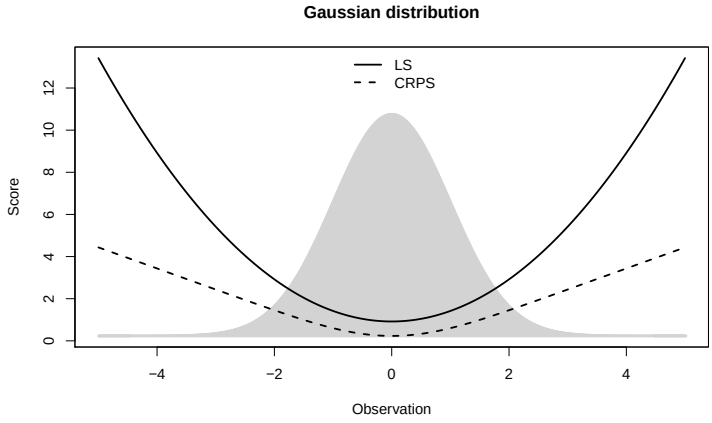

Figure 3: Continuous ranked probability score (CRPS, dashed) and log-score (LS, solid), evaluated at different (theoretical) observed values for an assumed Gaussian distribution with $\mu = 0, \sigma = 1$, with probability density values sketched as gray area.

derives from the statistical forecast consistency (calibration) which occurs if all forecasted probability bins count the same number of observations. However, due to the large number of individual PIT's, differences in those histograms will also be quantified by the reliability index (RI) which computes absolute differences from uniformity:

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

Herein, κ_k defines the relative number of observations in each bin k, and K 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 Equations 11 and 12.

To ensure independent test data for temperature forecasts, we perform a 10-fold cross-validation (CV). Therefore the data set is divided into 10 blocks and forecasts for each block are derived from models trained on the remaining 9 blocks. This leads to independent forecasts which are verified with PIW, PIC, RI, LS, CRPS. This approach is repeated for each lead time and station.

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 improve temperature forecasts. For simplicity, statistical models (Gaussian, logistic, Student-t) where CRPS or LS minimization is employed, will be referred to as CRPS or LS models, respectively.

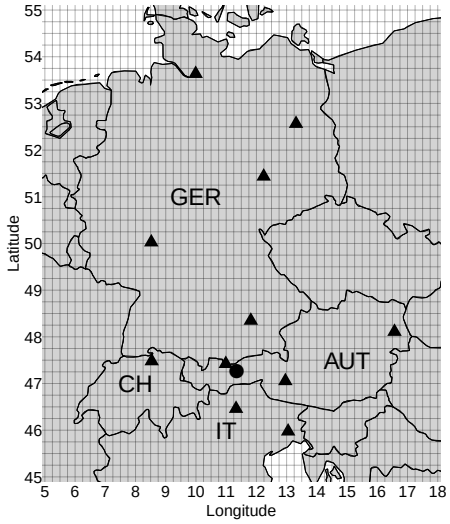

Figure 4: Study area with the sites in Austria (AUT), Italy (IT), Switzerland (CH), and Germany (GER): the filled circle represents the Alpine site which is used for the case study. The gray grid illustrates the underlying horizontal grid of the 50 member ECMWF ensemble forecasts.

3.1. Temperature data

Temperature records are used from 12 locations over Central Europe (Figure 4) for 3-hourly lead times from +6h to +96h in the time period between 2011–2016.

The corresponding ensemble forecasts of 2m air temperature are based on the 00 UTC initialization from the European Center of Medium Range Weather Forecasts (ECMWF). Overall, this yields 581076 observation/forecast-pairs to be validated, which include 311 different regression fits for different lead times at different stations.

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 lead to a cold bias, which can be seen when comparing observations with corresponding ensemble mean forecasts (Fig. 5, top left). 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, Vautard, and Strauss 1997; Hamill and Colucci 1998), which is shown for the bias-corrected ECMWF ensemble forecast for +24h in Fig. 5 (bottom left). Here, too many observations are counted below the lowest and above the highest member value (lowest and highest rank), indicating less forecast uncertainty than needed. Ideally the histogram should be uniformly distributed. These illustrated ensemble forecasts for +24h are the basis for later synthetic simulations, using the error characteristics for bias and underdisperision. 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.

