Decomposition and graphical portrayal of the quantile score

Sabrina Bentzien* and Petra Friederichs

Meteorological Institute, University of Bonn, and Hans-Ertel-Centre for Weather Research

Abstract

This study expands the pool of verification methods for probabilistic weather and climate predictions by a decomposition of the quantile score (QS). The QS is a proper score function and evaluates predictive quantiles on a set of forecast-observation pairs. We introduce a decomposition of the QS in reliability, resolution and uncertainty, and discuss the biases of the decomposition. Further, a reliability diagram for quantile forecasts is presented. Verification with the QS and its decomposition is illustrated on precipitation forecasts derived from the mesoscale weather prediction ensemble COSMO-DE-EPS of the German Meteorological Service. We argue that the QS is ready to become as popular as the Brier score in forecast verification.

Keywords: probabilistic forecasting, forecast verification, quantile score, score decomposition, reliability, resolution, ensemble forecasting

Version: 18.10.2013, submitted to Quarterly Journal of the Royal Meteorological Society

1 Introduction

Recently, forecast verification for probabilistic weather and climate predictions has seen great advances and interest. This is due to the increasing number of ensemble weather and climate prediction systems (EPS) on a variety of scales ranging from high-resolution limited area models of the atmospheric mesoscale to climate models on global scales. An EPS not only issues a deterministic future state of the atmosphere but a sample of possible future states. Ensemble postprocessing then translates such a sample of forecasts into probabilistic measures, e.g., in terms of predictive density or distribution functions, probabilities for threshold excesses or quantiles at given probability levels.

Predictive skill is generally assessed using score functions which apply to a set of forecast-observation pairs. An important characteristic of a score function for probabilistic forecasts

* *Corresponding author address:*
Sabrina Bentzien, Meteorological Institute, University of Bonn, Auf dem Huegel 20, 53121 Bonn, Germany
E-mail: bentzien@uni-bonn.de

is its propriety (Murphy and Epstein, 1967). Only proper score functions guarantee honesty and prevent hedging. The choice of an appropriate proper score function depends on the kind of probabilistic measure. Gneiting and Raftery (2007) reviews a variety of proper score functions, e.g., the Brier score (BS, Brier, 1950) or logarithmic score (Good, 1952) for probability forecasts, and the continuous ranked probability score (CRPS, Matheson and Winkler, 1976) to evaluate predictive distributions. Probabilistic predictions in terms of quantiles for a given probability level are evaluated using the quantile score (QS, Koenker and Machado, 1999; Gneiting and Raftery, 2007; Friederichs and Hense, 2007) which sums over a weighted absolute error between forecasts and observations. The advantage of a quantile representation of probabilistic forecasts is that probability levels can be defined without prior knowledge of the range of data. Further, they are more intuitive for end users.

It is shown that the CRPS corresponds to the integral of the BS over all thresholds, or likewise the integral of the QS over all probability levels (Laio and Tamea, 2007; Gneiting and Ranjan, 2011; Friederichs and Thorarinsdottir, 2012). The CRPS thus averages over the complete range of forecast thresholds and probability levels. Deficiencies in different parts of the distribution, e. g. the tail of a distribution, might be hidden. This has also been shown by Bentzien and Friederichs (2012). Here, a Gamma, log-normal and inverse-Gaussian distribution was fitted to precipitation data. The CRPS was similar for all three distributions. The evaluation of quantile forecasts, however, revealed the deficiencies of the different distributions with respect to the data. Thus, we highly recommend to extend the verification by considering the QS for different probability levels. This would give a more complete picture of forecast performance. In a similar sense, Gneiting and Ranjan (2011) proposed to compare density forecasts using threshold- and quantile-weighted scoring rules.

A comprehensive concept of forecast evaluation is proposed by A. H. Murphy and co-authors (e.g., Murphy, 1973; Murphy and Winkler, 1987; Murphy, 1993, 1996). Murphy and Winkler (1987) define a general framework for forecast verification based on the joint distribution of forecasts and observations. Murphy and Winkler (1987) and Murphy (1993) emphasize two aspects of prediction quality: reliability and resolution. For a perfectly reliable or calibrated forecast, the predictive distribution should equal the unknown distribution of the observation conditional on the forecast. The resolution, in turn, is related to the information content of a prediction system. It describes the variability of the observations under different forecasts, and indicates whether a probabilistic forecast can discriminate between different outcomes of an observation. Pointing in a similar direction, Gneiting et al. (2007) noted that "predictive performance (that) is based on the paradigm of maximizing the sharpness of the predictive distributions subject to calibration". It is thus of great interest to assess calibration and sharpness, or its related measures reliability and resolution.

Proper score functions are well suited to quantify these measures, since they can be decomposed into terms representing uncertainty, reliability, and resolution (e.g., DeGroot and Fienberg, 1983; Bröcker, 2009). They provide additional insight into the performance of a prediction system. Moreover, reliability diagrams may indicate important deficiencies. Up to date, decompositions of proper score functions have been proposed for the BS (Murphy, 1973), the CRPS (Hersbach, 2000; Candille and Talagrand, 2005), and the logarithmic score (Tödter and Ahrens, 2012).

This study presents the decomposition for the QS, and proposes an estimation procedure for the reliability, resolution and uncertainty parts. Moreover, we derive a reliability diagram

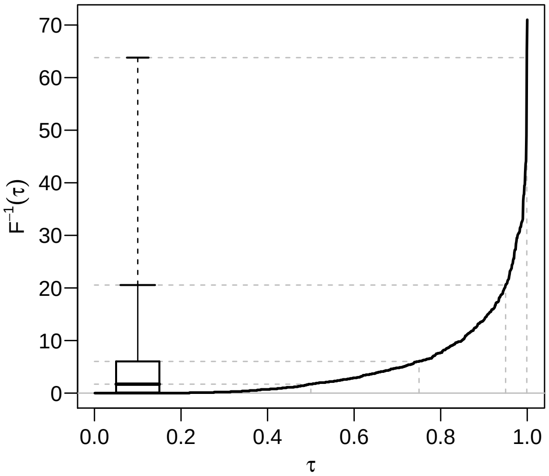

Figure 1: Empirical quantiles for precipitation [mm/12h] observed on June 6, 2011 between 12-24 UTC at 1079 rain gauges in Germany. A boxplot is presented in terms of the 0.25, 0.5, 0.75, 0.95 and 0.999 quantile.

for quantile forecasts and discus potential biases. We now dispose of a framework to evaluate the reliability and resolution components of a forecast system that issues predictive quantiles. In order to demonstrate the utility of the QS decomposition, we investigate quantile forecasts based on the mesoscale weather forecast ensemble COSMO-DE-EPS (Gebhardt et al., 2011; Peralta et al., 2012) operated by the German Meteorological Service (DWD).

The article is organized as follows. A brief description of quantile forecasts and their verification is given in Section 2. Section 3 reviews the concept of proper score functions and their decomposition. In Section 4 we derive the decomposition for the QS. Application to quantile forecasts for precipitation is given in Section 5. We summarize the study in Section 6, and discuss potential biases in an appendix.

2 Quantile forecasts

The τ-quantile of a random variable Y with cumulative distribution function $F(y) = Pr(Y \leq y)$ is given by

$$F^{-1}(\tau) = \inf \{y : F(y) \geq \tau\} \ ,$$

for all $0 \leq \tau \leq 1$. $q_\tau = F^{-1}(\tau)$ represents the τ-quantile at a probability level τ. Boxplots are very intuitive tools to graphically represent probabilistic information and uncertainty. They show quantile values as illustrated in Fig. 1. It shows the quantile function of precipitation observed on June 6, 2011 at 1079 weather stations in Germany. 78% measured precipitation above zero. Since precipitation cannot be negative, the quantiles at $\tau \leq 0.22$ are zero. In statistics it is common to say that precipitation is censored at zero. The distribution in Fig. 1 is largely skewed towards higher quantiles, reflecting the large variability of heavy precipitation on June 6, 2011. Note that a probability evaluation at a-priori selected thresholds might be inappropriate to describe the tail behaviour of heavy precipitation.

