DTU Library

Quantile forecast discrimination ability and value

Ben Bouallègue, Zied; Pinson, Pierre; Friederichs, Petra

[Link back to DTU Orbit](#)

Quarterly Journal of the Royal Meteorological Society

Q. J. R. Meteorol. Soc. **00:** 1–10 (2013)

Quantile forecast discrimination ability and value

Zied Ben Bouallègue[a,b*], Pierre Pinson[c], Petra Friederichs[b]

[a] *Deutscher Wetterdienst, Offenbach, Germany*
[b] *Meteorological Institute, University of Bonn, Germany*
[c] *Technical University of Denmark, Denmark*

[*]Correspondence to: Z. Ben Bouallègue, Deutscher Wetterdienst, Frankfurter Strasse 135 63067 Offenbach, Germany.
E-mail:zied.ben-bouallegue@dwd.de

While probabilistic forecast verification for categorical forecasts is well established, some of the existing concepts and methods have not found their equivalent for the case of continuous variables. New tools dedicated to the assessment of forecast discrimination ability and forecast value are introduced here, based on quantile forecasts being the base product for the continuous case. The relative user characteristic (RUC) curve and the quantile value plot allow analysing the performance of a forecast for a specific user in a decision-making framework. The RUC curve is designed as a user-based discrimination tool and the quantile value plot translates forecast discrimination ability in terms of economic value. The relationship between the overall value of a quantile forecast and the respective quantile skill score is also discussed. The application of these new verification approaches and tools is illustrated based on synthetic datasets, as well as for the case of global radiation forecasts from the high resolution ensemble COSMO-DE-EPS of the German Weather Service.

1. Introduction

Verification of probabilistic weather forecasts is an area of intensive research and growing interest as ensemble forecasting is becoming a standard approach in numerical weather prediction. Ensemble prediction systems (EPS) issue a sample of possible future states of the atmosphere (Lewis 2005; Leutbecher and Palmer 2008). The forecasts can be interpreted in the form of a predictive distribution and probabilistic products can be derived in order to support and optimize forecast-based decision-making (Krzysztofowicz 1983). Appropriate tools for the assessment of probabilistic products from this perspective are therefore essential.

Such tools already exist for probabilistic products expressed in the form a probability forecast for a defined event. The relative operating characteristic (ROC) curve is a common verification tool for the assessment of probability forecasts (Mason 1982). The ROC curve is related to decision-making analysis and the corresponding fundamental property of the forecast is called *discrimination*. Forecast discrimination assesses whether the forecast can be used to successfully discriminate between the observations (Murphy 1991) or, said differently, whether appropriate decisions can be taken based on a forecast. Discrimination is translated in terms of *economic value* using a simple cost-loss model that allows the specificity of a user to be taken into account through the definition of a *cost-loss ratio*. The derived quantitative measure is called *value score* or *relative value* and is usually represented in the form of a probability value plot showing the forecast value as a function

of the user's cost-loss ratio (Richardson 2000; Wilks 2001; Zhu *et al.* 2002). The value of a forecast is defined as the benefit to a user as a result of making decisions based on a forecast and has to be distinguished from forecast quality, the overall agreement between forecast and observation (Murphy 1993). In a verification process, value and quality can be seen as being from the point of view of the forecast user and from the point of view of the forecast provider, respectively. The distinction between the two *types of goodness*, value and quality, is crucial since a non-linear relationship between them can lead to situations where a large improvement in the forecast quality does not imply an increase in the forecast value, or conversely, a small improvement in forecast quality can bring a notable benefit in terms of forecast value (Chen *et al.* 1987; Buizza 2001; Pinson 2013).

Probabilistic products can be expressed in terms of a probability when the focus is on a particular event of interest, but also in terms of a quantile when the focus is on a particular probability level of interest. While a probability forecast first requires the definition of an event, i.e. the categorization of the original information, a quantile forecast is a 'single-valued' forecast expressed in the unit of the variable being forecast. Considering here probabilistic products derived from EPS simulations for continuous variables, such as temperature, wind speed or global radiation, quantile forecasts of the predictive distributions allow one to work with a continuous forecast as the original one by defining a nominal probability level. The choice of a probability level is directly related to the user's loss function: a quantile forecast at a given probability level is the optimal forecast for users with a specific

asymmetry in their loss function (Koenker and Machado 1999; Friederichs and Hense 2007; Gneiting 2011a). Asymmetric loss functions find numerous real-world applications, in particular in the renewable energy sector (Pinson *et al.* 2007; Pinson 2013). For example, consider solar energy producers who have to agree in advance about the amount of energy to be provided to their customer. If too much energy is produced, they will sell the extra at a reduced price, while if they do not produce enough energy they will be heavily penalised. The asymmetry in the producer's loss function is drawn from the different penalties associated to over- and under-forecasting.

Based on the relationship between user's loss function and quantile forecast level, the quantile score (QS) is the natural scoring rule for assessing the quality of quantile forecasts (Koenker and Machado 1999; Friederichs and Hense 2007; Gneiting 2011a). More recently, the verification of quantile forecasts has benefited from the tradition and concepts stemming from the probability forecast verification framework. It has been shown that QS is a *proper* scoring rule and a decomposition of the score has been proposed (Bentzien and Friederichs 2014). The QS decomposition provides information about *reliability* and *resolution*, two other fundamental attributes of a probabilistic forecast (Toth *et al.* 2003).

The aim of the paper at hand is to extend the range of verification methods dedicated to the assessment of quantile forecasts. In particular, the assessment of quantile forecasts from the user's perspective, in a decision-making framework, is explored here. Based on a simple cost-loss model, the concepts of forecast discrimination and forecast value are revisited focusing on a specific user rather than on an specific event. First, a new tool is proposed for the analysis of user-based discrimination. The so-called relative user characteristic (RUC) curve and the associated summary measure are shown to be adequate for the assessment of quantile forecast discrimination ability. Secondly, quantile forecast value is discussed as an application of the value score to quantile forecasts. The quantile value plot, showing the economic value of a forecast as a function of a range of events of interest, is proposed as a new tool for the visualization of quantile forecast performance. Finally, the relationship between quantile forecast value and quantile skill score is discussed in the same vein as the relationship between probability forecast value and Brier skill score (Murphy 1969). The concepts developed are first illustrated with the help of synthetic datasets and in a second step applied to probabilistic forecasts derived from an EPS.

The manuscript is organized as follows: Section 2 describes the datasets that are used to illustrate the discussion. Section 3 introduces definitions and notations and describes the relationship between quantile forecast and forecast user within a cost-loss model framework. Section 4 discusses the concept of discrimination and Section 5 the application of the economic value score to quantile forecasts. Section 6 presents the conclusions.

2. Data

2.1. Synthetic datasets

In order to illustrate the concepts discussed hereafter, we make use of synthetic and real datasets. The synthetic data are derived from a toy-model based on normal distributions often used to illustrate verification discussions (e.g. Hamill 2001; Weigel 2011). The toy-model is kept simple in order to facilitate the interpretation of the results.

We consider a signal s, normally distributed, written $s \sim \mathcal{N}(0, 1)$. We assume that the observations are randomly drawn from a distribution $\mathcal{N}(s, 1)$ and the associated predictive distribution described by $\mathcal{N}(s + \beta, \sigma)$ where β is the unconditional bias

parameter and σ the dispersion parameter. We define the following test-cases:

A_0 : $\beta = 0, \sigma = 1$ (a perfect probabilistic forecast) ,
A_1 : $\beta = -0.75, \sigma = 1$ (a biased forecast),
A_2 : $\beta = 0, \sigma = 1/3$ (an underdispersive forecast),
B : $\beta = \epsilon_B, \sigma = 1$ (a forecast with white noise),

where ϵ_B is derived from a uniform distribution defined on $[-5, 5]$. The first three datasets A_0, A_1 and A_2 differ only in terms of biases while the fourth dataset B corresponds to a forecast with a dynamically disturbed signal. The simulation setup, the values associated to β, σ and ϵ_B, has been chosen in order to make clear the reading of the Figures in Section 4 and 5 but does not affect the general interpretation of the simulation results.