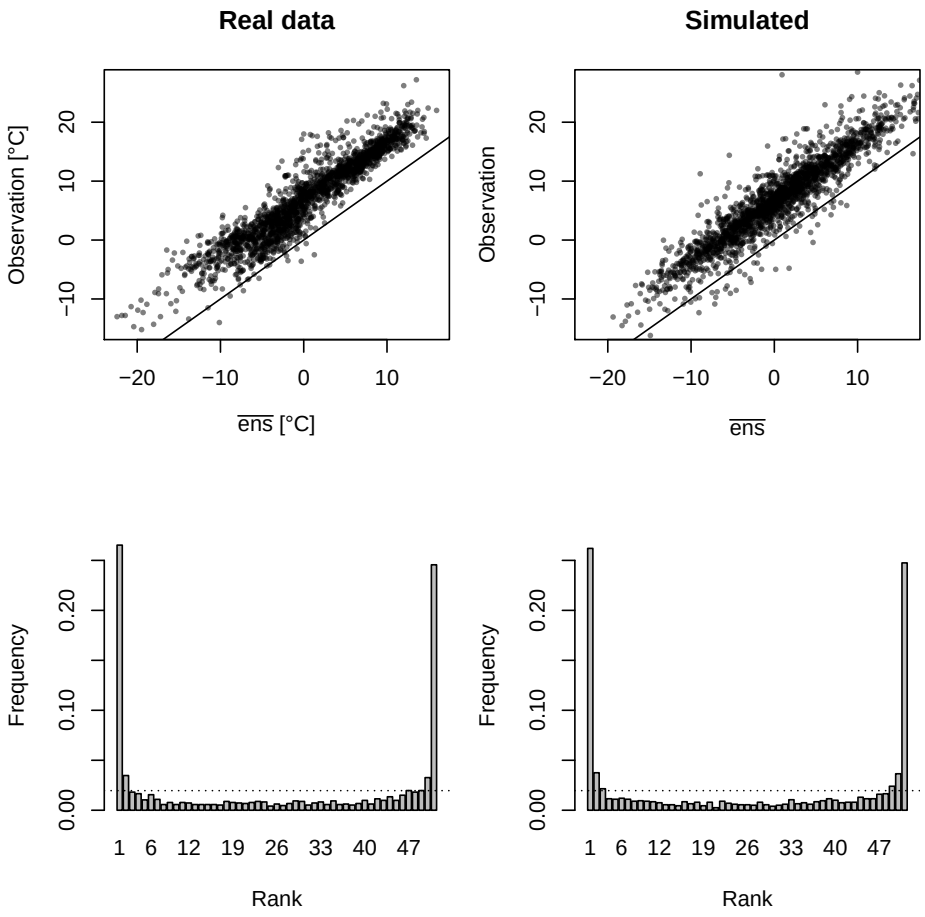

Figure 5: Error characteristics for real data at the Alpine site for +24h temperature forecasts (left) and simulated data (right). Top: Ensemble mean values $\overline{ens}$ against observed values. Bottom: Rank histograms of the bias-corrected 50 member ensembles. Dotted horizontal line indicates perfect calibration.

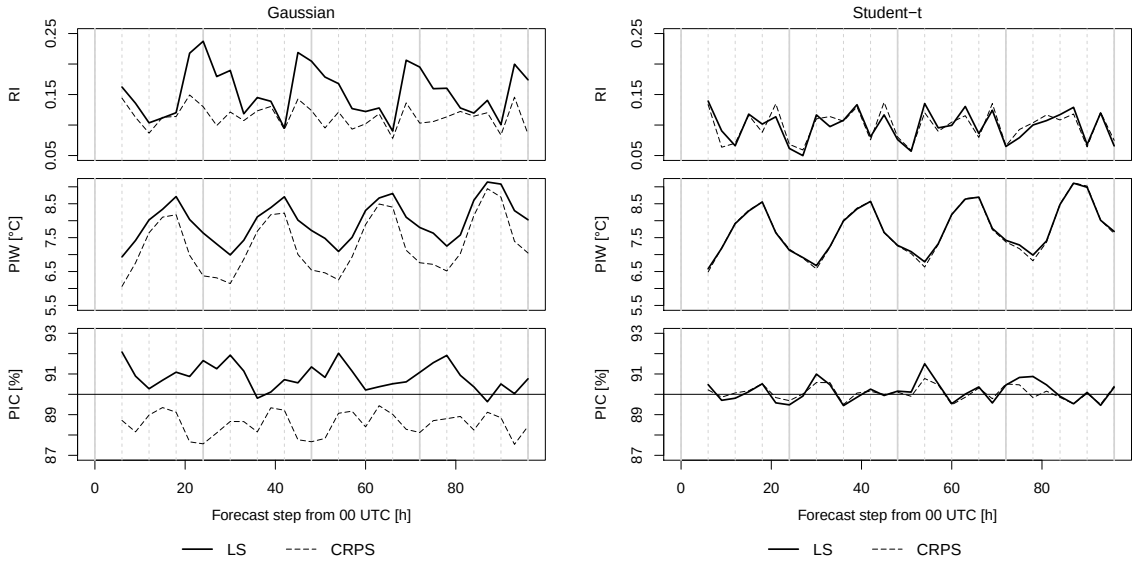

Figure 6: From top to bottom: Reliability index (RI), average width of the 90% prediction interval (PIW) and coverage of the 90% prediction interval (PIC), evaluated for Gaussian (left) and Student-t (right) models at the Alpine site from lead times +6h to +96h, estimated with LS (solid) or CRPS (dashed).