Following Koenker and Bassett (1978) and Koenker (2005), quantiles also arise from an optimization problem based on the check loss function

$$\rho_\tau(v) = v(\tau - I_{[v<0]}) = \begin{cases} v\tau & \text{if } v \geq 0 \\ v(\tau - 1) & \text{if } v < 0\,. \end{cases} \tag{1}$$

Here, $I_{[.]}$ is an indicator function which is 1 if the condition in brackets is true and zero otherwise. If v is the difference between a random event y and a quantile estimate $\hat{q}_\tau$, the $\rho_\tau(v)$ can be regarded as a distance measure. The absolute error $\mid y - \hat{q}_\tau \mid$ is then weighted with τ if the quantile estimate does not exceed y, and $1-\tau$ otherwise. The minimum of the expectation $E_F[\rho_\tau(Y - \hat{q}_\tau)]$ with respect to F is obtained if $\hat{q}_\tau = F^{-1}(\tau)$ is the true quantile of Y (see Koenker, 2005, pp. 5-6). This property makes the check function an ideal candidate for the estimation (e.g. quantile regression) or verification of quantiles (see Friederichs and Hense, 2008, appendix). For a more technical description as well as graphical representation of the check function the reader is referred to Koenker (2005).

3 Proper score functions

The concept of proper score functions goes back to Brier (1950), Good (1952) and Brown (1970), with a more technically description in Savage (1971) and Schervish (1989). It was reviewed and advanced by Matheson and Winkler (1976) and Gneiting and Raftery (2007). A *score function* projects the probabilistic forecast P from a set of possible probabilistic predictions $\mathcal{P}$ and an element of the sample space $\omega \in \Omega$ onto a real value $s(P, \omega) \in \mathbb{R}$. Probabilistic forecasts may be issued in terms of a cumulative distribution functions, probability density functions, probabilities of threshold excess or quantiles at given probability levels. This study defines a score function as a cost function that a forecaster intends to minimize, but it may equally be defined as a reward that a forecaster aims at maximizing (e.g., Gneiting and Raftery, 2007).

Consider now pairs of observations $\omega \in \Omega$ and forecasts $P \in \mathcal{P}$ issued by a probabilistic forecasting scheme. The score of the forecasting scheme is determined by the joint distribution $F(P, \omega)$, and its expected overall score is

$$\mathbf{S} = \int_{\mathcal{P}} \int_{\Omega} s(P, \omega)\, dF(P, \omega)\,.$$

Following the calibration-refinement factorization proposed by Murphy and Winkler (1987), the joint distribution is reformulated as $F(P, \omega) = F(\omega \mid P)\, F(P)$ where $F(P)$ is the marginal distribution of the forecasts. The distribution of ω for a fixed P is given by $Q(\omega) = F(\omega \mid P)$. The expected overall score of the forecasting scheme then reads

$$\begin{aligned} \mathbf{S} &= \int_{\mathcal{P}} \int_{\Omega} s(P, \omega)\, dQ(\omega)\, dF(P) \\ &= \int_{\mathcal{P}} S(P, Q)\, dF(P)\,, \end{aligned}$$

where $S(P, Q)$ is the *expected score* given a specific forecast P

$$S(P, Q) = \int_\Omega s(P, \omega)\, dQ(\omega)\,.$$

The score function is proper if $S(Q, Q) \leq S(P, Q)$ for all P, and strictly proper if equality is given if and only if $P = Q$ (Gneiting and Raftery, 2007).

3.1 Decomposition of proper scores

Following Gneiting and Raftery (2007), the entropy $e(Q) = S(Q, Q)$ is defined as the minimal achievable score under Q if $S(.,.)$ is proper. The divergence $d(P, Q) = S(P, Q) - S(Q, Q)$ is the non-negative difference between the expected score and the entropy. Note, that propriety of the score function implies propriety of the divergence (Thorarinsdottir et al., 2013). $d(P, Q)$ represents a measure for similarity between the probabilistic forecast P and the distribution Q, where smaller values indicate better correspondence. It is generally denoted as *reliability* (Murphy and Winkler, 1977). Given entropy and divergence, a proper score function can be decomposed into

$$S(P, Q) = e(Q) + d(P, Q)\,.$$

The entropy can be further decomposed using $e(Q) = S(\bar{Q}, Q) - d(\bar{Q}, Q)$, where $\bar{Q}$ denotes the marginal distribution of the observations, often denoted as climatology (Bröcker, 2009). $S(\bar{Q}, Q)$ represents the expected score function for a climatological forecast, and is called *uncertainty*. The *resolution* of a forecast P is expressed by the divergence between the climatological forecasts $\bar{Q}$ and Q. A large resolution indicates a good discrimination of Q under different forecasts P. The decomposition of the expected score is now given by

$$S(P, Q) = S(\bar{Q}, Q) - d(\bar{Q}, Q) + d(P, Q)\,. \tag{2}$$

Analogously, we obtain the overall score for the probabilistic forecasting scheme by integrating over the marginal distribution $dF(P)$

$$\begin{aligned} &\int_{\mathcal{P}} S(P, Q) dF(P) = \\ &\qquad \underbrace{S(\bar{Q}, \bar{Q})}_{uncertainty} - \underbrace{\int_{\mathcal{P}} d(\bar{Q}, Q) dF(P)}_{resolution} + \underbrace{\int_{\mathcal{P}} d(P, Q) dF(P)}_{reliability}\ . \end{aligned} \tag{3}$$

Since the uncertainty is solely a characteristic of the observations, changes in the probabilistic forecasting scheme affect only the reliability and resolution parts of the score, not the uncertainty.

4 Quantile score

The QS of a quantile forecast $p_\tau = P^{-1}(\tau)$ and an observation ω is defined by the check function ρ_τ in Eq. 1 (Koenker and Machado, 1999; Friederichs and Hense, 2007; Gneiting and Raftery, 2007)

$$s_\tau(P, \omega) = \rho_\tau(\omega - p_\tau)\,.$$

The expected score with respect to Q for a continuous variable $\omega \in \mathbb{R}$ is given by

$$\begin{aligned} S_\tau(P,Q) &= \int_{-\infty}^{\infty} \rho_\tau(\omega - p_\tau) dQ \\ &= \int_{-\infty}^{\infty} \tau(\omega - p_\tau) dQ - \int_{-\infty}^{p_\tau} (\omega - p_\tau) dQ\,. \end{aligned}$$

The propriety of this score is shown e.g. in Friederichs and Hense (2008). The decomposition into entropy and divergence yields

$$S_\tau(P,Q) = S_\tau(Q,Q) + \underbrace{\int_{q_\tau}^{p_\tau} (p_\tau - \omega) dQ}_{d_\tau(P,Q)}\,,$$

where q_τ is the τ-quantile of the distribution Q. The divergence measures the agreement between the quantile forecasts p_τ and the respective quantile of the distribution Q. For a perfect forecast ($p_\tau = q_\tau$) the integral is zero.

Let $\bar{q}_\tau$ be the τ-quantile of the climatology $\bar{Q}$. The entropy may be decomposed into uncertainty (i.e., the score of the climatology) and the divergence between Q and $\bar{Q}$

$$S_\tau(Q,Q) = S_\tau(\bar{Q},Q) - \underbrace{\int_{q_\tau}^{\bar{q}_\tau} (\bar{q}_\tau - \omega)\, dQ}_{d_\tau(\bar{Q},Q)}\ .$$

The decomposition in Eq. 2 for the QS is thus expressed as

$$\begin{aligned} & S_\tau(\bar{Q},Q) - d_\tau(\bar{Q},Q) + d_\tau(P,Q) \\ &\quad = \int_{-\infty}^{\infty} \tau(\omega - \bar{q}_\tau) dQ - \int_{-\infty}^{\bar{q}_\tau} (\omega - \bar{q}_\tau) dQ && (4) \\ &\quad - \int_{q_\tau}^{\bar{q}_\tau} (\bar{q}_\tau - \omega) dQ && (5) \\ &\quad + \int_{q_\tau}^{p_\tau} (p_\tau - \omega) dQ\,, && (6) \end{aligned}$$

where (4) represents the uncertainty, (5) the resolution, and (6) the reliability parts of the score. There exists no general solution of the integrals in Eq. 4-6.

