Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–Uncertainty Decomposition

STEVEN V. WEIJS, RONALD VAN NOOIJEN, AND NICK VAN DE GIESEN

Department of Water Resources Management, Delft University of Technology, Delft, Netherlands

(Manuscript received 17 September 2009, in final form 6 April 2010)

ABSTRACT

This paper presents a score that can be used for evaluating probabilistic forecasts of multicategory events. The score is a reinterpretation of the logarithmic score or ignorance score, now formulated as the relative entropy or Kullback–Leibler divergence of the forecast distribution from the observation distribution. Using the information–theoretical concepts of entropy and relative entropy, a decomposition into three components is presented, analogous to the classic decomposition of the Brier score. The information–theoretical twins of the components uncertainty, resolution, and reliability provide diagnostic information about the quality of forecasts. The overall score measures the information conveyed by the forecast. As was shown recently, information theory provides a sound framework for forecast verification. The new decomposition, which has proven to be very useful for the Brier score and is widely used, can help acceptance of the logarithmic score in meteorology.

1. Introduction

Forecasts are intended to provide information to the user. Forecast verification is the assessment of the quality of a single forecast or forecasting scheme (Jolliffe and Stephenson 2008). Verification should therefore assess the quality of the information provided by the forecast. It is important here to note the distinction between quality, which depends on the correspondence between forecasts and observations, and value, which depends on the benefits of forecasts to users (Murphy 1993). In this paper, we assume that the verification is intended to quantitatively measure quality. Several scores and visualization techniques have been developed that measure certain desirable properties of forecasts with the purpose of assessing their quality. One of the most commonly used skill scores (Stephenson et al. 2008) is the Brier score (BS) (Brier 1950), which is applicable to probabilistic forecasts of binary events. The Brier skill score (BSS) measures the BS relative to some reference forecast, which is usually climatology. Murphy (1973) showed that the BS can be decomposed into three components: uncertainty, resolution, and reliability. These components give insight into some different aspects of forecast quality. The first component, uncertainty, measures the inherent uncertainty in the process that is forecast. Resolution measures how much of this uncertainty is explained by the forecast. Reliability measures the bias in the probability estimates of the probabilistic forecasts. A perfect forecast has a resolution that is equal to (fully explains) the uncertainty and a perfect reliability.

Information theory provides a framework for measuring information and uncertainty (see Cover and Thomas 2006 for a good introduction). As forecast verification should assess the information that the forecaster provides to the user, using information theory for forecast verification appears to be a logical choice. A concept central to information theory is the measure of uncertainty named entropy. However, consulting two standard works about forecast verification, we noted that the word entropy is mentioned only thrice in Jolliffe and Stephenson (2003) and not one single time in Wilks (1995). This indicates that the use of information–theoretical measures for forecast verification is not yet widespread, although some important work has been done by Roulston and Smith (2002), Ahrens and Walser (2008), Leung and North (1990), and Kleeman (2002).

Leung and North (1990) used information–theoretical measures like entropy and transinformation in relation

Corresponding author address: Steven Weijs, Delft University of Technology, Stevinweg 1, P.O. Box 5048, 2600 GA Delft, Netherlands.
E-mail: s.v.weijs@tudelft.nl

DOI: 10.1175/2010MWR3229.1

to predictability. Kleeman (2002) proposed to use the relative entropy between the climatic and the forecast distribution to measure predictability. The applications of information theory in the framework of predictability are mostly concerned with modeled distributions of states and how uncertainty evolves over time. Forecast verification, however, is concerned with comparing observed values with the forecast probability distributions. Roulston and Smith (2002) introduced the ignorance score, a logarithmic score for forecast verification, reinterpreting the logarithmic score (Good 1952) from an information–theoretical point of view. They related their score to relative entropy between the forecast distribution and the "true" probability distribution function (PDF), which they defined as "the PDF of consistent initial conditions evolved forward in time under the dynamics of the real atmosphere." Ahrens and Walser (2008) proposed information–theoretical skill scores to be applied to cumulative probabilities of multicategory forecasts. Very recently, Benedetti (2009) showed that the logarithmic score is a unique measure of forecast goodness. He showed that the logarithmic score is the only score that simultaneously satisfies three basic requirements for such a measure. These requirements are additivity, locality (which he interprets as exclusive dependence on physical observations), and strictly proper behavior. For a discussion of these requirements, see Benedetti (2009). Furthermore, Benedetti (2009) analyzed the Brier score and showed that it is equivalent to a second-order approximation of the logarithmic score. He concludes that lasting success of the Brier score can be explained by the fact that it is an approximation of the logarithmic score. Benedetti also mentions the well-known and useful decomposition of the Brier score into uncertainty, resolution, and reliability as a possible reason for its popularity.

In this paper, we follow a route similar to Benedetti's, but from a different direction. From an analogy with the Brier score, we propose to use the Kullback–Leibler divergence (or relative entropy) of the observation from the forecast distribution as a measure for forecast verification. The score is named "divergence score." When assuming perfect observations, our measure is equal to the ignorance score or logarithmic score, and can be seen as a new reinterpretation of ignorance as the Kullback–Leibler divergence from the observation to the forecast distribution. Presenting a new decomposition into uncertainty, resolution, and reliability, analogous to the well-known decomposition of the Brier score (Murphy 1973), yields insight into the way the divergence score (DS) measures the information content of probabilistic binary forecasts. The decomposition can help acceptance and wider application of the logarithmic score in meteorology.

Section 2 of this paper presents the mathematical formulation of the DS and its components. Section 2 also shows the analogy with the Brier score components. Section 3 compares the divergence score with existing information–theoretical scores. It is shown that the DS is actually a reinterpretation of the ignorance score (Roulston and Smith 2002) and that one of the ranked mutual information scores defined by Ahrens and Walser (2008) is equal to the skill score version of DS, when the reliability component is neglected (perfect calibration assumed). A generalization to multicategory forecasts is presented in section 4. The inherent difficulty found in formulating skill scores for ordinal category forecasts is also analyzed and leads to the idea that this can be explained by explicitly distinguishing between information and useful information for some specific user. This distinction provides some insights in the roles of the forecaster and the user of the forecast. Section 5 presents an application to a real dataset of precipitation forecasts. Section 6 summarizes the conclusions and restates the main arguments for adopting the divergence score.

2. Definition of the divergence score

a. Background

By viewing the Brier score as a quadratic distance measure and translating it into the information–theoretical measures for uncertainty and divergence of one distribution from another, we formulate an information–theoretical twin of the Brier score and its components. First, some notation is introduced, followed by formulation of the Brier score. Then the information–theoretical concept of relative entropy is presented as an alternative scoring rule. In the second part of this section, it is shown how the new score can be decomposed into the classic Brier score components: uncertainty, resolution, and reliability.

b. Definitions

Consider a binary event, like a day without rainfall or with rainfall. This can be seen as a stochastic process with two possible outcomes. The outcome of the event can be represented in a probability mass function (PMF). For the case of binary events, the empirical PMF of the event after the outcome has been observed is a 2D vector, denoted by $\mathbf{o} = (1 - o, o)^{\mathrm{T}}$. Assuming certainty in the observations, $o \in \{0, 1\}$; therefore, $\mathbf{o} = (0, 1)^{\mathrm{T}}$ if it rained and $(1, 0)^{\mathrm{T}}$ otherwise. Now suppose a probabilistic forecast of the outcome of the binary event is issued in the form of a probability of occurrence f. This can also be written as a forecast PMF $\mathbf{f} = (1 - f, f)^{\mathrm{T}}$, with $f \in [0, 1]$. If for example an 80% chance of rainfall is forecast, this is denoted as $\mathbf{f} = (0.2, 0.8)^{\mathrm{T}}$.