3.2. Alpine site case study

In this subsection we apply the regression framework, as defined in Equations 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, PIC for the Gaussian and Student-t models which are estimated with both approaches (CRPS or LS minimization). For the Gaussian models (left panel figures), there is a clear difference between the LS and CRPS model for certain lead times (e.g., +24h) where calibration in terms of RI (top figure) 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 (middle panel). Both estimation approaches are not giving optimum calibration regarding the 90% interval (bottom panel). The LS model covers too many observations in the 90% interval and the CRPS too few.

The PIT histograms, which are shown in Figure 1 for the +24h 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 bin-wise sampling variations. Thus, as long as the PIT lies within this interval the forecasts can be regarded as calibrated.

Regarding the Gaussian models (Fig. 1, "Gaussian"), 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 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. Nevertheless, both approaches do not obtain best possible calibration and differ in the forecasted distribution

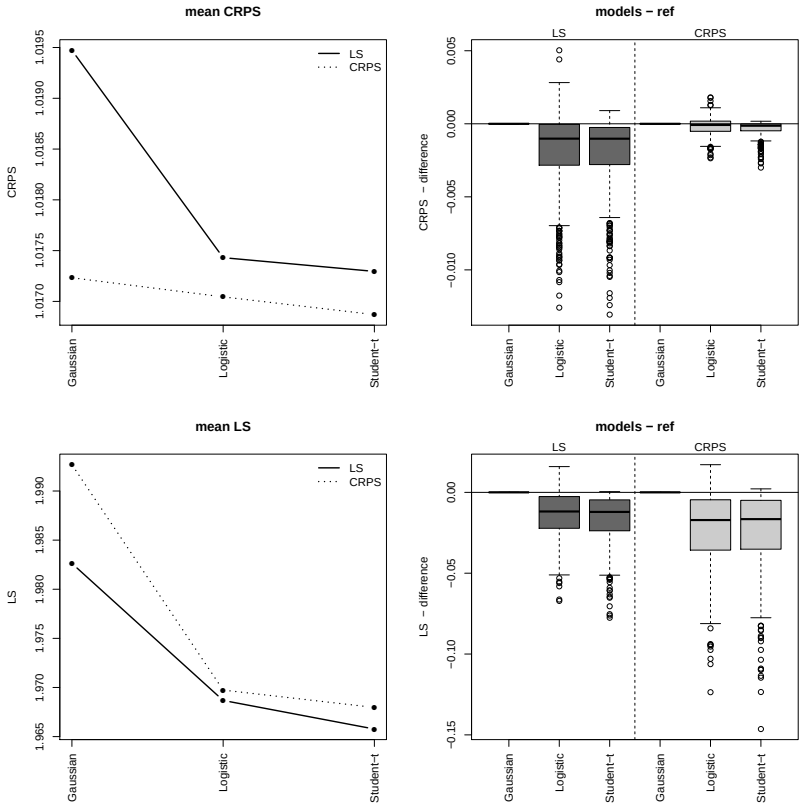

Figure 7: Mean scores and score differences (left to right) for LS-minimized (solid line and dark gray) and CRPS-minimized (dotted line and light gray) models, evaluated with the CRPS and LS (top to bottom). References are the Gaussian models for LS or CRPS minimization, respectively. Each boxplot contains 311 individual regression fits for each lead time and station separately.

parameters if the Gaussian distribution is assumed.

However, if the Student-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 Figure 6 (right panel figures). Regarding the overall calibration in terms of PIT, the example for +24h yields almost uniform histograms for the Student-t models for both minimization approaches (Fig. 1, "Student-t").

3.3. 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. Figure 7 summarizes 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 to the Gaussian CRPS models, respectively. Absolute differences are chosen rather than relative changes as skill scores cannot be computed for the LS (Gneiting *et al.* 2005).

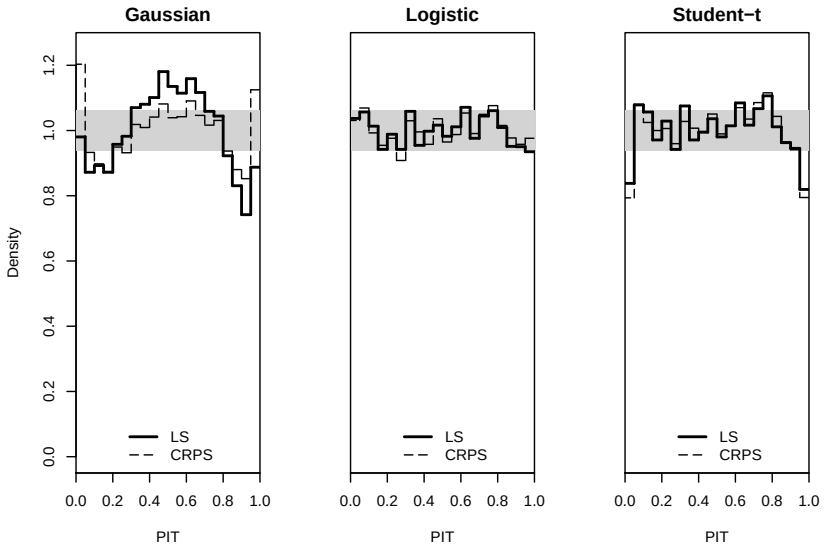

Figure 8: PIT value for Gaussian, logistic, and Student-t models (left to right) with LS (solid) or CRPS (dashed) minimization. Analysis includes 12 stations for lead times +18h. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

Figure 7 illustrates a clear difference for all individual regressions if heavy-tailed distribution models (logistic, Student-t) are applied (right panel figures). In terms of CRPS evaluation (Fig. 7, top), 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-t models.