4.1 Estimation of the QS and its decomposition

Consider a verification data set of $n = 1, ..., N$ pairs of forecasts $p_{\tau,n}$ and observations o_n. The average QS is given by

$$\mathcal{S}_N^P(\tau) = \frac{1}{N} \sum_{n=1}^{N} \rho_\tau(o_n - p_{\tau,n})\,. \tag{7}$$

The uncertainty component of the QS is estimated as the average score of the unconditional sample quantile $\bar{o}_\tau$, which is given as the τ-quantile of the N observations

$$\mathcal{S}_N^{\bar{Q}}(\tau) = \frac{1}{N} \sum_{n=1}^{N} \rho_\tau(o_n - \bar{o}_\tau)\,.$$

To estimate reliability and resolution we need to condition the observations on the forecasts. This requires a categorization of the continuous forecast values (Murphy and Epstein, 1967), so that for each category, the conditional quantile of the observations is estimated. For this we divide the data into $k = 1, \ldots, K$ subsamples $\mathcal{I}_k$. The subsamples are defined such that they represent similar forecasts. Each subsample $\mathcal{I}_k$ is represented by the average value $p_\tau^{(k)}$ of all forecasts $p_{\tau,n}$ with $n \in \mathcal{I}_k$. The $p_\tau^{(k)}$ represent the discretized quantile forecasts[1], and the conditional quantiles are estimated as the τ-quantile $o_\tau^{(k)}$ of all observations o_n with $n \in \mathcal{I}_k$. A quantile forecasting scheme is reliable, if for all bins $\mathcal{I}_k$ the observed quantile $o_\tau^{(k)}$ corresponds to the average forecast $p_\tau^{(k)}$.

Let N_k be the number of pairs in each bin with $\sum_{k=1}^{K} N_k = N$. The entropy $S_\tau(Q, Q)$ is estimated for each bin, separately, as

$$\widehat{S_\tau(Q,Q)}^{(k)} = \frac{1}{N_k} \sum_{n \in \mathcal{I}_k} \rho_\tau(o_n - o_\tau^{(k)})\,.$$

The reliability and resolution parts of the QS are estimated as the differences between the average scores for $p_\tau^{(k)}$ and $\bar{o}$, respectively, and the entropy in each bin.

$$\begin{aligned}
\widehat{d_\tau(\bar{Q},Q)}^{(k)} &= \widehat{S_\tau(\bar{Q},Q)}^{(k)} - \widehat{S_\tau(Q,Q)}^{(k)} \\
&= \frac{1}{N_k} \sum_{n \in \mathcal{I}_k} [\rho_\tau(o_n - \bar{o}_\tau) - \rho_\tau(o_n - o_\tau^{(k)})] \\
\widehat{d_\tau(P,Q)}^{(k)} &= \widehat{S_\tau(P,Q)}^{(k)} - \widehat{S_\tau(Q,Q)}^{(k)} \\
&= \frac{1}{N_k} \sum_{n \in \mathcal{I}_k} [\rho_\tau(o_n - p_\tau^{(k)}) - \rho_\tau(o_n - o_\tau^{(k)})]
\end{aligned}$$

[1] For the Brier score, the representing forecast value is often set to the mid of the interval. Particularly if the forecast values are not uniformly distributed the estimates of the reliability component are largely biased (see for instance Atger, 2004; Bröcker and Smith, 2007, amongst others). A better choice is using the average forecast value.

The decomposition (Eq. 3) of the average QS of the discretized forecasts $p_\tau^{(k)}$, $k = 1, ..., K$, is obtained by summation over all bins

$$\begin{aligned} &\sum_{k=1}^{K} \frac{N_k}{N} \, \widehat{S_\tau(P,Q)}^{(k)} \qquad (8) \\ &= \sum_{k=1}^{K} \frac{N_k}{N} \, \widehat{S_\tau(\bar{Q},Q)}^{(k)} \quad \text{uncertainty} \\ &- \sum_{k=1}^{K} \frac{N_k}{N} \, \widehat{d_\tau(\bar{Q},Q)}^{(k)} \quad \text{resolution} \\ &+ \sum_{k=1}^{K} \frac{N_k}{N} \, \widehat{d_\tau(P,Q)}^{(k)} \quad \text{reliability} \, . \end{aligned}$$

Note that, due to the discretization of the forecasts, this average is not equivalent to $\mathcal{S}_N^P(\tau)$ in Eq. 7.

4.2 Bias of the decomposition

Probabilistic forecasting schemes are subject to different kinds of biases. A *bias in the forecasts* results from small ensemble size (see for instance Ferro et al., 2008, and references herein), an insufficient representation of the underlying distribution, or other deficiencies in the generation of forecasts. These biases are quantified by proper score functions and their decomposition, and may be eliminated or reduced by postprocessing.

A *bias in score estimates* has two major sources: the discretization error, on the one hand, and the sampling uncertainty due to a finite sample size, on the other hand. The discretization error results from the categorization of the forecasts, which is necessary do estimate the decomposition. It affects the score as well as the reliability and resolution estimates, whereas the uncertainty remains unchanged.

Atger (2003), Atger (2004), Bröcker and Smith (2007), and Bröcker (2008) investigate the discretization error of the BS. A general procedure here is to categorize the forecast probabilities into intervals of equal width, and derive observed conditional relative frequencies. If the forecast are not uniformly distributed, this might result in undersampling of some categories and hence strong biases in the decomposition. (Atger, 2004) shows that the discretization error is reduced, if the categories are chosen such that they all contain an equal amount of forecast-observation pairs. They further note that the number of categories should be adjusted with respect to the sample size. However, any categorization automatically leads to errors even for perfectly reliable forecast systems (Bröcker, 2008).

The definition of the subsamples $\mathcal{I}_k$ is thus a critical point in the decomposition of the QS. To keep the discretization error as small as possible, we suggest a binning procedure which is similar to Atger (2004). The quantile forecasts are ordered and divided into K equally populated bins, i.e. the intervals are defined by the $1/K$-percentiles of the forecasts $\{p_{\tau,n}\}\,, n = 1, ..., N$. We discuss the discretization error in Sec. 5 and show that the findings of Atger (2004) hold for the QS.

Other biases in score estimates are due to the finite sample size and affect the reliability, resolution, and uncertainty estimates, but not the average score itself. Bröcker (2011) and Ferro and Fricker (2012) discuss these biases for the BS decomposition. They underline that unbiased estimators of reliability and resolution are unattainable. However, Ferro and Fricker (2012) were able to prove that the biases of their estimators converge faster to zero than those proposed by Bröcker (2011), which in turn converge faster to zero than those of the standard decomposition. It is thus of interest to assess at least qualitatively the biases of the QS decomposition. A detailed derivation is provided and discussed in an appendix.

4.3 Quantile reliability plot

A quantile reliability plot can be derived from the average forecasts $p_{\tau}^{(k)}$ and the conditional quantiles $o_{\tau}^{(k)}$. For each $k = 1, ..., K$, $o_{\tau}^{(k)}$ is plotted against $p_{\tau}^{(k)}$. For calibrated forecasts (i.e., reliability → 0), the points should lie on a diagonal line. The interpretation concerning over- or underforecasting of a quantile reliability diagram is analogue to the interpretation of a reliability diagram for probability forecasts of dichotomous events (see for example Wilks (2006), pp. 287–290).

Software routines for the calculation and decomposition of the QS as well as for plotting the quantile reliability will be available for the R statistical language (R Core Team, 2013) within the "verification" package (NCAR - Research Application Program, 2012), or upon request from the authors.

5 Results

5.1 Simulation Study

In order to demonstrate the performance of the QS and its decomposition, we assess quantile forecasts from a simulation with a well defined joint distribution between the random variables Y (observations or predictand) and X (forecasts or predictor). The joint distribution $F(x, y) = F_Y(y \mid X = x)F_X(x)$ describes the statistical dependence of the observations from the forecasts. If this distribution is known, perfect reliable forecasts can be generated from X. An analytical solution of the score function can be derived if the integral in Eq. (3) can be solved under the given joint distribution. The average score for a sample of realizations from the joint distribution can be analyzed against the analytical score.

Let Y be a gamma-distributed random variable with $Y \mid X \sim F_{Y|X}(y \mid X = x)$. Y depends on the random variable X through $E[Y|X = x] = 1/x$. For the shape parameter of $F_{Y|X}(y \mid X = x)$ we assume $\alpha_y = 1$, so that the rate parameter is $\beta_y = \alpha_y x = x$. The predictive distribution of Y for a given $X = x$ is given as

$$f_{Y|X}(y \mid X = x) = f_{\Gamma}(y; \alpha_y, \beta_y) = x \exp(-xy) \,. \tag{9}$$

X also follows a gamma-distribution with fixed shape $\alpha_x = 5$ and rate $\beta_x = 1$ parameter, and thus

$$f_X(x) = f_\Gamma(x; \alpha_x, \beta_x) = \frac{1}{24} x^4 exp(-x)\,. \tag{10}$$

Realizations from the joint distribution of X and Y can now be obtained by drawing x from Eq. (10) and plug into Eq. (9). Now the observation y is obtained by drawing from (9). We use this setup for two reasons. On the one hand, the gamma-distribution is a widely used distribution to statistically model precipitation and provides more complexity than a normal distribution. On the other hand, parameters and dependence of Y and X are kept simple such that conditional quantiles and scores can be calculated analytically.

