Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth

STEVEN V. WEIJS AND NICK VAN DE GIESEN

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

(Manuscript received 30 July 2010, in final form 18 February 2011)

ABSTRACT

Recently, an information-theoretical decomposition of Kullback–Leibler divergence into uncertainty, reliability, and resolution was introduced. In this article, this decomposition is generalized to the case where the observation is uncertain. Along with a modified decomposition of the divergence score, a second measure, the cross-entropy score, is presented, which measures the estimated information loss with respect to the truth instead of relative to the uncertain observations. The difference between the two scores is equal to the average observational uncertainty and vanishes when observations are assumed to be perfect. Not acknowledging for observation uncertainty can lead to both overestimation and underestimation of forecast skill, depending on the nature of the noise process.

1. Introduction

Information theory became a field of research with the mathematical theory of communication of Shannon (1948). Next to applications in data communication, there was a strong impact on computer science, ranging from data compression and cryptography to error correcting codes [e.g., the fact that a scratch on a compact disk (CD) is inaudible]. Other fields affected by information theory are statistics, gambling (Kelly 1956), and financial portfolio theory. Cover and Thomas (2006) give an extensive introduction into information theory and its applications. The basis of the theory is a measure of uncertainty represented by a probability distribution named entropy because of the mathematical similarity with the formulation of entropy in statistical thermodynamics.[1] The measure uniquely follows from a set of elementary desiderata for a measure of uncertainty to be useful (see Shannon 1948). Apart from entropy, information theory also defines measures like relative entropy (also Kullback–Leibler divergence), cross entropy, and mutual information. The theory defines information as the reduction in uncertainty. Various inequalities form certain ''laws of information,'' such as the data processing inequality, which says that we cannot increase the amount of information in data by processing it. In other words, we cannot produce information out of thin air, except when a meteorologist observes the atmosphere with measurement equipment and generates data. Although these measurements give some information, they also leave some uncertainty about the variables of interest, which is the topic of this paper.

Additive relations between various measures of information and uncertainty provide a useful basis for a framework for forecast verification. In Weijs et al. (2010a), it was argued that the Kullback–Leibler divergence from the observation to the forecast is a measure for forecast quality with a number of desirable properties (see also Benedetti 2010; Weijs et al. 2010b). The view was presented that forecasting can be seen as a communication problem, in which information is given by the forecaster to reduce the uncertainty of the user. The quality of such forecasts can therefore be evaluated using the information-theoretical measures relating to information and uncertainty. Examples of such scores are the average information gain from the climatological distribution to the forecast (Peirolo 2010) and the ignorance score (Roulston and Smith 2002). In Weijs et al. (2010a), it was found that the remaining uncertainty relative to the observations, as

[1] Note that we interpret probability as a carrier of incomplete information, which need not be associated with a stochastic process (see Jaynes 2003).

