Hydrol. Earth Syst. Sci. Discuss., 7, 4657–4685, 2010
www.hydrol-earth-syst-sci-discuss.net/7/4657/2010/
doi:10.5194/hessd-7-4657-2010

Hydrology and Earth System Sciences Discussions

Why hydrological forecasts should be evaluated using information theory

S. V. Weijs, G. Schoups, and N. van de Giesen

Section Water Resources, Delft University of Technology, Delft, The Netherlands

Received: 30 June 2010 – Accepted: 2 July 2010 – Published: 16 July 2010

Correspondence to: S. V. Weijs (s.v.weijs@tudelft.nl)

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

Abstract

Probabilistic forecasting is becoming increasingly popular in hydrology. Equally important are methods to evaluate such forecasts. There is still debate about which scores to use for this evaluation. In this paper we distinguish two scales for evaluation: information-uncertainty and utility-risk. We claim that the information-uncertainty scale is to be preferred for forecast evaluation. We propose a Kullback-Leibler divergence as the appropriate measure for forecast quality. Interpreting a decomposition of this measure into uncertainty, correct information and wrong information, it follows directly that deterministic forecasts, although they can still have value for decisions, increase uncertainty to infinity. We resolve this paradoxical result by proposing that deterministic forecasts are implicitly probabilistic or are implicitly assuming a decision problem. Although forecast value could be the final objective in engineering, we claim that for calibration of models representing a hydrological system, information should be the objective in calibration, because it allows to extract all information from the observations and avoids learning from information that is not there. Calibration based on maximizing value trains an implicit decision model, which inevitably results in a loss or distortion of information in the data and more risk of overfitting, possibly leading to less valuable and informative forecasts.

1 Introduction

Over the last decades, probabilistic forecasting has become increasingly important in the field of hydrology. Lacking enough information to completely eliminate uncertainty, probabilistic forecasts are intended to reduce uncertainty of the user about future events and communicate the remaining uncertainty (Krzysztofowicz, 2001; Montanari and Brath, 2004). In hydrology, the development of methods for evaluating such forecasts, however, has not kept pace with the development of methods of generating them (Laio and Tamea, 2007). This is an important problem, given the fact that science is

HESSD

7, 4657–4685, 2010

required to make testable predictions and therefore needs unambiguous methods for testing those predictions. Furthermore, the lack of methods for evaluation of hydrological forecasts may hinder acceptance of those forecasts by the public. In this paper we approach forecast evaluation from an information-theoretical point of view. By using a decomposition we developed recently (Weijs et al., 2010) in combination with some results from information theory, we provide insights into what evaluation scores measure and what, in our opinion, they should measure. The most important insights are that deterministic forecasts are not testable without additional assumptions and that the purpose of a model should not influence the measure that is used for its calibration.

1.1 What is a good forecast?

In general, the evaluation of forecasts can have several purposes. Evaluation may serve to assign a level of trust in the forecast, to reward good forecasters, to diagnose problems in forecasting models, and to provide an objective function for calibration of the forecasting models. All these purposes for evaluation have in common that the measures should allow comparisons between forecasts or between series of forecasts. Assigning a level of trust only makes sense if there are also alternatives; rewarding a good forecaster has no use if there is no other forecaster or no other period of forecasts to compare to; diagnosing problems is not possible if there is no reference of what the quality should be; optimization works by continuously comparing different models or parameter sets.

For directly comparing two (series of) forecasts, preferences must be complete (a forecast must either be better, worse, or equally good than another one) and transitive (preferences can not form a loop like $A>B>C>A$, where $>$ denotes "is better than"), which are the same requirements that are applicable to probability (Peterson, 2009). These two requirements naturally lead to measures that take the form of a scalar real number. In contrast to this requirement for a one-dimensional measure, however, Murphy (1993) argued that it is possible to distinguish three different dimensions of forecast

HESSD
7, 4657–4685, 2010

"goodness":

- Consistency: correspondence between forecasts and judgments;
- Quality: the correspondence between forecasts and observations;
- Value: incremental benefits of forecasts to users.

Consistency requires that what the forecaster communicates, the forecast, corresponds to his best judgment. This judgment is internal to the forecaster and should be a rational distillation of all information available to him. Because a forecaster has only limited access to information and is not completely rational, *his* best judgment may not be *the* best judgment, but by definition he can never knowingly let *his* best estimate diverge from *the* best estimate, or it would not be his best estimate.

Quality is the dimension that is most important in pure science, as it concerns putting the predictions to the test by comparing forecasts with observations. It is important to note in this respect that an observation is also just an estimate of the truth and therefore does not fundamentally differ from a forecast. In fact, we are comparing one estimate of truth with another. The estimate that we regard as most trustworthy, usually the one that is made in hindsight, is called observation, the other estimate is the prediction or forecast. In meteorology, the evaluation of quality is called verification (Latin: veritas = truthfulness). This term is somewhat misleading, because establishing that a model simulates the truth is impossible (Oreskes et al., 1994).

Value is related to a decision problem attached to the forecast and more closely related to engineering than to science. It is therefore not only dependent on the forecasts and the observations, but also on who is using the forecasts. Hydrological forecasts may, for example, have significant value for reservoir operation, evacuation decisions, and agriculture. Good forecasts for dam operation can for example lead to more hydropower, less flood damage, and, at the same time, fewer unnecessary pre-releases for flood protection. One could attempt to express these benefits in monetary terms, but from a decision-theoretical point of view, it is better to use the more general term

HESSD
7, 4657–4685, 2010

utility. This takes into account that not every unit of money necessarily has the same value and that other things than money might be important. By definition, the utility of an uncertain event is equal to the expected utility of that event (Von Neumann and Morgenstern, 1953). In engineering, risk is defined as expected damage or loss (disutility). Risk is therefore the opposite of utility. For adverse events, like floods, anticipation can reduce risk and the value of hydrological forecasts can thus be expressed as the reduction in risk they provide when used in decision making. At first sight, this seems to be an appropriate criterion for evaluation of real world forecasts.