The Brier score for a single forecast at time t measures distance between observation and forecast PMFs by the square Euclidean distance:

$$\mathrm{BS}_t = 2(f_t - o_t)^2 = (\mathbf{f}_t - \mathbf{o}_t)^{\mathrm{T}}(\mathbf{f}_t - \mathbf{o}_t). \quad (1)$$

For a series of forecasts and observations, the Brier score is simply the average of the Brier scores for the individual forecasts:

$$\mathrm{BS} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{f}_t - \mathbf{o}_t)^{\mathrm{T}}(\mathbf{f}_t - \mathbf{o}_t). \quad (2)$$

Note that this is the original definition by Brier (1950). Nowadays, the Brier score is almost always defined as half the value of (2) (Ahrens and Walser 2008).

In information theory, the difference between two probability distributions is measured by relative entropy or Kullback–Leibler divergence (D_{KL}). The D_{KL} is not symmetric, so as not to be confused with a true distance like the quadratic score. In this paper $D_{\mathrm{KL}}(\mathbf{x}\|\mathbf{y})$ will be referred to as the divergence from $\mathbf{x}$ to $\mathbf{y}$ (or of $\mathbf{y}$ from $\mathbf{x}$). It measures the missing information in case one assumes the distribution is $\mathbf{y}$ while the true distribution is $\mathbf{x}$.

c. The divergence score

We now define the divergence score (DS), replacing the quadratic distance from the BS with the Kullback–Leibler divergence. For one single forecast, the DS functions as a scoring rule. It is the Kullback–Leibler divergence of the forecast distribution from the observation distribution over the $n = 2$ possible events i:

$$\mathrm{DS}_t = D_{\mathrm{KL}}(\mathbf{o}_t\|\mathbf{f}_t) = \sum_{i=1}^{n}[\mathbf{o}_t]_i \log\left(\frac{[\mathbf{o}_t]_i}{[\mathbf{f}_t]_i}\right) \quad (3)$$

The DS over a series of forecast–observation pairs measures the average divergence of the forecast distribution from the observation:

$$\mathrm{DS} = \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t\|\mathbf{f}_t). \quad (4)$$

The divergence score can be interpreted as the information gain when one moves from the prior forecast distribution to the observation distribution. When this information gain is zero, the forecast already contained all the information that is in the observation and therefore is perfect. If the information gain from the forecast to the certain observation is equal to the climatological uncertainty, the forecast did not contain more information than the climate, and therefore was useless. Another way to view the divergence score is the remaining uncertainty about the true outcome, after having received the forecast (see Fig. 1).

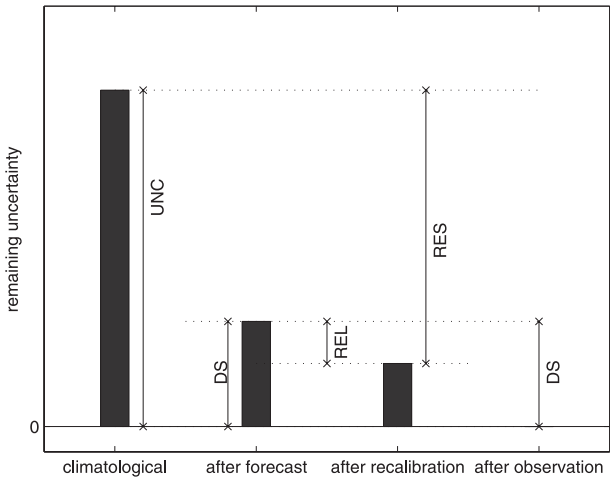

FIG. 1. The components measure the differences in uncertainty at different moments in forecasting process, judged in hindsight. The divergence score measures the remaining uncertainty after taking the forecast at face value.

d. Decomposition

The new score can be decomposed in a similar way as the Brier score. The classic decomposition of the BS into reliability (REL), resolution (RES), and uncertainty (UNC) is

$$\mathrm{BS} = \mathrm{REL}_{\mathrm{BS}} - \mathrm{RES}_{\mathrm{BS}} + \mathrm{UNC}_{\mathrm{BS}}, \quad (5)$$

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f}_k - \bar{\mathbf{o}}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_k - \bar{\mathbf{o}})^2 + \bar{\mathbf{o}}(1 - \bar{\mathbf{o}}), \quad (6)$$

where N is the total number of forecasts issued, K is the number of unique forecasts issued, $\bar{\mathbf{o}} = \sum_{t=1}^{N}\mathbf{o}/N$ is the observed climatological base rate for the event to occur, n_k is the number of forecasts with the same probability category, and $\bar{\mathbf{o}}_k$ is the observed frequency, given forecasts of probability $\mathbf{f}_k$. The reliability and resolution terms in (6) are summations of some distance measure between two binary probability distributions, while the last term measures the uncertainty in the climatic distribution, using a polynomial of degree two. We now present information–theoretical twins for each of the three quadratic components of the BS, using entropy and relative entropy. It is shown that they add up to the divergence score proposed earlier.

The first component, reliability, measures the conditional bias in the forecast probabilities. In the DS, it is

the expected divergence of the observed probability distribution from the forecast probability distribution, both stratified (conditioned) on all issued forecast probabilities. In the ideal case, the observed frequency is equal to the forecast probability for all of the issued forecast probabilities. Only in this case is the reliability 0. This is referred to as a perfectly calibrated forecast. Note that reliability is defined in the opposite direction to the meaning of the word in the English language. A perfectly reliable forecast has a reliability of 0. The reliability can be calculated with

$$\mathrm{REL}_{\mathrm{DS}} = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \mathbf{f}_k), \tag{7}$$

where N is the total number and K the number of unique forecasts issued, n_k is the number of forecasts with the same probability category, $\bar{\mathbf{o}}_k$ is the observed frequency distribution for forecasts in group k, and $\mathbf{f}_k$ is the forecast PMF for group k.

The second component, resolution, measures the reduction in climatic uncertainty. It can be seen as the amount of information in the forecast. In the DS, it is defined as the expected divergence of the conditional frequencies from the marginal frequency of occurrence. The minimum resolution is 0, which occurs when the climatological probability is always forecast or the forecasts are completely random. The resolution measures the amount of uncertainty in the observation explained by the forecast. In the ideal case the resolution is equal to the uncertainty, which means all uncertainty is explained. This is only the case for a deterministic forecast that is either always right or always wrong. In the last case, the forecast needs to be recalibrated:

$$\mathrm{RES}_{\mathrm{DS}} = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{o}}). \tag{8}$$

From (8) it becomes clear that the resolution term is the expectation over all forecast probabilities of the divergence from the conditional probability of occurrence to the marginal probability of occurrence. In information theory, this quantity is known as the mutual information I between the forecasts and the observation ($\underset{k}{E}$ denotes the expectation over k):

$$\mathrm{RES}_{\mathrm{DS}} = \underset{k}{E}[D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{o}})], \tag{9}$$

$$= \underset{k}{E}\{D_{\mathrm{KL}}[(\bar{\mathbf{o}} | \mathbf{f}_k) \| \bar{\mathbf{o}}]\} = I(\mathbf{f}; \mathbf{o}). \tag{10}$$