A similar picture is visible for LS evaluation (Fig. 7, bottom). 84% and 82% of the evaluated logistic models show smaller LS values for CRPS or LS minimization, respectively. Student-t models obtain smaller LS values than the Gaussian benchmark for 93% (CRPS minimization) and 97% (LS minimization) of all regressions.

The mean scores over all the individual LS and CRPS emphasize the benefit of the heavy-tailed models where score values are smallest. Not surprisingly, CRPS models perform better in CRPS evaluation and LS models in LS evaluation (Fig. 7, left panel figures).

On average CRPS and LS, the Student-t models perform best. However results also imply that the logistic models already improve the benchmark. Hence, there are situations where the logistic models might be good enough and where the tail-flexibility of the Student-t model is not necessary.

An example of the good calibration of logistic models is shown in Figure 8, which consists of predictions for all stations at lead time +18h. Similarly to the Alpine site as shown in Fig. 1, the Gaussian models 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, the heavy tail of the logistic distribution leads to an almost perfect and similar calibration for both approaches (middle). Additionally, the heavy tail created by the Student-t models seems to be too heavy for this particular lead time where too few events occur on the tails (right).

The Student-t models can clearly improve calibration compared to the Gaussian models, but the tails are not captured appropriately. Possibly, the assumption of a constant ν in Eq. 6 is too simple, and a seasonal variation of ν as in Eq. 7 might be more reasonable.

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.

4.1. Simulated dataset

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

$$\overline{ens}_i = \mathcal{N}(0.35, 6.91) \tag{14}$$
$$\log(SD_{ens,i}) = \mathcal{N}(-0.56, 0.43) \tag{15}$$

with the distribution parameters taken from the empirical means and standard deviations of the ECMWF ensemble at the Alpine site (Sec. 3.1). Observations are simulated from logistic distributions, which we found in the previous subsection to describe temperature data quite well. The location (μ_i^{true}) and scale (σ_i^{true}) parameters of these distributions are modeled as functions of the simulated ensemble statistics $\overline{ens}_i$ and $SD_{ens,i}$

$$\mu_i^{true} = \beta_0^{true} + \beta_1^{true} \cdot \overline{ens}_i \tag{16}$$
$$\log(\sigma_i^{true}) = \gamma_0^{true} + \gamma_1^{true} \cdot \log(SD_{ens,i}) \tag{17}$$

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

Thus, a data set of length 5000 is available with forecasts and corresponding observations that have similar properties as the real data used in Section 3.1. However, different to the real data the true coefficients $\beta_0^{true}, \beta_1^{true}, \gamma_0^{true}, \gamma_1^{true}$ are known and can directly be compared to estimated coefficients $\beta_0, \beta_1, \gamma_0, \gamma_1$ from non-homogeneous regression models of the form of Equations 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.

4.2. Simulation results

Figure 9 (left) 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 ($\boldsymbol{\beta}$) on median. However, differences occur in the scale submodel ($\boldsymbol{\gamma}$). Although the slope coefficient γ_1 expresses the true value on median, clear differences can be found for the intercept γ_0. Both approaches do not calculate the true coefficient of 0.9 and estimate

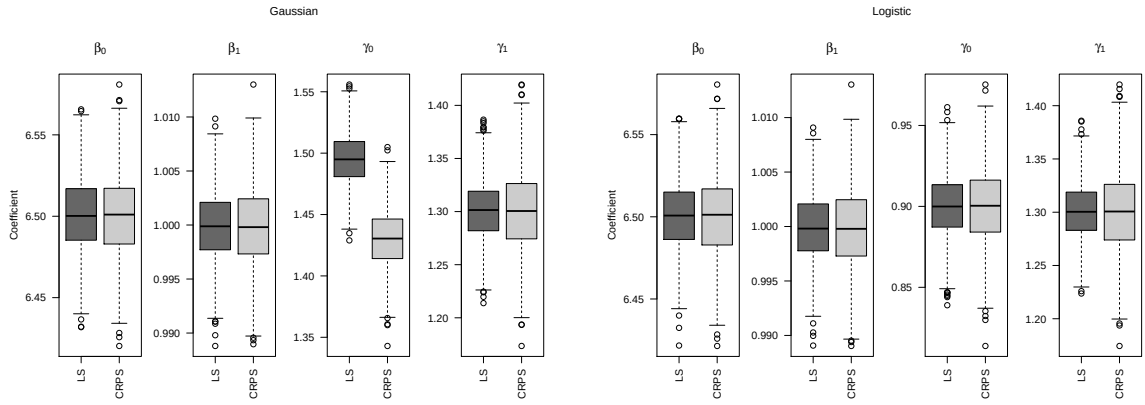

Figure 9: The estimated regression coefficients $(\beta_0, \beta_1, \gamma_0, \gamma_1)$ for the Gaussian models (left) and logistic models (right), estimated with LS (dark gray) or CRPS (light gray) minimization, respectively. Boxplots are based on the bootstrap procedure of repeating the simulation 1000 times and illustrate the interquartile range (0.25-0.75) in boxes, whiskers for +/- 1.5 times interquartile range, and outliers in solid circles.

a larger value. This is mainly the consequence of the scaling by approximately 1.8 since the standard deviation of the logistic distribution is e.g., $1.8 \cdot 0.9 = 1.62$.

