Verification tools for probabilistic forecasts of continuous hydrological variables

Hydrol. Earth Syst. Sci., 11, 1267–1277, 2007
www.hydrol-earth-syst-sci.net/11/1267/2007/

Hydrology and Earth System Sciences

Verification tools for probabilistic forecasts of continuous hydrological variables

F. Laio and S. Tamea

Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Torino, Italy

Received: 8 June 2006 – Published in Hydrol. Earth Syst. Sci. Discuss.: 8 August 2006
Revised: 19 October 2006 – Accepted: 26 March 2007 – Published: 3 May 2007

Abstract. In the present paper we describe some methods for verifying and evaluating probabilistic forecasts of hydrological variables. We propose an extension to continuous-valued variables of a verification method originated in the meteorological literature for the analysis of binary variables, and based on the use of a suitable cost-loss function to evaluate the quality of the forecasts. We find that this procedure is useful and reliable when it is complemented with other verification tools, borrowed from the economic literature, which are addressed to verify the statistical correctness of the probabilistic forecast. We illustrate our findings with a detailed application to the evaluation of probabilistic and deterministic forecasts of hourly discharge values.

1 Introduction

Probabilistic forecasts of hydrological variables are nowadays commonly used to quantify the prediction uncertainty and to supplement the information provided by point-value predictions (Krzysztofowicz, 2001; Ferraris et al., 2002; Todini, 2004; Montanari and Brath, 2004; Siccardi et al., 2005; Montanari, 2005; Tamea et al., 2005; Beven, 2006). However, probabilistic forecasts are still less familiar to many people than traditional deterministic forecasts, a major problem being the difficulty to correctly and univocally evaluate their quality (Richardson, 2003). This is especially true in the hydrological field, where the development of probabilistic forecast systems has not been accompanied by an analogous effort towards the proposition of methods to assess the performances of these probabilistic forecasts. In contrast, the usual choice when evaluating probabilistic predictions of hydrologic variables has been to adopt verification tools borrowed from the meteorological literature (e.g., Georgakakos et al., 2004; Gangopadhyay et al., 2005).

However, this meteorological-oriented approach has two drawbacks: first, most of the methods developed by the meteorologists were originally proposed for the probabilistic predictions of discrete-valued variables, and the adaptation of these techniques to deal with continuous-valued variables can reduce the discriminating capability of the verification tools (e.g., Wilks, 1995; Jolliffe and Stephenson, 2003). For example, a continuous-valued forecast can always be converted into a binary prediction by using a threshold filter (e.g., Georgakakos et al., 2004; Roulin, 2007): this allows one to use verification tools developed for binary variables, but it also reduces the amount of information carried by the forecast, and the usefulness of its verification. A second problem with the usual hydrological approach to probabilistic forecast evaluation is that it disregards some other available tools: more specifically, other verification methods exist, proposed in the last decade in the economic field (e.g., Diebold et al., 1998), but these methods have been usually ignored by the hydrologists, notwithstanding their relevance for the problem under consideration.

The purpose of this paper is to overcome these two problems and to provide an efficient approach to probabilistic forecast verification; in order to do that, we first need to describe some existing forecast verification tools. We do not have the ambition of fully reviewing the vast literature in the field, and we will limit ourselves to describe some methods, which in our opinion are the most suitable for application in the hydrological field (Sect. 2). This serves as a basis for developing, in Sect. 3.1, a simple cost-loss decision model which allows one to operationally evaluate a probabilistic forecast of a continuous-valued variable. We then consider in Sect. 3.2 the approach of the economists to forecast evaluation, and discuss its merits and drawbacks, with special attention to its applicability to hydrological predictions. The two approaches are compared in Sect. 4 through an example

Correspondence to: F. Laio
(francesco.laio@polito.it)

Published by Copernicus GmbH on behalf of the European Geosciences Union.

Table 1. Forecast verification tools, subdivided by the type of predicted variable (columns) and the forecast outcome (rows). We refer to Sect. 2 for details and references about the methods.

	Discrete predictands			**Continuous Predictands**
	Binary	**Multicategory**		
		Nominal	**Ordinal**	
Deterministic forecast	HIT RATE, THREAT SCORE, …	PEARSON'S COEFF. OF CONTINGENCY	GOODMAN AND KRUSKAL *G* STATISTIC	MSE MAE
Interval forecast	NOT APPLICABLE	NOT APPLICABLE	NOT APPLICABLE	TESTS FOR THE CONDITIONAL AND UNCOND. COVERAGE
Probabilistic forecast	BRIER SCORE	CONVERSION TO BINARY TABLES	RANKED PROBABILITY SCORE	THIS PAPER!

of application to the forecast of hourly discharge values. Finally, in Sect. 5 the conclusions are drawn, aimed at providing some guidelines for the use of probabilistic forecast evaluation methods in the hydrologic field.

2 General issues in forecast verification

Before describing the tools for verifying a probabilistic forecast, we need some definitions. Suppose that a time series of measurements of a variable x is available, sampled at regular intervals, $\{x_i\}$, i=1, .., N. A portion of the time series of size n, which we call "testing set", is forecasted, obtaining an estimate $\hat{x}_i$ of the actual value x_i. The predictions are carried out using the information available up to a time step $i-h$, where h is the lead time, or prediction horizon, of the forecast. Three different kinds of forecasts, with increasing level of complexity, can be carried out: if the result of the prediction is a single value for each predicted point, one has a deterministic forecast, $\tilde{x}_i$; if the prediction consists of an interval $[L_i(p), U_i(p)]$ wherein the future value x_i is supposed to lie with coverage probability p, one has an interval forecast (Chatfield, 2001; Christoffersen, 1998); finally, if the whole probability distribution of the predictands, $p_i(\hat{x}_i)$, is estimated, one has a probabilistic forecast (Abramson and Clemen, 1995; Tay and Wallis, 2000).

A second important discrimination regards the form of the variable under analysis: x can be a a continuous-valued variable, which is the most typical case in hydrology; or a discrete-valued variable, i.e. a variable that can take one and only one of a finite set of possible values (the typical case is the prediction of rainfall versus no rainfall events). When the predictands and forecasts are discrete but not binary variables, a further distinction occurs between ordinal and nominal events, depending on the presence of a natural order between the classes wherein x is partitioned (see Wilks, 1995, for details). The available verification tools depend upon the kind of forecast and predictands under analysis, as presented in Table 1. In all cases, the verification process requires that the obtained forecasts ($\tilde{x}_i$, or $\{L_i(p), U_i(p)\}$, or $p_i(\hat{x}_i)$) are compared to the real future values, x_i, for all points belonging to the testing set. We will now rapidly describe some of the verification tools available in the different situations, separating the cases when the predictand is a discrete variable from those when it is a continuous one.