2.2. COSMO-DE-EPS

Real datasets are provided by COSMO-DE-EPS, a regional ensemble prediction system run operationally at Deutscher Wetterdienst, Offenbach, Germany. The ensemble system is based on a 2.8 km grid resolution version of the COSMO model (Steppeler *et al.* 2003; Baldauf *et al.* 2011) with a model domain that covers Germany and parts of the neighbouring countries. The ensemble comprises 20 members including variations in initial conditions, physics parameterisations and boundary conditions (Gebhardt *et al.* 2011; Peralta *et al.* 2012).

COSMO-DE-EPS has been first developed focusing on high-impact weather events (Ben Bouallègue *et al.* 2013; Ben Bouallègue and Theis 2014) and is planned to be used for energy-applications. The focus in this paper is on global radiation (the sum of direct and diffuse shortwave radiations) which is the main weather variable affecting solar energy forecasts. Verification is applied to the 0300UTC run with a forecast horizon ranging between 5 and 15 hours. Two periods of 3 months are compared: winter (December, January, February) 2012/2013 and summer (June, July, August) 2013. The observation dataset consists of pyranometer measurements from 32 stations distributed over Germany and quality controlled (Becker and Behrens 2012).

Global radiation forecasts and observations are transformed into clearness index before verification. The clearness index is defined as the ratio between global radiation at ground and global radiation at the top of the atmosphere (Badescu 2008). This pre-processing of the data allows climatological effects and misinterpretation of the verification results to be avoided (Hamill and Juras 2006).

3. Definitions and framework

3.1. Quantile forecast, quantile score, and quantile skill score

We first consider the quantity to be forecast (or *observation*) $\Omega \in \Re$ that we assume to be a continuous random variable driven by a stochastic process. An observed event E is defined by a threshold ω as $E : \Omega \geq \omega$. The base rate π of an event E (or climatological frequency) corresponds to:

$$\pi = Pr(\Omega \geq \omega). \tag{1}$$

Consider now a predictive cumulative distribution $F(x)$. The probability forecast p_ω of event E is defined as:

$$p_\omega = 1 - F(\omega). \tag{2}$$

The quantile forecast q_τ at probability level τ ($0 \leq \tau \leq 1$) is defined as:

$$q_\tau := F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\} \tag{3}$$

such that the relationship between a probability forecast and a quantile forecast is expressed as:

$$p_{q_\tau} = 1 - \tau. \qquad (4)$$

Figure 1 shows an example of a cumulative distribution function $F(x)$. A threshold ω and the associated probability forecast $1 - p_\omega$ as well as a probability level τ and the associated quantile forecast q_τ are shown on the plot.

The quantile score (QS) is the scoring rule applied in order to assess the quality of a quantile forecast. QS is based on an asymmetric piecewise linear function ρ_τ called the check function. The check function was first defined in the context of quantile regression (Koenker and Bassett 1978):

$$\rho_\tau(u) = u[\tau - I(u < 0)] = \begin{cases} \tau u & \text{if} \quad u \geq 0 \\ (\tau - 1)u & \text{if} \quad u < 0 \end{cases} \qquad (5)$$

where $I(.)$ is an indicator function having value 1 if the condition in parenthesis is true and zero otherwise. QS results from the mean of the check function applied to the pairs $i = 1, ..., N$ of observation Ω_i and quantile forecast $q_{\tau,i}$ following

$$QS = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(\Omega_i - q_{\tau,i}), \qquad (6)$$

where N is the size of the verification sample. Developing Eq. (6) we can write

$$QS = \frac{1-\tau}{N} \sum_{i:\Omega_i < q_{\tau,i}} (q_{\tau,i} - \Omega_i) + \frac{\tau}{N} \sum_{i:\Omega_i \geq q_{\tau,i}} (\Omega_i - q_{\tau,i}) \qquad (7)$$

The scoring rule consists of penalties $1 - \tau$ and τ per unit of Ω associated with under-forecasting and over-forecasting, respectively.

Skill scores are computed in order to measure the relative benefit of using a forecast compared to a reference forecast (Wilks 2006). The quantile skill score (QSS) measures the skill of a quantile forecast compared to a reference quantile forecast. Considering the climatology as reference, QSS corresponds to:

$$QSS = \frac{QS_{\text{forecast}} - QS_{\text{climate}}}{QS_{\text{perfect}} - QS_{\text{climate}}} = 1 - \frac{QS_{\text{forecast}}}{QS_{\text{climate}}} \qquad (8)$$

where QS_{forecast}, QS_{perfect} and QS_{climate} represent the quantile scores of the forecast under assessment, of a perfect deterministic forecast and of a climatological τ-quantile forecast, respectively. QS_{perfect}, by definition, equals 0 and a climatological τ-quantile forecast, noted Ω_τ, is here defined as the τ-quantile of the observation distribution over the verification sample. Using the sample climatology for the reference forecasts, one should be aware, interpreting the results, that it produces a disadvantage for the forecasts, especially as the sample size is relatively small.

3.2. *Cost-loss model and optimal decision-making*

The framework used to discuss the concept of *user* and *decision-making* is based on a static cost-loss model (Thompson 1962; Katz and Murphy 1997). The cost-loss model describes situations of dichotomous decisions: a user has to decide whether or not to take protective action against potential occurrence of an event E. The decision is made based on a decision variable (or forecast) Λ. A decision criterion λ applied to the decision variable defines an action $A : \Lambda \geq \lambda$. Taking action implies a cost C. In the case of occurrence of the event E without preventive action, a loss L is encountered. The cost-loss ratio is denoted α:

$$\alpha = \frac{C}{L}. \qquad (9)$$

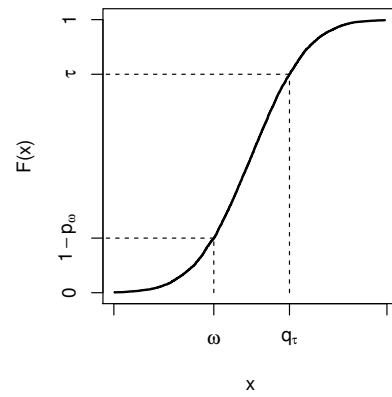**Figure 1.** Example of a predictive cumulative distribution function $F(x)$. Probabilistic products are derived either fixing a threshold ω and deriving the associated probability forecast p_ω, or fixing a probability level τ and deriving the associated quantile forecast q_τ.

A user with cost-loss ratio α is called hereafter an α-user. Based on this simple model the optimal decision strategy of an α-user can be discussed (e.g. Richardson 2011). The problem consists of finding, for a decision variable Λ, the *critical* decision criterion λ_α that minimizes the α-user mean expense if actions are taken when $\Lambda \geq \lambda_\alpha$.