The third component, uncertainty, measures the initial uncertainty about the event. This observational uncertainty is measured by the entropy of the climatological distribution $H(\bar{\mathbf{o}})$. It is a function of the climatological base rate only and does not depend on the forecast. The uncertainty is maximum if the probability of occurrence is 0.5 and 0 if the probability is either 0 or 1:

$$\mathrm{UNC}_{\mathrm{DS}} = H(\bar{\mathbf{o}}) = -\sum_{i=1}^{n} \{[\bar{\mathbf{o}}]_i \log[\bar{\mathbf{o}}]_i\}. \tag{11}$$

Like for the BS, for a single forecast–observation pair, uncertainty and resolution are 0 and the total score is equal to the reliability, which acts as a scoring rule. Over a larger number of forecasts uncertainty approaches climatic uncertainty and reliability should go to zero if the forecast is well calibrated. In the appendix it is shown that, just like in the Brier score, the relation DS = REL − RES + UNC holds. The relation between the components and the total score (DS) is

$$\begin{aligned} \mathrm{DS} &= \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t \| \mathbf{f}_t) \\ &= \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \mathbf{f}_k) \\ &\quad - \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{o}}) + H(\bar{\mathbf{o}}). \end{aligned} \tag{12}$$

Note that this decomposition is valid for forecasts of events with an arbitrary number of categories and is not restricted to the binary case.

e. Relation to Brier score and its components

For the binary case, the Brier score can be seen as a second-order approximation of the divergence score (also noted by Benedetti 2009). Both scores have their minimum only with a perfect forecast. When the forecast is not perfect, the Brier score is symmetric in the error in probabilities, while the divergence score is not, except for the case where the true forecast probability is 0.5. Therefore the divergence score is a double-valued function of the Brier score (Roulston and Smith 2002). Consequently, when two forecasting systems are compared, the forecasting system with the higher Brier score may have the lower divergence score.

The uncertainty component in the Brier score is a second-order approximation of the uncertainty term in the divergence score (entropy), with the same location of zero uncertainty (100% probability of one of the two events) and maximum uncertainty (equiprobable events with 50% probability, see Fig. 2). In the Brier score, the maximum of the uncertainty component is 0.5, while in the divergence score it is 1 (bit).

Resolution in the Brier score is the variance of conditional mean probabilities. It is a mean of squared deviations from the climatic probability. Resolution in the

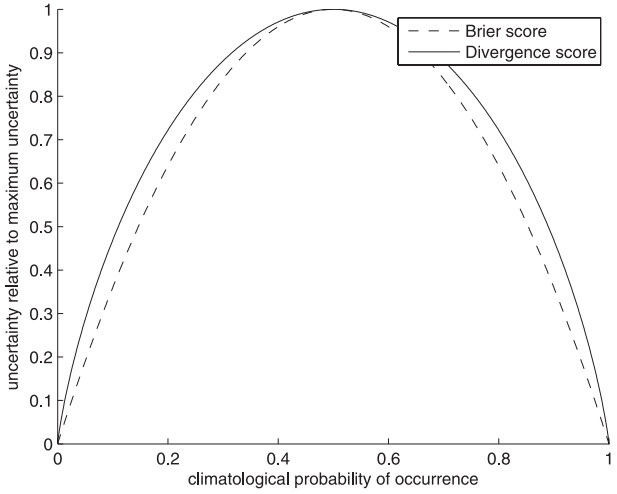

FIG. 2. The uncertainty component of the Brier score is a second-order approximation of the entropy, with coinciding minimum and maximum values. The uncertainty for the Brier score is divided by its maximum value of 0.5 to allow a clear comparison with the uncertainty term of the divergence score, which is measured in bits.

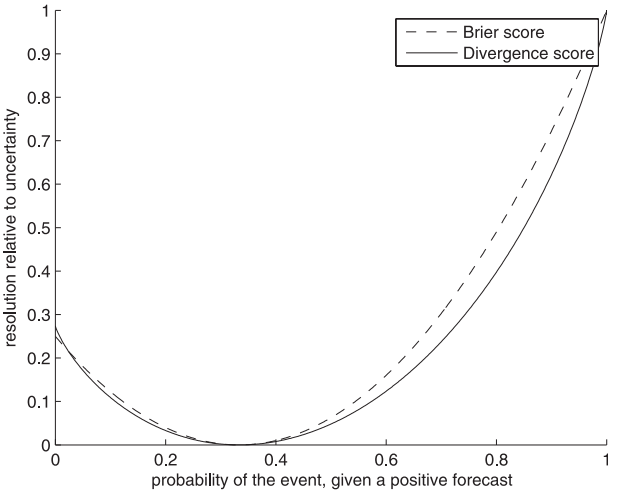

FIG. 3. The resolution term of the divergence score is asymmetric in probability while the Brier score resolution is not.

divergence score is a mean of divergences. Divergences are asymmetric in probabilities. The resolution in both the Brier and the divergence score can take on values between zero and the uncertainty. In both scores, it can be seen as the amount of uncertainty explained. The resolution in the Brier score is the second-order approximation of the resolution of the divergence score, satisfying the condition that the minimum is zero and in the same location (the climatic probability) and that the maximum possible value is equal to the inherent uncertainty of the forecast event (see Fig. 3).

Reliability in the Brier score is bounded between 0 and 1, whereas in the divergence score the reliability can reach infinity (see Fig. 4). This is the case when wrong deterministic forecasts are issued. Generally the reliability in the divergence score is especially sensitive to events with near-deterministic wrong forecasts. Overconfident forecasting is therefore sanctioned more heavily than in the Brier score.

f. Normalization to a skill score

Because the Brier score depends on the climatological probability, which is independent from the forecast quality, it is common practice to normalize it to the BSS with the climatology forecast as a reference. A perfect forecast is taken as a second reference. For the DS it is possible to use the same normalization procedure, yielding the divergence skill score (DSS):

$$\mathrm{DSS} = \frac{\mathrm{DS} - \mathrm{DS}_{\mathrm{ref}}}{\mathrm{DS}_{\mathrm{perf}} - \mathrm{DS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{DS}}{\mathrm{DS}_{\mathrm{ref}}}. \qquad (13)$$

The score for a perfect forecast $\mathrm{DS}_{\mathrm{perf}}$ is 0. In the climatological forecast $\mathrm{DS}_{\mathrm{ref}}$, both resolution and reliability are 0 (perfect reliability, no resolution). The DSS therefore reduces to

$$\mathrm{DSS} = 1 - \frac{\mathrm{UNC} - \mathrm{RES} + \mathrm{REL}}{\mathrm{UNC}} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}}. \qquad (14)$$

This leads to a positively oriented skill score that becomes one for a perfect forecast and zero for a forecast of always the climatological probability. Also, a completely random forecast that has a marginal distribution equal to climatology gets a zero score. Negative (i.e., ''worse than climate'') skill scores are possible if the