2.1 Discrete-valued predictands

Most of the methods for the analysis of discrete binary or multicategory predictands originate from the meteorological literature (see Wilks, 1995, or Jolliffe and Stephenson, 2003, for a detailed review). Consider a situation in which the variable x can be partitioned into k mutually exclusive classes, C_1, ..., C_k. Verification of deterministic forecasts of discrete predictands (row two, columns two to four in Table 1) requires the representation of the results through a contingency table, i.e. a table whose (r, c) cell contains the frequency of occurrence of the combination of a deterministic forecast falling in class C_r and an observed event in class C_c. Verification in this case is carried out by defining a suitable score to summarize in a single coefficient the information contained in the contingency table. Examples of these scores are the hit rate and the threat score for binary variables (Wilks, 1995), the so-called G statistic for multicategory ordinal variables (Goodman and Kruskal, 1954; Kendall and Stuart, 1977, p. 596), and the Pearson's coefficient of contingency (Goodman and Kruskal, 1954; Kendall and Stuart, 1977, p. 587) for multicategory nominal variables. As for the interval forecasts of discrete variables (row three, columns two to four in Table 1), these are seldom performed, due to inherent difficulty of combining the fixed coverage probability of the interval prediction and the coarse domain of the

discrete variable.

We now turn to the probabilistic forecast of discrete variables, and consider the case of a k-classes ordinal variable (row four, column four in Table 1). The probabilistic forecast of the i-th point in the testing set, x_i, has now the form of a vector $\{p_{i,1}, ..., p_{i,k}\}$, where $p_{i,j}>0$ (with $j=1, \ldots, k$) represents the probability assigned to the forecast $\hat{x}_i$ falling in class C_j. Analogously, one can define the vector $\{o_{i,1}, ..., o_{i,k}\}$, with $o_{i,j}=1$ if $x_i \in C_j$, and $o_{i,j}=0$ in the reverse case. A commonly adopted verification tool in this case is the Ranked Probability Score (Murphy, 1970, 1971; Epstein, 1969; Wilks, 1995) which takes the form

$$\mathrm{RPS} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{m=1}^{k-1}\left[P_{i,m} - O_{i,m}\right]^2\right\} \quad (1)$$

where $P_{i,m}=\sum_{j=1}^{m} p_{i,j}$ is the cumulative distribution function (cdf) of the forecasts $\hat{x}_i$, while $O_{i,m}=\sum_{j=1}^{m} o_{i,j}$ is the corresponding cdf of the observations x_i (which actually degenerates into a step function, taking only 0 and 1 values). The rationale behind the use of the RPS as a verification tool for ordered multicategory predictands lies in the fact that it is sensitive to distance, i.e. it assigns a higher score to a forecast which is "less distant" from the event, or class, which actually occurs (see Murphy, 1970). In the particular case when $k=2$ (binary predictand, row four, column two in Table 1) the ranked probability score reads

$$\mathrm{RPS}\big|_{k=2} = \frac{1}{n}\sum_{i=1}^{n}\left[P_{i,1} - O_{i,1}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[p_{i,1} - o_{i,1}\right]^2 \quad (2)$$

which is called the Brier score. Finally, the rather uncommon case of multicategory nominal variables is usually treated by converting the contingency table into binary tables (see Wilks, 1995).

2.2 Continuous-valued predictands

Consider now the situation when the variable to forecast is a continuous one (column five in Table 1). When the prediction is deterministic, the assessment of the quality of the forecast requires that a suitable discriminant measure between the forecasted and observed values is calculated, a good prediction being the one that minimizes the discrepancy. Commonly used measures are the mean squared error,

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}[\widetilde{x}_i - x_i]^2\,, \quad (3)$$

and the mean absolute error,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\widetilde{x}_i - x_i|\,. \quad (4)$$

Before considering the main point of the paper in Sect. 3 (verification of probabilistic forecasts of continuous variable), we consider the case of an interval forecast of the form $\{L_i(p), U_i(p)\}$ (Table 1, row three, column five). Define an indicator function I_i which is equal to 1 if $x_i \in \{L_i(p), U_i(p)\}$, while $I_i=0$ in the reverse case. Standard evaluation methods of interval forecasts consist in comparing the actual coverage $\frac{1}{n}\sum_{i=1}^{n} I_i$ of the interval, to the hypothetical coverage p. A likelihood ratio test for the hypothesis $\frac{1}{n}\sum_{i=1}^{n} I_i=p$ is proposed by Christoffersen (1998) to verify the (unconditional) coverage of the interval. However, this test has no power against the alternative that the events inside (or outside) the interval come clustered together. This shortcoming can be avoided by verifying that the I_i values form a random sequence in time; we refer to Christoffersen (1998) for a discussion of this problem and a description of an appropriate joint test of coverage and independence.

3 Verification tools for probabilistic forecasts of continuous variables

The main focus of the present paper is on the evaluation of probabilistic forecasts of continuous variables, which are frequently the object of investigation in the hydrological field. Two approaches to the problem are considered. The first one is adapted from analogous methods developed by the meteorologists when dealing with binary variables (Murphy, 1969; Wilks, 1995; Palmer, 2000; Richardson, 2003), and it is based on the comparative evaluation of the forecasts in terms of their operational value, or economic utility. This approach requires that the decision-making process of individual users is considered, and a cost-loss function is specified by the forecaster; the evaluation of the forecast involves a single statistic which measures the overall value of the prediction. Details on this approach are presented in Sect. 3.1. The other approach is preeminently used by the economists (e.g., Diebold et al., 1998; Berkovitz, 2001; Noceti et al., 2003), who avoid to measure the overall quality of the prediction and concentrate on the evaluation of the formal correctness of the uncertainty description provided by the probabilistic forecast. Suitable statistical tools are developed for this purpose, as detailed in Sect. 3.2.

3.1 Determining the operational value of probabilistic predictions

As mentioned, the approach of the meteorologists to probabilistic forecast evaluation requires the definition of a cost-loss function to determine the value of the forecast. This approach has been originally proposed by Murphy (1969) and Epstein (1969) for the evaluation of probabilistic forecasts of discrete-valued variables. The modification of this framework to deal with the evaluation of probabilistic forecasts of continuous-valued variables represents one of the purposes of this paper.

Suppose that the forecast user knows that the cost of the precautionary actions to guarantee protection against an

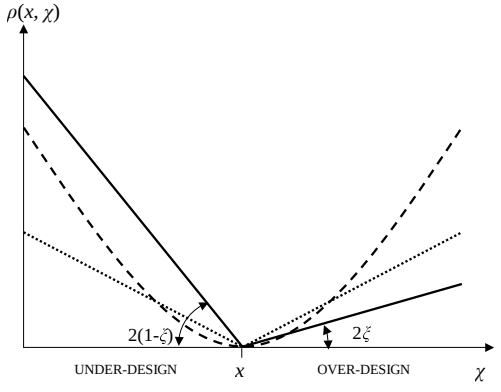

Fig. 1. Examples of quadratic (dashed line), absolute-value (dotted line) and asymmetric (continuous line, see Eq. 6) cost-loss functions. The variable χ on the horizontal axis is the "design" value, while x is the real future value.

hypothetical event χ is $C(\chi)$, where $C(\cdot)$ is an increasing function. The variable χ represents a sort of design value, that is fixed by the decision maker based on the forecast outcome: if the prediction is deterministic, χ is necessarily equal to the point forecast, $\chi=\tilde{x}$; if the prediction is probabilistic, then the χ value can be chosen among the possible forecast outcomes. In particular, the decision-maker will take a decision that minimizes the total expenditure of money. In order to do that, also the economic losses L, due to the actual occurrence of an event x, need to be defined: L is supposed to be zero if the observed event is lower than the design event, $x<\chi$ (in fact, in this case the precautionary actions guarantee protection), and to increase with $(x-\chi)$ when $x>\chi$. The overall cost-loss function is the sum of the cost and loss terms, and depends on both the observed and the design event, $CL(x, \chi)=C(\chi)+L(x, \chi)$.