Consider first the case of a probability forecast p_ω as a decision variable. Based on p_ω, does the user have to take action or not? In order to answer this question, the average expenses in the cases of positive and negative answers are compared. If the answer is yes, the user encounters a cost C on every occasion, so the average expense $\bar{E}_{\text{yes}}$ is simply

$$\bar{E}_{\text{yes}} = C. \qquad (10)$$

If the answer is no, the user has no cost but a loss L on each occasion where the event occurs, so on average the user's expense $\bar{E}_{\text{no}}$ is

$$\bar{E}_{\text{no}} = L Pr(\Omega \geq \omega \mid p_\omega), \qquad (11)$$

where $Pr(\Omega \geq \omega \mid p_\omega)$ is the probability that the event occurs when the probability forecast p_ω is issued. So, users with a cost-loss ratio $\alpha < Pr(\Omega \geq \omega \mid p_\omega)$ should take preventive action, while users with a greater cost-loss ratio should not. The critical decision criterion $p_\omega^\star$ associated with the decision variable p_ω is thus defined as

$$p_\omega^\star = \{p_\omega \mid Pr(\Omega \geq \omega \mid p_\omega) = \alpha\}. \qquad (12)$$

Thus, the action based on the probability forecast $A : p_\omega \geq p_\omega^\star$ optimizes the user's mean expense in the long term.

If the forecast is reliable, we have by definition $Pr(\Omega \geq \omega \mid p_\omega) = p_\omega$: the event actually happens with an observed relative frequency consistent with the forecast probability (Bröcker 2009). The optimal decision is then to take action if

$$p_\omega \geq \alpha. \qquad (13)$$

When the probability forecast is compared to the cost-loss ratio in order to decide whether or not to take action (without additional information about forecast reliability), we say that the probability forecast is taken at face value. For example, consider users who have to decide whether or not to take preventive action against precipitation occurrence. If the forecast probability of precipitation is 10%, users with cost-loss ratio lower than 10% take action. If the forecast is not reliable, the critical decision criterion is no longer α but has to be adjusted following Eq. (12). Statistical adjustments of the forecast based on past data is usually referred as *forecast calibration* (e.g. Gneiting *et al.* 2007).
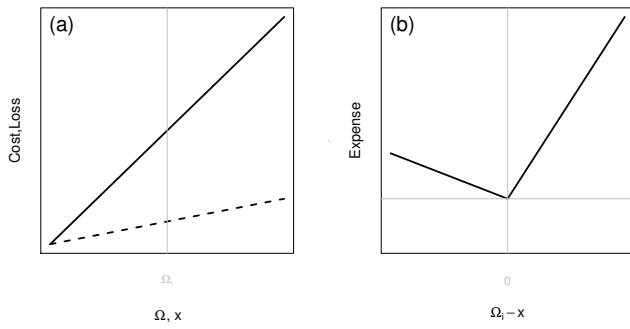
Figure 2. (a) Cost (dashed line) as a function of the level of protection x and loss (full line) as a function of the observation Ω. An observation Ω_i is represented by a vertical line. (b) Expense as a function of the difference between the observation Ω_i and the level of protection x. The horizontal line indicates the expense for a perfect level of protection.

Consider now a quantile forecast q_τ as a decision variable. We apply the same reasoning as for a probability forecast. The critical decision criterion $q_\tau^\star$ associated with q_τ is defined as

$$q_\tau^\star = \{q_\tau \mid Pr(\Omega \geq \omega \mid q_\tau) = \alpha\} \tag{14}$$

such that taking action when $q_\tau \geq q_\tau^\star$ minimizes the user mean expense. By definition, a quantile forecast is reliable if it satisfies

$$Pr(\Omega \geq \omega \mid q_\tau = \omega) = 1 - \tau, \tag{15}$$

i.e. the observed relative frequency of the event defined by the quantile forecast is consistent with the quantile forecast probability level. Eq. (14) has a straightforward solution

$$q_\tau^\star = \omega \tag{16}$$

when the decision variable is the quantile forecast at probability level τ defined as

$$\tau = 1 - \alpha. \tag{17}$$

Taking action when $q_\tau \geq \omega$ with $\tau = 1 - \alpha$ is equivalent to taking action when $p_\omega \geq \alpha$ since the cumulative probability distribution function $F(x)$ is by definition monotonically increasing (see e.g. Figure 1). Hence, a quantile forecast is taken at face value when the user's decision is made based on the comparison of the forecast with the event threshold ω. In our example, if the 90%-quantile forecast of precipitation is greater than zero, a user with cost-loss ratio $\alpha = 1 - 0.9 = 0.1$ takes preventive action.

In a general form, the critical decision criterion λ_α for an α-user is defined by

$$\lambda_\alpha = \{\lambda \mid Pr(\Omega \geq \omega \mid \lambda) = \alpha\} \tag{18}$$

where the decision variable could equally be the probability forecast p_ω or the quantile forecast q_τ with $\tau = 1 - \alpha$. Provided that the forecasts are reliable, the critical decision criteria are known and have a simple expression (Eqs (13,16)). In the following, we say that the decision variable is taken at *face value* when the user applies the decision criterion valid for a reliable forecast, irrespective of whether the forecast is actually reliable or not.

3.3. *Quantile forecast user*

The dichotomous decision problem is extended to a continuous decision problem considering the cost C and the loss L as unitary cost and unitary loss, respectively (Epstein 1969; Roulston *et al.* 2003). The cost of taking protection is a linear function of the level of protection x and the loss without protection is a linear function

of the observation Ω, as illustrated in Figure 2. The optimization problem consists of finding the level of protection that minimizes the expected user expense.

Considering a variable defined on $\Re+$ (the generalization to variables defined on $\Re$ is straightforward), the expense associated with a level of protection x corresponds to Cx. If the observation is Ω, then protection is perfect if $x = \Omega$. But if $x > \Omega$, then there is an unnecessary expense due to a larger level of protection than is actually needed. If the observation Ω is greater than the level of protection, then a loss $L(\Omega - x)$ is encountered and a cost $C(\Omega - x)$ avoided. Formally, we can write the expense function E as

$$E = \begin{cases} C(x - \Omega) & \text{if } \Omega < x \\ (L - C)(\Omega - x) & \text{if } \Omega \geq x. \end{cases} \tag{19}$$

The expense function is represented in Figure 2. If divided by L, the expense function is an asymmetric loss function equivalent to the check function defined in Eq. (5), where the asymmetry is given by $\tau = \frac{L-C}{L}$. Thus the optimal level of protection $x^\star$ which minimizes the user's mean expense corresponds to the $1 - \alpha$ quantile of the true predictive distribution of Ω.

This result is not new: quantile forecasts arise as an optimal solution for users with an asymmetric linear loss function (Koenker and Bassett 1978; Christoffersen and Diebold 1997). More recently, it has been shown that quantile forecasts are optimal forecasts in a stochastic optimization framework for a more general class of loss functions (Gneiting 2011b).

Asymmetric loss functions find a number of applications, in particular for operational decision-making problems related to the integration of renewable energies into the electricity grid. For example, asymmetric loss functions can be associated with market participants who want to optimize their bids or system operators who have to optimize their reserves. The different penalties associated to over- or under-forecasting draw the asymmetry in the user's loss function. The user's optimal forecast corresponds then to a specific quantile of the predictive distribution where the probability is defined by the user's cost-loss ratio (Pinson *et al.* 2007; Pinson 2013).

4. Discrimination

Based on the discussion developed in the previous Section, continuous decision making is seen in the following as a *continuum* of dichotomous decisions. For each threshold ω of the event spectrum, the question is whether to take action for the next unit of the variable. The adequate decision for a user in order to minimize the expected expense is a function of his (her) cost-loss ratio as defined in Eq. (18). Moreover, the relationship between cost-loss ratio and quantile probability level, $\tau = 1 - \alpha$, makes implicit the cost-loss ratio α of a user as soon as the level τ of the quantile forecast used as decision variable is selected.

