Nonlin. Processes Geophys., 15, 661–673, 2008
www.nonlin-processes-geophys.net/15/661/2008/

Nonlinear Processes in Geophysics

On reliability analysis of multi-categorical forecasts

J. Bröcker

Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 34, 01187 Dresden, Germany

Received: 6 December 2007 – Revised: 2 April 2008 – Accepted: 7 July 2008 – Published: 6 August 2008

Abstract. Reliability analysis of probabilistic forecasts, in particular through the rank histogram or Talagrand diagram, is revisited. Two shortcomings are pointed out: Firstly, a uniform rank histogram is but a necessary condition for reliability. Secondly, if the forecast is assumed to be reliable, an indication is needed how far a histogram is expected to deviate from uniformity merely due to randomness. Concerning the first shortcoming, it is suggested that forecasts be grouped or stratified along suitable criteria, and that reliability is analyzed individually for each forecast stratum. A reliable forecast should have uniform histograms for all individual forecast strata, not only for all forecasts as a whole. As to the second shortcoming, instead of the observed frequencies, the probability of the observed frequency is plotted, providing and indication of the likelihood of the result under the hypothesis that the forecast is reliable. Furthermore, a Goodness-Of-Fit statistic is discussed which is essentially the reliability term of the Ignorance score. The discussed tools are applied to medium range forecasts for 2 m-temperature anomalies at several locations and lead times. The forecasts are stratified along the expected ranked probability score. Those forecasts which feature a high expected score turn out to be particularly unreliable.

1 Introduction

Suppose the objective is to make forecasts about the possible values y of a random variable Y, the observation, which is unknown by the time the forecasts have to be made, but the actual value of which will be revealed at some future time. The observation is often referred to as the verification. A probability forecasts is a probability assignment to a collection of potential events. This paper focuses on forecasts for more than just two events. The events are typically chosen so as to be exhaustive (i.e. at least one event will happen for sure) and mutually exclusive (i.e. if one event happens, none of the others does). The events are defined in terms of the observation, that is, knowing the observation is sufficient to determine which of the events has occurred. Both the definition of the events as well as the assigned forecast probabilities can (but do not necessarily have to) vary between different forecast instances. As an example, consider the case where Y is a variable between zero and one, for example relative humidity. A probabilistic forecast might be specified by assigning varying probabilities to the ten sub-intervals with vertices $\frac{k}{10}$; k=0 . . . 10. Alternatively, a probabilistic forecast might be specified by defining ten intervals of varying length, each of which carries the forecast probability 0.1.

In very broad terms, this paper is about how to verify whether a probabilistic forecast has "come true". The appropriate answer to this question depends on the interpretation of probability. One common (but not the only possible) interpretation of probability is that of a long term observed relative frequency. If each individual forecast probability could be compared to a large number of observations, the degree of agreement between forecast probabilities and observed frequencies could be tested. If no significant disagreement occurs, the forecast is called reliable or calibrated. Reliability relates the forecast probabilities to actually observed frequencies and thus provides the forecast with a degree of objectivity. It is important to note that reliability is not the only desirable property of a probabilistic forecast, another one being resolution. Broadly speaking, a forecast shows resolution if different events are antedated by different forecast behavior. For example, a rain forecast has resolution if forecasts preceding rainy days are significantly different from those preceding sunny days. This paper deals exclusively with reliability. For extensive discussions on resolution and ways to quantify it, see for example Wilks (2006); Toth et al. (2005); Bröcker (2008).

A frequently employed tool to analyze forecast reliability is the histogram, in particular the rank histogram or Talagrand diagram. The idea of histograms is to divide the

Correspondence to: J. Bröcker
(broecker@pks.mpg.de)

observations among a limited number of categories, thereby defining a set of exhaustive and mutually exclusive events. Then the observed frequencies for these categories are compared with the corresponding forecast probabilities. Often the nature of the problem already suggests a set of categories, and sometimes it is even possible to define a set of categories with all corresponding forecast probabilities having the same numerical value. In the latter case, the histogram of observed frequencies is expected to be more or less uniform.

In this paper, I discuss how both information content and reliability of histograms can be improved. As currently employed, histograms suffer from two long noted conceptual problems. Firstly, a single histogram can only display the observed frequencies averaged over all forecasts. Reliability though means agreement of forecast probabilities and observed frequencies for each probability forecast individually. It might be objected that this definition of reliability is meaningless if for each forecast distribution, there is at most one observation. This is a practical difficulty, but not a conceptual one, as will be discussed later when reliability will be formalized using conditional probabilities. At any rate, there is agreement in the literature that a uniform (unconditional) histogram is only a necessary, but not a sufficient condition for reliability (Hamill and Colucci, 1998; Hamill, 2001; Gneiting et al., 2005). In order to alleviate this problem, the concept of forecast strata is introduced. Essentially, the forecasts get stratified along some descriptive variable, and subsequently reliability is analyzed individually for each stratum. Secondly, in the past, too little attention was paid to the question as to whether the obtained results are actually statistically significant. A forecast might show deviations from reliable behavior simply because of random fluctuations due to limited amounts of data, even if the forecast were reliable. As reliability analysis attempts to compare forecasts according to an allegedly objective criterion, any result should be treated with the utmost care in order to avoid conclusions which are in fact unwarranted by the data. An indication has to be provided of the observed frequencies' expected variations for a reliable forecast. Employing a Goodness-Of-Fit test (in one way or another) was suggested by Hamill and Colucci (1997); Hamill (2001), but it seems that only Elmore (2005) actually used them in connection with ensemble forecasts. Both problems will be revisited in this paper, and possible improvements will be suggested.

The remainder of this introduction gives an overview over the paper. In Sect. 1 the notion of reliability is discussed. A general definition of reliability is provided, from which a few commonly known conditions of reliability are derived. Special attention is paid to ensemble forecasts. Section 2 focuses on verifying reliability through histograms and related concepts. Stratification of forecasts is introduced as a means to check reliability on a more detailed level. The deviations of observed frequencies from ideal behavior are analyzed. A way to plot histograms is suggested that allows for an easier and immediate check as to whether the observed frequencies are consistent with reliability or not. A stratified Goodness-Of-Fit-test is introduced as a convenient summary statistics of the histogram. The statistic employed in this paper is similar to Pearson's classical statistic in that it is asymptotically χ^2-distributed, but in addition can be interpreted as the reliability term of the Ignorance score. In Sect. 3 the discussed tools are applied to a number of operational probabilistic weather forecasts. In these examples, the forecasts are stratified along the expected ranked probability score. The expected ranked probability score is not the true score of the forecast but the expectation value of the score if the verification was in fact distributed as specified by the forecast, and hence is a property of the forecast alone. The results indicate that the investigated forecasts are the less reliable, the better their expected ranked probability score. In this example, forecasts with small expected ranked probability score appear to suffer from both bias and insufficient spread. It seems that to rectify the problem, a de-biasing *conditioned* on the expected ranked probability score is necessary, since standard de-biasing would affect all forecasts at the same time. Section 4 concludes.

1.1 General considerations

In this subsection, a general definition of reliability along with its most important consequences is discussed. We will first fix some notation which will be employed throughout the paper. Let Y be the observation, which is unknown and to be forecast. For most of this section, in order to avoid unnecessarily technical language and concepts, the observation Y is assumed to take values in a (possibly infinite) interval E of the real numbers. For example, if Y is the relative humidity, then E would be the unit interval. It will be indicated how the stated results apply to other important types of observations. In most cases, this should be obvious. We will explicitely mention whenever the assumptions on E are essential.

Probability forecasts for Y might come in a variety of different forms. Again for simplicity's sake, I will assume for the beginning that the forecasts consist of distribution functions. A distribution function $G(y)$ is defined as the forecast probability for the event $Y < y$, for any arbitrary $y \in E$. A distribution function is sufficient to define the forecast probabilities for effectively[1] any event.

There are other forms of forecasts which can provide probabilistic information, most notably ensemble forecasts. Later in this section, it is discussed how the concepts outlined for distribution functions can be applied to ensemble forecasts.