The conditional τ-quantiles of the response variable Y for a given $X = x$ are given by

$$F^{-1}_{Y|X}(\tau \mid X = x) = -\frac{\log(1-\tau)}{x} = q_\tau\,. \tag{11}$$

The marginal density function of Y unconditional on X (i.e. the climatology) is obtained by marginalization of $f_{Y|X}(y|X = x)f_X(x)$ over x

$$f_{\bar{Y}}(y) = \int_0^\infty f_{Y|X}(y \mid X = x) f_X(x) dx = \frac{5}{(y+1)^6}.$$

The marginal distribution is obtained by integrating the density

$$F_{\bar{Y}}(y) = \int_0^y f_{\bar{Y}}(y')dy' = 1 - \frac{1}{(y+1)^5},$$

which in turn gives us the quantile function of the climatology

$$F^{-1}_{\bar{Y}}(\tau) = (1-\tau)^{-1/5} - 1 = \bar{q}_\tau\,.$$

The entropy $S_\tau(Q, Q)$ of this simulation and the uncertainty $S_\tau(\bar{Q}, \bar{Q})$ can be calculated analytically. Moreover, the resolution is equal to the difference between uncertainty and entropy, and the reliability part for this perfect forecast model is zero. The score decomposition can be derived analytically, and reads

$$\begin{aligned} \mathbf{S}_\tau &= \frac{\log(1-\tau)(\tau-1)}{4}\,, \\ E[S_\tau(\bar{Q}, \bar{Q})] &= \frac{5}{4}(1-\tau)\bar{q}_\tau\,, \\ E_X[d_\tau(\bar{Q}, Q)] &= E[S_\tau(\bar{Q}, \bar{Q})] - \mathbf{S}_\tau\,, \\ E_X[d_\tau(P, Q)] &= 0\,. \end{aligned}$$

5.1.1 Discretization error using true forecasts

In the following we concentrate on the results for median forecasts, but the results are similar for other probability levels (not shown). We generate $N = 5000$ random draws for X from

Table 1: Analytical values of QS and its decomposition for the median of the simulation study, and the respective estimates and uncertainty based on 1000 simulations of length $N = 5000$. Estimates of resolution and reliability depend on the binning procedure and are shown in Fig.2 .

$\tau = 0.5$	analytical	estimate	sd
QS	0.0866	0.0867	±0.0018
UNC	0.0929	0.0930	±0.0021
RES	0.0063	—	—
REL	0	—	—

the distribution given in Eq. (10) and $\tau = 0.5$. For each x, observations y are drawn from Eq. (9), and true quantile forecasts $p_\tau = q_\tau$ are obtained from the conditional quantile function in Eq. (11). True forecasts are perfectly reliable forecasts and thus have a reliability of zero. The average score $\mathcal{S}_N^{q_\tau}$ and uncertainty $\mathcal{S}_N^{\bar{q}_\tau}$ are estimated using Eq. (7). Table 1 shows the analytical QS and its decomposition for the true median forecasts. The mean value of the average score is estimated from 1000 re-samples, each with $N = 5000$ draws, and is close to the true scores and within the standard-deviation.

As aforementioned the estimates of reliability and resolution may strongly depend on the categorization. In order to investigate its influence, each data set of N forecast-observation pairs is divided into K equally populated bins, with $N_k = N/K$ data points within each bin. Figure 2 displays the QS decomposition for the median and a K ranging from 1 to 100. For $K = 1$, the conditional observed quantile equals the unconditional sample quantile, and the resolution part is zero. The respective discretized quantile forecast is given as the mean over all forecasts $1/N \sum_{i=1}^{N} p_{\tau,i}$. So even in the case of perfect quantile forecasts $p_{\tau,i}$, the average quantile forecast that represents the bin is not anymore perfect, and reliability is strongly biased. This effect is larger if the number of bins is small.

Increasing K leads to an increased resolution and a reduced QS estimate which approaches its analytical value. Although the QS becomes minimal for large K, we observe again an increase in the bias in the estimate of the reliability which we know is zero. Here, the effect of the sampling errors in the conditional quantile estimate as discussed in the appendix starts to become relevant. The larger K the fewer observations are within one bin and the larger are the sampling errors in the conditional quantile estimates. Another indication of this sampling bias is that the resolution part increases at the expense of reliability.

Fig. 2 also shows the results for an equidistant binning (grey letters). In general, a much larger number of bins is necessary to eliminate the discretization error in the QS. At lest 30 bins are needed to obtain a resolution term that is significantly larger than zero. Even for $K = 100$, the estimated resolution lies below the uncertainty range of the unbinned QS estimate. The sampling biases are much larger as indicated by the large reliability part, which results from an undersampling in the less populated bins. Very similar results are obtained for other probability levels (not shown).

A careful binning does not avoid biases, but helps to keep them as small as possible. We want to remind the reader, that the BS and its decomposition is subject to the same kind of biases. Atger (2004) has discussed ideas to estimate an optimal value of K. A re-sampling

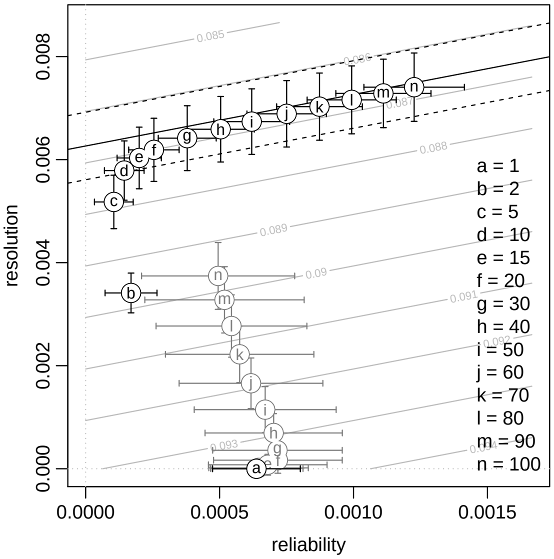

Figure 2: Reliability vs. resolution for a sample of $N = 5000$ median forecast, equally divided into K bins ranging from $K = 1$ to 100 indicated by black small letters. For a comparison, the grey small letters refer to a binning in equidistant intervals. The standard deviation for each point is estimated from 1000 re-samples. The grey contours denote lines of constant quantile score. The bold black line denotes the QS for the unbinned data set, the dashed lines the respective standard deviation. The legend refers to the number of bins K.

technique is proposed in order to assess significant differences in the conditional relative frequency estimate when two probability categories are merged together. However, as an ad-hoc solution, we suggest to use the smallest K where the gain in QS becomes smaller than the loss in reliability.

If different forecasting schemes are compared, the number of bins K should be kept constant for all schemes. K must be chosen such that the differences to the QS for unbinned data is small for all forecasting schemes. In this case we assume that resolution and reliability estimates can be compared honestly. We refrain from keeping the bin boundaries fixed for all forecasting schemes, since it cannot be guaranteed that fixed bins will be equally well represented for every scheme.

Table 2: Average values of QS and its decomposition for the true median forecasts, biased and stochastically disturbed median forecasts. Estimates are based on an equi-populate binning with $K = 15$. The uncertainty is a characteristic of the observations and obtained as 0.093 ± 0.002.

$\tau = 0.5$	true	biased	stochastic
QS	0.0871 ± 0.0018	0.0901 ± 0.0020	0.0908 ± 0.0018
RES	0.0060 ± 0.0006	0.0060 ± 0.0006	0.0041 ± 0.0005
REL	0.0002 ± 0.0001	0.0032 ± 0.0005	0.0019 ± 0.0004

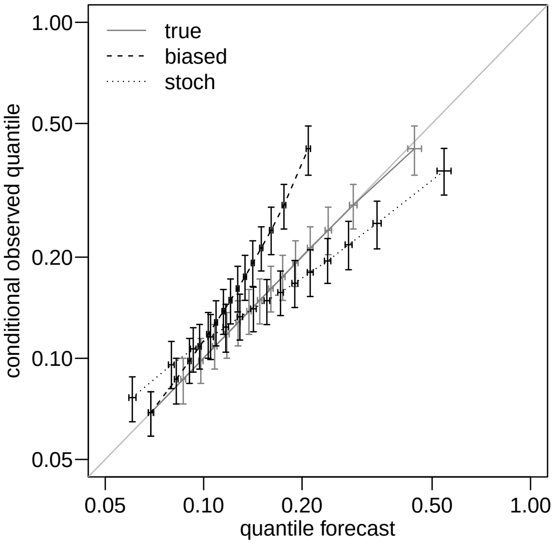

Figure 3: Q-REL plot for a sample of $N = 5000$ median forecasts, equally divided into $K = 15$ bins. Reliability curves are plotted for the true quantiles (solid line), and quantile forecasts derived with biased (dashed line) and stochastic (dotted line) rate parameter. The 95% confidence interval for each point is estimated from 1000 re-samples.