1.2 Problems with evaluation of hydrological forecasts

The current problem in defining a framework for evaluation of forecasts lies partly in that the distinction between the latter two dimensions, quality and value, is not always explicitly made. As most purposes of evaluation require a one-dimensional measure of goodness, a choice between value and quality must be made and if the latter is chosen, an unambiguous quality measure must be defined that can not rely on user preferences. The hydrological and meteorological literature, however, offers a wide range of verification measures. Although the properties of these measures are well-studied, it is not always clear what is actually measured. Laio and Tamea (2007) give an overview of some commonly used measures in meteorology that could be applicable in hydrology.

What is missing from this overview, and also in two standard works about forecast verification (Wilks, 2005) and (Jolliffe and Stephenson, 2003), are measures for forecast evaluation based on information theory (Weijs et al., 2010). We argue that information-theoretical scores are measures for quality par excellence, for forecasts stated in terms of probability.

Except for probabilistic forecasts, two other types of forecasts are commonly used and presented in the overview given in Laio and Tamea (2007): deterministic forecasts and interval forecasts. We think that these types of forecasts can in principle not be evaluated unambiguously without reference to external assumptions relating to

HESSD
7, 4657–4685, 2010

probability or utility. The result that the intervals contain 90% of the observations is meaningless if the intervals are not stated in terms of probability. The result that a deterministic forecast has an error of 10 m^3/s does not have meaning if it is not known what the implications are or how likely we think this error was.

Instead of seeing this as a problem of the evaluation methods, we argue that this should be seen as a problem of the forecasts themselves. They do not fulfill the requirement of testable predictions. Moreover, determininistic forecasts are not consistent with judgments, which, given that we know a model is an approximation, are better described in terms of probability.

Notwithstanding these problems with deterministic forecasts, they are still common in hydrology and are usually evaluated with measures like NSE and MSE, MAE. Actually, many of the methods for producing probabilistic forecasts make use of deterministic forecasts and their evaluation, for example Monte-Carlo based methods. Therefore, it is likely that there exists some reason that makes deterministic forecasts acceptable from a practical point of view. Also here the information-theoretical viewpoint could provide some new insights.

1.3 Outline

In this paper, we propose to use information theory as the central framework for forecast quality. By viewing the forecast evaluation problem from an information-theoretical perspective, we hope to shed some light on what is measured and what should be measured by verification scores.

In Sect. 2, we present an information-theoretical score for forecast quality along with its decomposition as recently presented by Weijs et al. (2010). In that paper it was also shown that the components of a commonly used Brier score decomposition are second order approximations to our information components. The information-theoretical divergence score can be interpreted as remaining uncertainty after receiving the forecast, which should be minimized. In Sect. 3 we analyse the seemingly paradoxical implication that deterministic forecasts increase the uncertainty to infinity and we offer two

HESSD
7, 4657–4685, 2010

interpretations to resolve this paradox. In Sect. 4, the question is addressed whether or not the utility a model provides for users should be considered in the calibration process. The conclusions are summarized in the last section, where we argue that issuing forecasts can best be considered a communication problem and that the information they provide is the most sensible measure for their evaluation.

2 Information-theoretical evaluation of forecasts

Information theory provides a number of measures relating uncertainty and information, within the framework of probability theory. Since forecasting can be seen as providing information to reduce uncertainty about future events, information theory appears to be an appropriate framework to evaluate forecasts. As we showed in our recent paper, Kullback-Leibler divergence, or relative entropy, can be used as a verification score and has a number of desirable properties. Starting from an analogy with the Brier score, we now introduce the divergence score and an insight-providing decomposition of it. For a more elaborate description and some other related discussions, see (Weijs et al., 2010)

2.1 Classical decomposition of the Brier score

The Brier score was introduced by (Brier, 1950) as a verification score for probabilistic forecasts. It is still the most widely used score for evaluating probabilistic forecasts of binary events. A binary event has two possible outcomes, e.g. exceedence or non-exceedence of a certain critical water level in a river. A probabilistic forecast for one such a binary event at time t can be represented by a probability mass function (PMF), which in this case is a two element vector, denoted by $\boldsymbol{f}_t$. The bold notation indicates a vector. For example, when a probabilistic flow forecast indicates that there is 20% chance that the critical flow will be exceeded, the forecast can be written as $\boldsymbol{f}_t=(1-f,f)^T=(0.8,0.2)^T$, where the scalar f denotes the probability of exceedence.

HESSD
7, 4657–4685, 2010

After the event is observed, the observation can also be written as a PMF, this time expressing the probabilities after the event has been observed. In case we assume perfect observations, and we observed exceedence of the critical level, the observation can be expressed as $\boldsymbol{o}_t = (1-o, o)^T = (0,1)^T$. In this paper, we assume perfect observations to allow for the decompositions we introduce, but in general, perfect observations are not a necessary assumption for the score to be meaningful. Given the preceeding definitions, the Brier score can now be defined as:

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

It must be noted that the Brier score is nowadays almost always defined as half this value (Ahrens and Walser, 2008). To make notation easier, we use the original definition of Brier (see Eq. 1). For a series of forecasts, the Brier score is defined as the average of Eq. (1) over all forecast instances. It can be interpreted as the mean squared error (MSE) in probabilities.

Murphy (1973) showed that the Brier score for such a series can be decomposed into three components: uncertainty, resolution and reliability:

$$\mathrm{BS} = \mathrm{REL}_{\mathrm{BS}} - \mathrm{RES}_{\mathrm{BS}} + \mathrm{UNC}_{\mathrm{BS}}, \quad (2)$$

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k (\boldsymbol{f}_k - \bar{\boldsymbol{o}}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\bar{\boldsymbol{o}}_k - \bar{\boldsymbol{o}})^2 + \bar{\boldsymbol{o}}^T (\mathbf{1} - \bar{\boldsymbol{o}}). \quad (3)$$

where N is the total number of forecasts and K the number of unique forecasts issued, $\bar{\boldsymbol{o}} = \sum_{t=1}^{N} \boldsymbol{o}_t / N$ the climatological (long term average) probability of occurrence of the event, n_k the number of forecasts within one category of unique forecasts, $\bar{\boldsymbol{o}}_k$ the observed frequency, given forecasts of probability $\boldsymbol{f}_k$ and $\mathbf{1}$ is a vector of ones of the same size as $\bar{\boldsymbol{o}}$.