Important for our discussion is that both the observation and the forecast are considered random quantities. The idea of considering the observation a random quantity should not

[1] The more mathematically inclined reader will be aware of the fact that a distribution function defines a measure on the σ-algebra of Borel sets, which, as far as I can see, contains all practically relevant events.

require any further comment, but taking the forecast a random variable probably does. The concept is presumably most easily understood for binary events. In this case, the forecast is just a single number (namely the forecast probability that the event occurs), hence it makes perfect sense to speak of it, in conjunction with the observation, as a random quantity (as done e.g. by Wilks, 2006; Toth et al., 2003, analyzing the Brier Score). Typically, both are strongly interdependent, and it is exactly this dependency we are after in reliability analysis. A random probabilistic forecast for a finite number of categories is almost as easily contemplated as for only two categories. Instead of a single probability, there is now a vector of probabilities. A distribution function can be thought of as an infinite collection of probabilities. The concept of random distribution functions obviously brings about all sorts of mathematical difficulties, like what a distribution of distribution functions should be etc, which we however need not to worry about here. It is of vital importance though to keep in mind that all quantities derived from the forecast distributions, such as the number $G(y)$ for a particular y or descriptive statistics like for example the variance or interquartile range of the forecast inherit randomness from G and are therefore random variables themselves.

Furthermore, averages over time are replaced with mathematical expectations. This of course imposes strong stationarity requirements on both the series of forecasts and the series of observations (since we are interested in the average behavior of forecasts and observations over time, not over parallel universes). Without such requirements though, the whole notion of observed frequency and hence reliability would cease to make sense. We can thus dispense of the notational inconvenience of a time index. Since we assume the probabilistic forecasts to be random quantities, we have to amend the notation to distinguish between random variables (which are functions) and a particular realization (which are function *values*). The probabilistic forecast as a random variable will henceforth be denoted as Γ. Realizations of Γ are distribution functions, denoted by Roman capitals F or G. The observation will be denoted by Y, while any particular realization of Y will be denoted by y.

As was already mentioned in the introduction, reliability means that for any individual forecast distribution, the limiting observed frequencies of the corresponding observation Y are equal to the forecast distribution (Toth et al., 2003, 2005). Within formulae, limiting observed frequencies are identified with probabilities and denoted with the symbol $\mathbb{P}$. Another interpretation of $\mathbb{P}$ is the probability measure on the space on which the random variables Γ and Y live. Using this convention, reliability can be formulated as

$$\mathbb{P}(Y < y|\Gamma = G) = G(y) \qquad \text{for any } y \in E. \tag{1}$$

The following remarks might help with the interpretation of Eq. (1).

1. To all intents and purposes, the notation $\mathbb{P}(Y<y|\Gamma=G)$ might be read as "The limiting observed frequency of the event $Y<y$, counted over all instances where the random forecast Γ is equal to the specific distribution function G".

2. The condition "$\Gamma=G$" on the left hand side of Eq. (1) involves the whole forecast distribution Γ, not only the probability it assigns to the event $Y<y$. One might think that the forecast Γ is already reliable if for any y held fixed, $\Gamma(y)$ is a reliable forecast for the binary event "$Y<y$" (i.e. for every y, the forecast $p=\Gamma(y)$ and the corresponding event "$Y<y$" form a diagonal reliability diagram). This condition is, in general, not sufficient, but gives a weaker form of reliability than Eq. (1).

3. A problem of Eq. (1) is that the relation $\Gamma=G$ might occur with probability zero only, in which case there are no observed frequencies to calculate. Although conditional probabilities are mathematically well defined even in this situation (Breiman, 1973), the fact remains that Eq. (1) cannot be verified pointwise, unless the forecast Γ assumes only a finite number of values G with positive probability (which is an important special case). This problem will be further discussed in Sect. 2.

4. Equation (1) is well known for the case of binary events, where it forms the basis for reliability diagrams (Murphy and Winkler, 1977; Toth et al., 2003; Wilks, 2006), the calibration-refinement-factorization (Murphy and Winkler, 1987) and the analysis of scoring rules, notably the Brier Score (Wilks, 2006; Toth et al., 2003). The essence of Eq. (1) in the general case can be found verbalized in Toth et al. (2003, 2005). Although he does not explicitely say so, Hamill (2001) also seems to think of his probabilities being conditioned on the forecast (as on the left hand side of Eq. 1), since otherwise it would not make sense to consider "expected values of probabilities", as for example in Eq. (1) of Hamill (2001).

1.2 Ensemble forecasts

As was mentioned before, distribution functions are not the only possible way to specify probabilistic forecasts, with ensemble forecasts being a very important alternative. An ensemble is a collection of random variables $\boldsymbol{X}=(X_1 \dots X_K)$, where $X_k \in E$ for all k. A realization of $\boldsymbol{X}$ will be denoted by $\boldsymbol{x}=(x_1 \dots x_K)$. Loosely speaking, the ensemble members are thought of as a collection of candidate values of the observation Y.

This subsection discusses how reliability and in particular Eq. (1) translate to the context of ensemble forecasts. The first question to be answered is how to interprete ensembles as probabilistic forecast. A common (but not the only possible) interpretation proposes that an ensemble constitutes a draw of independent samples from an underlying or "latent" probability forecast Γ (Hamill, 2001; Talagrand et al., 1997;

Anderson, 1996). This definition applies to multivariate ensembles as well.

A criterion of reliability similar to Eq. (1) using this interpretation of ensembles can be formulated as follows. Let Γ be the latent probability forecast, and by $X^{(k)}$ denote the ensemble member of rank k when sorting $\boldsymbol{X}$ in ascending order. Furthermore, by convention, $X^{(0)}$ and $X^{(K+1)}$ are set to the infimum and the supremum of E, respectively, whence always $\Gamma(X^{(0)})=0$ and $\Gamma(X^{(K+1)})=1$. It follows from Eq. (1) that

$$\mathbb{P}(Y<X^{(k)}|\Gamma=G)=\frac{k}{K+1} \qquad \text{for all } k=0\ldots K+1. \tag{2}$$

Equation (2) states that the rank of the observation Y among the ensemble members is a random variable which assumes the values $1\ldots K+1$ with equal conditional probability $\frac{1}{K+1}$. Equation (2) is in fact a slightly weaker condition than Eq. (1), and it is possible to construct a (somewhat pathological) case where criterion (2) is fulfilled but not criterion (1). To this end, suppose that $K=1$, that is, there is only one ensemble member $X^{(1)}$, drawn from the distribution G. Now suppose $F(y)$ is a distribution with median zero but otherwise arbitrary. If Y is drawn from the recentered distribution $F(y-X^{(1)})$, then an easy calculation shows that indeed

$$\mathbb{P}(Y < X^{(1)}|\Gamma = G) = \frac{1}{2},$$

so that Eq. (2) is fulfilled, while Eq. (1) is satisfied if and only if Y were drawn from G.

From Eq. (2) we get the identity

$$\mathbb{P}(Y < X^{(k)}) = \frac{k}{K+1}. \tag{3}$$

This follows from the general fact that if a conditional probability of an event does not depend on the condition, then it must be equal to the unconditional probability of the event. Equation (3) has been the basis for most studies on the reliability of ensemble forecasts so far. As already mentioned in the introduction, Eq. (3) represents but a necessary condition for reliability, as it follows from, but does not imply Eq. (2).

In Eq. (2), the conditioning involves the latent forecast distribution Γ, not the ensemble itself. This is an inconvenience, since typically the latent forecast distribution Γ is either unknown (at least to the person doing the reliability analysis) or a very unwieldy object, which is often the reason why ensembles are used in the first place. Equation (2) does *not* hold any longer if the condition "$\Gamma=G$" is replaced by something like "$\boldsymbol{X}=\boldsymbol{x}$". A simple counter-example is presented in the Appendix. In Sect. 2, I will return to this problem and discuss its ramifications.

Assuming a latent forecast distribution as done above is not the only way of establishing a connection between ensembles and probability forecasts. A different interpretation of ensembles states that for all k, the ensemble member $X^{(k)}$ represents the $\frac{k}{K+1}$ quantile of the forecast distribution. This definition renders Eq. (2) correct with the condition "$\Gamma=G$" being replaced by "$\boldsymbol{X}=\boldsymbol{x}$". In weather forecasting though, ensembles are produced in extremely high dimensional spaces, and it is hard to see how this interpretation should apply to such an ensemble when projected into one dimension. It must be said though that the standard interpretation of ensembles as a sample (as employed in this paper and in most studies elsewhere) is also but a highly idealized description of currently operational ensemble generation schemes.