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. In order 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 (Figure 9, right), 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, Figure 10 shows PIT histograms of the different models for different lengths of the simulated data sets. As expected and similar to the real data case study, the Gaussian "forecasts" humps at central PIT values show the lack of calibration (top left). 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 pronounced. In contrast, the difference between the estimation approaches becomes smaller if the correct (and known) logistic response distribution is assumed (Figure 10, top right).

As expected from estimation theory, the differences vanish with increasing sample size for the correct distribution assumption, and show a well defined W-shape for the wrong assumption (Figure 10, middle and bottom).

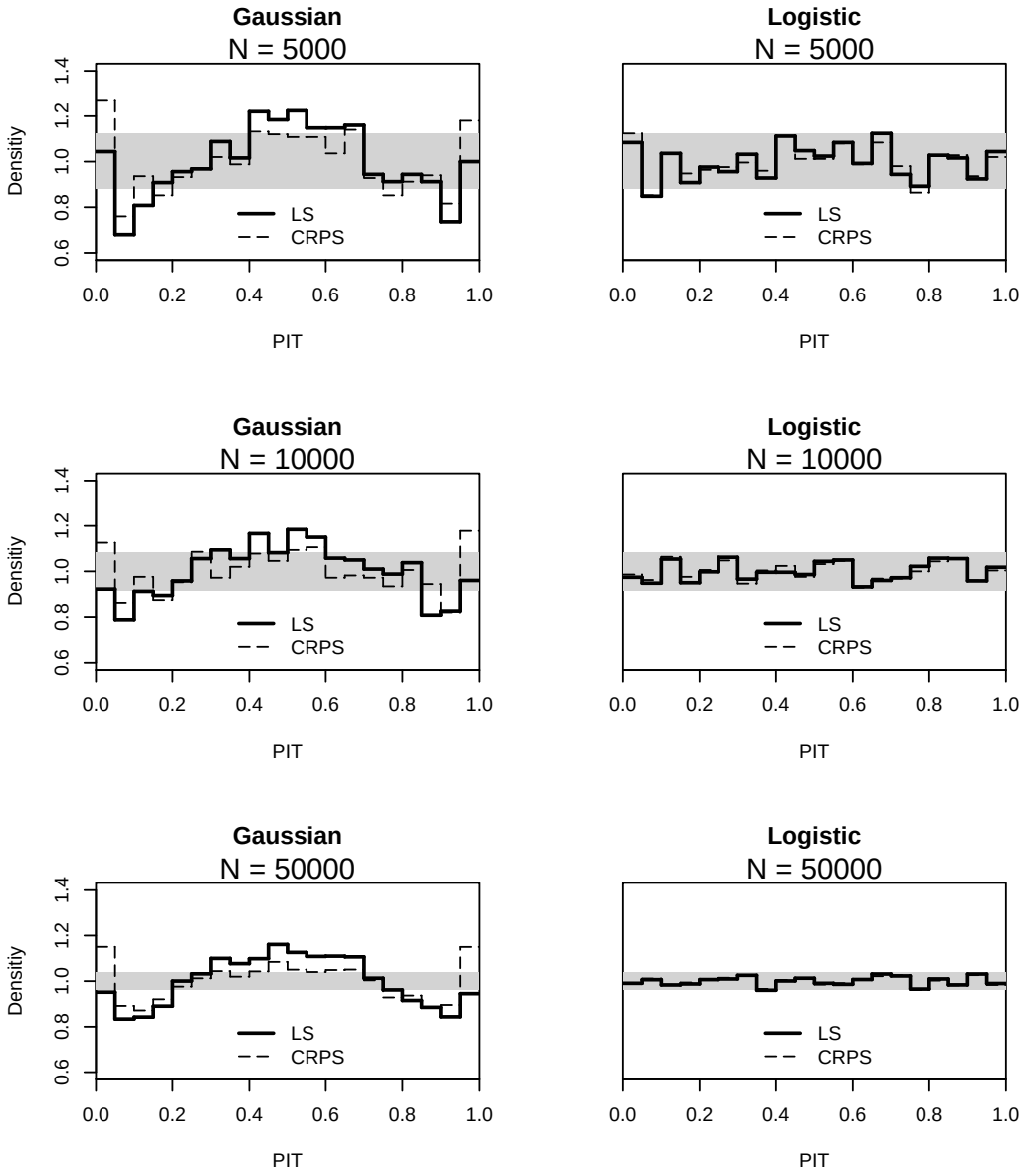

Figure 10: Calibration in terms of PIT values for one simulation with N=5000, 10000, and 50000 data (top to bottom) using the Gaussian or logistic model (left to right), estimated with LS (solid) or CRPS (dashed) minimization. The gray area illustrates the 95% consistency interval around perfect calibration, which should be 1. Binning is based on 5% intervals.