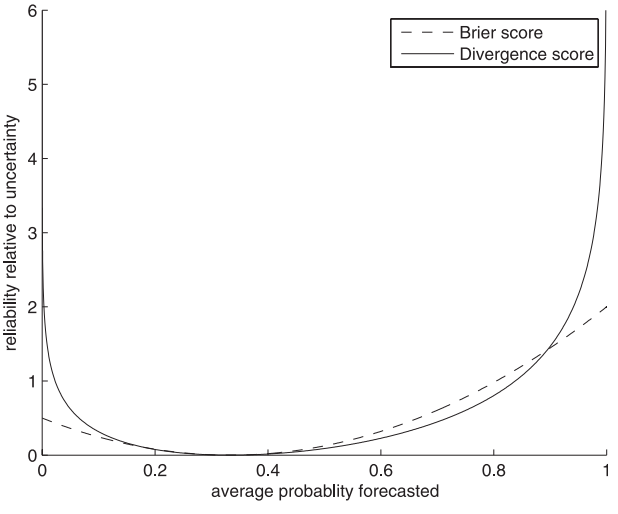

FIG. 4. The reliability is asymmetric in the DS, while symmetric in the BS. In the BS, it is bounded, while in the DS, it can reach infinity.

reliability is larger (worse) than the resolution. In case the resolution is significant, calibration of the forecast can yield a positive skill score, meaning that a decision maker using the recalibrated forecast is better off than a decision maker using climatology or a random strategy. This shows the importance of looking at the individual components when diagnosing a forecast system's performance.

Summarizing, the divergence score and its components combine two types of measures to replace the quadratic components in the Brier score decomposition. First, the quadratic distances between probability distributions are replaced by Kullback–Leibler divergences, which are asymmetric. Care should therefore be taken in which direction the divergence is calculated. Second, the polynomial uncertainty term is replaced by the entropy of the climatology distribution. The total scores and components are visualized in Fig. 1.

3. Relation to existing information–theoretical scores

a. Relation to the ranked mutual information skill scores

Ahrens and Walser (2008) proposed the ranked mutual information skill score (RMIS). The score is intended for use with multicategory forecasts, which will be treated later in this paper. For the special case of forecasts of binary events, the RMIS_O can be written as the mutual information between forecasts and observations divided by the entropy of the observations:

$$\mathrm{RMIS}_O = \frac{I(\mathbf{f}, \mathbf{o})}{H(\bar{\mathbf{o}})}. \tag{15}$$

When comparing (15) with (9), (11), and (14), it becomes clear that

$$\mathrm{RMIS}_O = \frac{\mathrm{RES}_{\mathrm{DS}}}{\mathrm{UNC}_{\mathrm{DS}}}. \tag{16}$$

This means that for the case of a binary forecast, RMIS_O equals the DSS in case the reliability is perfect (zero). In case the forecast is not well calibrated, RMIS_O neglects the reliability component and measures the amount of information that would be available to a user after calibration. The DSS measures the information conveyed to a user taking the forecasts at face value. The individual components of the DSS also indicate the potentially extractable information as measured by the RMIS_O.

b. Equivalence to the ignorance score

Roulston and Smith (2002) defined the ignorance score from the viewpoint of using the forecast probability distribution as basis for a data compression scheme. The scoring rule measures the ignorance or information deficit of a forecaster, compared to a person knowing the true outcome of the event (j). The ignorance scoring rule is defined as

$$\mathrm{IGN} = -\log_2 f_j, \tag{17}$$

where f_j is the probability that the forecaster had assigned to the event that actually occurred. The ignorance score is a reinterpretation of the logarithmic score by Good (1952).

By expanding the relative entropy measure that we use as a scoring rule, it becomes clear that divergence from the certain observation PMF ($\mathbf{o}$) to the forecast PMF ($\mathbf{f}$) is actually the same as the ignorance (IGN) or the logarithmic score:

$$D_{\mathrm{KL}}(\mathbf{o}_t \| \mathbf{f}_t) = \sum_{i=1}^{n} o_i \log\left(\frac{o_i}{f_i}\right) = o_{i\neq j} \log\left(\frac{o_{i\neq j}}{f_{i\neq j}}\right) + o_{i=j} \log\left(\frac{o_{i=j}}{f_{i=j}}\right). \tag{18}$$

Because $o_{i\neq j} = 0$ and $o_{i=j} = 1$, this reduces to

$$D_{\mathrm{KL}}(\mathbf{o}_t \| \mathbf{f}_t) = 0 \log\left(\frac{0}{f_{i\neq j}}\right) + 1 \log\left(\frac{1}{f_{i=j}}\right) = -\log f_j = \mathrm{IGN}. \tag{19}$$

This means that the divergence scoring rule presented in this paper (DS) is actually equal to the ignorance (IGN). The ignorance is therefore not only "A scoring rule ... closely related to relative entropy" as defined by Roulston and Smith (2002), but actually is also a relative entropy. The difference is in the distributions that are used to calculate the relative entropy. Roulston and Smith (2002) refer to a relation to the divergence between the unknown "true" distribution $\mathbf{p}$ and the forecast distribution $\mathbf{f}$ (see 20):

$$D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{f}) = E[\mathrm{IGN}] - H(\mathbf{p}). \tag{20}$$

The divergence that is used in the divergence score is calculated from the PMF after the observation $\mathbf{o}$ instead of $\mathbf{p}$. That makes the second term of the rhs vanish and IGN equal to the divergence.

Using the decomposition presented in (12), the ignorance score for a series of forecasts can now also be decomposed into a reliability, a resolution, and an uncertainty component. This decomposition, until now only applied to the Brier score, has proven very useful to gain insight into the aspects of forecast quality. Furthermore,

the new interpretation of the ignorance score as the average divergence of observation PMFs from forecast PMFs links to results from information theory more straightforwardly.

4. Generalization to multicategory forecasts

a. Nominal category forecasts

When extending verification scores from forecasts of binary events to multicategory forecasts, it is important to differentiate between nominal and ordinal forecast categories. In the case of nominal forecasts, there is basically one question that is relevant for assessing their quality: How well did the forecaster predict the category to which the outcome of the event belongs? In nominal category forecasts, there is no natural ordering in the categories into which the forecast event is classified. For this case of forecast verification, there is no difference between the categories in which the event did not fall. Although the probability attached to those events conveys information at the moment the forecast is received, the only probability relevant for verification, after the observation has been made, is the probability that the forecaster attached to the event that did occur. The quadratic score of Brier (1950) can also be used for multiple category forecasts. In that case, $\mathbf{f}_t$ and $\mathbf{o}_t$ are the PMFs of the event before and after the observation, now having more than two elements. The problem with this score is that it depends on how the forecast probabilities are spread out over the categories that did not occur. For nominal events this dependency is not desirable, as all probability attached to the events that did not occur is equally wrong.

The DS does not suffer from this problem because it only depends on the forecast probability of the event that did occur (19). The DS as presented in (3) can directly be applied on nominal category forecasts. A property of the score is that a high number of categories make it more difficult to obtain a good score. To compare nominal category forecasts with different numbers of categories, the DS should be normalized to a skill score (DSS); see (13).

b. Ordinal category forecasts

When dealing with forecasts of events in ordinal categories, there is a natural ordering in the categories. This means that the cumulative distribution function (CDF) starts to become meaningful. There are now two possible questions that can be relevant for verification of the probabilistic forecast.

1) How well did the forecaster predict the category to which the outcome of the event belongs?
2) How well did the forecaster predict the exceedence probabilities of the thresholds that define the category boundaries?