The present section will be finished with a few words on multidimensional ensembles. The tools discussed in this paper for analyzing reliability of ensemble forecasts crucially rely on the assumption that ensemble members and observations can be ranked. This is obviously not the case in higher dimensions. One possible solution is to project forecasts and observations into one dimension, thereby effectively restricting attention to the reliability of marginal distributions.

An alternative was suggested by Hansen and Smith (2004). To explain the general features of the idea, assume that $f(\boldsymbol{x})$ is a symmetric function of the ensemble, in other words a function that stays constant if the ensemble members are permuted. Hansen and Smith (2004) use the length of the minimum spanning tree. Using f, the variables f_0 and f_i; $i=1\ldots K$ are constructed, where $f_0=f(\boldsymbol{X})$, and f_i is similar but with the i-th ensemble member being replaced by Y. Hansen and Smith (2004) suggest that reliability be checked using standard tools but with f_0 and the f_i taking the roles of the observation and the ensemble members, respectively. It seems questionable though if reliability of the original ensemble implies that Eq. (2) or even only Eq. (3) holds for f_0 and the f_i. The difficulty is that although f_0 and the f_i all have the same distribution, they cannot be considered independent draws from a distribution. I have been unable to either prove or disprove Eq. (2) (or Eq. 3) in this situation, but numerical investigations suggest that Eq. (3) is not true for arbitrary symmetric f. The minimum spanning tree might be a fortunate exception though.

2 Verifying reliability

In this section, practical aspects of reliability analysis, or more specifically, ways to test Eq. (1) and display the results are discussed. Suppose we have available an archive of forecasts and corresponding observations $T:=\{(G_n, y_n), n=1\ldots N\}$, where the G_n are forecasts in the form of distribution functions and y_n are observations. In order to use these data for reliability tests, two difficulties need be addressed first. Firstly, Eq. (1), as it stands, cannot be employed directly for reliability tests in situations where the condition $\Gamma=G$ occurs with probability zero (i.e. there is a continuous range of possible forecasts), as was already mentioned. In practical terms, it is obviously impossible to calculate long term observed frequencies conditioned on a single forecast if no two forecasts will ever be the same. To obtain testable reliability criteria in this situation, the condition "$\Gamma{=}G$" on the forecast needs being relaxed by considering entire sets of forecasts. This point will be discussed in Sect. 2.1. Secondly, limiting observed frequencies will have to be estimated from only finite data amounts. Therefore, a forecast system might exhibit deviations from reliable behavior simply because of random fluctuations. How to take these fluctuations into account is the subject of Sect. 2.2.

2.1 Stratification of forecasts

In this section, the concept of forecast stratification is introduced. Forecasts are stratified by aggregating forecasts into different strata, where a stratum is simply a prescribed set of forecasts. Forecast strata can be specified by means of descriptive quantities, for example the forecasts' interquartile range, the mean, or the level of Gaussianity (according to some measure of Gaussianity), thereby delineating the forecasts along that particular quantity. Individual forecast strata will be denoted by sans serif capitals $\mathsf{A}, \mathsf{B}, \ldots$. The notation $G{\in}\mathsf{A}$ indicates that the distribution function G belongs to the forecast stratum A. The motivation for forecast stratification is to obtain a more detailed reliability assessment of the forecast system than by just a single histogram, but at the same time to aggregate enough forecast instances per forecast stratum to get sufficiently accurate frequency estimates. There is a price to pay for the advantages of forecast stratification. Since agreement between forecast probabilities and observed frequencies is still not required for each individual forecast, but only on average across a forecast stratum, we still end up testing a weaker form of reliability than required by Eqs. (1) or (2).

Let us start carrying out this program for Eq. (2), which covers the practically important case of ensemble forecasts. Supposing that A is a particular forecast stratum, we can average both sides of Eq. (2) over all $G{\in}\mathsf{A}$. The result is

$$\mathbb{P}(Y{<}X^{(k)}|\Gamma \in \mathsf{A}){=}\frac{k}{K{+}1} \qquad \text{for all } k{=}0\ldots K{+}1, \tag{4}$$

where as before $X^{(k)}$ denotes the ensemble member of rank k when sorting $\boldsymbol{X}$ in ascending order. The left hand side of Eq. (4) could be estimated by the corresponding observed frequency, that is by counting the fraction of instances in the forecast stratum for which the observation exceeds less than k ensemble members. What has been gained so far is that observed frequencies can be calculated over a larger number of instances (depending on the number of instances in the forecast stratum). But there remains a problem: as mentioned in Sect. 1, the latent forecast Γ is typically inaccessible in the case of ensemble forecasts, whence it is not clear how to stratify the latent forecast along different forecast strata. Two possible solutions to this problem will be suggested here. The first solution is to use the ensemble for stratification instead of the latent forecast. For example, if we want to delineate forecasts along the mean of Γ, the ensembles could be stratified along the ensemble mean. Thus the ensemble and the latent forecast are identified, thereby ignoring any error that might arise due to the fact that the ensemble is but a sample from the latent forecast. This should cause no significant error as long as the ensemble is not too small. The feasibility of this approach obviously depends on whether the stratum A allows for efficient estimators of the event $\Gamma{\in}\mathsf{A}$.

The second solution applies if certain aspects of the latent forecast distribution are known. Although the latent forecast Γ is not accessible as a whole, the ensembles could be stratified along various parameters which are important in the generation of the forecast. In weather forecasting for example, forecasts could be stratified along different weather regimes (as suggested by Hamill, 2001), which could be identified using the model analysis or even measurements at forecast time.

The convenient feature of ensemble forecasts is that the right hand side of Eq. (4) does not depend on the conditioning, and hence is the same for all forecast strata. The mathematical reason is that the events are defined in a particular way which renders all corresponding forecast probabilities constant. If the forecast is available in the form of a distribution function, this is not automatically the case. Under mild conditions though, it is possible to transform the observations so that all forecast distribution functions for the transformed variables are uniform, and in particular independent of G, which facilitates forecast stratification. The remainder of this section explains the "probability integral transform" (PIT), which can be employed to this effect. Suppose again that Γ is our probabilistic forecast, issued in the form of distribution functions. The PIT of Y is the random variable $\Gamma(Y)$ (Devroye, 1986; Gneiting et al., 2005). We are interested in the limiting observed frequencies of the PIT $\Gamma(Y)$. First note that

$$\mathbb{P}(\Gamma(Y) < z|\Gamma = G) = \mathbb{P}(G(Y) < z|\Gamma = G), \tag{5}$$

simply because if $\Gamma{=}G$, the events $G(Y){<}z$ and $\Gamma(Y){<}z$ are the same. If $F(y)$ is an invertible distribution function, the event $F(Y){<}z$ is the same as the event $Y{<}F^{-1}(z)$. Applying this to G in Eq. (5), we obtain

$$\mathbb{P}(\Gamma(Y) < z|\Gamma = G) = \mathbb{P}(Y < G^{-1}(z)|\Gamma = G). \tag{6}$$

But if we assume the forecast Γ to be reliable, we can employ Eq. (1) to write the right hand side of Eq. (6) as

$$\mathbb{P}(Y < G^{-1}(z)|\Gamma = G) = G(G^{-1}(z)) = z. \tag{7}$$

Combining Eqs. (6) and (7) we obtain

$$\mathbb{P}(\Gamma(Y) < z|\Gamma = G) = z. \tag{8}$$

Equation (8) reduces the reliability analysis of one-dimensional observations to checking whether the PIT has uniform conditional distributions.

An important assumption in the derivation of Eq. (8) was that all forecast distribution functions G are invertible. This seems a strong assumption at first sight, but it has to be kept in mind that distribution functions are by construction monotonous. If the distribution function G has jumps though, the equation $z{=}F(y)$ might not have a solution y, whence our derivation of Eq. (8) breaks down. In this case the PIT might in fact have a non-uniform distribution.