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: 10.1175/2011MWR3573.1

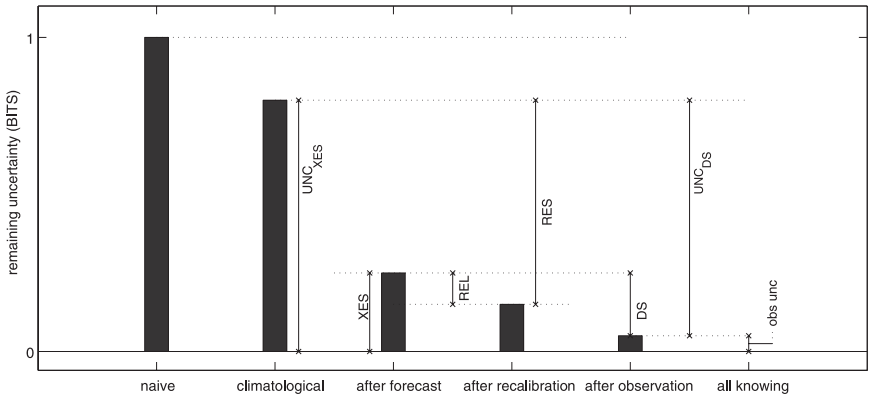

FIG. 1. The relations between the components and the scores presented in this article are additive. The bars give the average remaining uncertainty about the truth (measured in bits) for various (hypothetical) stages in the forecasting process. The naive forecast is always assigning 50% probability of precipitation (complete absence of information), the climatological forecast takes into account observed frequencies. This climatological uncertainty $\mathrm{UNC}_{\mathrm{XES}}$ can be reduced to XES by believing the forecasts **f**. If these are not completely reliable, the uncertainty can be further reduced with REL by recalibration. After observation, there is still some uncertainty (obs unc) about the hypothetical true outcome, given that observations are not perfect. Only for an all knowing observer, the uncertainty is reduced to 0. The resolution RES is the information that could maximally be extracted from the forecasts by perfect calibration. The divergence score DS and the new uncertainty component $\mathrm{UNC}_{\mathrm{DS}}$ measure the uncertainty after forecast and in the climate, relative to the observations.

measured by this Kullback–Leibler divergence from the observations to the forecasts, is an appropriate scoring rule, which we refer to as the divergence score.

Moreover, a decomposition was presented, analogous to the decomposition of the Brier score (Brier 1950; Murphy 1973), which decomposes the Brier score into (climatological) ''uncertainty'' minus ''resolution'' plus ''reliability'' (actually unreliability). The decomposition in Weijs et al. (2010a) is equal to the logarithmic case of a general decomposition for the proper scoring rules presented in Bröcker (2009). Note that in the latter, the asymmetric divergence measure was defined with a reverse order of the arguments compared to the Kullback–Leibler divergence common in information theory. A possible interpretation of the information-theoretical decomposition in Weijs et al. (2010a) is that ''the remaining uncertainty is equal to the missing information minus the correct information plus the wrong information'' (see Fig. 1). A forecast that is reliable but has no perfect resolution does not give complete information, but the information it does give is correct. In the decompositions of both the Brier score (BS) and the divergence score (DS), it was assumed that the observations are certain and correspond to the truth about the forecast variable.

In reality, no observation can be assumed to correspond to the true outcome with certainty. For example, in the evaluation of binary probabilistic precipitation forecasts of the Royal Netherlands Meteorological Office (KNMI), the observation that corresponds to ''no precipitation'' is defined as an observed precipitation of less than 0.3 mm on a given day. Given the observational errors in the exact precipitation amount, measured values close to the threshold would be best represented by a probabilistic observation, accounting for the uncertainties (see Fig. 2). Briggs et al. (2005) also noted that uncertainty in the observation must be taken into account to assess the true skill of forecasts. This requires either ''gold standard'' observations or subjective estimates of the observation errors. Bröcker and Smith (2007) proposed to use a noise model for the observation to transform the forecast, using this to define a generalized score.

We propose to define an ''uncertain observation'' as the retrospective conditional distribution of the true outcome of event or quantity that was forecast, given the reading on one or more measurement instruments. For example, when the spatial scale or location of the measurements and the forecasts differs, the distribution can be based on spatial statistics of various instruments. In another case, the distribution may be derived from a model of the observational noise (e.g., due to wind around a rain gauge, noise in the electronics, etc.). Note that the correctness and reliability of such uncertainty models cannot be verified, because the ''true'' value cannot be observed directly. Although the term verification suggests a comparison between forecasts and truth (Latin: veritas = truth), both the divergence and the Brier score are actually comparing the forecasts with observations, which are an estimate of the unknown

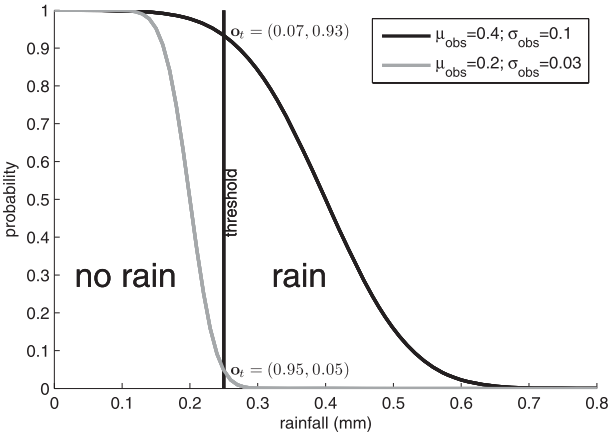

FIG. 2. The measurement uncertainty in the precipitation measurement leads to a probabilistic binary observation. In the example, a simple Gaussian measurement uncertainty is assumed. The measurement distributions are centered around the measurements and have a constant standard deviation.

truth about the forecast variable. An uncertain observation acknowledges this by representing the uncertainty explicitly by a probabilistic best estimate after the event has taken place.