The uncertainty term measures the inherent uncertainty in the climate. The uncertainty reaches a maximum for equiprobable outcomes and is zero if the outcome is always the same. The resolution and reliability terms in this decomposition can be seen as squared Euclidean distance measures between two probability distributions.

HESSD
7, 4657–4685, 2010

The resolution term measures how much of the climatic uncertainty can be resolved by the forecasts. This is expressed in the average distance of the conditional distributions of the observations from the marginal distribution of the observations. The reliability measures the average squared distance between the forecast distributions and the corresponding conditional distributions of observations. A perfect reliability of zero (a more accurate term would be unreliability) is attained when for all forecast probabilities, the observed conditional frequency matches that probability. In this case the forecast is said to be perfectly calibrated.

2.2 Information-theoretical equivalents: divergence score and decomposition

Information theory started with the paper of (Shannon, 1948), where he derived a measure of uncertainty (entropy) from three basic requirements for such a measure. The highly readable original paper is recommended for more background. The uncertainty of the climate (knowledge of long term frequencies but absence of other information) using this definition is

$$H(\bar{o}) = -\sum_{i=1}^{n} \{[\bar{o}]_i \log[\bar{o}]_i\}. \quad (4)$$

where $[\bar{o}]_i$ denotes the i^{th} element of vector $\bar{o}$. The logarithm has base 2, yielding the measure H in the unit bits. A related measure is relative entropy, also known as Kullback-Leibler divergence. This is a measure of the extra amount of uncertainty if one distribution is assumed, while the true distribution is different, this is the divergence from the true to the other distribution. In contrast to a distance like the Brier score, Kullback-Leibler divergence is not symmetric. The divergence depends on which of the two distributions is considered the true one.

HESSD
7, 4657–4685, 2010

We define the divergence score as the divergence *from* the observation PMF *to* the forecast PMF:

$$\mathrm{DS}_t = D_{\mathrm{KL}}(\boldsymbol{o}_t \| \boldsymbol{f}_t) = \sum_{i=1}^{n} [\boldsymbol{o}_t]_i \log\left(\frac{[\boldsymbol{o}_t]_i}{[\boldsymbol{f}_t]_i}\right). \tag{5}$$

where n is the number of possible outcomes (2 in the binary case). For a series of N forecasts and corresponding observations, the divergence score is

$$\mathrm{DS} = \frac{1}{N} \sum_{t=1}^{N} D_{\mathrm{KL}}(\boldsymbol{o}_t \| \boldsymbol{f}_t). \tag{6}$$

When replacing all quadratic distances in the Brier score decomposition by the appropriate divergences and replacing the uncertainty component by the information-theoretical definition of uncertainty, entropy, we obtain (see Table 1):

$$\mathrm{DS} = \frac{1}{N} \sum_{k=1}^{K} n_k D_{\mathrm{KL}}\left(\bar{\boldsymbol{o}}_k \| \boldsymbol{f}_k\right) - \frac{1}{N} \sum_{k=1}^{K} n_k D_{\mathrm{KL}}\left(\bar{\boldsymbol{o}}_k \| \bar{\boldsymbol{o}}\right) + H(\bar{\boldsymbol{o}}) \tag{7}$$

In the appendix of (Weijs et al., 2010) it was shown that this equation holds, and thus we have obtained an information-theoretical equivalent of the Brier score and its decomposition, which extends also to multiple category forecasts.

2.3 Relation between the divergence and Brier scores

The components of the Brier score are second order approximations of the components of the divergence score (see Table 1). The uncertainty has the same location of maximum and zero points. When scaled with its maximum value, the similarity becomes visible (see left figure in Table 1). The resolution (right figure in Table 1), can reach a maximum equal to the uncertainty term. When scaled with the uncertainty, again a similarity between the shapes of the the resolution components is visible. The reliability term, however, exhibits significant differences in the extremes. While the

HESSD
7, 4657–4685, 2010

reliability term of the Brier score is bounded, the analogous term in the divergence score can reach infinity. This happens when an outcome occurs that was given zero probability in the forecast.

2.4 Interpretations of divergence score and decomposition

The new information-theoretical equivalents of the components of the Brier score allow some additional interpretations. One of the interpretations of measures in information-theory starts from a definition of surprise. Surprise is something we feel when something unexpected happens. The lower the probability we assume something to have, the more surprised we are when observing it. Rain in a desert is surprising, rain in the Netherlands is less surprising and rain on the moon is a miracle yielding almost unbounded surprise. When the surprise of observing outcome x is defined as $S_x = \log(1/P(x))$, surprise can be measured in bits like information and uncertainty. Observing something that was a certain fact yields no surprise, heads on a fair coin yield one bit of surprise and observing a 1/1000 year flood in some year yields a surprise of approximately 10 bits. The entropy-measure for uncertainty can now be interpreted as the expected surprise about the truth: $H(X) = E_X\{S_x\}$, where E_X denotes the expectation operator with respect to the distribution of random variable X.

In general, uncertainty can now be interpreted as expected surprise about the true outcome. The fact that different expectations can be calculated according to different subjective probability distributions, reflects that uncertainty can be both something objective and subjective. The uncertainty a person thinks to have is the entropy of his subjective probability distribution. Kullback-Leibler divergence can be seen as the additional uncertainty one person estimates the other person to have compared to his own:

$$D_{\mathrm{KL}}(P(X)\|Q(X)) = E_{P(X)}\{S_{Q(x)} - S_{P(x)}\} \quad (8)$$

Because forecast verification is done in hindsight, the observation that is made can be used as a reference point to estimate the uncertainty in the forecast. The additional uncertainty (expected surprise about the truth), estimated from the viewpoint of

HESSD
7, 4657–4685, 2010

the observation is the best available estimate of the remaining uncertainty about the truth of the person having the forecast. Assuming perfect observations, the divergence score measures remaining uncertainty about the truth. In Fig. 1 it is shown how the components of the divergence score relate to this remaining uncertainty at different levels of informedness. Interpreting the figure, resolution can be seen as the correct information, that can be subtracted from the climatological uncertainty or missing information. The reliability term is added to the remaining uncertainty and represents the wrong information due to biased probability estimates. The wrong information can be reduced by calibration. It should be noted that the decomposition is only meaningful when enough data is available to properly calculate all conditionals (Weijs et al., 2010).