4.1. *General verification framework*

A general framework for forecast verification is based on the joint distribution of forecasts and observations (Murphy and Winkler 1987). The overall agreement between forecasts and observations is called quality and is measured by scoring rules, like QS for quantile forecasts. In order to access more information about the forecast performance, two factorizations of the joint distribution, into conditional and marginal distributions, can be applied: the *calibration-refinement* (CR) factorization when conditioning on the forecasts and the *likelihood-base rate* (LBR) factorization when conditioning on the observations. Summary measures based on these two factorizations are associated with attributes, fundamental characteristics of the forecast. Reliability and resolution are derived from the CR factorization while

discrimination is derived from the LBR factorization (Murphy and Winkler 1992).

Here the focus is on discrimination, the key forecast attribute for decision-making processes. A general definition of discrimination is "the ability of a forecasting system to produce different forecasts for those occasions having different realized outcomes" (Wilks 2006). Discrimination assumes calibration of a forecast and so does not account for reliability discrepancies. Though biases can strongly affect the skill of a forecast, this can be seen as an advantage since reliability, unlike discrimination, can be improved by recalibration. Investigating discrimination of a forecast means therefore focusing on the *necessary condition for skill* (Jolliffe and Stephenson 2005).

Discrimination assessment is here discussed in terms of *event* and *action* within the dichotomous decision framework. Regarding the LBR factorization, it is common practice to analyse discrimination in terms of hit rate H and false alarm rate F defined as

$$H = Pr(\Lambda \geq \lambda \mid \Omega \geq \omega) \tag{20}$$

and

$$F = Pr(\Lambda \geq \lambda \mid \Omega < \omega), \tag{21}$$

respectively. Actions $A : \Lambda \geq \lambda$ and events $E : \Omega \geq \omega$ are dichotomous, each presenting two alternatives, so H and F can be easily derived from the construction of a 2×2 contingency table. No discrimination corresponds to the case where:

$$H = F \tag{22}$$

for all $\lambda \in \Lambda$ and $\omega \in \Omega$, meaning that actions and event occurrence are independent (Bröcker 2014).

4.2. Event-based discrimination

We first focus on one particular event defined by a threshold ω, with event-specific hit rate H_λ and false alarm rate F_λ. A popular way to assess discrimination (Eq. (22)) is to plot the set of points (F_λ, H_λ) for a range of actions with $\lambda \in \Lambda$. The resulting curve is known as the relative operating characteristic (ROC) curve. When action and event occurrence are independent, the ROC curve is a diagonal line. Concavity of the curve indicates a discrimination ability in the forecast and the area under the curve (AUC) becomes a quantitative measure of forecast discrimination (Mason 1982). Figure 3 (a) shows an example of a ROC curve for the synthetic dataset A_0. The event of interest is $E : \Omega \geq 0$ with a base rate $\pi = Pr(\Omega \geq 0)$ of 0.5. The respective probability forecast $p_0 = 1 - F(0)$, the probability that simulation A_0 exceeds 0, is used as decision variable.

The interpretation of the ROC curve can be related to the dichotomous decision model described in Section 3.2 as discussed for example in Richardson (2011). In order to describe this relationship, we consider the slope of the ROC curve, defining first the gradient of a line joining two successive ROC points (F_λ, H_λ) and $(F_{\lambda+\Delta\lambda}, H_{\lambda+\Delta\lambda})$:

$$\frac{H_\lambda - H_{\lambda+\Delta\lambda}}{F_\lambda - F_{\lambda+\Delta\lambda}} = \frac{Pr(\Lambda \geq \lambda \mid \Omega \geq \omega) - Pr(\Lambda \geq \lambda + \Delta\lambda \mid \Omega \geq \omega)}{Pr(\Lambda \geq \lambda \mid \Omega < \omega) - Pr(\Lambda \geq \lambda + \Delta\lambda \mid \Omega < \omega)}. \tag{23}$$

The slope of the curve γ is obtained when $\Delta\lambda$ tends to 0:

$$\gamma(\lambda, \omega) = \frac{Pr(\Lambda = \lambda \mid \Omega \geq \omega)}{Pr(\Lambda = \lambda \mid \Omega < \omega)} \tag{24}$$

where the ratio is also know as the *likelihood ratio* (Bröcker 2011). Using the Bayes rule and the definition of the critical decision criterion of an α-user in Eq. (18), we can write

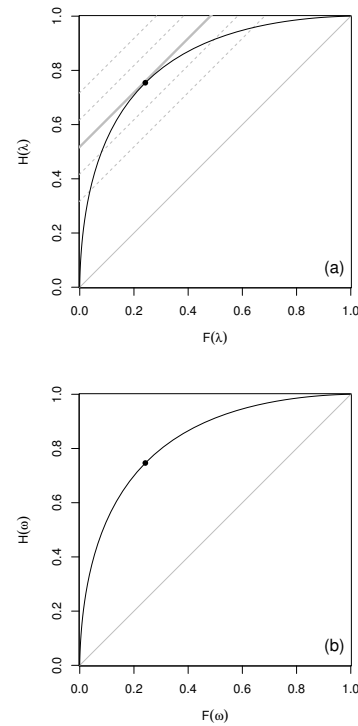$$\gamma(\lambda_\alpha, \omega) = \frac{1 - \pi}{\pi} \frac{\alpha}{1 - \alpha} \tag{25}$$

Figure 3. Discrimination curves for decision variables from the synthetic dataset A_0. The diagonal lines are the no discrimination lines. The points correspond to the (F,H) pair for the event $\Omega \geq 0$ and the action associated with the 50%-users. (a) ROC curve of the probability forecast p_0 for the event $E : \Omega \geq 0$, with base rate $\pi = 0.5$, and equi-cost lines (in grey) of slope $\gamma = 1$. (b) RUC curve of the quantile forecast $q_{0.5}$ for the user with cost-loss ratio $\alpha = 0.5$.

where $\pi = Pr(\Omega \geq \omega)$ is the base rate of an event $E : \Omega \geq \omega$ and λ_α the corresponding critical decision criterion of an α-user.

The range of decision criterion λ used to derive the ROC curve (F_λ, H_λ) corresponds to a range of critical decision criteria associated with users with different cost-loss ratios. Each point of the ROC curve is associated with a specific α-user that is identified by the slope of the curve at that point. The slope possibly ranges between 0 and $+\infty$ at the right-top and the bottom-left corners of the ROC plot respectively. Moving along the curve from the top to the bottom consists in varying the cost-loss ratio α between 0 and 1.

For example, consider a user with a cost-loss ratio $\alpha = 50\%$. In Figure 3, the point of the ROC curve with slope $\gamma = 1$ is highlighted ($\alpha = 0.5, \pi = 0.5$ in Eq. (25)). This point indicates the performance of the forecast in terms of H and F for this particular user. Conversely, the decision criterion applied to obtain this point corresponds to the critical decision criterion for the 50%-user.

The ROC curve applied to a decision variable, then, corresponds to testing whether actions and event occurrence are independent for one event and a range of users with different cost-loss ratios. The ROC curve is an *event specific* but *user unspecific* discrimination tool and is therefore well-adapted to probability forecast discrimination assessment.