When the uncertainty in the observations is accounted for by representing them with probability distributions with nonzero entropy (i.e., the observation assigns probability to more than one outcome), the decomposition that was presented in Weijs et al. (2010a) does not hold. The divergence score as a whole, however, is still a useful measure of correspondence between forecasts and observations. It would therefore be interesting to define a meaningful decomposition of the divergence score that is applicable in the case of uncertain observations. A second point is whether the quality of forecasts should be measured with respect to the known probabilistic observations or estimated with respect to the unknown truth.

In this article, we present a modification to the decomposition presented in Weijs et al. (2010a), which generalizes the case of uncertain observations. We discuss the interpretation of the new decomposition in terms of uncertainty and information. We furthermore present a second, related measure for forecast quality, in information theory often referred to as cross entropy, which in this case estimates the uncertainty relative to the truth instead of relative to the observation. A decomposition for this score is also presented. The scores are applied to a real dataset for illustration.

2. The concepts of surprise, uncertainty, and relative uncertainty

In information theory, uncertainty is related to surprise. Surprise is defined as minus the log of the prior probability assigned to the true outcome, $S = -\log P$. This forms the basis for the measure of uncertainty associated with a discrete probability distribution, entropy H, which is the expected surprise upon hearing the truth. For example, the uncertainty associated with a binary probabilistic forecast of 70% chance of precipitation, $\mathbf{f} = (0.3, 0.7)^{\mathrm{T}}$, is

$$H(\mathbf{f}) = E_{\mathbf{f}}\{S_{\mathbf{f}}\} = -\sum_{i=1}^{n} [\mathbf{f}]_i \log[\mathbf{f}]_i = 0.88 \text{ bits}, \qquad (1)$$

where $H(\mathbf{f})$ is the entropy of probability mass function (PMF) $\mathbf{f}$, calculated in the unit "bits" because the logarithm is taken to the base 2 (throughout this article). Here $E_{\mathbf{f}}$ denotes the expectation operator with respect to $\mathbf{f}$, and n is the number of categories in which the outcome can fall, in this case 2. The notation $[\mathbf{f}]_i$ will be used throughout this paper to denote the element i of vector $\mathbf{f}$.

Next to this entropy measure, introduced by Shannon (1948), there is also a definition of relative entropy, or Kullback–Leibler divergence D_{KL} (Kullback and Leibler 1951). This is a measure for the expected amount of additional surprise a person is expected to experience, compared to another person having a more accurate and reliable probability estimate. Therefore, it is a relative uncertainty. The divergence score is based on this idea and measures the expected extra surprise that a person having the forecast $\mathbf{f}_t$ for an instance t will experience, compared to a person knowing the observation $\mathbf{o}_t$, from the perspective of the latter:

$$\mathrm{DS}_t = D_{\mathrm{KL}}(\mathbf{o}_t || \mathbf{f}_t) = E_{\mathbf{o}_t}\{S_{\mathbf{f}_t} - S_{\mathbf{o}_t}\} = \sum_{i=1}^{n} [\mathbf{o}_t]_i \log\left[\frac{\mathbf{o}_t}{\mathbf{f}_t}\right]_i. \qquad (2)$$

3. Decomposition of the divergence score for uncertain observations

The decomposition of the divergence score DS for a series of N forecasts that was presented in Weijs et al. (2010a) was analogous to the decomposition of the Brier score by Murphy (1973) into uncertainty, resolution, and reliability [cf. Eqs. (3) and (6)]:

$$\begin{aligned}\mathrm{DS} &= \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t||\mathbf{f}_t) = \mathrm{REL}_{\mathrm{DS}} - \mathrm{RES}_{\mathrm{DS}} + \mathrm{UNC}_{\mathrm{DS}} \\ &= \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\overline{\mathbf{o}}_k||\mathbf{f}_k) \\ &\quad - \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\overline{\mathbf{o}}_k||\overline{\mathbf{o}}) + H(\overline{\mathbf{o}}), \end{aligned} \qquad (3)$$

in which N is the total number of forecasts issued, K is the number of unique forecasts issued, $\mathbf{o}_t$ is the observed outcome at time t, $\overline{\mathbf{o}} = \sum_{t=1}^{N} \mathbf{o}_t/N$ is the observed climatological frequency of the event, n_k is the number of forecasts with the same forecast probability, and $\overline{\mathbf{o}}_k$ is the observed frequency given forecasts of probability $\mathbf{f}_k$.