3 Deterministic forecasts are inconsistent

Can a forecaster be completely sure about something that in the end does not happen and still get credit for his forecast? This does not appear natural, but it often turns out to happen in practice. For example, a deterministic flow forecast of 200 m^3/s is considered quite good, when 210 m^3/s is observed. Apparently, it is already expected that some error will occur and a forecast that is 10 m^3/s off is considered to be not that bad. Hydrological models are per definition simplifications of reality. Often, they describe relations between macrostates, like averaged rainfall, mass of water in the groundwater reservoir, and flow through a river cross-section. Similar to problems in statistical thermodynamics, having limited information about what really goes on inside a hydrological system on a microscopical level, our forecasts on a macroscopical level can never be perfect (Weijs, 2009; Grandy Jr., 2008). What can be said about the real world on the basis of a model is therefore inherently erroneous to some extent, or should be stated in terms of probabilities.

How then, should deterministic forecasts be evaluated? Literally taken, a deterministic (point value) forecast states: "the outcome is x". Implicitly, such a forecast asks to be evaluated from a black and white view: the forecast is either wrong or right. The divergence score also reflects this. If the forecast were right, the perfect score of 0 would

HESSD
7, 4657–4685, 2010

be attained, if the forecast were wrong, however, a penalty of infinity would be given. If one such a forecast is given, the forecaster can look for another career, because even a future series of perfect forecasts can not average out the infinite penalty. The decomposition shows that the reliability component is responsible (Table 1, middle figure). Although the deterministic forecasts usually contain information about the observed outcomes, given that the resolution (correct information) is positive and removes some of the uncertainty, this is completely annihilated by the reliability term (the wrong information). The discrepancy between the information (reduction of uncertainty) that the forecasts contain and the information conveyed by the messages that constitute the forecasts is so large that the expected surprise about the truth of a person taking the forecast at face value goes to infinity. The fact that deterministic forecasts are still used in society (and unfortunately sometimes even preferred), while they explode uncertainty to infinity, seems to present a paradox. We propose two possible interpretations that offer a solution to this paradox.

3.1 Deterministic forecasts are implicitly probabilistic (information interpretation)

Fortunately, in reality, almost no person using deterministic forecasts takes them at face value. In fact, the forecast is implicitly recalibrated by the user, reducing the reliability term for the internal probability estimates the user bases his actions on. This can be seen as the user eliminating the wrong information from the forecast. The user can do the recalibration based on previous experience with the forecasts and common sense. The user of the forecast can think "if the forecaster says the water level will be 10 cm under the embankment, he implicitly also forecasts a little that overtopping will occur". Note that the example of Grand Forks in (Krzysztofowicz, 2001) shows that not all users do this. Mathematically this recalibration is equivalent to also attaching some probability to overtopping. However, it is not the task of a user to guess what the forecaster wanted to say. Consistency requires that the forecaster communicates his judgments to the user (Murphy, 1993).

HESSD
7, 4657–4685, 2010

The forecaster may also present the deterministic forecast as being an expected value or mean. This suggests an underlying probabilistic forecast. However, when taking the information-theoretical viewpoint, communicating an expected value means nothing without additional statements regarding the probability distribution. The principle of maximum entropy (PME) (Jaynes, 1957) states that when making inferences based on incomplete information, the best estimate for the probabilities is the distribution that is consistent with all information, but maximizes uncertainty. In this way, the uncertainty is reduced exactly by the amount the information permits, but not more. Maximizing entropy with known mean and variance, gives a Gaussian distribution, maximizing uncertainty about the velocities of gas molecules with known total kinetic energy gives the Boltzmann distribution (Jaynes and Bretthorst, 2003; Cover and Thomas, 2006). When PME is applied to expected value forecasts, however, the maximum entropy forecast distribution that is consistent with the information given by the forecaster is uniform between minus and plus infinity. It is the complete opposite end of the spectrum compared to the previous literal interpretation of the deterministic forecast: from claiming total certainty to claiming total uncertainty.

In the case of streamflow forecasts, the user can still get a less nonsensical forecast distribution by combining the information in the forecast with the common sense notion that streamflows in rivers are nonnegative. This extra constraint turns the PME forecast distribution for a known expected value into an exponential distribution (Cover and Thomas, 2006), see Eq. (9).

$$f(x;\mu) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}, \quad x \geq 0, \quad \text{and} \quad f(x;\mu) = 0, \quad x < 0 \tag{9}$$

This brings back the question who ought to specify these constraints, which in fact constitute information. The fact that the user can reduce the maximum entropy by adding this common sense constraint actually means that the forecaster failed to add this information.

HESSD
7, 4657–4685, 2010

As we argued in the introduction, predictions only make sense when they are testable, i.e. can be evaluated. One way to evaluate deterministic forecasts with information measures is to convert them to probabilistic forecasts by looking at the joint distribution of forecasts and observations. The conditional distributions of observations for each forecast value can then be seen as probabilistic forecast distributions. It is important to note however, that the probabilistic part of such a forecast is derived from data that includes the observations. When such forecasts are evaluated, the predictive performance is judged on the basis of an uncertainty model, that is derived from the same data that is used for its evaluation.

Also without explicit conversion to a probabilistic forecast, the uncertainty model becomes explicit when a series of deterministic forecasts is evaluated. An penalty (objective) function for a deterministic forecast can be interpreted as an uncertainty (information) measure for a corresponding probabilistic forecast. For example, a deterministic forecast evaluated with RMSE implicitly defines Gaussian forecast pdf. An important consequence of this insight is that the way to evaluate a deterministic model actually is the probabilistic part of model. The objective function (which is a likelihood measure) should therefore be stated a priori, as it forms part of the model that is put to the test against observations.