The first question is equal to the one that is of interest for nominal forecasts. However, in the ordinal case there is a difference between the categories in which the observed event did not fall. Forecasts of categories close to the one observed are preferred over categories more distant from the one observed. Therefore, skill scores for ordinal category forecasts are often required to be ''sensitive to distance'' (Epstein 1969; Laio and Tamea 2007; Murphy 1971, 1970). This requirement has led to the introduction of the ranked probability score (RPS; Epstein 1969), which is now widely used. The DS is not sensitive to distance in the sense of the RPS, because DS is insensitive to the forecasts for the nonoccurring events. However, there still is an apparent sensitivity to distance introduced through the forecast PMF. A forecaster will usually attach more probability to the categories adjacent to the one that is considered most likely, simply because they are also likely. Therefore, missing the exact category of the observation with the most likely forecast still leads to a relatively low penalty in the score, if the uncertainty estimation of the forecaster was correct and significant probability was forecast for the other likely (often neighboring) categories. Over a series of forecasts, the apparent distance sensitivity of the penalty given by the DS is therefore defined by the PMF of the forecaster alone, and independent of what the categories represent. In verification literature, the property of only being dependent on the probability assigned to the event that actually occurred is known as locality, which is often seen as a nondesirable property of a score. Whether or not locality is desirable can be questioned (Mason 2008). We argue that in absence of a context of the users of the forecast there is no justification for using nonlocal scores, which require some sensible distance measure to be specified apart from the natural distance sensitivity introduced by the forecast PMF. When assessing value or utility of forecasts, as opposed to quality, nonlocal scores can be used. However, in that case the distance measure should depend on the users and associated decision processes, as these determine the consequences of missing the true event by a certain number of categories distance. In nonlocal verification scores, the distance measure is often not explicitly specified in terms of utility, making it unclear what is actually measured. In those cases, the utility function of the users becomes more like an emerging property of the skill score instead of the other way around. Benedetti (2009) also presents locality as a basic requirement for a measure of forecast goodness, interpreting locality as ''exclusive dependence on physical

observations.'' He correctly states that it is a violation of scientific logic if two series of forecasts that assign the same probabilities to a series of observed events gain different scores, based on probabilities assigned to events that have never been observed. For a more elaborate treatment of this view on the fundamental discussion about locality, the reader is referred to Benedetti (2009) and Mason (2008).

The second question, regarding the forecast quality of exceedence probabilities, differs from the first because all the thresholds are considered at once. Therefore, the quality of a single forecast depends on the entire PMF of the forecast and not only on the probability forecast for the event that occurs. Therefore, scores that are formulated for cumulative distributions can never be local. This means that apart from the physical observations, the importance attached to the events influences the score. So some assumption about value is added and the score is not a pure measure of quality alone. The RPS evaluates the sum of squared differences in CDF values between the forecast and the observation of the event:

$$\mathrm{RPS} = \frac{1}{n-1}\sum_{m=1}^{n-1}\left[\left(\sum_{k=1}^{m} f_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2. \quad (21)$$

The RPS can be seen as a summation of the binary Brier scores over all $n - 1$ thresholds defined by the category boundaries. The summation implies that the Brier scores for all thresholds are weighted equally. Whether the BS for all thresholds should be considered equally important depends on the users. It has been shown that the RPS is a strictly proper scoring rule in case the cost–loss ratio is uniformly distributed over the users (Murphy 1970). In that case the RPS is a linear function of the expected utility.

c. The ranked divergence score

Now an information–theoretical score is presented for ordinal category forecasts, which are defined in terms of cumulative probabilities. An equivalent to the RPS would be the ranked divergence score (RDS), averaging of the DS over all $n - 1$ category thresholds m:

$$\mathrm{RDS} = \frac{1}{n-1}\sum_{m=1}^{n-1}\mathrm{DS}_m, \quad (22)$$

where DS_m denotes the divergence score for the forecast of the binary event $j \le m$. This assumes equally important thresholds. The RDS, just like the DS, is dependent on the climatological uncertainty. To make the score comparable between forecasts, the RDS can be converted into a skill score. Now, two intuitive options exist to do the normalization. The first is to normalize the individual DS_m scores for each threshold m to a skill score for that threshold, like (14), using the climatic uncertainty for the binary event defined by that threshold:

$$\mathrm{DSS}_m = 1 - \frac{\mathrm{DS}_m}{\mathrm{UNC}_m}, \quad (23)$$

and then averaging the resulting skill score over all thresholds:

$$\mathrm{RDSS}_1 = \frac{1}{n-1}\sum_{m=1}^{n-1}\mathrm{DSS}_m. \quad (24)$$

This means that the relative contributions to the reduction of climatic uncertainty about each threshold m are considered equally important. In other words, all skills of forecasts about the exceedence of the $n - 1$ thresholds are equally weighted.

The second option for normalization is the first to sum the DS_m and then normalizing with the climatic score for the sum:

$$\mathrm{RDSS}_2 = 1 - \frac{\sum_{m=1}^{n-1}\mathrm{DS}_m}{\sum_{m=1}^{n-1}\mathrm{UNC}_m}. \quad (25)$$

The formulation of the RDSS_2 according to (25) does not normalize the scores for the different thresholds individually, but applies the same normalization to every DS_m. This means that the merits of the forecaster for all thresholds are implicitly weighted according to the inherent uncertainties in the climate. In this way, the forecast of extreme (nearly certain) events are hardly contributing to the total score, while they could in fact be most important for the users.

d. Relation to ranked mutual information

An alternative skill score defined in terms of cumulative probabilities is the RMIS_O in (15) as defined by Ahrens and Walser (2008). The version of the RMIS_O for multiple category forecasts can be written as

$$\mathrm{RMIS}_O = 1 - \frac{\sum_{m=1}^{n-1} I(\mathbf{f}_m, \mathbf{o}_m)}{\sum_{m=1}^{n-1} H(\bar{\mathbf{o}}_m)}, \quad (26)$$

where $\mathbf{f}_m$ denotes the series forecast probabilities of exceedence of threshold m, $\mathbf{o}_m$ denotes the corresponding

series of observations, and $\bar{o}_m$ is the average observed occurrence. For a perfectly reliable forecast, the RMIS_O is therefore equal to the RDSS_2 formulated in (25). For forecasts that are not well calibrated, the RMIS_O measures the amount of information that would be available after calibration, while the RDSS measures the information as presented by the forecaster. By using the decomposition presented in (14) it is possible to write

$$\mathrm{RDSS}_1 = \frac{1}{n-1}\sum_{m=1}^{n-1}\frac{\mathrm{RES}_m}{\mathrm{UNC}_m} - \frac{1}{n-1}\sum_{m=1}^{n-1}\frac{\mathrm{REL}_m}{\mathrm{UNC}_m}. \qquad (27)$$

By presenting both the resolution and the reliability terms of (27) separately, the potential information and the loss of information due to imperfect calibration are visible.