The uncertain observation $\mathbf{o}_t$ is the probability mass function (PMF) of the true outcome of the uncertain event that is forecast, given the available information after it occurred. Note that when perfect data assimilation is performed, this also includes all information from $\mathbf{f}_t$. Because measurements are usually indirect, we can regard the observation as a (usually subjective) conditional distribution of the true outcome, given the information from the measurement equipment.

The decomposition as formulated in Eq. (3) relies on the assumption that observations are certain [i.e., $\mathbf{o}_t = (0, 1)^{\mathrm{T}}$ or $\mathbf{o}_t = (1, 0)^{\mathrm{T}}$]. In the appendix of Weijs et al. (2010a) this assumption was used to rewrite the closing term of the decomposition $1/N\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t||\overline{\mathbf{o}})$ as the uncertainty component $H(\overline{\mathbf{o}})$. When we want the decomposition to be valid for uncertain observations, the last step of the derivation in Weijs et al. (2010a) can be omitted. We thus replace the uncertainty component, the last term in Eq. (3), by the average Kullback–Leibler divergence from the uncertain observations to the average observation (i.e., the observed climatic distribution), the last term in Eq. (4):

$$\mathrm{DS} = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\overline{\mathbf{o}}_k||\mathbf{f}_k) - \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\overline{\mathbf{o}}_k||\overline{\mathbf{o}}) + \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}(\mathbf{o}_t||\overline{\mathbf{o}}). \tag{4}$$

This represents the expected climatological uncertainty relative to the observation, which is depicted in Fig. 1 as $\mathrm{UNC}_{\mathrm{DS}}$. By writing the uncertainty term of the divergence score decomposition in this way, it remains valid for uncertain observations. The original uncertainty term, the entropy $H(\overline{\mathbf{o}})$, can be seen as representing the estimated climatological uncertainty relative to the truth, which we from now on will denote as $\mathrm{UNC}_{\mathrm{XES}}$, because it is part of a decomposition of XES we will introduce shortly.

Likewise, DS represents the average remaining uncertainty relative to the observations, which in the case of uncertain observations can become different from the estimated remaining uncertainty relative to the truth. This latter measure, XES, will be introduced in section 4.

Analogy for the Brier score decomposition

Analogously to the new decomposition of the DS, the Brier score decomposition introduced by Murphy (1973) can be modified in a similar manner to remain valid for uncertain observations. This can be achieved by replacing the uncertainty term in the original decomposition by the average squared Euclidean distance from the observations to the average observation. The original decomposition and the modified one are shown in Eqs. (6) and (7), respectively. For perfect observations, Eqs. (6) and (7) are the same. When observational uncertainty is considered, Eq. (6) does not hold, but Eq. (8) does. For ease of notation, we write $(\mathbf{f}_t - \mathbf{o}_t)^2$ for $(\mathbf{f}_t - \mathbf{o}_t)^{\mathrm{T}}(\mathbf{f}_t - \mathbf{o}_t)$:

$$\mathrm{BS} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{f}_t - \mathbf{o}_t)^2 = \mathrm{REL}_{\mathrm{BS}} - \mathrm{RES}_{\mathrm{BS}} + \mathrm{UNC}_{\mathrm{BS}}, \tag{5}$$

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

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f}_k - \overline{\mathbf{o}}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{\mathbf{o}}_k - \overline{\mathbf{o}})^2 + \frac{1}{N}\sum_{t=1}^{N}(\mathbf{o}_t - \overline{\mathbf{o}})^2. \tag{7}$$

4. Expected remaining uncertainty about the truth: The cross-entropy score

We now present the cross-entropy score XES. The expected uncertainty relative to the unknown truth can be expressed by taking the expectation, with respect to the PMF that represents the uncertain observation, of the Kullback–Leibler divergence from the hypothetical truth to the forecast distribution:

$$\mathrm{XES} = \frac{1}{N}\sum_{t=1}^{N} E_{\mathbf{o}_t} D_{\mathrm{KL}}(\mathbf{v}_t||\mathbf{f}_t) = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{[\mathbf{o}_t]_j[\mathbf{v}_t]_i \log\frac{[\mathbf{v}_t]_i}{[\mathbf{f}_t]_i}\right\}. \tag{8}$$