While this approach may under some conditions be acceptable for calibration to train the error model, for evaluation of forecasts it is unacceptable, because it uses the data against which it is evaluated. A correct approach would be to explicitly formulate and train an error model in the calibration, and use that model to make probabilistic predictions for the evaluation period, that can subsequently be evaluated with the divergence score. The error models are not restricted to Gaussian distributions, but can take more flexible forms. Such an approach is taken in Schoups and Vrugt (2010).

As a last consideration, we want to stress that even if an error model is properly formulated and added to the deterministic "physical" part, the resulting model still represents a false dichotomy between true behaviour of the system and the error, as was argued by Koutsoyiannis (2010). A more consistent approach would be to explicitly

HESSD
7, 4657–4685, 2010

make the probabilistic part of the model an integrated part of the physical reality it is supposed to simplify. Such approaches can lie in studying the time-evolution of chaotic systems (Koutsoyiannis, 2010) or in applying the principle of maximum entropy in combination with macroscopic constraints, as for example suggested by Weijs (2009) and Koutsoyiannis (2005).

Concluding, from the information-theoretical viewpoint several reasons come to light why deterministic forecasts should in fact be considered to be implicitly probabilistic. The problem with these forecasts is that they leave too much of the probabilistic interpretation to the user. It might be considered ironic that the users who are claimed to not be able to handle probabilistic forecasts and are for that reason provided with deterministic forecasts are the ones who have to rely most on their ability to subconsciously make probability estimates based on the limited information in the deterministic forecast.

3.2 Deterministic forecasts can still have value for decisions (utility interpretation)

A second, independent interpretation of deterministic forecasts that justifies their existence is their usefulness, even to users who do not make subconscious probability estimates. Even though a reservoir operator might be infinitely surprised if he has taken a deterministic inflow forecast of 200 m^3/s at face-value and he finds out the inflow was 210 m^3/s, his loss is not infinite. The operator might spill some water, but not all is lost.

The difference between surprise and loss is due to the fact that most decision problems are not equal to placing stakes in a series of horse races. Such a horse race is the classical example where information can be directly related to utility, see Kelly (1956) and Cover and Thomas (2006) for more explanation. Kelly showed that when betting on a series of horse races, where the accumulated winnings can be reinvested in the next bet, the stakes the gambler should put on each horse should be proportional to the estimated winning probabilities. In a single instance of such a horse race, all money not bet on the winning horse is lost, so the only probability that is important for the results

HESSD
7, 4657–4685, 2010

is the one attached to the winning horse. In contrast, for decision problems like reservoir operation, optimally preparing for 200 m^3/s automatically implies also preparing for 210 m^3/s to some extent. This makes the loss function non-local (locality is discussed in Sect. 4.1).

Another difference with the horse race is that the total amount of value at stake in hydrological decision making usually does not depend on the previous gains, while the results for the horse race assume that the gambler invests all his previously accumulated capital in the bets. The gambler therefore wants to maximize the product of rates of return over the whole series of bets, while for a reservoir operator, each period offers a new opportunity to gain something from the water, even though he spilled all his water in the previous month. This is comparable with a gambler whose wife allows him to bet a fixed amount of money each week (Kelly, 1956) and then spends it all in the bar on the same evening without possibility of reinvesting in the next bet. Assuming a utility linear in the beer he buys with the winnings, the best decision is to bet all money on the one horse with the best expected return. Again, one loss is not fatal for the whole series of bets. He just hopes for better luck next week. The evaluation of the value of deterministic forecasts is therefore not as black and white as evaluation of the information they contain.

The evaluation of deterministic forecasts in this interpretation is thus connected to a decision problem. Decisions can be taken as if the forecasts are really certain, and still be of value. The loss functions for evaluating forecasts can be seen as functions that somehow map the discrepancy between forecast value and observed value to a loss of the decision based on the wrong forecast, compared to a perfect forecast. In the utility interpretation, evaluating deterministic forecasts with mean squared error implicitly defines a decision process in which the disutility is a quadratic function of the distance between forecast and observation. In that case, a series of forecasts that has the smallest MSE has most utility or value for the user.

HESSD
7, 4657–4685, 2010

4 Information versus utility as calibration objective

Value-based forecast evaluation is inevitably connected to a particular user with a decision problem and therefore cannot be done without explicit consideration of the user base of forecasts. Moreover, an obvious question that arises is whether it is desirable to base the evaluation on the value to a particular user or group of users. In that case, the evaluation becomes an evaluation of decisions rather than of the forecasts themselves or of the hydrological model that produced them (see Fig. 3). This difference is particularly important if the results of the evaluation are used in a learning or calibration process. In that case, two effects can occur by using value instead of information as a calibration objective:

- The model learns from information that is not there (treated in Sect. 4.1).
- The model fails to learn from all information that *is* there (treated in Sect. 4.2).

4.1 Locality and philosophy of science

Locality is a property of scores for probabilistic forecasts. A score is said to be local if the score only depends on the probability assigned to (a small region around) the event that occurred, and does not depend on how the probability is spread out over the values that did not occur. In contrast to this, non-local scores do depend on how that probability is spread out (see Fig. 2 for a comparison). Usually they are required to be sensitive to distance, which means that probability attached to values far from the observed value is punished more heavily than forecast probability that was assigned to values close to the observation. This concept of distance only plays a role in forecasts of continuous and ordinal discrete predictands. For both these types of predictands, an extension of the Brier score exists: the Ranked Probability Score (RPS) and the continuous RPS (CRPS) (see Laio and Tamea, 2007 for description and references). Both these scores are non-local, while the divergence score is local.

HESSD
7, 4657–4685, 2010

For most decision problems, expected utility is a non-local score: a reservoir operator that attached most probability to values far from the true inflow is worse off than one that used a forecast with most probability close to the true value, even if the probability (density) attached to the true value was the same. Therefore, non-local scores are sometimes considered to have more intuitive appeal than local scores.