As in the derivation of Eq. (4), averaging both sides of Eq. (8) over a specific forecast stratum A gives

$$\mathbb{P}(\Gamma(Y) < z | G \in \mathsf{A}) = z, \tag{9}$$

Again, Eq. (9) represents a weaker reliability condition than Eq. (8), for the same reasons that Eq. (4) presents a weaker form of reliability than Eq. (1). A special but practically important case of Eq. (9) obtains by choosing only a single "forecast stratum" which in fact encompasses all possible forecasts. The resulting equation is

$$\mathbb{P}(\Gamma(Y) < z) = z, \tag{10}$$

which is the PIT-version of Eq. (3) and amounts to checking a single distribution only. Again, it is well known that Eq. (10) is but a necessary condition for reliability (Gneiting et al., 2005), and examples of forecasts and observations for which Eq. (10) holds but not Eq. (8) are easily constructed.

2.2 Estimating observed frequencies

After having stratified the forecasts, we have to compare observed frequencies with forecast probabilities independently for each forecast stratum. If the PIT is employed, then according to Eq. (9) this amounts to checking whether the transformed observation $G_n(y_n)$ exhibits a uniform distribution. This is a standard problem of statistics and will not be considered any further in this paper. If ensemble forecasts are considered, then according to Eq. (4) we have to verify that the observed frequencies

$$f_{k,\mathsf{A}} := \frac{\#\{x_{k-1} \leq y_n < x_k; n \in I_{\mathsf{A}}\}}{\# I_{\mathsf{A}}} \tag{11}$$

agree with the corresponding forecast probabilities $\frac{1}{K+1}$. Here, x_k are the ensemble members for $k{=}1 \ldots K$, and per definition $x_0{=}{-}\infty$, $x_{K+1}{=}\infty$. Furthermore, I_{A} is the set of instances n for which the ensemble forecast falls into forecast stratum A, and the symbol "#" in front of a set denotes the number of elements in that set. A widely applied special form of this procedure results by choosing the trivial forecast stratification of considering a single stratum encompassing all forecasts.

This subsection is devoted to testing and displaying the agreement between forecast probabilities and observed frequencies. This is one of the oldest problem of statistics, if not the oldest. In meteorology, the most widely applied tool for this purpose is the histogram, presumably because of its exceeding simplicity (Hamill, 2001; Hamill and Colucci, 1997, 1998; Toth et al., 2003; Talagrand et al., 1997). The histogram comprises a plot of the observed frequencies f_k over k. The observed frequencies (or "height of the histogram bars") is subsequently compared to the corresponding forecast probabilities by visual inspection. If the forecasts are stratified, each forecast stratum requires its own histogram, since there is a set of observed frequencies and corresponding forecast probabilities per forecast stratum.

The problem with interpreting histograms is to decide when a forecast probability and a corresponding observed frequency should be considered "similar". If the forecast is reliable, a large archive of forecasts and corresponding verifications is expected to yield better agreement between forecast probabilities and observed frequencies than a small archive, as in the latter case larger random variations are expected. Hence one and the same histogram has to be interpreted differently depending on the size of the archive. To allow for unambiguous interpretation of the histogram, the graphical presentation should provide guidance as to whether the deviations from ideal behavior are within the expected range of fluctuations. In Bröcker and Smith (2007b), this problem was considered for the case of forecasts for binary observations. In this particular situation, reliability is often investigated by means of reliability diagrams. Ideally, reliability diagrams should be diagonal, but in practice, random fluctuations can cause the reliability diagram to exhibit deviations from this behavior even if the forecast system was reliable. In Bröcker and Smith (2007b), it was suggested how to modify reliability diagrams so as to visualize whether fluctuations are still consistent with reliability or not. The aim of the present discussion is to develop similar tools for the more general forecasts considered in this paper.

The idea is to plot the histogram "on probability paper": instead of the actual observed frequency, we show how probable that observed frequency would be if the forecast was reliable. To explain what this means, assume first the general situation in which there are L distinct events possible. (The predominant example we have in mind is that there are K ensemble members and the events are defined as the possible ranks of the verification among the ensemble members; in this particular case, there are $K{+}1$ possible events, whence $L{=}K{+}1$.) Suppose the forecast probability for the event labeled l is p_l, and the total number of trials is N, then – assuming that the forecast probability represents the true chance of events – the number n_l of instances exhibiting the event is of binomial distribution with parameters p_l and N. If $B(n_l, p_l, N)$ denoted the cumulative binomial distribution function, then the number

$$\nu_l = B(n_l, p_l, N) \tag{12}$$

gives the probability that, for a reliable forecast, an observed frequency *smaller* than n_l/N occurs. The interpretation of the ν_l is that for reliable forecasts, the ν_l are expected to be smaller than a given number q with a probability q. In other

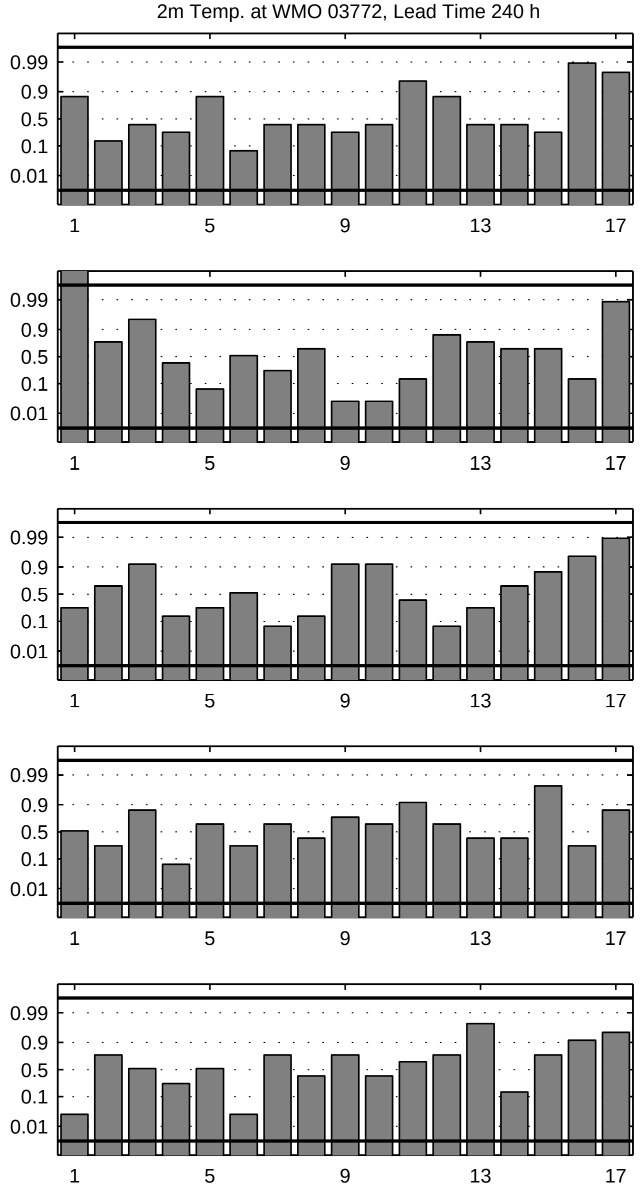

Fig. 1. Stratified ν_l-diagram for London Heathrow. Forecasts are stratified along the ERPS and distributed among five strata so that each stratum contains 20% of all instances. Going from the top to the bottom viewgraph, the ERPS increases (i.e. the expected skill becomes worse). The ν_l-diagrams are fairly uniform for all strata, except for the second one.