In which $n = 2$ is the number of categories in which the event can fall. Here $\mathbf{v}_t$ denotes the hypothetical distribution of the truth at instance t, which, like a perfect observation, is either $(1, 0)^{\mathrm{T}}$ if the event in fact did not occur or $(0, 1)^{\mathrm{T}}$ if the event truly occurred. The $E_{\mathbf{o}_t}$ is the expectation operator with respect to the probability distribution $\mathbf{o}_t$. In this case, the Kullback–Leibler divergence $D_{\mathrm{KL}}(\mathbf{v}_t||\mathbf{f}_t)$ reduces to the logarithmic score (Good 1952),

which is also known as the ignorance score (Roulston and Smith 2002). These scores are simply minus the logarithm of the probability attached to the event that truly occurred:

$$D_{\mathrm{KL}}(\mathbf{v}_t||\mathbf{f}_t) = -\log[\mathbf{f}_t]_{k(t)}, \tag{9}$$

where $k(t)$ is the category in which the true outcome of the event fell at instance t. Because $\mathbf{o}_t$ is the best estimate of the unknown true outcome, we can use the expectation $E_{\mathbf{o}_t} D_{\mathrm{KL}}(\mathbf{v}_t||\mathbf{f}_t)$ to evaluate the forecast, which can also be written as the right-hand expression in Eq. (10). In information theory, this expression is often defined as the cross entropy between $\mathbf{o}_t$ and $\mathbf{f}_t$, hence we refer to it as cross-entropy score XES_t:

$$\mathrm{XES}_t = E_{\mathbf{o}_t} D_{\mathrm{KL}}(\mathbf{v}_t||\mathbf{f}_t) = -\sum_{i=1}^{n} [\mathbf{o}_t]_i \log[\mathbf{f}_t]_i. \tag{10}$$

This measure can be interpreted as the expected remaining uncertainty relative to the truth, for a single forecast $\mathbf{f}_t$ that is evaluated in the light of the observation $\mathbf{o}_t$. In the following interpretations, it is reasonable to assume that this observation is a reliable probability estimate of the truth, because there is no way to establish its unreliability without having access to the truth. The difference with the divergence score becomes clear from Fig. 1. For a series of forecasts, the cross-entropy score is defined as $\mathrm{XES} = 1/N \sum_{t=1}^{N} \mathrm{XES}_t$.

Decomposition of cross entropy

From the figure, we can see that the relation between all the components allows for several decompositions. The relation between DS and XES can be written as

$$\mathrm{XES} = \mathrm{DS} + \frac{1}{N}\sum_{t=1}^{N} H(\mathbf{o}_t). \tag{11}$$

The estimated remaining uncertainty in the forecasts relative to the truth (XES) is equal to the average uncertainty relative to the observations (DS) plus the average uncertainty that the observations represent, relative to the estimated truth [the second term on the right-hand side of Eq. (11)].

Another natural decomposition for the XES is the original decomposition of DS for perfect observations as presented in Weijs et al. (2010a). For uncertain observations, the three components presented there add up to the XES instead of to the DS (see also Fig. 1). The decomposition of the cross-entropy score XES therefore reads as

$$\mathrm{XES} = \mathrm{REL}_{\mathrm{XES}} - \mathrm{RES}_{\mathrm{XES}} + \mathrm{UNC}_{\mathrm{XES}}, \tag{12}$$

$$-\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{n} [\mathbf{o}_t]_i \log[\mathbf{f}_t]_i = \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k||\mathbf{f}_k) - \frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{\mathbf{o}}_k||\bar{\mathbf{o}}) + H(\bar{\mathbf{o}}). \tag{13}$$

Note that the resolution and reliability components are equal to those of the DS decomposition in Eq. (4).