4.3. User-based discrimination

We focus now on a user with cost-loss ratio α. The critical decision criterion λ_α defines the action of this specific user with respect to an event. We define then the user-specific hit rate H_ω and false alarm rate F_ω as in Eqs (20) and (21) for a fixed α. In order to test Eq. (22), the set of points (F_ω, H_ω) are plotted for a range of events. We call the resulting curve a *relative user characteristic* (RUC) curve because it is a comparison of two user characteristics (F_ω and H_ω) as the event definition varies. As for the ROC curve, the no discrimination line corresponds to
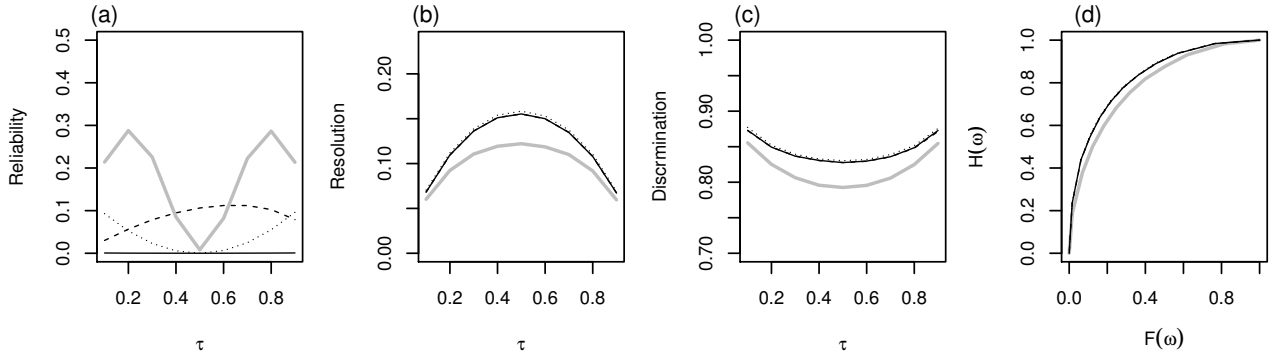
Figure 4. (a) Reliability, (b) resolution and (c) discrimination as a function of the probability level τ of the τ-quantile forecasts and (d) RUC curves for the 50%-quantile forecasts ($\tau = 0.5$). The results are shown for the simulation test cases A_0 (full lines), A_1 (dashed lines), A_2 (dotted lines) and B (full grey line).

the diagonal line and concavity of the curve indicates forecast discrimination ability.

Figure 3 (b) shows an example of a RUC curve valid for a user with cost-loss ratio $\alpha = 50\%$. In this example, the decision variable is the 50%-quantile forecast from the synthetic dataset A_0. Moving along the RUC curve from the bottom left corner to the top right corner involves varying the event under focus, the event's base rate varying from 0 to 1, respectively. The point with slope $\gamma = 1$ corresponds to the event $E : \Omega \geq 0$ with base rate $\pi = 0.5$. This point is obviously the same as in Figure 3 (a).

In order to produce a RUC curve, critical decision criteria have to be known for a range of events. They can be estimated resolving Eq. (14) numerically. In practice, critical decision criteria can also be estimated by means of a reliability diagram. For example, a reliability diagram for quantile forecasts plots the conditional observed quantile as a function of quantile forecast categories (Bentzien and Friederichs 2014). With regard to Eq. (15), we can deduce that the mean forecast in each forecast category (horizontal axis of the diagram) is an estimation of the critical decision criteria associated with the events defined by the corresponding conditional observed quantile (vertical axis of the diagram).

The RUC curve is user specific (and event unspecific) and therefore well-adapted to quantile forecast discrimination. A summary measure of quantile discrimination ability is obtained mimicking the ROC framework: the area under the RUC curve, noted here AUC', is proposed as a quantitative measure of discrimination for quantile forecasts. Considering n_E events $E_i : \Omega \geq \omega_i$, $i = 1, ..., n_E$ with increasing base rate, AUC' is estimated by a trapezoidal approximation as

$$AUC' = \sum_{i=0}^{n_E} 0.5(H_{\omega_{i+1}} + H_{\omega_i})(F_{\omega_{i+1}} - F_{\omega_i}) \quad (26)$$

with the trivial points $H_{\omega_0} = F_{\omega_0} = 0$ (for an event of base rate 0) and $H_{\omega_{n_E+1}} = F_{\omega_{n_E+1}} = 1$ (for an event of base rate 1). In order to reduce the biases introduced by the limited number of RUC points, the RUC curve can be fitted under a bi-normal assumption. The procedure involves considering F_ω and H_ω as both expressed as integrations of the standard normal distribution (Mason 1982). The bi-normal model has been shown to be valid in most cases when applied in the ROC framework (Mason and Graham 2002; Atger 2004).

The properties of the RUC curve and AUC' are discussed with the help of illustrative examples based on 4 simple simulation test cases (see Section 2.1). In Figure 4, the forecast attributes reliability, resolution and discrimination are shown as a function of the probability level τ of the τ-quantile forecast under assessment. RUC curves for the 50%-quantile forecasts are also shown. Quantile forecast reliability and resolution are estimated using the decomposition of the quantile score (Bentzien and

Friederichs 2014) while discrimination curves and summary measures are estimated based on the bi-normal assumption.

Figures 4 (a) shows the lack of reliability, which occurs by construction in the simulations A_1, A_2 and B. In Figures 4 (b) and 4 (c), resolution and discrimination measures deliver a similar message comparing the different simulations which illustrates the idea that "resolution and discrimination are the two faces of the same coin" (Bröcker 2014). Resolution and discrimination exhibit however different behaviours as a function of the probability level reflecting the fact that the first takes the forecaster's perspective and the second the user's perspective. Moreover, discrimination ability is identical for the simulations A_0, A_1 and A_2: they are unaffected by biases and dispersion errors. Indeed, AUC' is by construction insensitive to conditional and unconditional biases. In contrast, the forecast derived from simulation B with a perturbed signal presents less discrimination ability than forecasts from the other simulations, in particular for the 50%-quantile forecast. Focusing on users with cost-loss ratio $\alpha = 0.5$ ($\tau = 0.5$), RUC curves for the 50%-quantile forecasts of simulations A_0, A_1, A_2, and B are shown in Figure 4 (d). The largest discrepancies between simulations A and B are visible at the centre of the RUC curves, so for events with intermediate base rates, while for events with small or large base rates the RUC curves tend to overlap.

5. Value of quantile forecasts

5.1. Economic value

The cost-loss model described in Section 3.2 has been used to develop the concept of economic value of a probabilistic forecast. The forecast value is assessed considering decision-making made by an α-user about the occurrence of an event. The value of a forecast (also called value score or relative value) is defined as

$$V = \frac{\bar{E}_{\text{climate}} - \bar{E}_{\text{forecast}}}{\bar{E}_{\text{climate}} - \bar{E}_{\text{perfect}}}, \quad (27)$$

where the mean expense $\bar{E}$ of an α-user is estimated when decisions are based on a forecast ($\bar{E}_{\text{forecast}}$), on a perfect deterministic forecast ($\bar{E}_{\text{perfect}}$), or on climatological information ($\bar{E}_{\text{climate}}$) (Richardson 2000; Wilks 2001; Zhu *et al.* 2002). V is a measure of the economic gain (or reduction of mean expense) when using a forecast relative to the gain when using a perfect deterministic forecast.