5. Conclusion

Non-homogeneous regression is a commonly used post-processing strategy to statistically correct NWP ensemble forecasts. This approach predicts full parametric forecast distributions for the weather quantities of interest. In order 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 data from 12 stations in Central Europe and in a simulation study.

In principle, LS and CRPS minimization differently penalize 'extreme' events or events which deviate larger from the mean forecast, respectively. Consequently, the assumed forecast distribution plays a crucial rule 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 at the expenses of a decreased calibration.

CRPS minimization apparently improves calibration, but only when looking 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, and if differences between the approaches occur.

To account for a potentially heavier tail, this study introduces and compares the logistic and Student-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-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.

6. Acknowledgments

This work is part of a PhD - project and funded by Autonome Provinz Bozen - Abteilung Bildungsförderung, Universität und Forschung (Nr. ORBZ110725). We thank the Austrian weather Service (ZAMG) for access to ECMWF EPS data.

A. Computational Details

The estimation of regression coefficients is performed in *R* (R Core Team 2017) using the *crch* package (Messner, Mayr, and Zeileis 2016), which is able to perform minimization of the CRPS or LS. Closed expressions of the CRPS for the Gaussian, logistic, and Student-t distribution are based on the *scoringRules* package (Jordan, Krüger, and Lerch 2017).

References

Aldrich J (1997). "R. A. Fisher and the making of maximum likelihood 1912-1922." *Statistical Science*, **12**(3), 162–176. doi:10.1214/ss/1030037906.

Anderson JL (1996). "A method for producing and evaluating probabilistic forecast from ensemble model integration." *Journal of Climate*, **9**, 1518–1530. ISSN 0894-8755. doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Bauer P, Thorpe A, Brunet G (2015). "The quiet revolution of numerical weather prediction." *Nature*, **525**, 47–55. doi:10.1038/nature14956.

Bröcker J, Smith La (2007). "Increasing the Reliability of Reliability Diagrams." *Weather and Forecasting*, **22**, 651–661. doi:10.1175/WAF993.1.

Casella G, Berger RL (2002). *Statistical inference*. 2 edition. Thomson Learning.

Feldmann K, Scheuerer M, Thorarinsdottir TL (2015). "Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression." *Monthly Weather Review*, **143**(2005), 955–971. doi:10.1175/MWR-D-14-00210.1.

Gneiting T, Balabdaoui F, Raftery AE (2007). "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 243–268. doi:10.1111/j.1467-9868.2007.00587.x.

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation." *Monthly Weather Review*, **133**, 1098–1118. ISSN 1520-0493. doi:10.1175/MWR2904.1.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006). "The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification." *Quarterly Journal of the Royal Meteorological Society*, **132**, 2925–2942. doi:10.1256/qj.05.235.

Hagedorn R, Hamill T, Whitaker J (2008). "Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: two-meter temperatures." *Monthly Weather Review*, **136**(7), 2608–2619. doi:10.1175/2007MWR2410.1.

Hamill TM, Colucci SJ (1998). "Evaluation of Eta RSM ensemble probabilistic precipitation forecasts." *Monthly Weather Review*, **126**, 711–724. doi:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

Hemri S, Haiden T, Pappenberger F (2016). "Discrete Postprocessing of Total Cloud Cover Ensemble Forecasts." *Monthly Weather Review*, **144**(7), 2565–2577. doi:10.1175/MWR-D-15-0426.1.

Hersbach H (2000). "Decomposition of the continuous ranked probability score for ensemble prediction systems." *Weather and Forecasting*, **15**, 559–570. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Huber PJ (1967). "The behavior of maximum likelihood estimates under nonstandard conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 221–233.

Jordan A, Krüger F, Lerch S (2017). *Evaluating probabilistic forecasts with the scoringRules package*. URL https://cran.r-project.org/web/packages/scoringRules/vignettes/crpsformulas.html.

Klein N, Kneib T, Lang S, Sohn A (2015). "Bayesian structured additive distributional regression with an application to regional income inequality in Germany." *The Annals of Applied Statistics*, **9**(2), 1024–1052. doi:10.1214/15-AOAS823.

Leith C (1974). "Theoretical skill of Monte Carlo forecasts." *Monthly Weather Review*, **102**, 409–418. doi:10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lorenz EN (1963). "Deterministic nonperiodic flow." *Journal of the Atmospheric Sciences*, **20**, 130–141. doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Messner JW, Mayr GJ, Wilks DS, Zeileis A (2014a). "Extending extended logistic regression: Extended versus separate versus ordered versus censored." *Monthly Weather Review*, **142**, 3003–3014. doi:10.1175/MWR-D-13-00355.1.