5.1.2 Quantile reliability

We now consider the effect of imperfect median forecasts. One forecast is systematically biased by constant errors ϵ_0 and ϵ_1 with $\beta_y = \epsilon_0 + \epsilon_1 x$. For the other imperfect forecast x is disturbed by a multiplicative noise ξ. Figure 3 shows a quantile reliability (Q-REL) plot for the perfect forecasts, the biased forecasts with $\epsilon_0 = 2$ and $\epsilon_1 = 0.8$, and the stochastically disturbed forecast with ξ being uniformly distributed on the interval $[-0.5, +0.5]$. We choose $K = 15$ equally populated bins to estimate the reliability and resolution parts of the QS. For each bin the conditional quantile of the observations is plotted against the average forecast value. Note that in case of our simulation study, the variable of interest takes values between 0 and 1, and is not a probability. The axes of a Q-REL plot resemble the range of forecast/observation values, which can in general be any interval on the real line depending on the variable of interest (temperature, pressure, precipitation, etc.). Since we keep only the number of bins K constant (and not the bin boundaries), the re-sampling also modifies the binning intervals. Confidence intervals are obtained for the conditional quantiles (y-axis) as well as for the average forecast values (x-axis).

The true quantiles are perfectly calibrated. The biased forecasting scheme shows strong deviations in reliability particularly for large quantile forecasts, with a systematic underforecasting of the observed quantiles. However, the resolution of the biased forecasts is not affected, since the linear transformation does not change the information content given by the predictor variable X (see Tab. 2). The increase of the QS of about 3.4% is thus solely due to an increase in the reliability term.

The reliability curve of the stochastically perturbed forecasting scheme is significantly flatter than the 45° line, indicating an underforecasting of the lower and overforecasting of

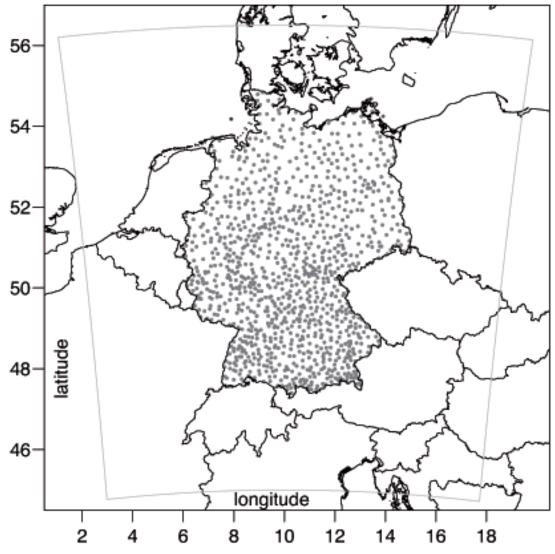

Figure 4: Location of 1079 observational sites (points) and model domain of COSMO-DE (grey lines).

the higher quantiles. The outcomes of Y have a weaker dependence on the predictor variable X (Wilks, 2006, p. 288-289), and the resolution is reduced (see Tab. 2). The QS of the stochastically perturbed forecasts is about 4.2% larger than the QS of the perfect forecast. This increase is related in almost equal parts to an increased reliability term and a loss in resolution.

5.2 Real forecast data

We will now assess quantile forecasts for precipitation that are derived from the German-focused COSMO-DE-EPS (Gebhardt et al., 2011; Peralta et al., 2012). COSMO-DE-EPS is a 20 members ensemble system operated by the German Meteorological Service (DWD)[2]. Forecasts are initialized 8 times a day for a lead time of 21 hours. The model domain of COSMO-DE-EPS is illustrated in Fig. 4 by the grey lines. It contains 421 × 461 grid points with a horizontal grid spacing of 2.8 km. A study on the general performance and the calibration of COSMO-DE-EPS can be found in Ben Bouallègue et al. (2013), and with special emphasis on precipitation in Ben Bouallègue (2013) and Scheuerer (2013).

In the following, we concentrate on daily precipitation accumulations between 12-24 UTC for the year 2011. Observations are taken from the observational network of DWD with 1079 observational sites located in Germany (Fig. 4).

Due to the rapid update cycle of COSMO-DE-EPS, we use the time-lagging method to inexpensively increase the ensemble size (Hoffman and Kalnay, 1983). Forecasts initialized at 12/9/6/3 UTC corresponds to the observational period 12-24 UTC within 0-12/3-15/6-18/9-21 hours of the forecast. Using four time-lagged forecasts for each of the 20 ensemble members, we obtain an 80 members time-lagged COSMO-DE-EPS (TLE). Time-lagging

[2] Please note that forecasts are taken from the pre-operational period, which has the same setup as the operational system that started in May 2012.

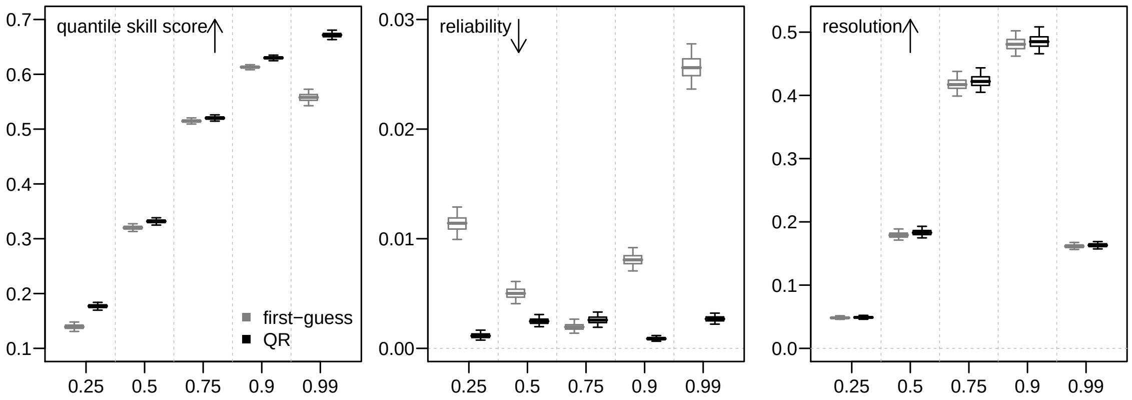

Figure 5: QSS, reliability and resolution for different probability levels. Forecasts are derived from the TLE as first-guess forecasts (grey boxes) and via QR (black boxes). The 95% confidence interval is estimated via block-bootstrapping. For reliability and resolution estimates, forecasts are categorized into $K = 30$ equal-populated bins.

increases the ensemble spread and has a positive impact especially for precipitation forecasts (Ben Bouallègue et al., 2013).

We pursue two approaches to derive predictive quantiles from the TLE. The first approach is to derive the predictive quantiles directly from the precipitation forecasts at each grid point. Quantiles are estimated from the order statistics of the ensemble members. To improve the estimates a spatial neighbourhood of 5×5 grid boxes is included (compare e.g. Bentzien and Friederichs, 2012), so that quantiles are estimated from a total number of $25 \times 80 = 2000$ values. We denote the predictive quantiles as first-guess quantiles Q_τ^{fg}.

The skill of the first-guess quantiles can be largely improved by an additional statistical postprocessing. This second approach uses censored quantile regression (QR) as presented e.g. in Friederichs and Hense (2007). Based on historical data, QR estimates a statistical model between the ensembles first-guess estimates and the observations. The postprocessed quantile forecasts are denoted as Q_τ^{qr}. For an out-of-sample verification we employ cross-validation. More details on the definition of the predictors, the censored QR, and the cross-validation procedure is given in Bentzien and Friederichs (2012).