Following *e.g.* Richardson (2011), the mean expense of a forecast user can be written as

$$\bar{E}_{\text{forecast}} = F(1 - \pi)C - H\pi(L - C) + \pi L, \quad (28)$$

where H and F are the hit rate and false alarm rate as defined in Eqs (20) and (21), respectively, and π the base rate of the event of
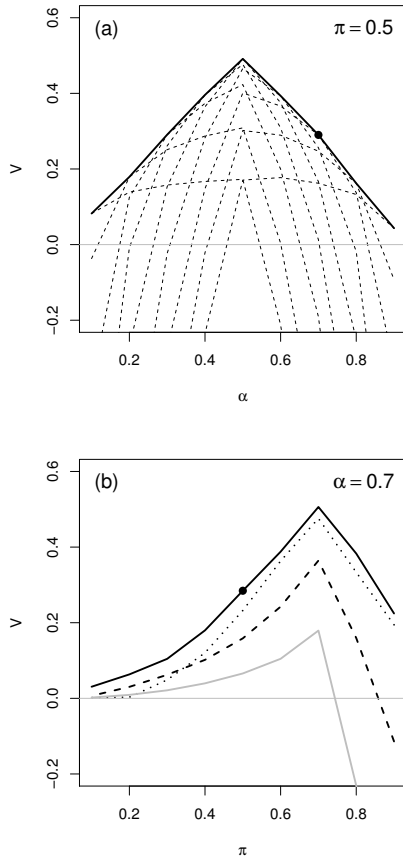
Figure 5. (a) Potential value V of the probability forecast from simulation A_0 for the event defined as $E : \omega \geq 0$ with base rate $\pi = 0.5$. The dashed lines represent the forecast value when the probability levels $0.1, 0.2, ..., 0.9$ are chosen as decision criterion. The full line represents the envelope of the dashed lines. (b) Value V for users with cost-loss ratio $\alpha = 0.7$ of the 30%-quantile forecasts taken at face value from the 4 synthetic datasets: A_0 (full black line), A_1 (dashed line), A_2 (dotted line) and B (full grey line). The black point is the common point of the two plots: value of the simulation A_0 for the event with base rate $\pi = 0.5$ and a user with cost loss ratio $\alpha = 0.7$.

interest. A user with a perfect deterministic forecast at hand has to face costs only. The user mean expense corresponds in this case to:

$$\bar{E}_{\text{perfect}} = \pi C. \tag{29}$$

For a user who bases his (her) decision on climatological information, the optimal mean expense is expressed as

$$\bar{E}_{\text{climate}} = \begin{cases} C & \text{if} \quad \alpha < \pi \\ \pi L & \text{if} \quad \alpha \geq \pi, \end{cases} \tag{30}$$

depending on the relationship between cost-loss ratio and base rate. Combining Eqs (28)-(30), the value of a forecast can finally be written as:

$$V = \begin{cases} (1 - F) - \left(\dfrac{\pi}{1 - \pi}\right)\left(\dfrac{1 - \alpha}{\alpha}\right)(1 - H) & \text{if} \quad \alpha < \pi \\ H - \left(\dfrac{1 - \pi}{\pi}\right)\left(\dfrac{\alpha}{1 - \alpha}\right) F & \text{if} \quad \alpha \geq \pi. \end{cases} \tag{31}$$

So, the economic value V is defined for an event with base rate π and a user with cost-loss ratio α. V depends on the forecast performance in terms of H and F.

Applied to a probability forecast, the event's base rate is fixed and the value of a probability forecast is generally represented in the form of a probability value plot showing V as a function of α. An example is provided in Fig. 5 (a), applied to simulation A_0 considering the event $E : \Omega \geq 0$. The forecast value curves are plotted for a range of probabilities as decision criterion, then

the optimal values for each α-user (the upper envelope of the relative value curves) is selected to represent the value of the probabilistic forecast system (e.g. Richardson 2000; Wilks 2001). The probability value plot is related to the ROC framework since the pairs (F, H) of Eq. (31) are the ones used to draw the ROC curve. It has also been shown that the overall value of a probability forecast, considering all potential users, corresponds to the Brier skill score of the forecast if the distribution of cost-loss ratio is uniform over all users (Murphy 1969; Richardson 2011).

5.2. Quantile value plot

Applied to a quantile forecast, so focusing on a α-user, the value score is evaluated for a range of events of interest defined for example by their base rate π. A new tool is therefore proposed for the assessment of quantile forecast performance: the quantile value plot which represents how V varies as a function of π. This is illustrated in Figure 5 (b). The value of the 30%-quantile forecasts is plotted when the quantile forecasts derived from simulations A_0, A_1, A_2, and B are taken at face value. Taking a quantile at face value means using it as it is, so for each event it implies considering the event threshold as decision criterion (see Section 3.2). An alternative is to apply the critical decision criteria, i.e. to use the (F, H) pairs from the RUC curve to estimate the value in Eq. (31). We talk then about *potential* value since it corresponds to the maximum value of the forecast, i.e. the maximum that could be potentially reached if an adequate calibration is applied to the forecast. Indeed, value and *potential* value are by definition identical if the forecast is reliable.

A parallel between probability value plot and quantile value plot can be draw. In a probability value plot, the decision variable is a probability forecast, the base rate π of the event under focus is fixed and the forecast value V is then plotted for a range of cost-loss ratios. The role of α and π are inverted in order to produce a quantile value plot rather than a probability value plot. The cost-loss ratio is defined by the quantile probability level and a range of events of interest are scanned. It results that the cost-loss ratio of the end-user does not appear explicitly in a quantile value plot while it corresponds to the horizontal axis in a value plot for probability forecasts.

The fundamental properties of V are however the same when focusing on one event or on one user. These properties (demonstrations can be found e.g. in Richardson 2011) are recalled here. First, the forecast value reaches its maximum when $\pi = \alpha$ (or noted differently when $\pi = 1 - \tau$). For instance, a forecast user with a cost-loss ratio of $\alpha = 0.1$ draws a maximum benefit from a forecast if his (her) event of interest has a climatological probability of occurrence of 10%. Secondly, the value of a reliable forecasts (full line in Figure 5 (b)) is always greater than the value of the same forecast with biases (dashed and dotted lines in Figure 5 (b)). The value of the reliable forecast corresponds to the potential value of the two other datasets. Finally, the potential value is by definition always non-negative.

5.3. A real example

The tools introduced for the assessment of quantile forecast discrimination and value are here applied to a real dataset. Quantile forecasts of global radiation are derived from COSMO-DE-EPS and assessed for two periods of the year 2013. Results for the winter period are shown in Figure 6 and results for the summer period in Figure 7. Quantile discrimination is estimated with the area under the RUC curve (AUC') for probability levels $\tau = 0.1, 0.2, ..., 0.9$. A deeper analysis is performed for the 10%-, 50%- and 90%-quantile forecasts with the help of quantile value plots.
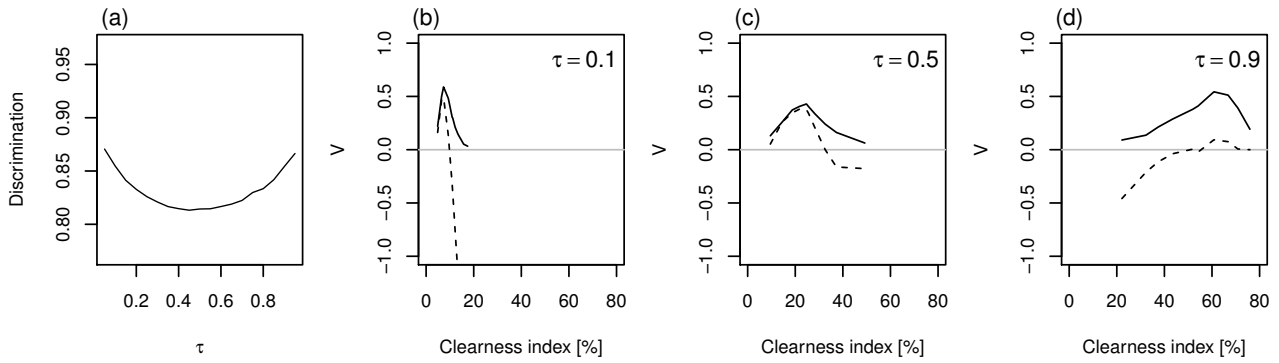
Figure 6. Verification results for COSMO-DE-EPS global radiation forecasts during winter 2012/2013: quantile discrimination ability (AUC') as a function of the probability level (a), potential value (full line) and actual value (dashed line) of the 10%-quantile forecast (b), 50%-quantile forecast (c) and 50%-quantile forecast (d) as a function of the event of interest defined by thresholds of the clearness index in %.
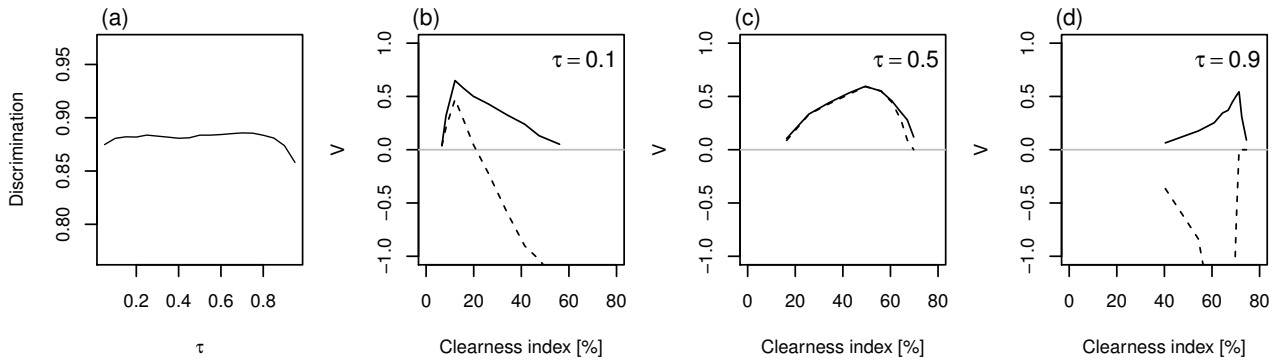