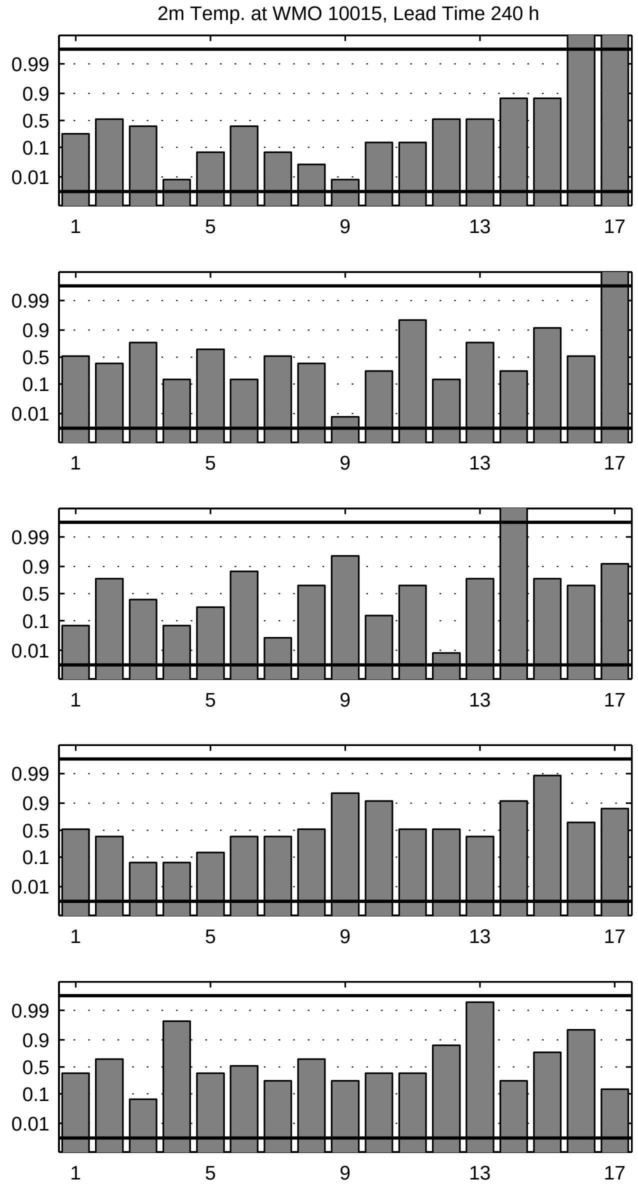

Fig. 2. Stratified ν_l-diagram for Helgoland Düne. Forecasts are stratified along the ERPS and distributed among five strata so that each stratum contains 20% of all instances. Going from the top to the bottom viewgraph, the ERPS increases (i.e. the expected skill becomes worse). The ν_l-diagrams are generally less uniform for smaller ERPS (upper viewgraphs).

words, if the forecast was reliable and we could repeat the reliability test an infinite number of times with new data each time, then for any q, a fraction q of all test runs would exhibit a ν_l smaller than q. Therefore, the ν_l provide direct quantitative information as to whether the deviations from reliable behavior are systematic or merely random. Hence the ν_l are more easy to interprete than the n_l as used in standard histograms. In all other respects, the interpretation of the ν_l is exactly the same as that of the n_l. For these reasons, it is suggested to plot diagrams of the ν_l rather than the n_l. Note that if the forecasts have been stratified, then $N{=}\#I_{\mathsf{A}}$ (as in Eq. 11). If furthermore the events are defined by ranking the verification in an ensemble forecasts, then $f_{l,\mathsf{A}}{=}n_l/N$ (again as in Eq. 11).

I found that the readability of the ν_l-diagram is further improved by scaling the ordinate by the logit-transformation $\log(\frac{\nu}{1-\nu})$. This has the effect of displaying the small probabilities 0.1, 0.01, 0.001, ... as well as the large probabilities

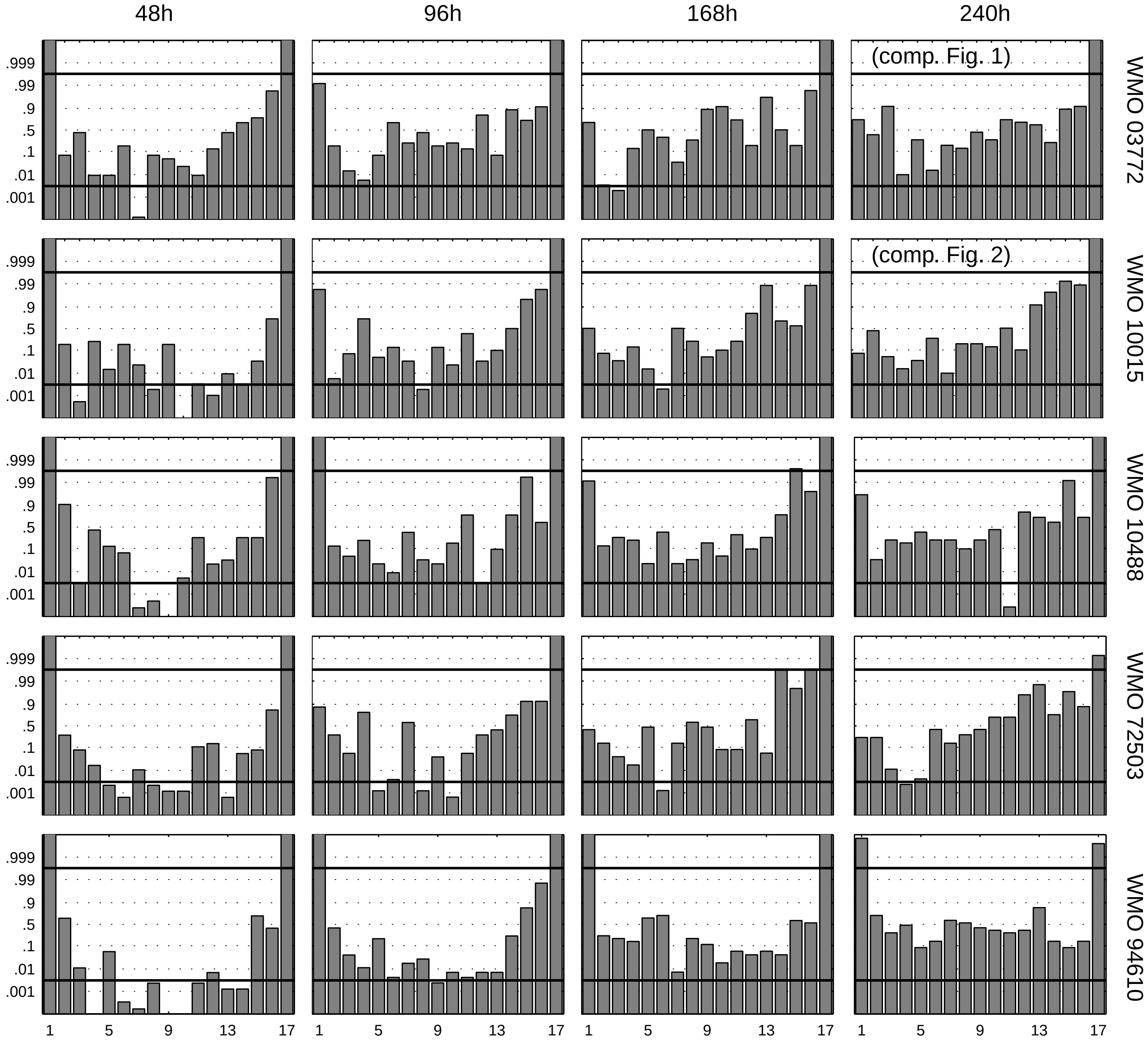

Fig. 3. Unstratified ν_l-diagrams for 2 m-temperature ensemble forecasts for various locations (rows) and lead times (columns). The ordinates display the probability of the observed frequency (see Eq. 12) on a $\log(\frac{\nu}{1-\nu})$-scale. The abscissa shows the ranks, binned into 17 bins (3 ranks per bin).

0.9, 0.99, 0.999, ... equidistantly. In Sect. 3, the reliability of some ensemble weather forecasts will be analyzed using the ν_l-diagram. In Figs. 1 and 2, ensemble forecasts have been stratified along five forecast strata. The corresponding ν_l-diagrams are displayed in five individual viewgraphs. The verification was distributed among 17 possible events, whence there are 17 bars in each viewgraph. Unstratified ν_l-diagrams are shown in Fig. 3. These plots are part of the results to be discussed in Sect. 3. When interpreting these diagrams, it is important to note that while the probability of any particular ordinate being below the value q is indeed q, the probability of all ordinates being below the value q is smaller, namely q^{17} (since there are 17 bins). This so-called Bonferroni-correction needs to be applied if the whole histogram is considered. In all ν_l-diagrams shown in this paper, Bonferroni-quantiles for 5% and 95% are shown as black horizontal lines. A detailed explanation as to the data underlying these figures is given in Sect. 3.