5.2.1 Quantile verification

The overall skill of the quantile forecasts is assessed using the quantile skill score QSS=1-QS/QS$_{clim}$. It gives the percental improvement of a forecasting scheme over a reference forecast, here climatology, which is estimated at each station independently (Hamill and Juras, 2006). The sampling uncertainty of the QSS is assessed via 7-day block-bootstrapping (Efron and Tibshirani, 1993) with 1000 replicates simultaneously for all stations, thereby preserving spatial and temporal correlations. We restrict the presentation to the interquartile range and two higher quantiles, since the lower quantiles are generally censored at zero. Positive skill is obtained for all quantile forecasts (Fig. 5). Median and 0.75-quantiles are already well represented by the first-guesses. The first-guess forecasts are significantly improved by postprocessing, where the benefits are largest for lower and higher probability

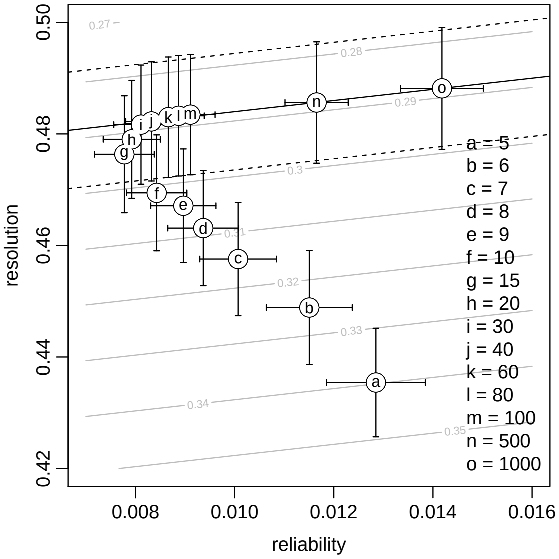

Figure 6: Reliability vs. resolution for the QS decomposition of $Q_{0.9}^{fg}$. Forecasts are equally divided into K bins ranging from 5 to 1000 (indicated by small letters). The standard deviation is estimated from 1000 bootstrap samples. The grey contours denote lines of constant quantile score. The bold black line denotes the QS for the unbinned data set, the dashed lines the respective standard deviation.

levels.

To derive the decomposition of the QS, the data are binned with respect to the quantile forecasts. The first bin consists of all data where the quantile forecast equals zero. The non-zero forecasts are regrouped into equally populated intervals. The boundaries of intervals are defined as the k/K-quantiles of the non-zero forecasts, with $k = 1, \ldots, K$ and K the number of intervals.

Fig. 6 shows the QS decomposition for the discretized forecast-observation pairs $Q_{0.9}^{fg}$ as a function of K ranging between 5 and 1000. For small K, the bias in QS for the discretized forecasts is very large. Increasing K leads to a larger resolution and thus reduces the QS. For $K > 30$, the gain in QS becomes small compared to the loss in reliability. Very similar results are obtained for the other quantiles (not shown). In the remainder of this section we thus use equally populated bins with $K = 30$.

5.2.2 Score decomposition

We will now discuss in more detail the decomposition of the QS (Fig. 5). Table 3 gives the number of observations in each bin for $K = 30$. Since lower probability levels are more often censored, the average number of observations in each bin varies between 2500 and 3100 for $\tau = 0.25$, and amounts between 7000 to 12000 for $\tau = 0.99$.

Fig. 5 shows reliability and resolution parts of the QS. Note that a smaller reliability represents a better agreement between forecasts and observations, whereas the resolution is better the larger it is. We already noticed that, except for the 0.75-quantile, statistical postprocessing via QR significantly improves the QS. Due to the decomposition of the QS,

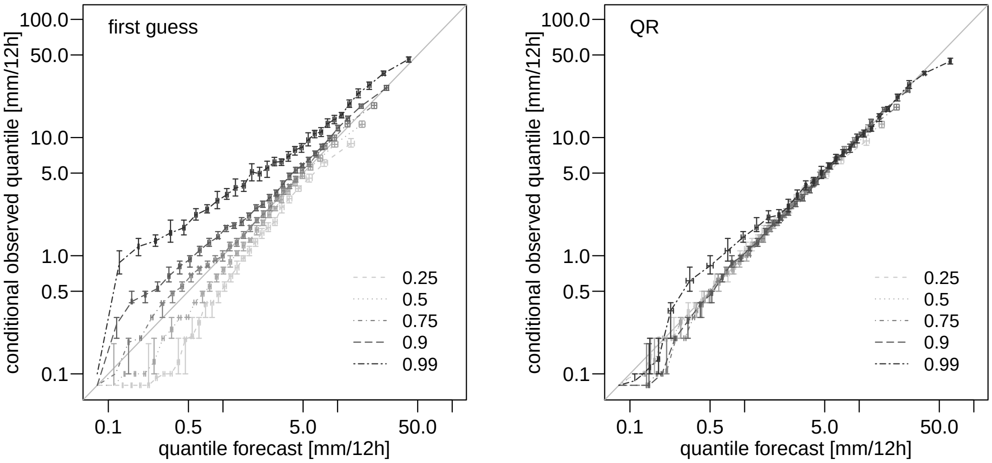

Figure 7: Q-REL plots on double-logarithmic scale for first-guess and QR forecasts derived from the TLE. The precipitation quantiles are measured in $mm/12h$. The 95% confidence interval is estimated via block-bootstrapping.

we can now relate this gain in performance to an improvement of reliability, which is largely decreased after postprocessing especially for the outer quantiles. In contrast to the reliability, the resolution is not improved by postprocessing.

Q-REL plots for the first-guess forecasts are shown in Fig. 7. It is well seen in the double-logarithmic representation that for $\tau \leq 0.5$, the first-guess quantiles are significantly overestimated. The high quantiles, $\tau \geq 0.9$, in turn are largely underestimated. Only the zero quantile forecasts are well calibrated for almost all probability levels. The $Q^{fg}_{0.75}$ is close to the diagonal, and the potential for further improvements is small. The miscalibration of the quantile forecasts is a consequence of the underdispersiveness of COSMO-DE-EPS (Ben Bouallègue et al., 2013), and/or an insufficient ensemble size.

We now turn to the postprocessed quantile forecasts. Quantile forecasts above 0.5 mm are almost perfectly calibrated for all probability levels except $\tau = 0.99$. The Q-REL plot reveals an underestimation of the smaller forecast values, whereas the zero quantile forecasts correspond well to the observations. Although the underestimation of 0.99-quantile is largely reduce after postprocessing, it still shows deviations from the diagonal. Higher forecast values are well calibrated. Only for the 0.99 quantile the forecasts in the highest category above 40 mm/12h are significantly overestimated.

5.2.3 Forecast communication

The decomposition of the QS shows us the strengths and weaknesses of this forecast system. But how can forecasts be communicated to users? As mentioned in Sec. 2 a very intuitive way is to display quantiles in form of boxplots. Fig. 8 shows time series of daily quantile forecasts for June 2011 at three selected stations. The boxes denote the interquartile range (IQR). In 50% of the days the observation is expected to lie within the IQR. Since lower quantiles are generally censored, only the upper whisker are shown for the 0.9- and 0.99-quantile.

Table 3: Number of quantile forecasts equal to zero N_0 and number of quantile forecasts within each bin N_k for $K = 30$. The total number of forecast-observation pairs is $N = 384\,679$ (with $257\,431$ observations equal to zero).

TLE		FG		QR
τ	N_0	$\bar{N}_k$	N_0	$\bar{N}_k$
0.25	291447	3108	306072	âĹij2498
0.50	264107	4019	264570	âĹij3961
0.75	234389	5010	211131	âĹij6090
0.90	206956	5924	169848	âĹij9632
0.99	171413	7109	115353	12443

The forecast system is able to distinguish between days with and without rainfall. On June 5-6, and June 22 convective rainfall events occurred in many parts of Germany, and more local convective events occurred e.g. in Munich between June 16-18 and in Berlin on June 8. The forecast systems is able to capture these events by higher quantile forecasts. Especially the 0.99-quantile exceeds the IQR multiple times on such days, indicating a high risk of extreme precipitation. Note, that there is still a chance of 1% that the 0.99-quantile will be exceeded, as it happened in Bonn on June 5.

Quantile forecasts present the possible range of observations without limitation, as is the case for a-priori selected threshold excesses. The range of events can change with location, season, or climate conditions. A specific quantile, e. g. the 0.99-quantile, is always "extreme" regardless of such conditions. However, quantile forecasts are only statistical meaningful if they are calibrated, so that the user can rely on the forecast system and make "good" decisions. Calibration of quantile forecasts is checked properly by the decomposition of the quantile score as proposed in this paper.