An example can help to follow the reasoning: consider the case when x is the water stage at a given point along a river, and χ is the design value selected by the decision-maker on the basis of the information provided by the forecaster. The larger is χ, the more impactive and expensive are the necessary precautionary actions (emission of flood warnings, closure of roads and bridges, temporal flood proofing interventions, people evacuation, etc.); this explains why $C(\chi)$ is taken as an increasing function of χ. If x overcomes χ, some losses will also occur; as the distance between the observed and hypothesized values, $(x-\chi)$, increases, the losses become more and more relevant, including disruption of cultivated areas, inundation of civil infrastructures, flooding of inhabited areas, loss of human lives, etc. As a consequence, $L(x, \chi)$ is an increasing function of $(x-\chi)$ when $x>\chi$. Once the cost-loss function is defined, it is still necessary to determine the optimal design value, χ^*, i.e. the value that minimizes the total expenses. However, the future value x is obviously not known, which complicates the optimization problem. This is where the probabilistic prediction turns out to be useful: in fact, the decision maker can use the probabilistic forecast $p(\hat{x})$ to represent the probability distribution of the future events, $f(x)$. Under this hypothesis, he/she will be able to calculate the expected expenses $\overline{CL}(\chi)=\int_{\text{all}\hat{x}} CL(\hat{x}, \chi)p(\hat{x})\mathrm{d}\hat{x}$, and to take the decision χ^* that minimizes $\overline{CL}(\chi)$ (e.g., Diebold et al., 1998; Palmer, 2000; Richardson, 2003). The decision χ^* will depend upon the probabilistic forecast through $p(\hat{x})$, and a better prediction will decrease the actual expenditure of money $CL(x, \chi^*)=C(\chi^*)+L(x, \chi^*)$. This provides a general framework for the comparison of probabilistic forecasts based upon their operational value.

We proceed in our description by specifying the above procedure for the case of a simple cost-loss function, which we propose here to evaluate probabilistic forecasts of hydrologic variables. We suppose $C(\chi)$ is a linear function, $C(\chi)=c\cdot\chi$, where c is a constant, and $L(x, \chi)$ is a stepwise linear one, $L(x, \chi)=H(x-\chi)\cdot l\cdot(x-\chi)$. Here $H(\cdot)$ is the Heavyside function, which is equal to one for positive arguments and zero otherwise, and l is a constant (note that $l>c$, since otherwise one would spend more money to guarantee protection than what is eventually lost, which is an anti-economic principle). The cost-loss function reads

$$CL(x, \chi) = c\cdot\chi + H(x-\chi)\cdot l\cdot(x-\chi). \quad (5)$$

A linear transformation of Eq. (5), obtained by subtracting $c\cdot x$ and dividing by $l/2$,

$$\begin{aligned}\rho_\xi(x, \chi) &= 2\xi(\chi-x) + 2H(x-\chi)\cdot(x-\chi)\\ &= |\chi-x| + 2(\xi-0.5)(\chi-x). \quad (6)\end{aligned}$$

is a completely equivalent cost-loss function (a similar function is used by Epstein (1969) and by Murphy (1970) when dealing with binary or multicategory variables), but it is more suitable to evaluating predictions. In fact, it depends on a single parameter, the cost-loss ratio $\xi=c/l<1$, and it attains a null value when $\chi=x$, i.e. when the hypothetical value is equal to the actually occurred one (perfect forecast).

An example of such cost-loss function is reported in Fig. 1, continuous line, where it is compared to an absolute value cost loss-function, $\rho_{\text{abs}}(x, \chi)=|x-\chi|$, and to a quadratic cost-loss function, $\rho_{\text{quad}}(x, \chi)=(x-\chi)^2$. The main difference is in the fact that the ρ_ξ function assigns different weights to under-design and to over-design, which is more appropriate when environmental (hydrological) variables are predicted. In this case, ξ values lower than 0.5, giving rise to cost-loss functions similar in shape to the one in Fig. 1, are to be preferred: in fact, the losses are expected to be much greater than the costs of protection. Also note that the ρ_ξ function is the generalization of the absolute value cost-loss function, as ρ_ξ converges to ρ_{abs} when $\xi=0.5$ (this is another reason why it is convenient to use ρ_ξ rather than CL from Eq. 5).

Once the loss function is defined, one can search for the optimal design value χ^*. By taking the expected value of Eq. (6), one obtains

$$\overline{\rho}_\xi(\chi) = 2\xi\left(\chi - \int_{\text{all}\hat{x}} \hat{x}\, p(\hat{x})\mathrm{d}\hat{x}\right) + 2\int_\chi^\infty (\hat{x}-\chi)\, p(\hat{x})\mathrm{d}\hat{x}, \quad (7)$$

whose derivative with respect to χ, equated to zero, provides the optimal decision χ^*

$$P(\chi^*) = 1-\xi \Rightarrow \chi^* = P^{-1}(1-\xi) \quad (8)$$

that depends only on the cumulative distribution function of the forecasts, $P(\cdot)$, and on the cost loss ratio $\xi<1$. Of course, the same result would have been obtained by using Eq. (5) as the cost-loss function (this is why the two formulations are equivalent). In contrast, if a similar procedure is adopted with the absolute value or the quadratic cost-loss function (Fig. 1), the median and the mean of the forecasts distribution are respectively selected as the design values χ^*.

The total expenses will now amount to $\rho_\xi(x, \chi^*)=|\chi^* - x|+2(\xi-0.5)(\chi^* - x)$, and the operational value of different predictions will be found from the averaged $\rho(x, \chi^*)$ values over the n points in the testing set,

$$EC(\xi) = \frac{1}{n}\sum_{i=1}^n \rho_\xi(x_i, \chi_i^*) = \frac{1}{n}\sum_{i=1}^n \left\{|P_i^{-1}(1-\xi) - x_i| + 2(\xi-0.5)(P_i^{-1}(1-\xi) - x_i)\right\}. \quad (9)$$

The lower is the obtained $EC(\xi)$ value (EC stands for "expected cost"), the more valuable is the forecast. Note that, when the prediction is deterministic, $P(x)=H(x-\tilde{x})$, and, as mentioned, $\chi^*=\tilde{x}$ for any ξ. In this case Eq. (9) reads

$$EC_{\text{det}}(\xi) = \frac{1}{n}\sum_{i=1}^n \{|\tilde{x}_i - x_i| + 2(\xi-0.5)(\tilde{x}_i - x_i)\}, \quad (10)$$

which is a discrepancy measure similar to the mean squared error and mean absolute error defined in Eqs. (3) and (4).