Figure 7. Same as Figure 6 but for summer 2013.

The discrimination ability of the EPS quantile forecasts varies as a function of the probability level but is greater than 0.80 which can be interpreted as good performance. For the winter season, discrimination is higher for high and low probability levels than intermediate ones whereas for the summer season, discrimination is approximately constant over the probability levels with a tendency to decrease for high levels. Inspection of the quantile value plot allows a deeper insight into the forecast potential performance. This could be relevant for quantile users with a specific interest in only one part of the event spectrum. For example, consider a user with a non linear loss function, the loss becoming zero if the outcome is below (above) a given threshold T. The user can then interpret the quantile value plot focusing on the range of events below (above) T disregarding the rest of the plot. The potential value and the actual value of the COSMO-DE-EPS quantile forecasts are plotted as a function of event in terms of the clearness index in % to simplify the reading of the plots. An event has a different base rate for each season which complicates a direct comparison of the quantile value plot in Figures 6 and 7. The comparison of the potential and actual values (full lines and dashed lines) shows the impact of the reliability discrepancies on the forecast value.

5.4.   *Overall value and Quantile Skill Score*

As a final step in drawing a parallel between probability forecast verification and quantile forecast verification, the relationship between value and skill score with climatology as a reference is explored. It has been shown that the overall value of a probability forecast is equivalent to its Brier Skill Score (BSS) when the users have a uniform distribution of cost-loss ratio (Murphy 1969; Richardson 2011). Similarly, we now investigate the relationship between the overall value of a quantile forecast and its QSS.

For this purpose, we extend the cost-loss model to more than two observation categories assuming that the cost C and the loss L of the cost-loss model are the unitary increment of cost and loss per unit of variable, respectively, as discussed in Section 3.3. Following Richardson (2011), the overall value is defined as the ratio

$$V_{all} = \frac{T_C - T_F}{T_C - T_P} \qquad (32)$$

where the total mean expense T of a user is estimated when decisions are based on a climatological forecast (T_C), on a perfect deterministic forecast (T_P) or on a given forecast (T_F) so that Eq. (32) is the extension of Eq. (27) to all possible events.

The total expense for a perfect deterministic forecast corresponds to the sum of the costs C associated with each observation. The total mean expense T_P can then be expressed as

$$T_P = \frac{1}{N} \sum_{i=1}^{N} C\Omega_i. \qquad (33)$$

For a climatological quantile forecast Ω_τ, the total expense corresponds to the sum of the costs associated with Ω_τ and the losses encountered when the observations are greater than the climatological forecast ($\Omega_i \geq \Omega_\tau$). The total mean expense for a climatological forecast T_C is written as

$$T_C = \frac{1}{N} \sum_{i=1}^{N} C\Omega_\tau + \frac{1}{N} \sum_{i:\Omega_i \geq \Omega_\tau} L(\Omega_i - \Omega_\tau). \qquad (34)$$

Considering now a sample of quantile forecasts $q_{\tau,i}$ and the corresponding observations Ω_i, the total expense of a forecasts user corresponds in that case to the sum of the costs associated with each forecast $q_{\tau,i}$ and the losses encountered when $\Omega_i \geq q_{\tau,i}$,

given by

$$T_F = \frac{1}{N}\sum_{i=1}^{N} C q_{\tau,i} + \frac{1}{N}\sum_{i:\Omega_i \geq q_{\tau,i}} L(\Omega_i - q_{\tau,i}). \quad (35)$$

Combining Eqs (33)-(35), it is shown in the Appendix that the overall value V_{all} corresponds to QSS (Eq. (8)) with the climatology as a reference based on the assumption of constant cost-loss ratio for all outcomes. In other words, extending the dichotomous event-action framework to a continuous framework allows one to turn back to the 'classical' or 'natural' measure of performance for quantile forecast. Conversely, using the dichotomous framework provides the keys to making a deeper analysis of the quantile performance at the event level.

6. Conclusion

Verification measures and tools related to users' decision-making are provided here for quantile forecasts as decision variables. Drawing a parallel with the verification of probability forecasts, the new verification tools allow the scuite of verification methods for quantile forecasts to be completed. In particular, the concepts of forecast discrimination and forecast value are discussed based on a simple cost-loss model.

First, the RUC curve is shown to be the counterpart of the ROC curve when the focus is on a given user rather than on a given event. The areas under the RUC and ROC curves are summary measures of discrimination adapted to quantile and probability forecasts, respectively. Both measures share the same properties, such as non-sensitivity to calibration.

Second, the translation of discrimination ability into value is explored with the help of the value score. The definition of the forecast value is directly adopted from the probability forecast verification framework. Forecast value and forecast potential value are estimated when the decision variable is a quantile forecast, so focusing on a user with a specific cost-loss ratio. The first is obtained when the forecast is taken at face value and the second when critical decision criteria are applied. The value of a quantile forecast can then be plotted as a function of a range of events of interest, defined for example in terms of base rates. The derived plot is called a quantile value plot and provides a valuable insight into the performance of a quantile forecast. As a real example, the discrimination ability and value of global radiation forecasts from COSMO-DE-EPS are demonstrated over a summer and a winter period.

Finally, it is shown that the overall value of a quantile forecast corresponds to the quantile skill score with climatology as reference when a constant cost-loss ratio for all outcomes is assumed. In the same spirit as the weighted version of the continuous ranked probability score proposed by Gneiting and Ranjan (2011), a weighted version of the quantile skill score could be envisaged in order to take into account specific use of quantile forecasts.