5. Example application

As an illustration of the new term in the decomposition, the scores were calculated for a real dataset of binary probabilistic rainfall forecasts of the Royal Netherlands Meteorological Office (KNMI). The observed rainfall amounts were transformed into probabilistic uncertain observations using a very simple uncertainty model. The purpose of this exercise is merely to illustrate the concepts in this article. The forecasts that are evaluated are the forecast probabilities of a daily precipitation of 0.3 mm or more. This is the same dataset that was used in Weijs et al. (2010a). In that paper the rainfall amounts x_t, which were given with a precision of 0.1 mm, were converted to binary observations with a simple threshold filter: if $x_t \geq 0.3 \rightarrow \mathbf{o}_t = (0, 1)$, if $x < 0.3 \rightarrow \mathbf{o}_t = (1, 0)$. In this article, we assume random measurement error to make $\mathbf{o}_t$ probabilistic and account for the uncertainty in the observation.

The model of the uncertainty in the observation is Gaussian. The observed rainfall amount becomes a random variable with a normal probability density function:

$$g_{\mathrm{obs}}(x) = N(\mu_{\mathrm{obs}}, \sigma_{\mathrm{obs}}),$$

with mean μ_{obs} and standard deviation σ_{obs}. Because in this case we deal with a binary predictand, the probability distribution function (pdf) of the observation can be converted to a binary probability mass function $\mathbf{o}_t = (1 - o_t, o_t)$ by using

$$o_t = 1 - G_{\mathrm{obs}}(T) = 1 - \int_{-\infty}^{T} g_{\mathrm{obs}}(x)\, dx,$$

in which $G_{\mathrm{obs}}(T)$ is the cumulative distribution function of the observation, evaluated at threshold T. This conversion is illustrated in Fig. 2.

The decompositions of the DS and XES scores for the entire dataset were calculated for a range of different standard deviations σ_{obs} for the measurement uncertainty. In Fig. 3 it can be seen that while the average observation uncertainty grows, the divergence score improves (decreases), but the cross-entropy score XES deteriorates.

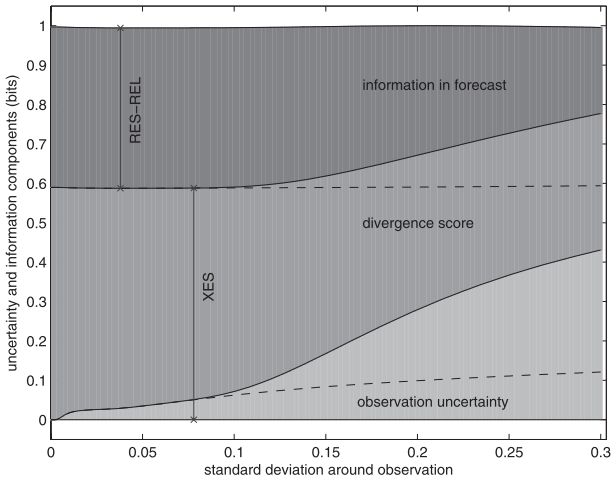

FIG. 3. The resulting decompositions (reliability not shown) as a function of standard deviation measurement. The growing XES with standard deviation, indicates that for the homoscedastic Gaussian observation uncertainty model, forecast quality is lower than would be estimated assuming perfect observations. The dashed lines show the decomposition for the same observation uncertainty model, with the exception that measurements of 0 mm are assumed to be certain. The almost constant XES in that case indicates that the estimation of forecast quality is robust against those observation uncertainties.

This indicates that if the observation uncertainty is reflected by this model, the best estimate for the information loss compared to the truth is higher than would be assumed neglecting observation uncertainty. Not taking the observation uncertainty in account would lead to an overestimation of the forecast quality in this specific case. A closer analysis reveals that most of the deterioration is caused by observations of 0 mm, which start to give significant probability to rain when the standard deviation grows beyond 0.1 mm. As many of the 0-mm observations are during cloudless days and in fact almost certain, we might reconsider the simple Gaussian uncertainty model.

When the uncertainty model is changed to have no uncertainty for 0-mm observations, the decomposition changes significantly (see dashed lines in Fig. 3), leading to a cross-entropy score that is almost constant (sometimes even slightly decreasing) with increasing standard deviation. For this particular error model, the uncertainty in the observations thus hardly affects the estimation of the forecast quality. This gives us confidence that as long as the 0-mm observations are certain, the estimate of forecast quality is robust against Gaussian observation errors with standard deviation up to 0.3 mm. Although in that case there is significant observation uncertainty (bottom dashed line, Fig. 3) that lowers the divergence score, the changes in the cross-entropy score for individual forecasts cancel each other out. Not surprisingly, the robustness of forecast quality estimates depends very much on the characteristics of the observation uncertainty. Further experiments are necessary to determine how to formulate realistic observation uncertainty models and how this can benefit verification practice.