A difficulty with Eq. (9) is that the expected cost depends on the cost-loss ratio ξ; different predictions can thus be ranked in different manners by different users, implying that there cannot be an universally accepted "best" probabilistic prediction. This can be especially problematic, since the cost-loss function is seldom known, and, even when it is simplified as in Eq. (6), it may be difficult to set a specific value for the cost-loss ratio ξ. Our preferred solution is therefore to avoid fixing a ξ value, but rather to graphically represent how the expected costs, associated to different forecasting systems, change with ξ. Special attention should be paid to the $EC(\xi)$ curves in the part of the diagram where $\xi<0.5$, corresponding to situations where the losses are very relevant compared to the costs of the precautionary actions. We also propose to re-scale the $EC(\xi)$ curves with respect to the cost of a "climatologic" mean-value deterministic prediction, $\tilde{x}_i=\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. By setting this value in Eq. (10) one obtains that the expected cost of the climatologic prediction is the mean deviation $\delta=\frac{1}{n}\sum_{i=1}^n |x_i - \overline{x}|$. Our proposal is to plot $EC(\xi)/\delta$ versus ξ, in order to be able to directly determine the value of the forecast compared to the mean-value prediction: if $EC(\xi)/\delta$ is lower (larger) than one, the forecast is more (less) valuable than the climatologic prediction. An example of application of this procedure is reported in Sect. 4.

The idea of plotting the $EC(\xi)$ curve is new (however, Palmer (2000), Richardson (2003) and Roulin (2007) use a similar graph for determining the value of probabilistic predictions of discrete variables); more frequently, the meteorologists face the difficulty of setting an exact value for ξ by supposing that ξ is a random variable with a uniform $U(0, 1)$ distribution (e.g., Murphy, 1969), and then taking the average value of $EC(\xi)$ over the possible ξ values. This corresponds to calculating the area $\overline{EC}$ below the $EC(\xi)$ curves,

$$\overline{EC} = \int_0^1 EC(\xi)\mathrm{d}\xi = \frac{1}{n}\sum_{i=1}^n \int_0^1 \rho_\xi(x_i, \chi_i^*)\mathrm{d}\xi. \quad (11)$$

Since the integral and summation terms interchange, we can concentrate on a single addendum in the summation and elide the subscripts i for simplicity:

$$T = \int_0^1 \rho_\xi(x, \chi^*)\mathrm{d}\xi = \int_0^1 \{|P^{-1}(1-\xi) - x| + 2(\xi-0.5)(P^{-1}(1-\xi) - x)\}\mathrm{d}\xi. \quad (12)$$

Substituting $y=P^{-1}(1-\xi)$ one has

$$T = \int_{-\infty}^\infty \{|y-x| + [1-2P(y)](y-x)\}\, p(y)\mathrm{d}y = \int_{-\infty}^\infty 2[H(y-x) - P(y)](y-x)p(y)\mathrm{d}y. \quad (13)$$

Using the formula for integration by parts, and considering that $H(y-x)-P(y)=0$ when $y\to\pm\infty$, one obtains

$$T = \int_{-\infty}^\infty [H(y-x) - P(y)]^2\mathrm{d}y, \quad (14)$$

i.e. that $\int_0^1 \rho_\xi(x, \chi^*)\mathrm{d}\xi$ is equivalent to the continuous ranked probability score, CRPS $=\int_{-\infty}^\infty [H(y-x)-P(y)]^2\mathrm{d}y$, which is sometimes used to assess the performances of probabilistic forecasts of continuous variables (Hersbach, 2000). As a consequence, $\overline{EC}$ in Eq. (11) is also equivalent to $\overline{\text{CRPS}}$ (Hersbach, 2000, Eq. 5). This equivalence is not surprising: in fact, the CRPS is the limit of the ranked probability score in Eq. (1) for an infinite number k of zero-width classes (see Hersbach, 2000), and the RPS was obtained by applying to discrete variables a cost-loss function which is similar to ρ_ξ in Eq. (6) (Murphy, 1969, 1970). However, the manner how we obtained the CRPS in Eq. (14) is novel, and allows one to better understand what are its qualities and drawbacks. In particular, Eqs. (12) to (14) demonstrate that the CRPS is averaged over different cost-loss ratios, and, as such, its

indications can be misleading, due to the excessive weight assigned in its calculation to expenses correspondent to ξ values larger than 0.5, which are rather unrealistic in the hydrologic field. In our opinion, it is better to evaluate the different predictions by plotting the $EC(\xi)/\delta$ curves, rather than trying to summarize all information in a single statistic.

3.2 Statistically-oriented evaluation of probabilistic forecasts

The economists criticize the approach based on the evaluation of the forecasts through the use of cost-loss functions for the fact that the evaluation turns out to be user-dependent rather than objective: in fact, two users with different cost-loss functions may rank in a different manner two forecasts. Moreover, they argue that the cost-loss function is seldom known, which introduces an undesired element of uncertainty in the evaluation (Diebold et al., 1998). The followed approach is therefore to leave aside considerations on the operational value of the probabilistic forecast, and simply verifying if the forecast is correct under a statistical viewpoint. A correct probabilistic forecast of x_i is one whose probability density function $p_i(\hat{x}_i)$ coincides with the true distribution of x_i, $f_i(x_i)$. Even if $f_i(x_i)$ is not known (the distribution changes with i, and only one sampled value is available), it is feasible to build up a test of the hypothesis $H_0 : p_i(\hat{x}_i) \equiv f_i(x_i)$. The test is based on the probability integral transform, $z_i = P_i(x_i)$, that consists in evaluating the cumulative distribution function of the predictions in correspondence to the observed value x_i (Berkovitz, 2001). Under the hypothesis H_0, the distribution of z_i is uniform, $U(0, 1)$. If one applies the probability integral transform to all points in the testing set, a sample of z_i values is obtained. If the probability forecast is correct, the z_i values are mutually independent and identically $U(0, 1)$ distributed. The test of the hypothesis H_0 can therefore be split into an independence test and a goodness-of-fit test of the $U(0, 1)$ hypothesis.

As for the independence, the usual suggestion is to look at the autocorrelation function of the z_i's and of their powers $z_i^2, ..., z_i^m$ (e.g., Diebold et al., 1998). This produces some proliferation of the test statistics (one for each considered power), with possible problems of interpretation of the results. Our proposal is to use instead the Kendall's τ test of independence (Kendall and Stuart, 1977). Consider the sequence $z_1, ..., z_n$, and their associated ranks $R_1, ..., R_n$, i.e. their position in the ordered vector of the z_i's. Kendall's τ test of independence is based on the statistic

$$\tau = 1 - \frac{4N_d}{(n-1)(n-2)}, \tag{15}$$

where N_d is the number of discordances, i.e. the number of pairs (R_i, R_{i+1}) and (R_j, R_{j+1}) that satisfy either $R_i < R_j$ and $R_{i+1} > R_{j+1}$, or $R_i > R_j$ and $R_{i+1} < R_{j+1}$.