Appendix

Overall value and Quantile Skill Score

From Eqs (33) and (34), the difference in expense between climatological and perfect deterministic forecasts can be written

as

$$T_C - T_P = \frac{1}{N}\sum_{i=1}^{N} C(\Omega_\tau - \Omega_i) + \frac{1}{N}\sum_{i:\Omega_i \geq \Omega_\tau} L(\Omega_i - \Omega_\tau) \quad (36)$$

Considering the relationship $\tau = 1 - \dfrac{C}{L}$ and setting L equal to 1 in the following demonstration without loss of generality, we obtain

$$T_C - T_P = \frac{(1-\tau)}{N}\sum_{i=1}^{N} (\Omega_\tau - \Omega_i) + \frac{1}{N}\sum_{i:\Omega_i \geq \Omega_\tau} (\Omega_i - \Omega_\tau) \quad (37)$$

and with some algebra

$$T_C - T_P = \frac{(1-\tau)}{N}\sum_{i:\Omega_i \leq \Omega_\tau} (\Omega_\tau - \Omega_i) + \frac{\tau}{N}\sum_{i:\Omega_i \geq \Omega_\tau} (\Omega_i - \Omega_\tau) \quad (38)$$

This mean expense difference, $T_C - T_P$, corresponds to the definition of the quantile score for a climatological forecast (QS_{climate}).

In the same manner, from Eqs (35) and (34), the difference between climatological forecast expense and the quantile forecast expense is written as

$$T_C - T_F = \frac{1}{N}\sum_{i=1}^{N} C\Omega_\tau + \frac{1}{N}\sum_{i:\Omega_i \geq \Omega_\tau} L(\Omega_i - \Omega_\tau) \\ - \frac{1}{N}\sum_{i=1}^{N} C q_{\tau,i} - \frac{1}{N}\sum_{i:\Omega_i \geq q_{\tau,i}} L(\Omega_i - q_{\tau,i}) \quad (39)$$

which becomes after some algebra

$$T_C - T_F = \frac{(1-\tau)}{N}\sum_{i:\Omega_i \leq \Omega_\tau} (\Omega_\tau - \Omega_i) + \frac{\tau}{N}\sum_{i:\Omega_i \geq \Omega_\tau} (\Omega_i - \Omega_\tau) \\ - \left(\frac{(1-\tau)}{N}\sum_{i:\Omega_i \leq q_{\tau,i}} (q_{\tau,i} - \Omega_i) + \frac{\tau}{N}\sum_{i:\Omega_i \geq q_{\tau,i}} (\Omega_i - q_{\tau,i})\right) \quad (40)$$

where the first term corresponds to the definition of the quantile score for a climatological forecast (QS_{climate}, Eq. (38)), and the second term to the quantile score (QS_{forecast}, Eq. (7)). With regard to the definition of the quantile skill score and of the overall value (Eqs (8) and (32), respectively), we end up with:

$$V_{all} = QSS \quad (41)$$

References

Atger F. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Q.J.R. Meteorol.Soc* **130**: 627–646.

Badescu V. 2008. *Modelling solar radiation at the earth surface: Recent advances.* Springer-Verlag Berlin Heidelberg, 517 pp.

Baldauf M, Seifert A, Förstner J, Majewski D, Raschendorfer M, Reinhardt T. 2011. Operational convective-scale numerical weather prediction with the COSMO model. *Mon. Wea. Rev.* **139**: 3887–3905.

Becker R, Behrens K. 2012. Quality assessment of heterogeneous surface radiation network data. *Adv. Sci. Res.* **8**: 93–97.

Ben Bouallègue Z, Theis S. 2014. Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Met. Apps.* **21**: 922–929.

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

Bentzien S, Friederichs P. 2014. Decomposition and graphical portrayal of the quantile score. *Q.J.R. Meteorol. Soc.* **140**: 1924–1934.

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Q.J.R. Meteorol. Soc.* **35**: 1512–1519.

Bröcker J. 2011. Probability forecasts. *In: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edition* .

Bröcker J. 2014. Resolution and discrimination two sides of the same coin. *Quarterly Journal of the Royal Meteorological Society* doi:10.1002/qj.2434.

Buizza R. 2001. Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.* **129**: 2329–2345.

Chen YS, Ehrendorfer M, Murphy AH. 1987. On the relationship between the quality and value of forecasts in the generalized cost-loss ratio situation. *Mon. Wea. Rev.* **115**: 1534–1541.

Christoffersen PF, Diebold FX. 1997. Optimal prediction under asymmetric loss. *Econometric Theory* **13**(6): 808–817.

Epstein E. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.* **8**: 985–987.

Friederichs P, Hense A. 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.* **135**: 2365–2378.

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

Gneiting T. 2011a. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**: 746–762.

Gneiting T. 2011b. Quantiles as optimal point forecasts. *International Journal of Forecasting* **27**: 197–207.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration, and sharpness. *J. Roy. Stat. Soc.* **69B**: 243–268.

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

Hamill TM. 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.* **129**: 550–560.

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real or is it the varying climatology? *Q.J.R. Meteorol. Soc.* **132**: 2905–2923.

Jolliffe I, Stephenson DB. 2005. Comments on discussion of verification concepts in forecast verification: A practitioners guide in atmospheric science. *Wea. Forecasting* **20**: 796–800.

Katz R, Murphy A. 1997. *Forecast value: prototype decision-making models*. Richard W. Katz, Allan H. Murphy (eds.), Economic Value of Weather and Climate Forecasts, Cambridge University Press, Cambridge.

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

Koenker R, Machado J. 1999. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**: 1296–1310.

Krzysztofowicz R. 1983. Why should a forecaster and a decision maker use bayes theorem. *Water Resour. Res.* **19**: 327–336.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**: 3515–3539.

Lewis JM. 2005. Roots of Ensemble Forecasting. *Mon. Weather Rev.* **133**: 1865–+, doi:10.1175/MWR2949.1.

Mason I. 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.* **30**: 291–303.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* **128**: 2145–2166.

Murphy AH. 1969. Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteor.* **8**: 863–873.

Murphy AH. 1991. Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.* **119**: 1590–1601.

Murphy AH. 1993. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather Forecasting* **8**: 281–293.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Wea. Rev.* **115**: 1330–1338.

Murphy AH, Winkler RL. 1992. Diagnostic verification of probability forecasts. *Int. J. Forecasters* **7**: 435–455.

Peralta C, Ben Bouallègue Z, Theis SE, Gebhardt C. 2012. Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research* **117**, doi:10.1029/2011JD016581.

Pinson P. 2013. Wind energy: Forecasting challenges for its operational management. *Statistical Science* **28**(4): 564–585.

Pinson P, Chevallier C, Kariniotakis G. 2007. Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Trans. on Power Systems* **22**(3): 1148–1156.

Richardson DS. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **126**: 649–667.

Richardson DS. 2011. Economic value and skill. *In: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edition* .

Roulston M, Kaplan D, Hardenberg J, Smith L. 2003. Using medium-range weather forecasts to improve the value of wind power production. *Renewable Energy* **28**: 585–602.

Steppeler J, Doms G, Schättler U, Bitzer HW, Gassmann A, Damrath U, Gregoric G. 2003. Meso-gamma scale forecasts using the nonhydrostatic model LM. *Met. Atm. Phys.* **82**: 75–96.

Thompson J. 1962. Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteor.* **1**: 13–17.

Toth Z, Talagrand O, Candille G, Zhu Y. 2003. Probability and ensemble forecasts. *in Jolliffe I T, Stephenson D B. 2003. In Forecast Verification: A Practitioner's Guide in Atmospheric Science. Jolliffe IT and Stephenson DB (ed.). John Wiley and Sons, Chichester, UK* : 137–164.

Weigel AP. 2011. Ensemble verification. *In: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edition* .

Wilks DS. 2001. A skill score based on economic value for probability forecasts. *Met. App.* **8**: 209–2019.

Wilks DS. 2006. Statistical methods in the atmospheric sciences. 2nd ed. *Academic Press* : 627 pp.

Zhu Y, Toth Z, Wobus R, Richardson D, Mylne K. 2002. The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.* : 73–83.