Obviously, plotting a ν_l-diagram for each forecast stratum requires plenty of space and might be unnecessarily detailed, whence it would be convenient to have a summary statistic which allowed for condensing a ν_l-diagram into one number. Thus, a single graph can contain summary statistics for different forecast strata. This section will be finished with discussing such a summary statistic. The general idea of Goodness-Of-Fit tests is to measure the similarity of the forecast probabilities p_l and the observed frequencies n_l/N by means of a suitable contrast function. The contrast function should vanish if $p_l{=}n_l/N$ for all l but otherwise be positive. Furthermore, in order to decide when the contrast should be considered large, the distribution of the constrast function has to be known (at least asymptotically for large N). A particular example is Pearson's original χ^2 Goodness-Of-Fit test. For unstratified forecasts, employing the χ^2 Goodness-Of-Fit test was suggested by Hamill (2001) and Anderson (1996) in the context of ensemble weather forecasts. In the present paper, it is proposed to perform a Goodness-Of-Fit test individually for each forecast stratum. For the contrast function, the choice taken here is the statistic

$$R := \sum_l -\log\left(\frac{p_l}{n_l/N}\right) n_l/N, \tag{13}$$

henceforth referred to as R-statistic, which is motivated by a coincidence of two interesting facts. The first fact about the R-statistic is that $2N{\cdot}R$ is asymptotically of χ^2-distribution with $K-1$ degrees of freedom (see e.g. Mood et al., 1974). Hence, rather than R itself, we consider $P_{\chi^2}(R)$, where P_{χ^2} is the cumulative χ^2-distribution function with $K-1$ degrees of freedom. The second fact of the R-statistic is an interesting connection to the Ignorance score. The quality of probabilistic forecasts is most appropriately measured by means of proper scores (Gneiting and Raftery, 2007; Bröcker and Smith, 2007a), of which the Ignorance is one example. The Ignorance score is defined as follows: for each observation y_n there is a corresponding verifying probability, which is the forecast probability assigned to the event which eventually occurs. The verifying probability is denoted by p_{y_n}. Note that in our case, due to the stratification, the forecast probabilities do not depend explicitely on time.

The Ignorance is defined as

$$\frac{1}{N}\sum_n -\log(p_{y_n}) = \sum_l -\log(p_l)n_l/N. \tag{14}$$

The Ignorance is a proper score, which means that for any two probability assignments $p_l, q_l, l{=}1\ldots L$,

$$\sum_l -\log\left(\frac{p_l}{q_l}\right) q_l \geq 0 \tag{15}$$

with equality if and only if $p_l{=}q_l$ for all $l{=}1\ldots L$. Setting $q_l{=}n_l/N$ demonstrates that the R-statistic is positive definite, which justifies interpreting it as a contrast function quantifying the discrepancy between forecast probabilities and observed frequencies. The overall score of the forecast as defined by Eq. (14) can be decomposed as follows

$$\sum_l -\log(p_l)n_l/N = \sum_l -\log(n_l/N)n_l/N + R, \tag{16}$$

where the first term on the right hand side is the skill of the "forecast" n_l/N. Since R is positive definite, Eq. (16) seems to suggest a fool-proof way of improving forecast skill: We simply adopt n_l/N as our new forecast probability for category l. Equation (16) ensures that a thus "recalibrated" forecast cannot have a skill worse than the original forecast. This conclusion is faulty though, since the recalibrated forecast is evaluated "in sample", which means on the same data already used to re-calibrate the forecast. This is a grave violation of the principles of statistical good practice. The conclusion is justified though if R is not merely positive but unusually large, that is larger than would normally be expected if the forecast was reliable. What should be considered an unusually large value for R is quantified by the χ^2-distribution.

A problem of the discussed Goodness-Of-Fit test (and, in fact, of any other Goodness-Of-Fit test) is that it is always possible to construct histograms which will perfectly pass the test, despite exhibiting obvious pathologies that are unlikely to arise by mere randomness. For example, histograms often display a clear trend or are convex. A trend upwards indicates that higher categories are assigned too small forecasts probabilities to, or in other words, they verify too often. In the case of ensemble forecasts, this indicates under-forecasting of the ensembles. Convex histograms can arise for two reasons. Either the ensembles exhibit systematically too small spread, and thus the extreme ranks verify too often. Or the histogram in fact confounds two forecast strata, one containing over-forecasting and one containing under-forecasting ensembles. Since the R-statistic is invariant against re-ordering of the categories though, forecasts with convex or tilted histograms might pass the Goodness-Of-Fit test undetected.

There are numerous other statistics suitable for Goodness-Of-Fit tests which, other than Pearson's classical statistic, are sensitive to the ordering of the categories (Elmore, 2005), such as the Cramér-von-Mises statistic. Apart from the R-statistic though, I have been unable to relate any of the common Goodness-Of-Fit statistics to the reliability term of a proper score.

3 Numerical examples

In this section, the discussed tools are applied to ensemble forecasts of two-meter temperature anomalies. Results are presented for five different locations (see Table 1 for details). The forecasts consist of the 50 (perturbed) member ensemble produced by the ECMWF ensemble prediction system. Station data of two-meter temperature was kindly provided by ECMWF as well. Forecasts were available for the years 2001–2005, featuring lead times from one to ten days. All

Table 1. Locations and beginning of the data record for the studied data sets.

WMO Number	Name	Location	Data Record starting
03772	London Heathrow AP	51°29′ N 000°27′ W	1 Jan 1981
10015	Helgoland Düne	54°11′ N 007°54′ E	1 Jan 1981
10488	Dresden Klotzsche AP	51°08′ N 013°47′ E	1 Sep 1991
72503	NY La Guardia AP	40°46′ N 073°54′ W	1 Apr 1981
94610	Perth Intl. AP	31°56′ S 115°57′ E	1 Jan 1981

data verified at noon. The observations from years previous to 2001 were used to fit a temperature normal, consisting of a fourth order trigonometric polynomial. The normal was subtracted from both ensembles and observations. Furthermore, the ensembles were de-biased, using the years 2001 and 2002.

Generally, the ensembles display significant deviations from reliability, in particular for short lead times. Plots were produced for three different diagnostics: Stratified ν_l-diagrams (Figs. 1, 2), unstratified ν_l-diagrams (Fig. 3), and stratified R-statistics (Fig. 4). Due to lack of space, only two stratified ν_l-diagrams (Figs. 1 for London Heathrow and 2 for Helgoland Düne), lead time 240 h, are shown here. The stratified ν_l-diagrams and R-statistics were produced by grouping the forecasts according to their expected ranked probability score (ERPS), as will be explained below. Although there are 51 ranks, only 17 bins were used rather than 51, in order to avoid overly cluttered plots. Since 17 divides 51, there are no aliasing effects (i.e. each bin contains the observed frequencies of exactly 3 of the 51 possible ranks). Unstratified ν_l-diagrams are shown in Fig. 3. Without exception, the corresponding (unstratified) R-statistics (not shown) exceeded the 95% quantile, indicating significant deviation from reliable behavior. Finally, stratified R-statistics are shown in Fig. 4. The configuration for the bins and forecast strata for the R-statistics was exactly like for the unstratified and stratified ν_l-diagrams.

A general result of the present study is that ensembles which “pretend” to have a good score are particularly unreliable, as will be demonstrated by stratifying the forecasts along the expected ranked probability score. To define the expected ranked probability score (ERPS), consider the ranked probability score (Epstein, 1969; Murphy, 1971) (RPS), defined via the scoring rule

$$S(y, G) := \int (G(\eta) - H(\eta - y))^2 \mathrm{d}\eta, \tag{17}$$

where G is the forecast distribution and H is the Heaviside function, which is one for positive arguments and zero otherwise. Note that the scoring rule (Eq. 17) gives a small value if the forecast is concentrated near y, which implies that a small RPS indicates a good score. The RPS is known to be a proper scoring rule (Gneiting et al., 2005; Bröcker and Smith, 2007a). The expected RPS (ERPS) is the mathematical expectation value of $S(Y, G)$ when Y is assumed to be of distribution G, that is

$$\mathrm{ERPS}(G) := \int S(y, G) \mathrm{d}G(y). \tag{18}$$

In other words, the ERPS is the score we would obtain on average if the verification were in fact drawn from the forecast. The ERPS is a function of the forecasts alone. To compute the ERPS, no observations are required. Of course, the ERPS does not provide the true score of the forecast, but rather a self-rating of the forecast distribution G, similar to, but more comprehensive than for example the variance of G. In the present situation G is not available explicitely, whence the ERPS has to be estimated from the ensemble. This is done here by first computing $S(x_i, F_{-i})$ for all ensemble members x_i, where F_{-i} is the empirical distribution function of the ensemble without the i'th member. The eventual estimate of the ERPS is obtained by averaging thus:

$$\mathrm{ERPS}(\boldsymbol{x}) := \frac{1}{K} \sum_i S(x_i, F_{-i})$$

It turns out that for the forecasts used in this study, the ERPS correlates very strongly with the ensemble standard deviation. In general, this need not be so. All forecasts were stratified according to their EPRS and ranked so that each forecast stratum contained 20% of all time instants.