Under the null hypothesis of independence and with $n > 10$, the standardized statistic

$$\tau_{\mathrm{st}} = \frac{\tau}{\sigma_\tau} = \tau \cdot \sqrt{\frac{9n(n-1)}{2(2n+5)}} \tag{16}$$

has a normal distribution with null mean and unitary variance (Kendall and Stuart, 1977), which allows one to easily determine the limit values for the independence test. For example, the 95% test of independence will be passed if τ_{st} is below 1.645 (one-tail test).

Consider now the uniformity hypothesis: many goodness-of-fit tests for this hypothesis exist (D'Agostino and Stephens, 1986; Noceti et al., 2003). However, Diebold et al. (1998) argue that it is better to adopt a less formal graphical method, based on an histogram representation of the density of the z_i's. We agree that the graphical representation is more revealing, but prefer a probability plot representation that does not require a subjective binning of the data. The probability plot is a plot of the z_i values versus their empirical cumulative distribution function, R_i/n. The shape of the resulting curve reveals if the data are approximatively uniform, in which case the $(z_i, R_i/n)$ points are close to the bisector of the diagram. Kolmogorov confidence bands can also be represented on the same graph in order to provide a more formal test of uniformity. The Kolmogorov bands are two straight lines, parallel to the bisector and at a distance $q(\alpha)/\sqrt{n}$ from it, where $q(\alpha)$ is a coefficient, dependent upon the significance level of the test α (e.g., $q(\alpha = 0.05) = 1.358$, see D'Agostino and Stephens, 1986). The test is passed when the curves remain inside these confidence bands.

The probability plot representation does not only tell if the uniformity test is passed or not, but also provides a tool to investigate the causes behind deviations from uniformity. In fact, the shape of the curves in the probability plot (see Fig. 2) is suggestive of the encountered problem, since the steepness of the curves is larger where more z_i points concentrate. In the case of the continuous line in Fig. 2, for example, the z_i points are concentrated in the vicinity of the end points 0 and 1. This corresponds to having the real x_i values that fall, more frequently than expected, on the tails of the distribution of the forecasts. As a consequence, the probabilistic prediction is "narrow". Similar considerations apply to the other curves in Fig. 2. The probability plot representation has already been used by De Gooijer and Zerom (2000); in contrast, it should not be confused with the apparently similar attributes diagram (Wilks, 1995), which is a tool for the verification of probabilistic predictions of binary variables.

When using this approach to forecast verification, one ends out with results concerning with the formal correctness of the probabilistic prediction; however, these results do not imply that the prediction is good: there can exist a prediction that passes the independence and uniformity tests, but has no operational value. In our opinion, the method should therefore necessarily be used together with some other method,

like those described in Sect. 3.1, allowing one to understand if the prediction is really valuable or not.

A final comment is necessary regarding multi-step-ahead predictions (i.e., characterized by a prediction horizon $h \neq 1$). In this case, serial correlation in the z_i series is expected up to a lag $h-1$ (Box and Jenkins, 1970), and the independence and goodness-of-fit tests should be applied separately to the h subseries $\{z_1, z_{1+h}, z_{1+2h}, ...\}$, $\{z_2, z_{2+h}, z_{2+2h}, ...\}$, ..., $\{z_h, z_{2h}, z_{3h}, ...\}$ (Diebold et al., 1998). One obtains h τ_{st} statistics and h probability plots for each prediction. The global tests will be obtained from the combination of the tests performed on each of the subseries: however, the combination is complicated by the fact that the h subseries are mutually (not internally!) dependent. When the samples being tested are correlated, the correct significance level to have a global α-level test should be between α (linearly dependent samples) and α/h (independent samples). In our opinion the correlation between the subseries is strong, and it is thus better to perform the tests on the h subseries with a significance level α, instead of using a level α/h for each sub-test as suggested by Diebold et al. (1998).

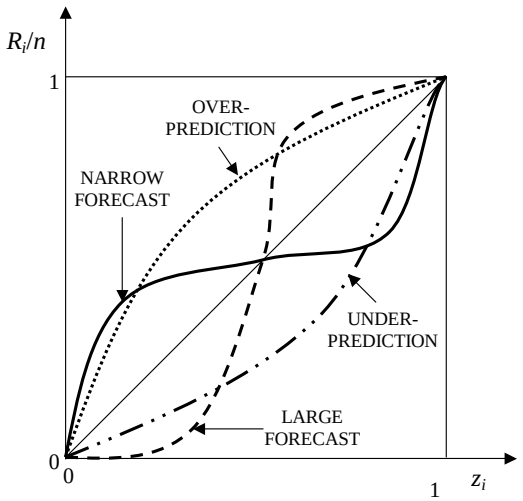

Fig. 2. Examples of the possible outcomes of a probability plot representation of the $z_i = P_i(x_i)$ values versus their corresponding ranks R_i (divided by the sample size n). If the points lie close to the bisector, the forecast is deemed reliable; otherwise, a problem with the spread of the probabilistic forecast, or a prediction bias, are detected.

4 Application and discussion

The verification tools described in the previous sections are applied to the probabilistic forecasts of a discharge time series, obtained with a prediction method developed by Tamea et al. (2005) and Laio et al. (2007), and based on local polynomial regression techniques (Farmer and Sidorowich, 1987; Fan and Gijbels, 1996; Cleveland and Loader, 1996; Porporato and Ridolfi, 1997; Regonda et al., 2005). We use this prediction method as a mean to exemplify the described verification techniques; we therefore refer to Tamea et al. (2005) and Laio et al. (2007) for a detailed description of the prediction method, which is here briefly introduced. A time series of past values of discharge (and concurring average precipitation over the basin) is required as input. The functional relation between the values to be forecasted, x_i, and the regressors $\mathbf{x}_{i-h}$ (vector of past discharge and precipitation values) is locally approximated by polynomials, whose coefficients are estimated on a neighborhood of size k of each query point. The regressions obtained, different from point to point, are applied to the respective query points to give the deterministic prediction $\hat{x}_i$. The method produces forecasts for the points in the testing set, provided that a set $\mathcal{S}$ of model parameter values is assigned by the forecaster.

We use in our verification exercise four different types of predictions, all based on the mentioned local polynomial regression method. Two forecasting techniques are deterministic and two are probabilistic, as detailed hereafter.

1. *Best deterministic prediction*: it is the point forecast obtained by selecting the parameter set $\mathcal{S}_{best}$ that produces the "best" deterministic predictions when the method is applied to the calibration set, i.e. to a set of discharge values selected for cross-validation purposes (see Tamea et al., 2005).

2. *Ensemble forecast*: it is a probabilistic forecast obtained by selecting q parameter sets rather than a single one (we use in the following example the $q = 100$ sets that minimize the mean absolute error over the calibration set). Each of these sets is separately used to obtain q different predictions for each point x_i in the testing set. The empirical distribution function of this sample of q predictions is taken as representative of the distribution characterizing the ensemble forecast.