6 Summary

This article expands the forecast verification framework by introducing a decomposition of the quantile score. We present an estimation procedure of the reliability, resolution and uncertainty parts. A graphical representation of the reliability in form of a quantile reliability diagram is provided, with a similar interpretation as the reliability diagram for probability forecasts.

The estimation of forecast attributes like reliability and resolution requires the categorization of continuous forecast values. In order to obtain statistical meaningful results, each category should be sufficiently represented by the forecast-observation pairs. A binning approach similar to Atger (2004) divides the data set in equally populated subsamples conditioning on the forecast values. However, even for perfectly reliable forecasts, every kind of categorization leads to biases in the QS and its decomposition. This is also the case for the Brier score. Within a simulation study, we show the sensitivity of the biases to the number of subsamples K. A careful choice of K can keep the bias from the categorization as small as possible.

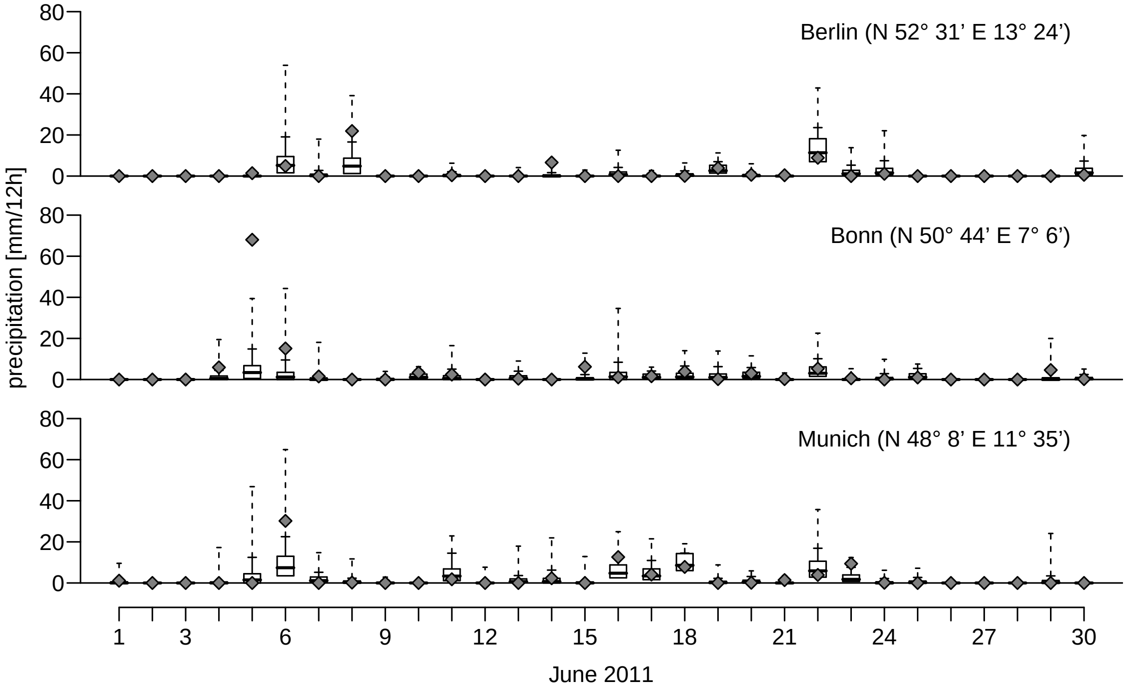

Figure 8: Time series of daily quantile forecasts for June 2011 at three selected stations located in the northeast (Berlin), west (Bonn), and south (Munich) of Germany. The boxplots show the interquartile range, the upper whiskers refer to the 0.9-quantile (solid line) and 0.99-quantile (dashed line). Observations are plotted as squares.

The QS decomposition is illustrated for precipitation forecasts derived from a mesoscale weather prediction ensemble. The challenge in probabilistic quantitative precipitation forecasting is the discrete-continuous character of precipitation and its highly skewed distribution. First-guess quantile forecasts are estimated from the order statistics of the ensemble members, enlarged by time-lagging and the neighbourhood method. Although these first-guess quantile forecasts have positive skill with respect to climatology, the decomposition of the QS reveals significant deficiencies in the reliability of the first-guess quantile forecasts. A statistical postprocessing in terms of quantile regression largely improves the reliability, and provides skilful forecasts even for very high probability levels. The effect of the post-processing on the resolution in turn is small.

The QS decomposition provides valuable insights in the performance of predictive quantiles, no matter how they are derived. A verification of predictive distributions should assess the performance of its quantiles, as well as its exceedance probabilities over fixed thresholds. In a similar sense, Gneiting and Ranjan (2011) proposed to compare density forecasts using threshold- and quantile-weighted scoring rules. We think that with this decomposition the quantile score is ready to become as popular in forecast verification as the Brier score.

Acknowledgments

This work was funded by Deutscher Wetterdienst in Offenbach, Germany, within the Extramural Research Program and the Hans-Ertel Centre for Weather Research. The Hans Ertel Centre is a network of Universities, Research Institutes and the Deutscher Wetterdienst which is funded by the BMVBS (Federal Ministry of Transport, Building and Urban Development). The authors thank Tilmann Gneiting, Michael Weniger, and Jochen Bröcker for helpful discussions and Martin Göber for providing the rain gauge data. Special thank is due to the editor and two anonymous reviewers for their valuable suggestions.

Appendix

Motivated by the studies of Bröcker (2011) and Ferro and Fricker (2012) we finally want to assess sample biases in the estimates of reliability and resolution. We start from discretized forecasts $p_\tau^{(k)}$ assuming that each forecast is equally probable (i.e., discretization using equally populated binning). In other words, we assess expectations with respect to the joint distribution $F(p_\tau^{(k)}, o_n)$ of the discretized forecasts $p_\tau^{(k)}$ and the observations o_n. The biases thus only relate to the estimation of reliability and resolution, not to the effect that results from the binning of the forecasts. We also assume that the observations given a certain forecast are identically and independently distributed.

The expectation of the reliability estimate (REL, Eq. 8) with respect to the joint distribution $F(p_\tau^{(k)}, o)$ is calculated as

$$\begin{aligned} E[REL] &= E\left[\sum_{k=1}^{K} \frac{N_k}{N}\, \widehat{d_\tau(P,Q)}^{(k)}\right] \\ &= \Phi \sum_{k=1}^{K} E\left[\widehat{d_\tau(P,Q)}^{(k)}\right], \end{aligned}$$

where $\Phi = N_k/N$ is the relative frequency of $p_\tau^{(k)}$. Since the $p_\tau^{(k)}$ represent equally populated bins $\mathcal{I}_k$, the relative frequency only depends on the number of bins K. Consequently, we only need to assess the biases of the reliability and resolution within each bin, independently. For the expectation of the reliability for an interval $\mathcal{I}_k$ we can write

$$\begin{aligned} &E[REL^{(k)}] \\ &\quad = \frac{1}{N_k} \sum_{n \in \mathcal{I}_k} \left(E\left[\rho_\tau(o_n - p_\tau^{(k)})\right] - E\left[\rho_\tau(o_n - o_\tau^{(k)})\right] \right) \\ &\quad = S(P^{(k)}, Q^{(k)}) - E\left[\rho_\tau(o_n - o_\tau^{(k)})\right] \;,\; n \in \mathcal{I}_k \,. \end{aligned}$$

The bias in the reliability component results from the fact that we need to estimate the conditional quantile $o_\tau^{(k)}$. Let us replace $o_\tau^{(k)} = q_\tau^{(k)} + o_\tau^{(k)} - q_\tau^{(k)}$, where $q_\tau^{(k)}$ is the true quantile, and use a slightly different representation of the check function, namely

$$\begin{aligned}&\rho_\tau\left((o_n - q_\tau^{(k)}) - (o_\tau^{(k)} - q_\tau^{(k)})\right)\\ &\quad = \tau(o_n - q_\tau^{(k)}) - \tau(o_\tau^{(k)} - q_\tau^{(k)})\\ &\quad - \left((o_n - q_\tau^{(k)}) - (o_\tau^{(k)} - q_\tau^{(k)})\right) I_{o_n < o_\tau^{(k)}}.\end{aligned}$$