Apart from including the reliability or not, another question is how to weight the scores for the different thresholds, to come to one aggregated score. As every binary decision by some user with a certain cost–loss ratio can be associated with some threshold, the weighting reflects the importance of the forecast to the various users. No matter what aggregation method is chosen, there will always be an implicit assumption about the user's importance and stakes in a decision-making process. This is inherent to summarizing forecast performance in one score. A diagram that plots the two skill score components against the thresholds contains the relevant information characteristics for different users. In this way each user can look up the score on the individual threshold, that is relevant for his decision problem, and compare it with the performance of some other forecasting system on that threshold.

e. Information and useful information

Forecasting is concerned with the transfer of information about the true outcome of uncertain future events that are important to a given specific user. The information in the forecast should reduce the uncertainty about the true outcome. It is important to note the difference between two estimations of this uncertainty. First, there is the uncertainty a receiver of a forecast has about the truth, estimated in hindsight, knowing the observation. This uncertainty is measured by the divergence score. Second, there is the perceived uncertainty about the truth in the eyes of the user after adopting the forecast, which is measured by the entropy of the forecast. The first depends on the observation, while the second does not. We note that information–theoretical concepts measure information objectively, without considering its use. The usefulness of information is different for each specific user. The amount of useful information in a forecast can explicitly be subdivided into two elements:

1) reduction of uncertainty about the truth (the information–theory part); and
2) the usefulness of this uncertainty reduction (the user part).

The first element is only dependent on the user's probability estimate of the event's outcome before and after the forecast is received and on the true outcome. If the (subjective) probability distribution of the receiver does not change upon receiving the forecast, no information is transferred by it. If the probability distribution changed, but the divergence to the observation increased, the forecast increased the uncertainty about the truth as estimated from the observation, which is in itself an estimation of the unknown truth (although for the decomposition we assume it to be a perfect estimation). A forecast is less informative to a user already having a good forecast. To make the information–theoretical part of useful information in a forecast independent of the user, remaining uncertainty is estimated instead of its reduction.

The second element of useful information in a forecast, usefulness, is user and problem specific. A forecast is useful if it is about a relevant subject. Communicating the exceedence probability of a certain threshold that is not a threshold for action for a specific user does not help him much. Usefulness also depends on how much importance is attached to events. This can be, for example, the costs associated with a certain outcome–action combination, typically reflected in a so-called payoff matrix. Implicitly, also information–theoretical scores make some assumption on the usefulness of events. The assumption is that the user attaches his own importance to the events by placing repeated proportional bets, each time reinvesting his money. This is referred to as Kelly betting [for a more detailed explanation, see Kelly (1956) and Roulston and Smith (2002)]. In other words, the assumption is that the user maximizes the utility of his information in a fair game by strategically deciding on his importance or stakes.

The explicit consideration of usefulness of information brings up an interesting question about the roles of the forecaster and the user of forecasts. The divergence score measures the remaining uncertainty after adopting the forecast, which is completely independent of the user. This focuses the score on evaluating a main task of the forecaster, which is to give the best possible estimate of probabilities. It might also be argued, however, that a forecaster should not just reduce uncertainty, but also deliver utility for users' decisions. To be able to judge forecasts on that criterion, assumptions need to be made about the users and their decision problems. When scores based on these objectives are used to improve forecasting

procedures, maximizing these two objectives does not always lead to the same answer. In such cases an improvement of the utility of forecasts may coincide with a reduction in informativeness. More research is needed to investigate the nature of this trade-off, which is strongly related to model complexity, overfitting, and calibration versus validation.

5. An example: Rainfall forecasts in the Netherlands

As an illustration of the practical application of the divergence score and its decomposition, it was applied to a series of probabilistic rainfall forecasts and corresponding observations for the measurement location De Bilt, Netherlands. The forecast series consist of daily forecasts of the probability of a precipitation amount equal or larger than 0.3 mm, for 0–5 days ahead. They are derived using output of both the European Centre for Medium-Range Weather Forecasts (ECMWF) deterministic numerical weather prediction model and the ECMWF ensemble prediction model. Predictors from both models are used in a logistic regression to produce probability forecasts for precipitation. The data cover the period from December 2005 to November 2008, in which the average probability was 0.4613, leading to an initial uncertainty of 0.9957 bits.

Figure 5 confirms the expectation that both the Brier skill score and the divergence skill score show a decline with increasing lead time. It also shows that the forecasts possess skill over annual climatology up to a lead time of at least 5 days. The dashed lines show the potential skill that could be attainable after recalibration. The estimation of this potential skill, however, is dependent on the correct decomposition. The decompositions of both the Brier and the divergence score need enough data (large enough n_k) to calculate the conditional distributions $\mathbf{f}_k$ for all unique forecasts $k \in (1, K)$. To be able to calculate the contribution to reliability of all the 99% forecasts, for example, at least 100 of such forecasts are necessary to not surely overestimate reliability. Also for a larger number of forecasts, there is a bias toward overestimation of the reliability, which decreases with the amount of data available per conditional to be estimated.

A solution for estimating the components with limited data is rounding the forecast probabilities to a limited set of values. In this way, less conditional distributions $\mathbf{f}_k$ need to be estimated and more data per distribution are available.

Figure 6 shows that for these 3 yr of data, using finer-grained probabilities as forecasts leads to an increasing overestimation of reliability. The skill scores themselves are not sensitive to this overestimation because the lack of data causes a compensating overestimation of resolution.

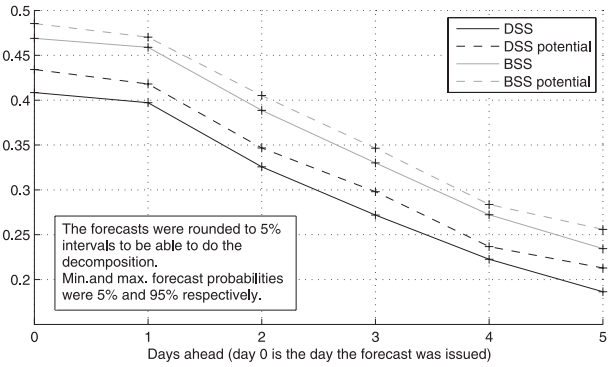

FIG. 5. The skill according to both scores decreases with lead time. The potential skills, which should be interpreted with caution, indicate the part of uncertainty that would be explained after recalibration.

The potential skill, however, should be interpreted with caution, as solving reliability issues by calibration on the basis of biased estimates of reliability does not lead to a real increase in skill. From the figure it can also be noted that too coarse-grained probabilities lead to a real loss of skill. In this case, giving the forecasts in 5% intervals seems the minimum needed to convey the full information that is in the raw forecast probabilities.

Figure 7 sheds more light on the relation between the Brier skill score and the divergence skill score, based on 5-day-ahead forecasts from a second dataset, which covers February 2008 to December 2009. For this set, the forecast probabilities ranged from 1% to 99%. The black dots indicate the scores that were attained for single forecast observation pairs. The dots show that the BSS and DSS have a monotonic relation as scoring rules. The limits of this relation are at (1, 1) for perfect forecasts and, in this case, $(-\infty, -3.095)$ for certain but wrong forecasts. The worst forecast was 98%, while no rain fell.