3. *Probabilistic forecast*: the same as before, but with a suitable parameter uncertainty representation attached to each member in the ensemble; this is obtained by using the k residuals of the local polynomial regressions (k is the so called "number of neighbors"). The residuals are converted into out-of-sample errors by resampling and inflating them with a multiplying factor accounting for the prediction uncertainty, according with Kendall and Stuart (1977); finally they are summed up to the point predictions in the ensemble (see Laio et al., 2007, for more details). A large sample of $\hat{x}_{i,j}$, $j=1, .., q \cdot k$ values is obtained, whose empirical distribution function is taken as the estimate of $p_i(\hat{x}_i)$.

4. *Median prediction*: it is a deterministic prediction obtained by taking,for each point in the testing set, the median of the above defined probabilistic prediction $p_i(\hat{x}_i)$ as the estimator of $\tilde{x}_i$.

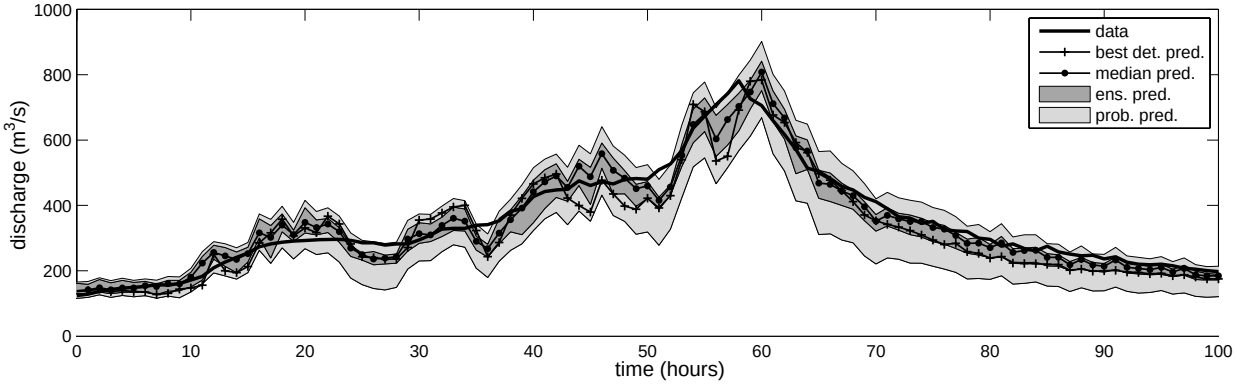

Fig. 3. An example of forecasts of an hourly discharge time series: portion of the testing set with predictions at h=6 h, showing the outcomes of the four prediction methods described in Sect. 4 (see Tamea et al. (2005) and Laio et al. (2007) for all the details).

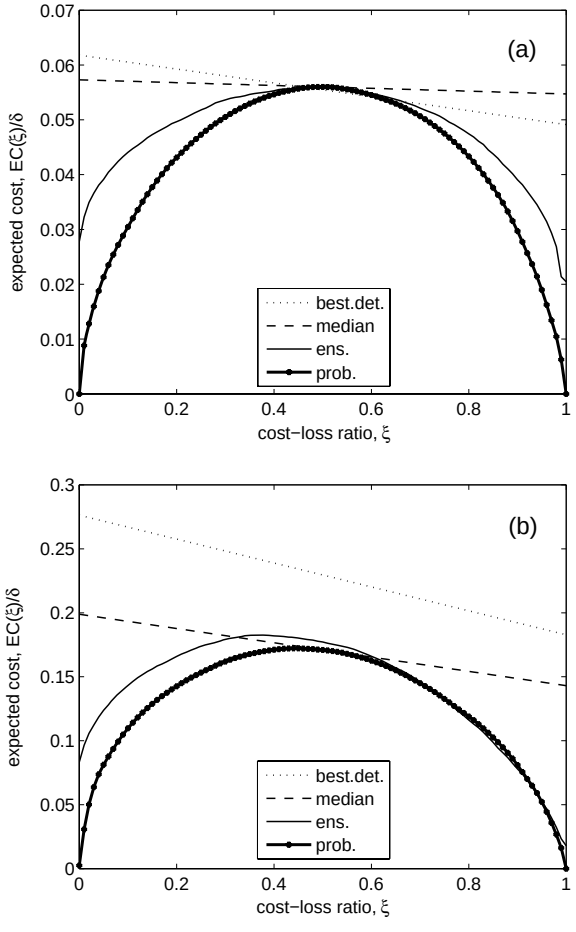

Fig. 4. Representation of the expected cost from Eq. (9) (re-scaled by the mean deviation δ) as a function of the cost-loss ratio ξ, for a 1 step-ahead **(a)** and a 6-steps ahead **(b)** hourly discharge prediction. The four lines in each graph refer to four different forecasting methods, described at the beginning of Sect. 4.

The prediction methods have been applied to the discharge time series of the Tanaro river, in the northwest of Italy. The catchment basin at the gauge station of Farigliano has an extension of 1522 km^2 and an elevation ranging from 235 to 2651 m above the sea level. The hourly discharge time series has been measured from 1997 to 2002. The testing set covers the period between 14 November 2002 and 27 November 2002, and corresponds to an important flood event. The mean rainfall over the catchment is used as an endogenous variable for the prediction. The mean rainfall is determined from the data collected by eleven rain gauges located on the basin. Both hydrometric and pluviometric data have been collected by the Regional Agency for the Protection of the Environment (ARPA-Piemonte), and are the same already used by Tamea et al. (2005). Prediction horizons of one and six hours (corresponding to h=1 and h=6) are considered in the following examples. A portion of the testing set with the corresponding four types of prediction at h=6 is displayed in Fig. 3, where the two series of point forecasts (1. and 4.) are displayed together with the 90% bands from the two probabilistic prediction methods (2. and 3.).

In Fig. 4 the expected cost from Eq. (9), re-scaled by using the mean deviation δ, is represented as a function of the cost-loss ratio ξ for the four predictions listed above. Note that the $EC(\xi)/\delta$ values are much lower than 1, both for the 1-h ahead prediction (Fig. 4a) and for the 6-h ahead prediction (Fig. 4b), demonstrating that all forecasting methods are very competitive with respect to the climatological prediction. The quality of the four prediction methods can now be comparatively evaluated: the lower is the expected cost of a forecast, the higher is its operational value. It is clear from Fig. 4 that the two probabilistic methods outperform the deterministic ones, in particular in the part of the diagram that is more important when dealing with flood events (large expected losses compared to the costs, i.e. low ξ values).

It is also interesting to comment on the shape of the four curves in the diagram: the relation between the expected cost and ξ turns out to be linear when the prediction is deterministic; in fact, by setting $P_i^{-1}(1-\xi)=\tilde{x}_i$ in Eq. (9), one obtains the equation of a straight line, whose slope is two times the bias of the prediction, $\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - x_i)$, and whose intercept with the $\xi=0.5$ vertical line is the MAE of the forecast, a measure of the spread of the prediction errors. As expected, both the (negative) bias and the spread of the errors increase when the prediction horizon passes from 1 to 6 h. The median prediction is better than the best deterministic prediction for $\xi<0.5$, which is mainly due to the beneficial effect of taking an ensemble of predictions rather than a single one (see Georgakakos et al., 2004; Tamea et al., 2005; Regonda et al., 2005).