In Figs. 1 and 2, stratified ν_l-diagrams are shown for London Heathrow and Helgoland Düne, respectively, for lead time 240 h. In both figures, the five viewgraphs represent ν_l-diagrams for the five forecast strata. The uppermost ν_l-diagram corresponds to forecast with very small ERPS (i.e. the self-rating is high), while lower diagrams correspond to forecasts with increasingly larger ERPS (i.e. the self-rating is low). Observed frequencies for forecasts with large ERPS are in general more uniform than for the smaller ERPS forecasts. Furthermore, for Helgoland Düne the diagrams for small ERPS seem to be tilted down to the left, unlike as for large ERPS, for which no such trend is apparent. The corresponding unstratified ν_l-diagrams (Fig. 3) or stratified R-statistics (Fig. 4) provide less specific information, albeit focusing on different aspects. As to the precise deficiency of the high ERPS forecasts, the full stratified ν_l-diagrams would need to be consulted (similar to Figs. 1 and 2, but for shorter lead times).

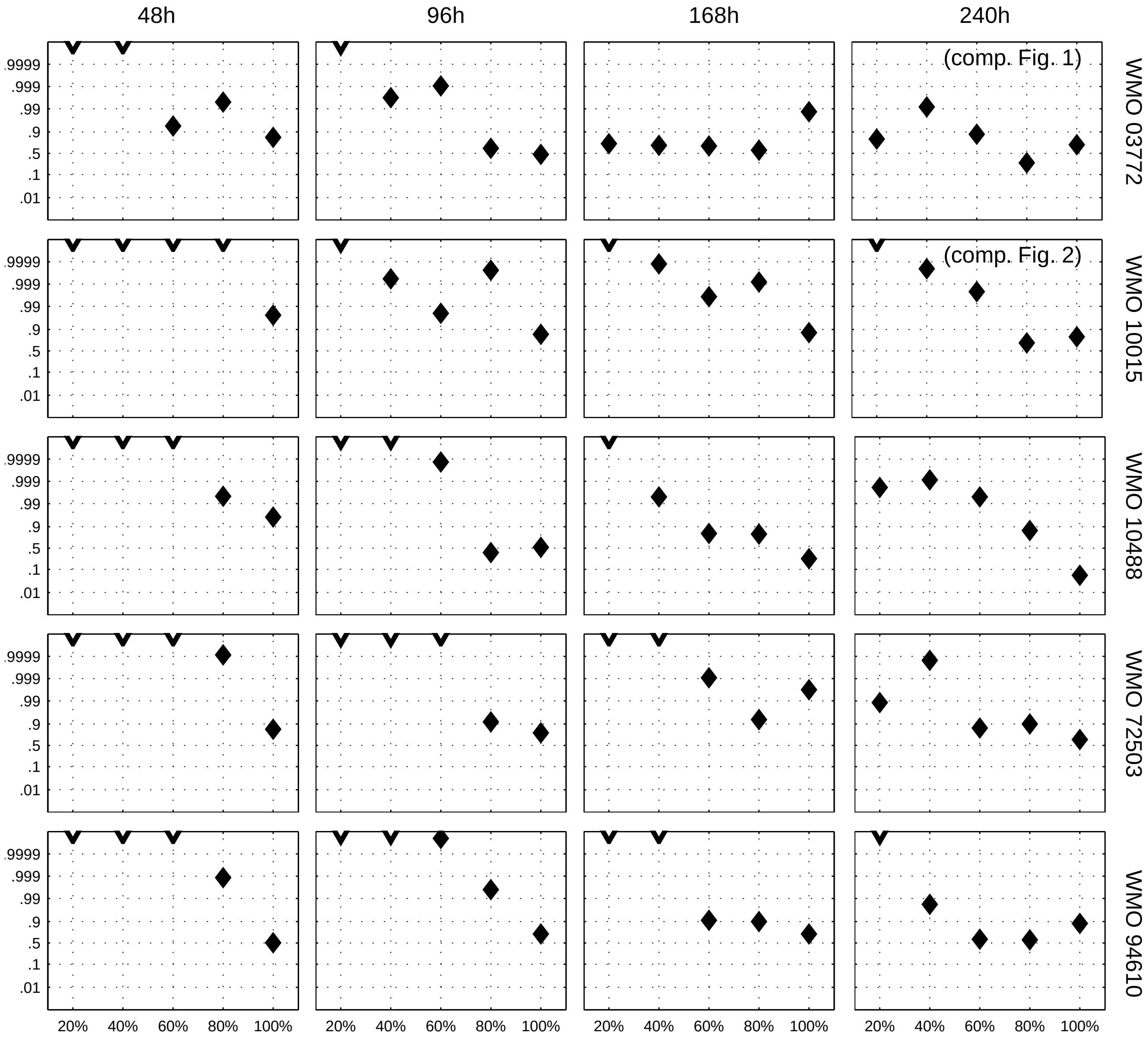

Fig. 4. Stratified R-statistic for 2 m-temperature ensemble forecasts for various locations (rows) and lead times (columns). The ordinates display the p-values of the R-statistic (which has a χ^2-distribution) on a $\log(\frac{\nu}{1-\nu})$-scale. The abscissa shows the ERPS, binned into 5 bins with an equal number of instances in each bin. The ERPS increases when going from left to right (i.e. the expected skill of the forecasts decreases).

Figure 3 shows the unstratified ν_l-diagrams for lead time 48 h, 96 h, 168 h, and 240 h (in columns 1–4, respectively), while diagrams of the stratified R-statistics are shown in Fig. 4. As can be discerned from both diagnostics, the reliability seems to improve for higher lead times. The unstratified ν_l-diagrams indicate too frequent occurrences of the extreme ranks (the diagram is convex). An overall bias seems to be present as well, as the diagrams seem to be tilted down to the left. Unstratified ν_l-diagrams can confound bias and insufficient spread and do not allow to discern any possible connections between these deficiencies and ERPS. For lead time 240 h, the unstratified ν_l diagram for London Heathrow (first row, last column of Fig. 3) indicates an overpopulation of the highest ranks, which, at first sight, seems not to be present at the corresponding stratified ν_l diagram (Fig. 1). It should be noted though that the stratified observed frequencies of the highest rank all happen to be comparably high at the same time, causing the unstratified observed frequency (just being the average of the stratified ones) to be significantly too high.

The stratified R-statistic (Fig. 4) helps to clarify the situation somewhat. Forecasts with small ERPS are, in general, significantly less reliable than forecasts featuring large values of ERPS. Forecasts generally become more reliable with increasing lead time. For short lead times though, almost all R-statistics are beyond the 90% quantiles, with some of them even being off the axis scale, indicated by triangles pointing down. For London Heathrow, this phenomenon is evident for small lead times, while for large lead times, forecasts of different ERPS seem to be more or less equally reliable. For location WMO10015, again there is a significant discrepancy in reliability between forecasts with large and small ERPS, but here this phenomenon is present even at lead time 240h (second row of Fig. 4).

The results are in general confirmed by the other investigated locations. Location WMO10488 is rather similar to WMO03772 in that for longer lead times, all forecasts become more reliable, while for WMO94610 and WMO72503, low ERPS forecast tend to be unreliable throughout.

We can conclude that the small ERPS forecasts have a tendency to under-forecast. This problem cannot be removed by a simple bias correction, as this would affect all forecasts equally. It seems that small ERPS forecasts need different (stronger) de-biasing than large ERPS forecasts. As an operational recommendation, the present study suggests that forecasts be stratified first, with different de-biasing being subsequently applied to each stratum. If for this purpose forecasts should be stratified along the ERPS or rather somehow else requires further investigation.