The total scores for different weeks of forecasts are plotted as gray dots. They are averages of sets of seven black dots. Because the relation of the single forecast scores is not a straight line, a scatter occurs in the relation of the weekly average scores, which is therefore no longer monotonic. The scatter implies that two series of forecasts can be ranked differently by the Brier score and the divergence score. In this example, the scatter is relatively small ($r^2 = 0.9938$) and will probably have no significant implications, but it would be larger if many overconfident forecasts were issued. An interesting example of differently ranked forecast series are the two weeks indicated by the triangles, where the scores disagree on which of the two weeks was better forecast than climate. The downward-pointing triangle marks the score for forecasts in week A, where performance according to the divergence score was worse than the climatological forecast (DSS = -0.0758), but according to the Brier score was slightly

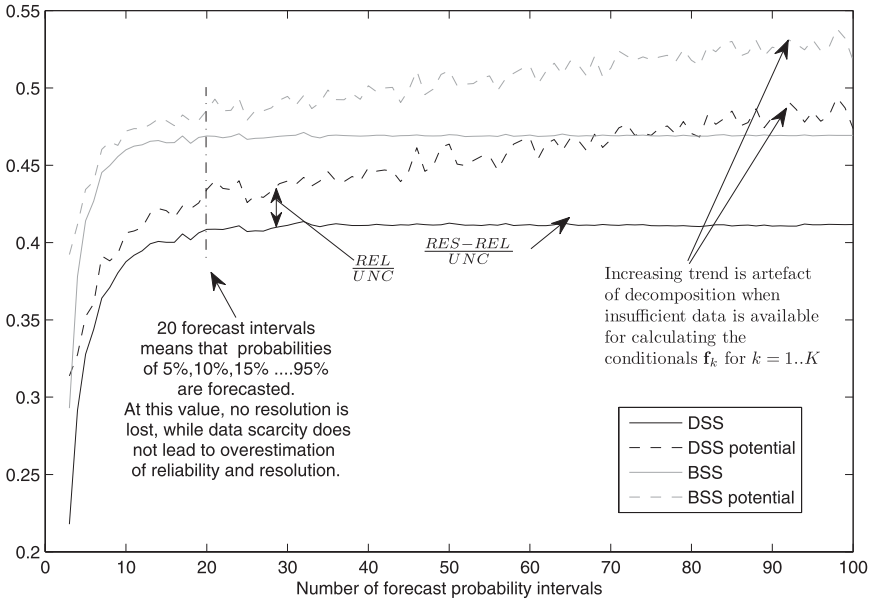

FIG. 6. The calculation of reliability is sensitive to the rounding of the forecasts. If not enough data are available, rounding is necessary to estimate the conditionals. Too coarse rounding causes an information loss.

better than climate (BSS = 0.0230). Conversely, the upward-pointing triangle marks week B, where the forecasts according to Brier were worse than climate (=−0.0355) but still contained more information than climate according to the DSS (=0.0066).

Given that the scatter in the practical example is small, the Brier score appears to be a reasonable approximation of the divergence score and is useful to get an idea about the quality of forecasts. More practical comparisons are needed to determine if the approximation can lead to significantly different results in practice. These are mostly expected in case extreme probabilities are forecast.

The severe penalty the divergence gives for errors in the extreme probabilities, which is sometimes seen as a weakness, should actually be viewed as one of its strengths. As the saying goes, "you have to be cruel to be kind." It is constructive to give an infinite penalty to a forecaster who issues a wrong forecast that was supposed to be certain. This is fair because the value that a user would be willing to risk when trusting such a forecast is also infinite.

6. Conclusions

Analogous to the Brier score, which measures the squared Euclidean distance between the distributions of observation and forecast, we formulated an information–theoretical verification score, measuring the Kullback–Leibler divergence between those distributions. More precisely, our score measures the divergence from the distribution of the event after the observation to the distribution that is the probability forecast. Our divergence score is a reinterpretation of the ignorance score or logarithmic score, which was previously not defined as a Kullback–Leibler divergence. Extending the analogy to the useful and well-known decomposition of the Brier score, the divergence score can be decomposed into uncertainty − resolution + reliability. For binary events, Brier score and its components are second-order approximations of the divergence score and its components.

The divergence score and its decomposition generalize to multicategory forecasts. A distinction can be made

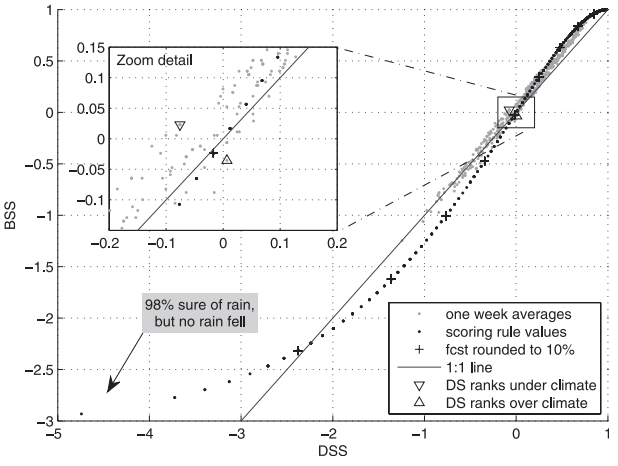

FIG. 7. Relation between the Brier skill score and the divergence skill score. For single forecasts they have a monotonic relation, but for averages of series of forecasts, a scatter in the relation can cause forecasts to be ranked differently by both scores.

between nominal and ordinal category forecasts. Scores based on the cumulative distribution over ordinal categories can be seen as combinations of binary scores on multiple thresholds. How the scores for all thresholds should be weighted relative to each other depends on the user of the forecast. Scores on cumulative distribution are therefore not exclusively dependent on physical observations, but contain subjective weights for the different thresholds. Two possible formulations of a ranked divergence skill score have been formulated. The first equally weighs the skill scores relative to climate, while the second equally weights the absolute scores. The second-ranked divergence skill score is equal to the existing ranked mutual information skill score for the case of perfectly calibrated forecasts, but additionally includes a reliability component, measuring miscalibration.

In forecasting, a distinction can be made between information and useful information in a forecast. The latter cannot be evaluated without a statement about context in which the forecast will be used. The first is only dependent on how the forecasts relate to the observations and is objective. Therefore, in the authors' opinion, information should be the measure for forecast quality. It can be measured using the logarithmic score, which now can be interpreted as the Kullback–Leibler divergence of the forecast from the observation. Useful information or forecast value, on the other hand, is a different aspect of forecast ''goodness'' (Murphy 1993), which should be evaluated while explicitly considering the decision problems of the users of the forecast.

The Brier score can be used as an approximation for quality or as an exact measure of value under the assumption of a group of users with uniformly distributed cost–loss ratios. In our opinion, these two applications should be clearly separated. In case one wants to assess quality, information–theoretical scores should be preferred. If an approximation is sufficient, the Brier score could still be used, with the advantage that it is well understood and extensive experience exists with the use of it. However, when extreme probabilities have to be forecast, the differences might become significant and the divergence score is to be preferred on theoretical grounds.

In case value is to be measured, an inventory of the users of the forecasts should be made to asses the total utility. When explicitly investigating the user base, a better estimator for utility than the Brier score can probably be defined. Using the Brier score as a surrogate for forecast value, implicitly assuming the emergent utility function is appropriate for a specific type of forecasts, is clearly unsatisfactory. In this respect, it is important to also stress that the divergence score does not measure value, but quality. Only in a very unrealistic case (a bookmaker offering fair odds) does a clear relation exist between the two. It might be argued that for practitioners in meteorology, quality is most likely of more concern than value, because the latter is in fact evaluating decisions rather than forecasts. (To facilitate the use of the score and its decomposition, scripts that can be used in Matlab and Octave are available online at http://divergence.wrm.tudelft.nl.)

Acknowledgments. The authors thank the Royal Netherlands Meteorological Institute (KNMI) and Meteo Consult for kindly providing the forecast and observation data that was used for this research. We also thank both anonymous reviewers for their constructive comments.