There is, however, a serious philosophical problem with non-local scores if used in a learning process. In principle, the knowledge a model embodies comes from observations or prior information (which in the end also comes from observation, see Fig. 3). By calibrating a model, the information in the observations is merged with the prior information, through a feedback of the objective function value to the search process. It is therefore a violation of scientific logic if the score that is intended to evaluate the quality of forecasts depends on what is stated about things that are not observed. Changes in the objective function would cause the model to learn something from an evaluation of what is stated about a non-observed event. In an extreme case, two series that forecast the same probabilities for all the events that were observed, can obtain different scores based on differences in the probabilities assigned to unobserved events (Benedetti, 2010). A similar argument in the context of experimental design was made by Bernardo (1979). If these non-local scores are used as objectives in calibration or inference (see for example Gneiting et al., 2005), things are thus inferred from non-observed outcomes, i.e. information that is not there.

4.2 Utility as a data filter

The use of utility in calibration can, next to using non-existing information, also lead to learning only from part of the information that is in the observations. In that sense, the decision problem that specifies the utility acts like a filter on the information. The information-theoretical data processing inequality tells us that this filter can only decrease information (see Cover and Thomas, 2006). For example, when a binary evacuation decision is coupled to a conceptual rainfall-runoff model for flood forecasting,

HESSD
7, 4657–4685, 2010

the calibration towards maximum utility of the system will train the hydrological model to optimally distinguish flood-evacuation events. This implies that in the training, all that the hydrological model sees from the continuous observed discharges is a binary signal: flood or no flood. This constitutes at most one bit of information per observation (in the unlikely case that a 50% of the observations is above the flood threshold, i.e. the climatic uncertainty is 1 bit), while the original signal contained far more information (see Fig. 3). The hydrological model will therefore have far less information to learn from. Given the fact that there is a balance between the available information for calibration and the complexity that a model is allowed to have (see Schoups et al., 2008), hydrological models that are trained on this kind of utility functions are likely to become overly complex relative to the data . They will surely achieve better utility results on the calibration data (because there is less information to fit), but are likely to perform worse on an independent validation dataset. The model that has been trained with maximum information as an objective is likely to yield better results for the validation set, even in terms of utility. Because it has the unfiltered information from the observations to learn from, it is less prone to overfitting: the complexity of a conceptual hydrological model is better warranted by the full information. Training for optimal classification of flood events would benefit from more parsimonious data-driven models that make a mapping directly from predictors to decisions, but this complicates the use of prior information on the workings of the hydrological system, which can be another valuable source of information to improve forecasts. Examples are constraints on mass balance and energy limits for evaporation.

The third route of information to the model, the input observations, can also be affected by the information filter. For example, if a binary decision problem (e.g. to be or not to be in the flood zone tomorrow) is considered, the information from input observations travels through the model and subsequently through the decision model, which maps the model input to a binary signal (to be or not to be). The binary signal is all that enters the evaluation and can be learned from the input observations. When a model is evaluated based on a cost-loss model of a two action- two state of the world

HESSD
7, 4657–4685, 2010

decision problem, the maximum amount of information that can be learned from each input-output observation pair is thus 2 bits.

This framework shows some similarity with the ideas presented in (Gupta et al., 2009, 1998, 2008). In those papers it is also argued that information can be lost the evaluation. However, the important difference of this framework compared to those ideas is that we argue that information is lost by using measures other than information (in other words, measures that do not reflect likelihood), while Gupta et al. (2008) argue that information is lost because of the low dimensionality of the evaluation measure. In our information-theoretical viewpoint, we can in principle learn all we need from the observations through a single measure (a real number can contain infinitely many bits of information). This can only be achieved if the mapping of the information in prior knowledge, input observation and output observations to the scoring rules reflecting the likelihood of the model connecting these sources is reliable. In principle, this is equivalent to endorsing the likelihood principle, which states that all information that the data contains about a model is in the likelihood function (as argued by Robert (2007) p.14, Jaynes (1957), p.250 and Berger and Wolpert (1988).

The information-theoretical logarithmic scoring rules are the only scoring rules that are both local and proper (proofs can be found in Bernardo, 1979 and Benedetti, 2010). Where propriety is the requirement that the scoring rule can only be optimized when the forecaster does not lie. Scoring rules that are not proper can be hedged, meaning that the expected score is maximized by forecasting probabilities that are not consistent with the best estimates of the forecaster. A utility function that includes the importance of the outcomes can be hedged by attaching more forecast probability to important events. A model that is trained on such a measure is thus encouraged to "lie". All utility functions that are not linear functions of information violate either locality or propriety, which makes them doubtful objectives for calibration.

HESSD
7, 4657–4685, 2010

5 Conclusions

The difficulties and debate about the evaluation of forecasts can be significantly clarified using results from information theory. When information is seen as a measurable quantity, like energy, a sort of "information intuition" develops, similar to the "energy intuition" that is used to detect logical flaws in claims for perpetuum mobiles. Science is required to make testable predictions. Forecasts should therefore be stated in terms that make it clear how to evaluate them. Deterministic and interval forecasts fail this criterion. Probabilistic forecasts can be evaluated using information theory. The decomposition of the divergence score that we presented can provide additional insight in the interaction between uncertainty, correct information and wrong information.

Starting from the observation that deterministic forecasts are still commonly used and evaluated, but are worthless from an information-theoretical viewpoint, we draw the conclusion that these forecasts are either implicitly probabilistic or should be viewed in connection to a decision problem. In both interpretations, the evaluation depends on external information that is not provided in the forecast. Deterministic forecasts leave too much interpretation to the user, if seen as implicit probabilistic forecasts or make too many assumptions on the user if they are evaluated using another utility measure.

On the one hand, forecasting can be seen as a communication problem in which uncertainty about the outcome of a random event is reduced by delivering an informative message to a user. On the other hand, forecasting can be seen as an addition of value to a decision problem. Any measure that is not information only becomes meaningful when it is interpreted in terms of utilities. When addressing forecast value, it is important to see that in fact we are evaluating decisions based on forecasts and not the correspondence between the observations and the forecasts themselves.

This is especially important in calibration, where a model has to learn from observations. When calibration objectives are used that are not information-measures, the model either learns from information that is not there or uses only part of the information in the observations, or both. Because the amount of available information is

HESSD
7, 4657–4685, 2010

related to optimal model complexity, hydrological models trained for user specific utilities are more prone to overfitting, which might lead to worse results in an independent validation test.

Acknowledgements. The authors thank Ronald van Nooijen and Luciano Raso for fruitful discussions about the manuscript.