Messner JW, Mayr GJ, Zeileis A (2016). "Heteroscedastic Censored and Truncated Regression with crch." *The R Journal*, **8**(1), 173–181.

Messner JW, Mayr GJ, Zeileis A, Wilks DS (2014b). "Heteroscedastic extended logistic regression for postprocessing of ensemble guidance." *Monthly Weather Review*, **142**, 448–456. ISSN 0027-0644. doi:10.1175/MWR-D-13-00271.1.

Mohammadi SA, Rahmani M, Azadi M (2015). "Optimization of continuous ranked probability score using PSO." *Decision Science Letters*, **4**, 373–378. doi:10.5267/j.dsl.2015.4.001.

Möller A, GroßJ (2016). "Probabilistic temperature forecasting based on an ensemble autoregressive modification." *Quarterly Journal of the Royal Meteorological Society*, **142**(696), 1385–1394. doi:10.1002/qj.2741.

Mullen SL, Buizza R (2002). "The Impact of Horizontal Resolution and Ensemble Size of Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System." *Weather and Forecasting*, **17**, 173–191. doi:10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005). "Using Bayesian model averaging to calibrate forecast ensembles." *Monthly Weather Review*, **133**, 1155–1174. ISSN 0027-0644. doi:10.1175/MWR2906.1.

Roulston MS, Smith LA (2003). "Combining dynamical and statistical ensembles." *Tellus*, **55A**, 16–30. doi:10.1034/j.1600-0870.2003.201378.x.

Scheuerer M (2014). "Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics." *Quarterly Journal of the Royal Meteorological Society*, **140**, 1086–1096. doi:10.1002/qj.2183.

Scheuerer M, Büermann L (2014). "Spatially adaptive post-processing of ensemble forecasts for temperature." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 405–422. doi:10.1111/rssc.12040.

Scheuerer M, Hamill TM (2015). "Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions." *Monthly Weather Review*, **143**, 4578–4596. doi:10.1175/MWR-D-15-0061.1.

Scheuerer M, Möller D (2015). "Probabilistic wind speed forecasting on a grid based on ensemble model output statistics." *Annals of Applied Statistics*, **9**(3), 1328–1349. doi:10.1214/15-AOAS843.

Selten R (1998). "Axiomatic characterization of the quadratic scoring rule." *Experimental Economics*, **1**(1), 43–62. doi:10.1007/BF01426214.

Stauffer R, Mayr GJ, Messner JW, Umlauf N, Zeileis A (2017). "Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model." *International Journal of Climatology*, **37**(7), 3264–3275. ISSN 1097-0088. doi:10.1002/joc.4913.

Stigler SM (2007). "The epic story of maximum likelihood." *Statistical Science*, **22**(4), 598–620. doi:10.1214/07-STS249.

Student (1908). "The probable error of a mean." *Biometrika*, pp. 1–25.

Taillardat M, Mestre O, Zamo M, Naveau P (2016). "Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics." *Monthly Weather Review*, **144**, 2375–2393. doi:10.1175/MWR-D-15-0260.1.

Talagrand O, Vautard R, Strauss B (1997). "Evaluation of probabilistic prediction systems." In *Proceeding of workshop on predictability, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, UK*, pp. 1–25.

Thorarinsdottir TL, Gneiting T (2010). "Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173**, 371–388. doi:10.1111/j.1467-985X.2009.00616.x.

Vrugt JA, Clark MP, Diks CGH, Duan Q, Robinson BA (2006). "Multi-objective calibration of forecast ensembles using Bayesian model averaging." *Geophysical Research Letters*, **33**(19), 2–7. doi:10.1029/2006GL027126.

White H (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press. doi:10.1017/CCOL0521252806.

Wilks D, Hamill T (2007). "Comparison of Ensemble-MOS Methods Using GFS Reforecasts." *Monthly Weather Review*, **135**, 2379–2390. doi:10.1175/MWR3402.1.

Wilks DS (2009). "Extending logistic regression to provide full-probability-distribution MOS forecasts." *Meteorological Applications*, **16**, 361–368. doi:10.1002/met.134.

Winkelmann R, Boes S (2006). *Analysis of Microdata*. Springer. doi:10.1007/3-540-29607-7.

Yuen R, Stoev S (2014). "CRPS M-estimation for max-stable models." *Extremes*, **17**(3), 387–410. doi:10.1007/s10687-014-0185-x.