APPENDIX

The Decomposition of the Divergence Score

First we use the definition of the Kullback–Leibler divergence to define the total score and the resolution and reliability components and the entropy for the uncertainty component:

$$D_{\mathrm{KL}}(\mathbf{v} \,\|\, \mathbf{w}) = \sum_{i=1}^{n} v_i \log \frac{v_i}{w_i},$$

$$\mathrm{DS} = \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t \| \mathbf{f}_t),$$

$$\mathrm{REL} = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{f}}_k),$$

$$\mathrm{RES} = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{o}}),$$

$$\mathrm{UNC} = \frac{1}{N}\sum_{t=1}^{N} H(\bar{\mathbf{o}}) = -\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n}\{[\bar{\mathbf{o}}]_i \log[\bar{\mathbf{o}}]_i\}.$$

Now we can simplify the expression for

$$\begin{aligned}\mathrm{REL} - \mathrm{RES} &= \sum_{k=1}^{K} n_k \{D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{f}}_k) - D_{\mathrm{KL}}(\bar{\mathbf{o}}_k \| \bar{\mathbf{o}})\} \\ &= \sum_{k=1}^{K} n_k \sum_{i=1}^{n} [\bar{\mathbf{o}}_k]_i \left\{ \log \frac{[\bar{\mathbf{o}}_k]_i}{[\bar{\mathbf{f}}_k]_i} - \log \frac{[\bar{\mathbf{o}}_k]_i}{[\bar{\mathbf{o}}]_i} \right\} \\ &= \sum_{k=1}^{K} n_k \sum_{i=1}^{n} [\bar{\mathbf{o}}_k]_i \left\{ \log \frac{[\bar{\mathbf{o}}]_i}{[\bar{\mathbf{f}}_k]_i} \right\}.\end{aligned}$$

Note that

$$\mathrm{DS} = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n} [\mathbf{o}_t]_i \left\{ \log \frac{[\mathbf{o}_t]_i}{[\mathbf{f}_t]_i} \right\},$$

where K is equal to the number of different $\mathbf{f}_t$ and one bin for each $\mathbf{f}_t$. We can label both bins and outcomes by k. We label the outcomes in a bin by

$$[\mathbf{o}_t]_{k,m_k},$$

where $m_k = 1 \ldots n_k$ so

$$\mathrm{DS} = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{m_k=1}^{n_k}[\mathbf{o}_{k,m_k}]_i\left\{\log\frac{[\mathbf{o}_{k,m_k}]_i}{[\bar{\mathbf{f}}_k]_i}\right\},$$

which can be written as

$$\begin{aligned}\mathrm{DS} &= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\sum_{m_k=1}^{n_k}[\mathbf{o}_{k,m_k}]_i\left\{\log\frac{[\mathbf{o}_{k,m_k}]_i}{[\mathbf{f}_t]_i} - \log\frac{[\mathbf{o}_{k,m_k}]_i}{[\overline{\mathbf{o}}]_i} + \log\frac{[\mathbf{o}_{k,m_k}]_i}{[\overline{\mathbf{o}}]_i}\right\}\right)\\ &= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\sum_{m_k=1}^{n_k}[\mathbf{o}_{k,m_k}]_i\left\{\log\frac{[\overline{\mathbf{o}}]_i}{[\mathbf{f}_t]_i} + \log\frac{[\mathbf{o}_{k,m_k}]_i}{[\overline{\mathbf{o}}]_i}\right\}\right)\\ &= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(n_k[\overline{\mathbf{o}}_k]_i\left\{\log\frac{[\overline{\mathbf{o}}]_i}{[\bar{\mathbf{f}}_k]_i}\right\}\right) + \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\sum_{m_k=1}^{n_k}[\mathbf{o}_{k,m_k}]_i\left\{\log\frac{[\mathbf{o}_{k,m_k}]_i}{[\overline{\mathbf{o}}]_i}\right\}\right).\end{aligned}$$

We can now recognize the first term as REL − RES, so

$$\begin{aligned}&\mathrm{DS} - (\mathrm{REL} - \mathrm{RES})\\ &= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\sum_{m_k=1}^{n_k}[\mathbf{o}_{k,m_k}]_i\left\{\log\frac{[\mathbf{o}_{k,m_k}]_i}{[\overline{\mathbf{o}}]_i}\right\}\right)\\ &= \frac{1}{N}\sum_{t=1}^{N}\left(\sum_{i=1}^{n}[\mathbf{o}_t]_i\left\{\log\frac{[\mathbf{o}_t]_i}{[\overline{\mathbf{o}}]_i}\right\}\right) = \frac{1}{N}\sum_{t=1}^{N}D_{\mathrm{KL}}(\mathbf{o}_t\|\overline{\mathbf{o}}).\end{aligned}$$

Note that, with

$$\lim_{x\to 0} x\,\log x = 0$$

and for $n = 2$, $\mathbf{o}_t \in \{(1, 0)^{\mathrm{T}}, (0, 1)^{\mathrm{T}}\}$, we find

$$\begin{aligned}\sum_{i=1}^{n}[\mathbf{o}_t]_i\left\{\log\frac{[\mathbf{o}_t]_i}{[\overline{\mathbf{o}}]_i}\right\} &= \sum_{i=1}^{n}\{[\mathbf{o}_t]_i\log[\mathbf{o}_t]_i - [\mathbf{o}_t]_i\log[\overline{\mathbf{o}}]_i\}\\ &= -\sum_{i=1}^{n}[\mathbf{o}_t]_i\log[\overline{\mathbf{o}}]_i,\end{aligned}$$

so

$$\sum_{t=1}^{N}\left(\sum_{i=1}^{n}[\mathbf{o}_t]_i\left\{\log\frac{[\mathbf{o}_t]_i}{[\overline{\mathbf{o}}]_i}\right\}\right) = N\sum_{i=1}^{n}[\overline{\mathbf{o}}]_i\log[\overline{\mathbf{o}}]_i,$$

so

$$\mathrm{DS} - (\mathrm{REL} - \mathrm{RES}) = -\sum_{i=1}^{n}[\overline{\mathbf{o}}]_i\log[\overline{\mathbf{o}}]_i = H(\overline{\mathbf{o}}) = \mathrm{UNC}.$$

REFERENCES

Ahrens, B., and A. Walser, 2008: Information-based skill scores for probabilistic forecasts. *Mon. Wea. Rev.,* **136,** 352–363.

Benedetti, R., 2009: Scoring rules for forecast verification. *Mon. Wea. Rev.,* **138,** 203–211.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Cover, T., and J. Thomas, 2006: *Elements of Information Theory*. 2nd ed. Wiley-Interscience, 776 pp.

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

Good, I., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **14B,** 107–114.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 254 pp.

——, and ——, 2008: Proper scores for probability forecasts can never be equitable. *Mon. Wea. Rev.,* **136,** 1505–1510.

Kelly, J., Jr., 1956: A new interpretation of information rate. *IRE Trans. Info. Theory,* **2,** 185–189.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.,* **59,** 2057–2072.

Laio, F., and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.,* **11,** 1267–1277.

Leung, L., and G. North, 1990: Information theory and climate prediction. *J. Climate,* **3,** 5–14.

Mason, S., 2008: Understanding forecast verification statistics. *Meteor. Appl.,* **15,** 31–40.

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

——, 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Stephenson, D. B., C. A. S. Coelho, and I. T. Jolliffe, 2008: Two extra components in the brier score decomposition. *Wea. Forecasting,* **23,** 752–757.

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.