References

Ahrens, B. and Walser, A.: Information-based skill scores for probabilistic forecasts, Mon. Weather Rev., 136, 352–363, 2008. 4664

Benedetti, R.: Scoring rules for forecast verification, Mon. Weather Rev., 138, 203–211, 2010. 4675

Berger, J. and Wolpert, R.: The Likelihood Principle, 2nd edn., Institute of Mathematical Statistics, Hayward, CA, 1988. 4677

Bernardo, J.: Expected information as expected utility, Ann. Stat, 7, 686–690, 1979. 4675

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, 1950. 4663

Cover, T. and Thomas, J.: Elements of Information Theory, Wiley-Interscience, New York, 2006. 4670, 4672, 4675

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, 2005. 4675

Grandy Jr., W.: Entropy and the Time Evolution of Macroscopic Systems, Oxford University Press, New York, 2008. 4668

Gupta, H., Kling, H., Yilmaz, K., and Martinez, G.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009. 4677

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, 1998. 4677

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, 2008. 4677

HESSD

7, 4657–4685, 2010

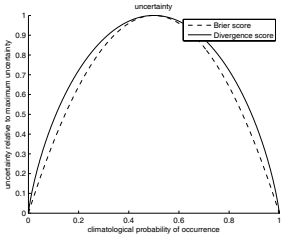
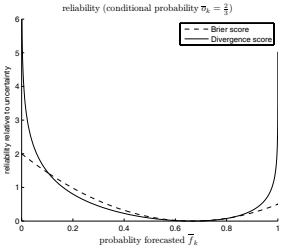
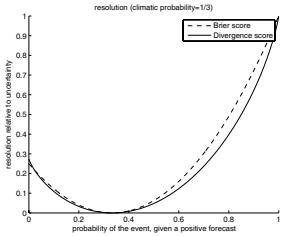
Jaynes, E. and Bretthorst, G.: Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK, 2003. 4670
Jaynes, E. T.: Information theory and statistical mechanics, Phys. Rev., 106, 620–630, 1957. 4670, 4677
Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, Chichester, UK, 2003. 4661
Kelly, J.: A new interpretation of information rate, IEEE T. Inform. Theory., 2, 185–189, 1956. 4672, 4673
Koutsoyiannis, D.: Uncertainty, entropy, scaling and hydrological statistics. 1. Marginal distributional properties of hydrological processes and state scaling, Hydrolog. Sci. J., 50, 381–404, 2005. 4672
Koutsoyiannis, D.: *HESS Opinions* "A random walk on water", Hydrol. Earth Syst. Sci., 14, 585–601, doi:10.5194/hess-14-585-2010, 2010. 4671, 4672
Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, 2001. 4658, 4669
Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007. 4658, 4661, 4674
Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004. 4658
Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973. 4664
Murphy, A. H.: What is a good forecast?: An essay on the nature of goodness in weather forecasting, Weather Forecast., 8, 281–293, 1993. 4659, 4669
Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation, and confirmation of numerical models in the earth sciences, Science, 263, 641–646, 1994. 4660
Peterson, M. B.: An Introduction to Decision Theory, Cambridge University Press, Cambridge, UK, 2009. 4659
Robert, C.: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Springer Verlag, New York, 2007. 4677
Schoups, G. and Vrugt, J.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, Water Resour.

HESSD
7, 4657–4685, 2010

Res., in press, doi:10.1029/2009WR008933, 2010. 4671
Schoups, G., van de Giesen, N. C., and Savenije, H. H. G.: Model complexity control for hydrologic prediction, Water Resour. Res., 44, W00B03, doi:10.1029/2008WR006836, 2008. 4676
Shannon, C. E.: A mathematical theory of communication, Bell System Technical J., 27, 379–423, 1948. 4665
Von Neumann, J. and Morgenstern, O.: Theory of Games and Economic Behavior, 3rd edn., Princeton University Press, USA, 1953. 4661
Weijs, S.: Interactive comment on "HESS Opinions "A random walk on water"" by D. Koutsoyiannis, Hydrology and Earth System Sciences Discussions, 6, C2733–C2745, 2009. 4668, 4672
Weijs, S., Van Nooijen, R., and Van de Giesen, N.: Kullback-Leibler divergence as a forecast skill score with classical reliability-resolution-uncertainty-decomposition, Mon. Weather Rev., early online release, 2010. 4659, 4661, 4662, 4663, 4666, 4668
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd edn., Academic Press, San Diego, CA, 2006. 4661

HESSD
7, 4657–4685, 2010

Table 1. Comparison between the expressions and behaviour of the decompositions of the Brier score and the divergence score for the case of binary events.

	UNC	REL	RES
Brier Score	$\bar{o}(1-\bar{o})$	$\frac{1}{N}\sum_{k=1}^{K} n_k(f_k-\bar{o}_k)^2$	$\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k-\bar{o})^2$
Divergence Score	$H(\bar{o})$	$\frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{o}_k \Vert f_k)$	$\frac{1}{N}\sum_{k=1}^{K} n_k D_{\mathrm{KL}}(\bar{o}_k \Vert \bar{o})$
			

HESSD
7, 4657–4685, 2010

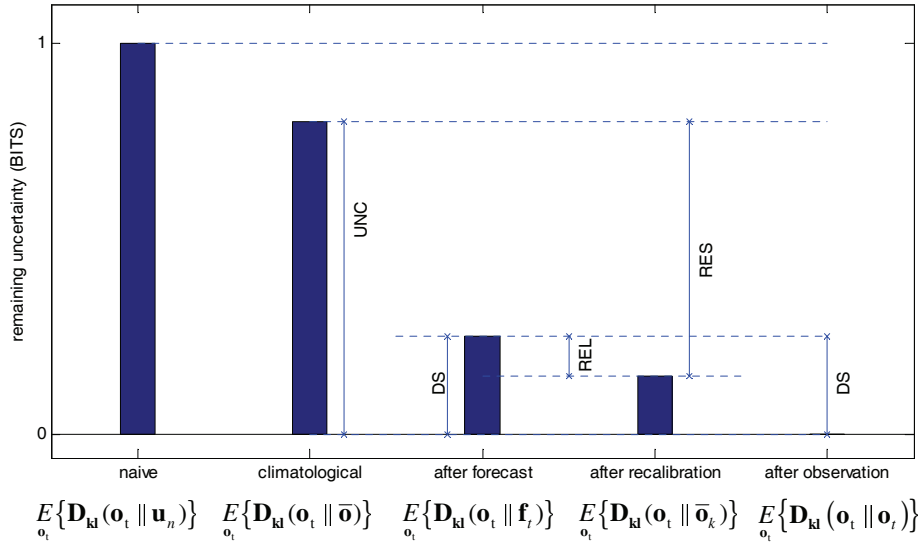


Fig. 1. The remaining uncertainty for different distributions in the forecasting process can be measured by the average Kullback-Leibler divergence from the observations. These uncertainties have some additive relations.