6. Discussion and conclusions

a. Discussion: Divergence versus cross entropy

For the divergence score DS, worse observations lead to better scores for forecast quality, because the quality is evaluated relative to the observations. This might be considered undesirable, especially when the performances for two locations with similar climates are compared, while the quality of the observations is not the same. On the other hand, the divergence score has the advantage of not making explicit reference to a truth beyond the observations, which might be philosophically more appealing.

If the cross-entropy score XES is used as a scoring rule, the score estimates the quality of the forecasts at reducing uncertainty about the truth. This quality may be estimated differently in the light of observation uncertainty, but should not be relative to it. The skill might both be overestimated and underestimated in the presence of observation uncertainty. This depends on the nature of the errors, which should be modeled to the best possible extent. The XES allows a better comparison between the quality of different forecasts. In other words, the benchmark to compare the forecasts to is the truth. Because the uncertainty of the forecasts relative to this benchmark can only be evaluated if we would know the truth, we can only estimate its expected value. In contrast, in the divergence score DS, the benchmark to which the forecasts are compared are the probabilistic estimations of the truth. The remaining uncertainty with respect to these estimates, the observations, can be calculated exactly. Summarizing, the divergence score is the exact divergence from an estimate of the truth (the observation), while the cross-entropy score is an estimated (expected) divergence from the exact truth.

b. Conclusions

When extending the use of the divergence score to the case of uncertain observations, the cross-entropy score is a more intuitive measure for intercomparison of forecasts at locations with different observational uncertainty. The divergence score can be interpreted as a measure for the remaining uncertainty relative to the observation. The cross-entropy score can be seen as the expected remaining uncertainty with respect to a hypothetical true outcome. Both scores can be decomposed

into uncertainty, resolution and reliability. The difference in the decompositions is in the uncertainty component. For the case of the cross-entropy score, it represents the climatic uncertainty relative to the truth, for the case of the divergence score it represents the climatic uncertainty relative to the observation. The difference between the two uncertainty components is equal to the difference between the cross-entropy and divergence scores, and corresponds to the average observational uncertainty. If the observations are assumed perfect, which is usually the case in verification practice, both scores and decompositions are equal.

New Matlab/Octave scripts for the decompositions, calculating all information-theoretical terms presented here, can be freely downloaded (see online at http://divergence.wrm.tudelft.nl).

Acknowledgments. The authors thank the Royal Netherlands Meteorological Office (KNMI) and Meteo Consult for kindly providing the forecast and observation data that was used for this research. We also thank the three anonymous reviewers for their thoughtful comments and interesting philosophical discussions.

REFERENCES

Benedetti, R., 2010: 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.

Briggs, W., M. Pocernich, and D. Ruppert, 2005: Incorporating misclassification error in skill assessment. *Mon. Wea. Rev.,* **133,** 3382–3392.

Bröcker, J., 2009: Reliability, sufficiency, and the decomposition of proper scores. *Quart. J. Roy. Meteor. Soc.,* **135** (643), 1512–1519.

——, and L. Smith, 2007: Scoring probabilistic forecasts: The importance of being proper. *Wea. Forecasting,* **22,** 382–388.

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

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

Jaynes, E. T., 2003: *Probability Theory: The Logic of Science*. Cambridge University Press, 758 pp.

Kelly, J., 1956: A new interpretation of information rate. *IEEE Trans. Info. Theory,* **2** (3), 185–189.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann. Math. Stat.,* **22,** 79–86.

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

Peirolo, R., 2010: Information gain as a score for probabilistic forecasts. *Meteor. Appl.,* **18,** 9–17, doi:10.1002/met.188.

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

Shannon, C. E., 1948: A mathematical theory of communication. *Bell Syst. Tech. J.,* **27** (3), 379–423.

Weijs, S., R. van Nooijen, and N. van de Giesen, 2010a: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Mon. Wea. Rev.,* **138,** 3387–3399.

——, G. Schoups, and N. van de Giesen, 2010b: Why hydrological predictions should be evaluated using information theory. *Hydrol. Earth Syst. Sci.,* **14** (12), 2545–2558.