4 Conclusions

In this paper, the reliability of probabilistic forecasts, in particular ensemble forecasts, was revisited. A general mathematical definition of reliability was given, formalizing definitions of reliability given earlier by several authors. A frequently employed tool for reliability analysis, the rank histogram or Talagrand diagram, was discussed, and two shortcomings were pointed out. A long noted fact is that a uniform rank histogram is but a necessary condition for reliability. To obtain a more detailed picture of the reliability of the forecasting systems in different situations, it was suggested that forecasts be grouped, forming so-called forecast strata, and that individual histograms be plotted for all forecast strata. For a reliable forecast system, all forecast strata should exhibit uniform histograms. Secondly, histograms computed from limited data amounts are never exactly uniform, even for reliable forecast systems. Hence, an indication is needed how far a histogram is expected to deviate from uniformity merely due to randomness. One possible solution is to plot the probability of the observed frequency, instead of the observed frequency itself, thereby providing an indication of the likelihood of the result under the hypothesis that the forecast is reliable. Another advantage is that this plot is expected to be uniform even if the forecast probabilities are different for different categories. Furthermore, a slightly nonstandard Goodness-Of-Fit statistic was discussed. The employed contrast function relates directly to the reliability term of the Ignorance score. Again, Goodness-Of-Fit tests can be separately applied to individual forecast strata. The discussed tools are applied to 2 m-temperature anomalies for several locations and lead times. In addition to demonstrating the tools at work, the results suggest that the forecasts are particularly unreliable if they are expected to have high skill. It seems that the forecasts are both biased and under-disperse. To rectify this problem, different amounts of bias correction would need to be applied to different forecast strata, as a standard (indiscriminate) bias correction would affect all forecasts similarly.

Appendix A

In this appendix, it is shown (by means of a simple counter-example) that Eq. (2) generally does not hold if the condition "$\Gamma=G$" is replaced by "$\boldsymbol{X}=\boldsymbol{x}$". To recall the statement, if Y is the observation and $\boldsymbol{X}$ is an ensemble forecast for Y with K members, then in general

$$\mathbb{P}(Y < X^{(k)}|\boldsymbol{X}) \neq \frac{k}{K+1}, \tag{A1}$$

even if Y and $\boldsymbol{X}$ are independent draws from the same underlying distribution. Here is a simple example. Suppose the underlying or latent forecast distribution G is a normal distribution with standard deviation σ and a mean μ. We assume μ to be random too, with a standard normal distribution (i.e. with mean zero and standard deviation one), thus giving rise to randomness of the forecast G, while σ is known and fixed. In other words, for the verification Y and the ensemble members $X_1 \ldots X_K$, we assume the model

$$\begin{aligned} Y &= \mu + \sigma r_0, \\ X_1 &= \mu + \sigma r_1, \\ &\vdots \\ X_K &= \mu + \sigma r_K, \end{aligned} \tag{A2}$$

where $\mu; r_0 \ldots r_K$ are independent random variables with standard normal distribution. For simplicity, let us assume there is only one ensemble member, that is, $K=1$. We claim that even though the verification is smaller than the ensemble member with probability 0.5 on average, this is not true for each individual forecast instance. To see this, we investigate the distribution

$$\Phi(y|x) = \mathbb{P}(Y < y|X_1 = x),$$

which is the distribution of the verification given the ensemble (or the single ensemble member in this case). We will demonstrate that $\Phi(y|x)$ is not equal to 0.5 for $y=x$, or in other words that the median of $\Phi(y|x)$ is not equal to x.

An easy calculation shows that $\Phi(y|x)$ is a normal distribution as a function of y with mean (and also median) equal to $E(Y|X_1{=}x)$. But it follows from the relations (A2) that

$$E(Y|X_1 = x) = E(\mu|X_1 = x) = \frac{1}{1+\sigma^2}x.$$

Hence the median of $\Phi(y|x)$ is equal to $\frac{1}{1+\sigma^2}x$, which is different from x. In other words,

$$\mathbb{P}(Y < X_1|X_1 = x) = \Phi(x|x) \neq 0.5.$$

For the unconditional probability $\mathbb{P}(Y{<}X_1)$ however, we get

$$\begin{aligned}\mathbb{P}(Y < X_1) &= E(\mathbb{P}(Y < X_1|\mu)) \\ &= E(\mathbb{P}(r_0 < r_1)) \\ &= \int_{-\infty}^{\infty} F(r)\mathrm{d}F(r) \\ &= \frac{1}{2}\left(F^2(\infty) - F^2(-\infty)\right) \\ &= \frac{1}{2}, \end{aligned} \tag{A3}$$

where $F(r)$ is the distribution of the r_i in Eq. (A2). This calculation shows that indeed on average, the verification is smaller than the ensemble member with probability 0.5. This is, of course, a particular case of Eq. (3).

Acknowledgements. The author gratefully acknowledges fruitful discussions with the members of the Centre for the Analysis of Time Series, London School of Economics, in particular Liam Clarke, Milena Cuellar and Leonard A. Smith, as well as members of the Time Series Analysis group at the Max-Planck-Institute for the Physics of Complex Systems, in particular Thomas Laubrich. Two anonymous referees and one editor provided valuable comments and suggestions, including the term "forecast stratification". The data was kindfully provided by the European Centre For Medium Range Weather Forecasts (ECMWF).

Edited by: Z. Toth
Reviewed by: two anonymous referees

MAX-PLANCK-GESELLSCHAFT

The publication of this article is
financed by the Max Planck Society.

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996.

Breiman, L.: Probability, Addison-Wesley-Publishing, 1973.

Bröcker, J.: Decomposition of Proper Scores, Tech. rep., Max-Planck-Institut für Physik komplexer Systeme, Dresden, arXiv:0806.0813 [physics.ao-ph], 2008.

Bröcker, J. and Smith, L. A.: Scoring Probabilistic Forecasts: The Importance of Being Proper, Weather and Forecasting, 22, 382–388, 2007a.

Bröcker, J. and Smith, L. A.: Increasing the Reliability of Reliability Diagrams, Weather and Forecasting, 22, 651–661, 2007b.

Devroye, L.: Non-Uniform Random Variate Generation, Springer Verlag, 1986.

Elmore, K. L.: Alternatives to the Chi-Square Test for Evaluating Rank Histograms from Ensemble Forecasts, Weather and Forecasting, 20, 789–795, 2005.

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

Gneiting, T. and Raftery, A.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Statist. Assoc., 102, 359–378, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration, and Sharpness, Tech. rep., Department of Statistics, University of Washington, 2005.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001.

Hamill, T. M. and Colucci, S. J.: Verification of Eta–RSM Short Range Ensemble Forecasts, Mon. Weather Rev., 125, 1312–1327, 1997.

Hamill, T. M. and Colucci, S. J.: Evaluation of Eta–RSM Ensemble Probabilistic Precipitation Forecasts, Mon. Weather Rev., 126, 711–724, 1998.

Hansen, J. and Smith, L.: Extending the Limits of Forecast Verification with the Minimum Spanning Tree, Mon. Weather Rev., 132, 1522–1528, 2004.

Mood, A. M., Graybill, F. A., and Boes, D. C.: Introduction to the Theory of Statistics, McGraw-Hill Series in Probability and Statistics, McGraw-Hill, 1974.

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

Murphy, A. H. and Winkler, R. L.: Reliability of Subjective Probability Forecasts of Precipitation and Temperature, Appl. Statist., 26, 41–47, 1977.

Murphy, A. H. and Winkler, R. L.: A General Framework for Forecast Verification, Mon. Weather Rev., 115, 1330–1338, 1987.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of Probabilistic Prediction Systems, in: Workshop on Predictability, pp. 1–25, ECMWF, 1997.

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and Ensemble Forecasts, in: Forecast Verification, edited by: Jolliffe, I. T. and Stephenson, D. B., chap. 7, pp. 137–163, John Wiley & Sons, Ltd., Chichester, 2003.

Toth, Z., Talagrand, O., and Zhu: The attributes of forecast systems: A framework for the evaluation and calibration of weather forecasts, in: Predictability of Weather and Climate, edited by: Palmer, T. N. and Hagedorn, R., pp. 584–595, Cambridge University Press, 2005.

Wilks, D. S.: Statistical Methods in the Athmospheric Sciences, vol. 59 of International Geophysics Series, Academic Press, second edn., 2006.