Affiliation:

Manuel Gebetsberger, Georg J. Mayr
Institute of Atmospheric and Cryospheric Sciences
Faculty of Geo- and Atmospheric Sciences
Universität Innsbruck
Innrain 52
6020 Innsbruck, Austria
E-mail: Manuel.Gebetsberger@uibk.ac.at, Georg.Mayr@uibk.ac.at

Achim Zeileis, Jakob W. Messner
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@uibk.ac.at,jwmm@elektro.dtu.dk

Jakob W. Messner
Department of Electrical Engineering
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark
E-mail: jwmm@elektro.dtu.dk

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

http://uibk.ac.at/eeecon/wopec/

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

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

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

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

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

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

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

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

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

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

2017-13 **Thomas Kneib, Nikolaus Umlauf:** A Primer on Bayesian Distributional Regression

2017-12 **Susanne Berger, Nathaniel Graham, Achim Zeileis:** Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R

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

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

2017-09 **Andreas Exenberger:** The Logic of Inequality Extraction: An Application to Gini and Top Incomes Data

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

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

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

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

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

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

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

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

2016-33 **Alexander Razen, Stefan Lang, Judith Santer:** Estimation of spatially correlated random scaling factors based on Markov random field priors

2016-32 **Meike Köhler, Nikolaus Umlauf, Andreas Beyerlein, Christiane Winkler, Anette-Gabriele Ziegler, Sonja Greven:** Flexible Bayesian additive joint models with an application to type 1 diabetes research

2016-31 **Markus Dabernig, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Simultaneous ensemble post-processing for multiple lead times with standardized anomalies

2016-30 **Alexander Razen, Stefan Lang:** Random scaling factors in Bayesian distributional regression models with an application to real estate data

2016-29 **Glenn Dutcher, Daniela Glätzle-Rützler, Dmitry Ryvkin:** Don't hate the player, hate the game: Uncovering the foundations of cheating in contests

2016-28 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Tricks for improving non-homogeneous regression for probabilistic precipitation forecasts: Perfect predictions, heavy tails, and link functions

2016-27 **Michael Razen, Matthias Stefan:** Greed: Taking a deadly sin to the lab

2016-26 **Florian Wickelmaier, Achim Zeileis:** Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models

2016-25 **Michel Philipp, Carolin Strobl, Jimmy de la Torre, Achim Zeileis:** On the estimation of standard errors in cognitive diagnosis models

2016-24 **Florian Lindner, Julia Rose:** No need for more time: Intertemporal allocation decisions under time pressure

2016-23 **Christoph Eder, Martin Halla:** The long-lasting shadow of the allied occupation of Austria on its spatial equilibrium

2016-22 **Christoph Eder:** Missing men: World War II casualties and structural change

2016-21 **Reto Stauffer, Jakob Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies *published in Monthly Weather Review*

2016-20 **Christina Bannier, Eberhard Feess, Natalie Packham, Markus Walzl:** Incentive schemes, private information and the double-edged role of competition for agents

2016-19 **Martin Geiger, Richard Hule:** Correlation and coordination risk

2016-18 **Yola Engler, Rudolf Kerschbamer, Lionel Page:** Why did he do that? Using counterfactuals to study the effect of intentions in extensive form games

2016-17 **Yola Engler, Rudolf Kerschbamer, Lionel Page:** Guilt-averse or reciprocal? Looking at behavioural motivations in the trust game

2016-16 **Esther Blanco, Tobias Haller, James M. Walker:** Provision of public goods: Unconditional and conditional donations from outsiders

2016-15 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** Predictive bookmaker consensus model for the UEFA Euro 2016

2016-14 **Martin Halla, Harald Mayr, Gerald J. Pruckner, Pilar García-Gómez:** Cutting fertility? The effect of Cesarean deliveries on subsequent fertility and maternal labor supply

2016-13 **Wolfgang Frimmel, Martin Halla, Rudolf Winter-Ebmer:** How does parental divorce affect children's long-term outcomes?

2016-12 **Michael Kirchler, Stefan Palan:** Immaterial and monetary gifts in economic transactions. Evidence from the field

2016-11 **Michel Philipp, Achim Zeileis, Carolin Strobl:** A toolkit for stability assessment of tree-based learners

2016-10 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** Affirmative action or just discrimination? A study on the endogenous emergence of quotas *published in Journal of Economic Behavior and Organization*

2016-09 **Loukas Balafoutas, Helena Fornwagner:** The limits of guilt

2016-08 **Markus Dabernig, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Spatial ensemble post-processing with standardized anomalies

2016-07 **Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model

2016-06 **Michael Razen, Jürgen Huber, Michael Kirchler:** Cash inflow and trading horizon in asset markets

2016-05 **Ting Wang, Carolin Strobl, Achim Zeileis, Edgar C. Merkle:** Score-based tests of differential item functioning in the two-parameter model

2016-04 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Non-homogeneous boosting for predictor selection in ensemble post-processing

2016-03 **Dietmar Fehr, Matthias Sutter:** Gossip and the efficiency of interactions

2016-02 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Rankings and risk-taking in the finance industry

2016-01 **Sibylle Puntscher, Janette Walde, Gottfried Tappeiner:** Do methodical traps lead to wrong development strategies for welfare? A multilevel approach considering heterogeneity across industrialized and developing countries

University of Innsbruck

Working Papers in Economics and Statistics

2017-23

Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis

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

Abstract
Non-homogeneous regression models are widely used to statistically post-process numerical ensemble weather prediction models. Such regression models are capable of forecasting full probability distributions and correct 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 non-homogeneous regression, where results show discrepancies for the classical Gaussian assumption. The heavy-tailed logistic and Student-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.

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