An aggravating factor is that the estimate of the conditional quantile $o_\tau^{(k)}$ is a function of the sample of observations within one bin. Other than the true quantile $q_\tau^{(k)}$ it depends on o_n. We may split the index function $I_{o_n < o_\tau^{(k)}}$ into

$$\begin{aligned}I_{o_n < o_\tau^{(k)}} &= I_{o_n < q_\tau^{(k)}}\\ &+ I_{q_\tau^{(k)} < o_n < o_\tau^{(k)}} I_{q_\tau^{(k)} < o_\tau^{(k)}}\\ &- I_{o_\tau^{(k)} < o_n < q_\tau^{(k)}} I_{o_\tau^{(k)} < q_\tau^{(k)}} ,\end{aligned}$$

and obtain

$$\begin{aligned}E[\rho_\tau(o_n - o_\tau^{(k)})] &= S(Q^{(k)}, Q^{(k)}) - \underbrace{E[\rho_\tau(o_\tau^{(k)} - q_\tau^{(k)})]}_{\text{bias I}}\\ &- \underbrace{E[(o_n - o_\tau^{(k)}) I_{q_\tau^{(k)} < o_n < o_\tau^{(k)}} I_{q_\tau^{(k)} < o_\tau^{(k)}}]}_{\text{bias II}}\\ &+ \underbrace{E[(o_n - o_\tau^{(k)}) I_{o_\tau^{(k)} < o_n < q_\tau^{(k)}} I_{o_\tau^{(k)} < q_\tau^{(k)}}]}_{\text{bias II}} . \qquad (12)\end{aligned}$$

$S(Q^{(k)}, Q^{(k)})$ is now the true expected score or entropy. Bias term I in (12) resembles the expected score for $q_\tau^{(k)}$ if evaluated on the estimates $o_\tau^{(k)}$. Bias I is always positive, and as it comes with a negative sign, bias I always leads to an underestimation of the conditional entropy estimate. A part of bias I is due to the loss in degrees of freedom by the estimation of $o_\tau^{(k)}$, since both the estimation and the evaluation are performed on the same sample. It is the same effect that leads to the biased variance estimator if one has to estimate the expectation value. Bias II in turn is always positive and counteracts bias I. It only acts on those sample members that lie in between $q_\tau^{(k)}$ and its estimator $o_\tau^{(k)}$, and thus is particularly large if $o_\tau^{(k)}$ is strongly biased. It is responsible for the increased bias in reliability when K becomes large (compare Sec. 5.1.1 and Fig. 2).

Unbiased estimates of quantiles other than the median are hard to obtain. Hyndman and Fan (1996) suggest to use a median-unbiased quantile estimator, their type 8 estimator, which is approximately median-unbiased regardless of the underlying distribution. For normally distributed data there exists an approximately unbiased estimator which might be preferred in this case. Note that biases may be particularly large for high quantiles and large K (i.e. small subsamples).

For the expectation of the resolution (RES) we obtain a similar expression

$$\begin{aligned}
&E[RES^{(k)}] \\
&\quad = \frac{1}{N_k}\sum_{n\in\mathcal{I}_k}\left(E[\rho_\tau(o_n-\bar{o}_\tau)]-E[\rho_\tau(o_n-o_\tau^{(k)})]\right) \\
&\quad = S(\bar{Q},Q^{(k)})-S(Q^{(k)},Q^{(k)}) \\
&\qquad -\underbrace{E[\rho_\tau(\bar{o}_\tau-\bar{q}_\tau)]}_{\text{bias III}} \\
&\qquad -\underbrace{E[(o_n-\bar{o}_\tau^{(k)})I_{\bar{q}_\tau<o_n<\bar{o}_\tau^{(k)}}I_{\bar{q}_\tau<\bar{o}_\tau^{(k)}}]}_{\text{bias IV}} \\
&\qquad +\underbrace{E[(o_n-\bar{o}_\tau)I_{\bar{o}_\tau^{(k)}<o_n<\bar{q}_\tau}I_{o\bar{o}_\tau^{(k)}<\bar{q}_\tau}]}_{\text{bias IV}} \\
&\qquad +\text{bias I}-\text{bias II}.
\end{aligned}$$

The biases now result from the estimation of $\bar{q}_\tau$ and $q_\tau^{(k)}$, and the two components $S(Q^{(k)},Q^{(k)})$ and $S(\bar{Q},Q^{(k)})$, respectively. Since the estimates for the climatology are based on a significantly larger data set, the biases III and IV are small. We thus conclude, that a bias in the quantile estimator (I) leads to an overestimation of the resolution, whereas the bias resulting from the entropy estimates (II) reduces the resolution.

Bias terms III and IV also affect the uncertainty, as they result from the estimation of the climatological quantile

$$\begin{aligned}
E[UNC^{(k)}] &= \frac{1}{N_k}\sum_{n\in\mathcal{I}_k}E[\rho_\tau(o_n-\bar{o}_\tau)] \\
&= S(\bar{Q},Q^{(k)})-\text{bias III}+\text{bias IV}\,.
\end{aligned}$$

It is important to note that in contrast to the discretization error, the biases do not effect the overall score. As for the BS, unbiased estimators are unavailable, and at this stage, we are unable to propose a correction for the biases. The only solution to avoid large biases is to carefully define the binning intervals.

References

Atger F. 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review* **131**(8): 1509–1523.

—. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society* **130**(597): 627–646.

Ben Bouallègue Z. 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting* **28**(2): 515–524.

Ben Bouallègue Z, Theis SE, Gebhardt C. 2013. Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift* **22**(1): 49–59.

Bentzien S Friederichs P. 2012. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting* **27**(4): 988–1002.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1): 1–3.

Bröcker J. 2008. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review* **136**(11): 4488–4502.

—. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* **135**: 1512–1519.

—. 2011. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics* **39**: 655–667.

Bröcker J Smith LA. 2007. Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22**(3): 651–661.

Brown TA 1970. Probabilistic forecasts and reproducing scoring systems. Tech. Rep. RM-6299-ARPA, Santa Monica, California: RAND Corporation.

Candille G Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131**(609): 2131–2150.

DeGroot MH Fienberg SE. 1983. The comparison and evaluation of forecasters. *The statistician* **32**: 12–22.

Efron B Tibshirani RJ. 1993. *An Introduction to the Bootstrap*, Chapman&Hall/CRC.

Ferro CAT Fricker TE. 2012. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society* **138**(668): 1954–1960.

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications* **15**: 19–24.

Friederichs P Hense A. 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review* **135**: 2365–2378.

—. 2008. A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting* **23**(4): 659–673.

Friederichs P Thorarinsdottir TL. 2012. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23**: 579–594.

Gebhardt C, Theis SE, Paulat M, Ben Bouallègue Z. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variations of lateral boundaries. *Atmospheric Research* **100**: 168–177.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B (Methodological)* **69**: 243–268.

Gneiting T Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**: 359–378.

Gneiting T Ranjan R. 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics* **29**(3): 411–422.

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**(1): 107–114.

Hamill TM Juras J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* **132**: 2905–2923.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570.

Hoffman RN Kalnay E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**: 100–118.

Hyndman RJ Fan Y. 1996. Sample quantiles in statistical packages. *The American Statistician* **50**(4): 361–365.

Koenker R. 2005. *Quantile Regression*, vol. 38 of *Econometric Society Monographs*, Cambridge University Press.

Koenker R Bassett G. 1978. Regression quantiles. *Econometrica* **46**(1): 33–50.

Koenker R Machado JAF. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**(448): 1296–1310.

Laio F Tamea S. 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences* **11**(4): 1267–1277.

Matheson JE Winkler RL. 1976. Scoring rules for continuous probability distributions. *Management Science* **22**(10): 1087–1096.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

—. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**(2): 281–293.

—. 1996. General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Monthly Weather Review* **124**: 2353–2369.

Murphy AH Epstein ES. 1967. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology* **6**(5): 748–755.

Murphy AH Winkler RL. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **26**(1): 41–47.

—. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**(7): 1330–1338.

NCAR - Research Application Program 2012. *verification: Forecast verification utilities.*, r package version 1.35.

Peralta C, Ben Bouallègue Z, Theis SE, Gebhardt C, Buchhold M. 2012. Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research* **117**: D07108.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Savage LJ. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**(336): 783–801.

Schervish MJ. 1989. A general method for comparing probability assessors. *The Annals of Statistics* **17**(4): 1856–1879.

Scheuerer M. 2013. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society* **in press**.

Thorarinsdottir TL, Gneiting T, Gissibl N. 2013. Using proper divergence functions to evaluate climate models. *ArXiv e-prints* **1301.5927**.

Tödter J Ahrens B. 2012. Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review* **140**(6): 2005–2017.

Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences*, vol. 91 of *International Geophysics Series*, Academic Press, 2nd ed.