HESSD
7, 4657–4685, 2010

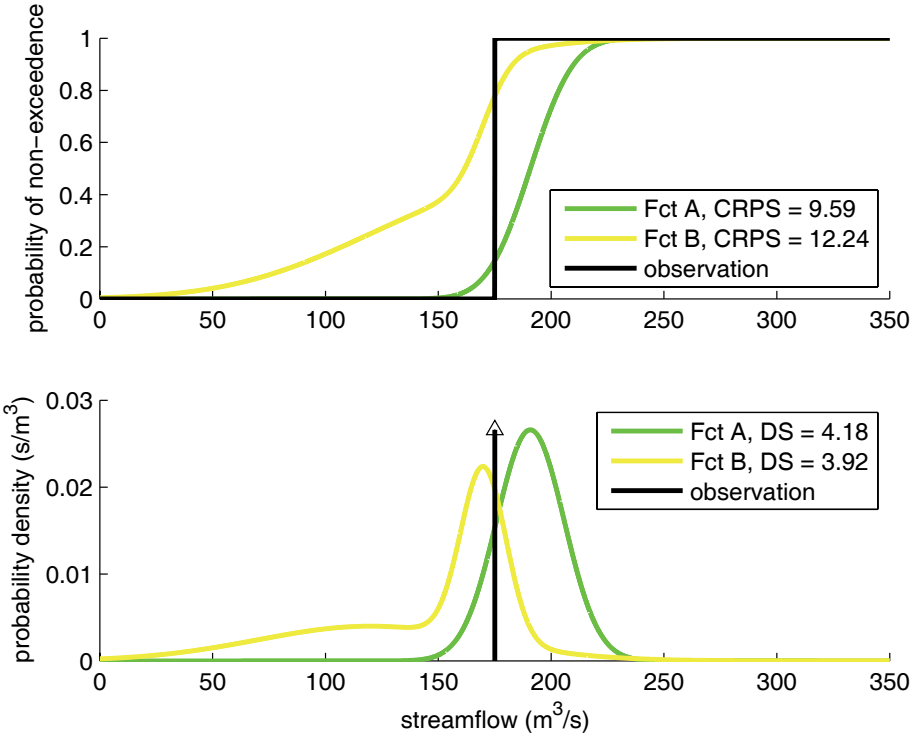


Fig. 2. The RPS and CPRS scores measure the sum of squared differences in CDFs. Therefore they depend on probabilities assigned to events that where not observed. The divergence score only depends on the value of the PDF (the slope of the CDF) at the value of the observation. In the example, forecast A has a better CPRS than forecast B, even though it assigned a higher probability to what was observed.

HESSD
7, 4657–4685, 2010

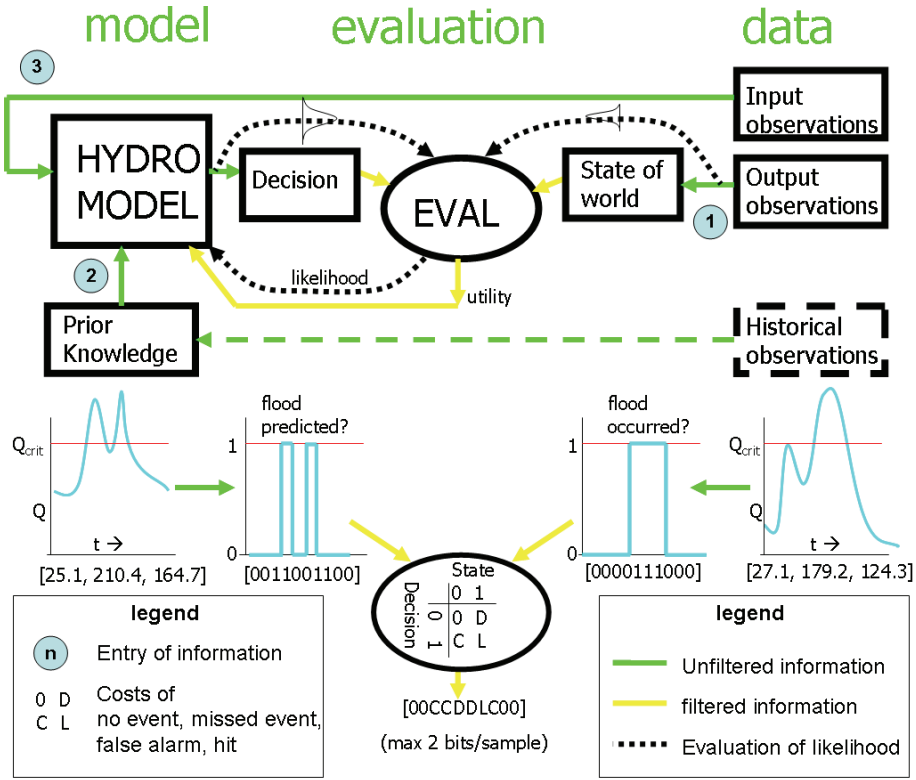


Fig. 3. There are three routes through which information can enter the model in a learning process. When evaluating a model based on value, the decision model that is implicitly defined by the loss function acts as a filter on the information in the observations. For example, all that the model can see from the a training on a binary decision model is 2 bits of information per input-output observation pair, which are contained in the sequence of costs fed back to the model.

HESSD
7, 4657–4685, 2010