On the same diagram the probabilistic predictions tend to have a parabolic shape, with null (or very low) values at the extremes and a maximum for $\xi\approx0.5$. The reason for the low values at the extremes is the following: when $\xi=0$ the cost of the precautionary actions is null, and one can therefore always take an action that protects against any possible occurring flood. Analytically, when $\xi=0$, one has $P_i^{-1}(1-\xi)=\max(\hat{x}_i)$ and $EC(\xi=0)=\frac{1}{n}\sum_{i=1}^{n}\left\{|\max(\hat{x}_i)-x_i|-(\max(\hat{x}_i)-x_i)\right\}$; the only terms contributing to the expected cost are therefore those when the actually occurred value x_i is greater than the maximum predicted value, $\max(\hat{x}_i)$, which never happens for the more reliable probabilistic prediction, and only rarely for the ensemble prediction.

When $\xi=1$ the cost of the precautionary action is equal to that of the eventually occurring losses; there is thus no convenience to take any action, i.e. the design value χ in Eq. (6) can be set to zero (actually, to $\min(\hat{x}_i)$). As a consequence, $\rho_{\xi=1}$ and $EC(\xi=1)$ are also null (or very low). In this second case the total cost would in reality be different from zero, due to the losses, but the passage from Eq. (5) to Eq. (6) produces this fictitious result. However, this is not a relevant incongruence, since both extremes $\xi=0$ and $\xi=1$ correspond to unrealistic situations when the decision to be taken is obvious, and the forecast is useless. It is not then the shape of the single curve on the diagram that is of interest, but the relations between the curves for fixed ξ values. Considering this aspect, it can be noted how the probabilistic prediction provides more valuable results than the ensemble prediction, in particular for the more relevant low ξ values. As a further detail, the continuous ranked probability score values are the following: at $h=1$, $\overline{\text{CRPS}}=10.1\,\text{m}^3/\text{s}$ for the best deterministic prediction, $\overline{\text{CRPS}}=8.8\,\text{m}^3/\text{s}$ for the ensemble prediction, $\overline{\text{CRPS}}=7.7\,\text{m}^3/\text{s}$ for the probabilistic prediction, and $\overline{\text{CRPS}}=10.2\ \text{m}^3/\text{s}$ for the median prediction. At $h=6$, the corresponding values are $41.7\,\text{m}^3/\text{s}$ (best), $25.8\,\text{m}^3/\text{s}$ (ensemble), $23.6\,\text{m}^3/\text{s}$ (probabilistic), and $31.1\,\text{m}^3/\text{s}$ (median). These values correspond to the areas below the curves in Fig. 4, multiplied by the mean deviation $\delta=181.7\,\text{m}^3/\text{s}$. Also these results confirm the superiority of the probabilistic method, even if, as mentioned, the relevance of the $\overline{\text{CRPS}}$ index is doubtful when dealing with hydrological applications.

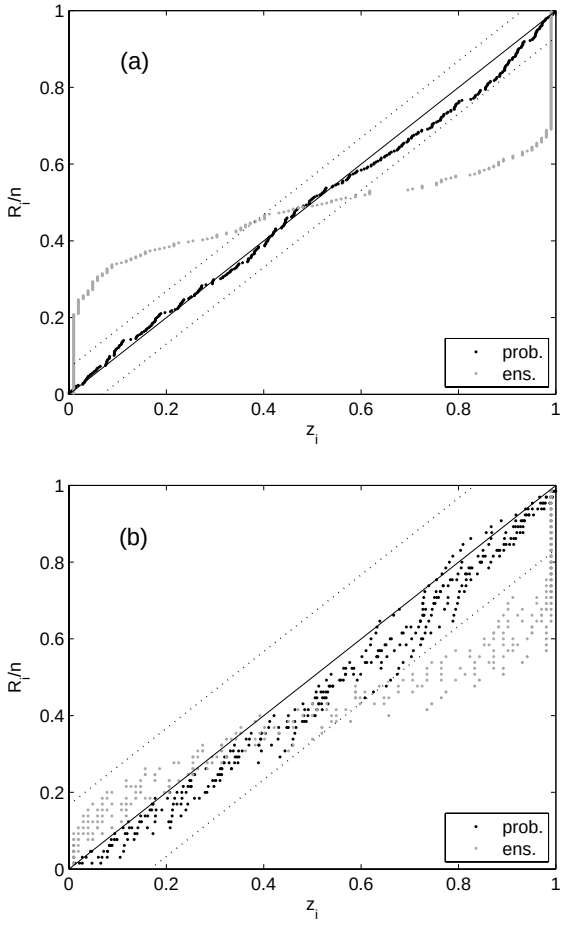

Fig. 5. Probability plot representation (see Fig. 2) of the ensemble (gray circles) and probabilistic (black circles) forecasts of an hourly discharge time series. Each point in the diagram corresponds to a point in the testing set. Panel **(a)** refers to 1 step-ahead predictions, panel **(b)** to 6-steps ahead forecasts. The Kolmogorov 5% significance bands are also reported as dashed lines.

We now turn to the application of the statistically-oriented forecast verification tools: since these methods are targeted at evaluating probabilistic predictions only, the comparison will be limited to the ensemble and probabilistic predictions methods. As mentioned, the verification of the probabilistic forecast is a two step process, requiring to apply the transformation $z_i=P_i(x_i)$ and then separately test the independence and the uniformity of the z_i's. The standardized Kendall's τ_{st} statistic in Eq. (16) is calculated, obtaining $\tau_{\text{st}}=1.38$ (ensemble) and $\tau_{\text{st}}=-3.17$ (probabilistic) for $h=1$. Both values are not significant at the 5% level,

i.e. the independence test is passed. For h=6, six subseries $\{z_1, z_7, z_{13}, ...\}$, $\{z_2, z_8, z_{14}, ...\}$, ..., $\{z_6, z_{12}, z_{18}, ...\}$ are constructed, and six different τ_{st} values (for each prediction) are obtained. The independence test is passed if the maximum among these values is not significant at the α level. The obtained values are τ_{st}=3.11 (ensemble) and τ_{st}=0.90 (probabilistic), i.e. the independence test is passed at the 5% level by the probabilistic prediction, but not by the ensemble prediction (note that the test would not be passed even when the significance level was reduced to α/h=0.008, as suggested by Diebold et al., 1998).

The uniformity of the z_i's is then verified by plotting the data versus their empirical cumulative distribution around the central value (see Fig. 5). Using Fig. 2 as a guide to evaluate the results, it is clear that the ensemble method provides predictions that are very narrow around the central value. In contrast, the forecasts obtained through the probabilistic method are very reliable (the points remain inside the Kolmogorov bands with 5% significance). A (slight) negative bias is detected at h=6 (Fig. 5b), as apparent from the fact that the points lie below the bisector of the diagram (compare to Fig. 2). The results for h=6 refer to six different curves, due to the mentioned separation of the testing set in six subseries. The Kolmogorov bands are larger in Fig. 5b with respect to Fig. 5a for that same reason; in fact, when testing separately the 6 sub-series, the actual size of the samples is $n/6$ and the acceptability limits become larger.

5 Conclusions

We have here compared different strategies for evaluating the performances of probabilistic prediction methods of continuous variables. All analyzed methods have some interesting characteristics, but none of them, taken alone, allows a complete and fair evaluation of the quality of the forecast. Our suggestion is to use two methods together, as each carries a fundamental information about the prediction quality. More in detail, the expected cost diagram (Fig. 4) is a very useful tool to understand the operational value of the forecast, especially when comparing different (deterministic and probabilistic) predictions. This approach, however, does not provide sufficient information for a complete evaluation: in fact, the reliability of the forecast, i.e. the fact that the distribution of the predictions, $p_i(\hat{x}_i)$, is equal to the real distribution, $f_i(x_i)$, is hypothesized rather than verified (see Sect. 3). Moreover, the definition of a cost-loss function always demands some subjective choice: for example, we have taken ρ_ξ in Eq. (6) to be piecewise linear, but a quadratic (asymmetric) function would also be a proper choice. We do not think that the outcomes of the forecast verification would qualitatively change when using a quadratic cost-loss function, but in any case we believe that it is necessary to complement the expected cost curve with other tools, aimed at verifying the statistical congruence of the forecast, i.e. the hypothesis that $p_i(\hat{x}_i)=f_i(x_i)$. More in detail, we found that suitable tools, based on the probability integral transform $z_i=P_i(x_i)$, require the application of the Kendall's independence test and the representation of the z_i's through a probability plot (Fig. 5), which allows one to assess the uniformity of the z_i's. The combination of these two approaches, respectively based on the concept of operational value of the forecast and on the formal statistical verification of its reliability, provides the basis for an exhaustive and effective probabilistic forecast evaluation.

Acknowledgements. We are grateful to L. Ridolfi for his helpful suggestions and comments during the course of this research.

Edited by: R. Rudari

References

Abramson, B. and Clemen, C.: Probability forecasting, Int. J. Forecast., 11, 1–4, 1995.

Berkovitz, J.: Testing density forecasts, with applications to risk management, J. Bus. Econ. Stat., 19, 465–474, 2001.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Box, G. E. P. and Jenkins, G. M.: Time series analysis: forecasting and control, Holden Day, San Francisco, CA, 1970.

Chatfield, C.: Prediction intervals for time series forecasting, in: Principles of forecasting: a handbook for researchers and practitioners, edited by Armstrong, J., pp. 475–494, Kluwer Academic Publishers, Norwell, MA, 2001.

Christoffersen, P. F.: Evaluating interval forecast, Int. Econ. Rev., 39, 841–862, 1998.

Cleveland, W. S. and Loader, C. L.: Smoothing by Local Regression: Principles and Methods, in: Statistical Theory and Computational Aspects of Smoothing, edited by: Haerdle, W. and Schimek, M. G., 10–49, Springer, New York, 1996.

D'Agostino, R. B. and Stephens, A. M., eds.: Goodness-of-fit techniques, Dekker, New York, 1986.

De Gooijer, J. G. and Zerom, D.: Kernel-based multistep-ahead predictions of the US short-term interest rate, J. Forecast., 19, 335–353, 2000.

Diebold, F. X., Gunther, T. A., and Tay, A. S.: Evaluating density forecasts with applications to financial risk management, Int. Econ. Rev., 39, 863–883, 1998.

Epstein, E. S.: A scoring system for probability forecasts of ranked categories, J. Appl. Meteorol., 8, 985–987, notes and correspondence, 1969.

Fan, J. and Gijbels, I.: Local polynomial modelling and its applications, Chapman and Hall, London, UK, 1996.

Farmer, J. D. and Sidorowich, J. J.: Predicting chaotic time series, Phys. Rev. Lett., 59, 845–848, 1987.

Ferraris, L., Rudari, R., and Siccardi, F.: The uncertainty in the prediction of flash floods in the Northern Mediterranean Environment, J. Hydrometeorol., 3, 714–727, 2002.

Gangopadhyay, S., Clark, M., and Rajagopalan, B.: Statistical downscaling using K-nearest neighbors, Water Resour. Res., 41, doi:10.1029/2004WR003444, 2005.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298, 222241, doi:10.1016/j.jhydrol.2004.03.037, 2004.

Goodman, L. A. and Kruskal, W. H.: Measures of association for cross classifications, J. Am. Stat. Assoc., 49, 732–764, 1954.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–569, 2000.

Jolliffe, I. T. and Stephenson, D. B., eds.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, New York, USA, 2003.

Kendall, M. G. and Stuart, A.: The advanced theory of statistics, Griffin Press, London, 2nd ed., 1977.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, 2001.

Laio, F., Ridolfi, L., and Tamea, S.: Probabilistic prediction of real-world time series: a local regression approach, Geophys. Res. Lett., 34, L03403, doi:10.1029/2006GL028776, 2007.

Montanari, A.: Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 41, W08406, doi:10.1029/2004WR003826, 2005.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004.

Murphy, A. H.: Measures of the utility of probabilistic predictions in cost-loss ratio decision situation in which knowledge of the cost-loss ratio is incomplete, J. Appl. Meteorol., 8, 863–873, 1969.

Murphy, A. H.: The ranked probability score and the probability score: a comparison, Mon. Weather Rev., 98, 917–924, 1970.

Murphy, A. H.: A note on the ranked probability score, J. Appl. Meteorol., 10, 155–156, 1971.

Noceti, P., Smith, J., and Hodges, S.: An evaluation of tests of distributional forecasts, J. Forecast., 22, 447–455, 2003.

Palmer, T. N.: Predicting uncertainty in forecasts of weather and climate, Rep. Progr. Phys., 63, 71–116, 2000.

Porporato, A. and Ridolfi, L.: Nonlinear analysis of river flow time sequences, Water Resour. Res., 33, 1353–1367, 1997.

Regonda, S., Rajagopalan, B., Lall, U., Clark, M., and Moon, Y.: Local polynomial method for ensemble forecast of time series, Nonlin. Proc. Geophys., 12, 397–406, 2005.

Richardson, D. S.: Economic value and skill, in: Forecast verification: a practitioner's guide in atmospheric science, edited by: Jolliffe, I. T. and Stephenson, D. B., Wiley, New York, USA, 2003.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737, 2007, http://www.hydrol-earth-syst-sci.net/11/725/2007/.

Siccardi, F., Boni, G., Ferraris, L., and Rudari, R.: A hydrometeorological approach for probabilistic flood forecast, J. Geophys. Res., 110, D05101, doi:10.1029/2004JD005314, 2005.

Tamea, S., Laio, F., and Ridolfi, L.: Probabilistic nonlinear prediction of river flows, Water Resour. Res., 41, W09421, doi:10.1029/2005WR004136, 2005.

Tay, A. S. and Wallis, K. F.: Density forecasting: a survey, J. Forecast., 19, 235–254, 2000.

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrol. Process., 18, 2743–2746, 2004.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic press, 1995.