

A Score for Probability Forecasts of Binary Events Based on the User's Cost–Loss Relations

KARL-IVAR IVARSSON

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

(Manuscript received 25 January 2019, in final form 22 October 2019)

ABSTRACT

A score for verifying probabilistic forecasts is presented. It is called the continuous specific score (CSS) and is intended for binary events only. The score is based on the user's cost–loss relations and their relative importance. The relative importance is determined by a continuous function of the user's density of loss for various cost–loss relations. One may also consider CSS as the result of the expected mean value based on a probability function of the various loss values for all possible cost–loss relations for one single user. The CSS is a negatively oriented score and has the following properties: perfect forecasts yield the score zero, and 100% probability for adverse weather (AW) when AW does not occur leads to the score one. The result of a 0% forecast of AW when AW occurs is the inverted value of the ratio of the average cost to the average loss minus one. Different possible usages of the score are discussed. An effective cost–loss ratio (ECLR) is defined. It measures how important low cost–loss ratios are compared to higher ones.

1. Introduction

The verification of probability forecasts and the discussion regarding the best way to do this goes back to the middle of the last century or perhaps even longer. The most widely used score for verifying probability scores is the Brier score, which in its simplest form is just the mean squared difference between the forecast probability and the observed outcome. The observation is either zero or one (Brier 1950). Extensions of this score are also commonly used such as the ranked probability score (RPS), which is based on the sum of Brier scores. The Brier score and the RPS are intended for discrete events, but the Brier score is also extended to continuous variables in which the sum is replaced by an integral, the continuous ranked probability score. There are also other scores in the literature. The second most common one seems to be the logarithmic score, probably first introduced by Good (1952), but for forecast verification by Raiffa (1969). This score is described later in this paper. The spherical score is also mentioned (see Bickel

2007; Bröcker 2009; Gneiting and Raftery 2007), but this score has received relative little attention or use in meteorology or in other fields (Murphy and Katz 1985).

The relation between the scores for probability forecasts and the economic value of the forecasts was first explored in the 1960s and the early 1970s (e.g., Murphy 1966, 1969; Staël von Holstein 1970). This relation is interesting, since it makes it possible to show if a verification score, at least in some sense, really measures the usefulness of the forecasts. In the paper Murphy (1993) the usefulness of the forecasts is called “type 3 goodness” or “value.” It is also possible to examine the inherent assumptions behind a particular verification score with respect to user's sensitivity of different forecast errors. Murphy (1966) discovered that the Brier score can be interpreted as a linear combination of a score based on a user with a uniform probability distribution for all cost–loss ratios between zero and one. The “cost” refers to the expenses due to a protection against certain adverse weather (AW) and the “loss” to the damage caused by AW when not protecting. The cost–loss ratio is described in more detail later. The probability distribution derived by Allan Murphy is referred to as the “expected kernel of utility,” which is an expected mean value. Another interpretation is that all users within a large group really have this uniform distribution. In that case, the score becomes the “real”

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Karl-Ivar Ivarsson, karl-ivar.ivarsson@smhi.se

mean value. This interpretation of the Brier score is adopted by Richardson (2001). Roulston and Smith (2002) showed that the logarithmic score can also be derived in a similar way, but with a slightly more complex distribution. This distribution consists of two terms. One of those terms implies a uniform density function of the user's cost.

The methods of generating a general score based on cost-loss relations in the papers mentioned earlier, all suffer from being rather complex. In this paper, the score is further refined by

- 1) Making it easier to calculate from the density function of the user's loss.
- 2) Letting it be zero for perfect forecasts and one for an incorrect 100% forecast of AW. A forecast of 0% of AW that is incorrect results in a number dependent on the structure of the density function of the user's loss.

Since the score is based on a continuous function of the user's cost-loss, it will be referred to as the continuous specific score (CSS). The derivation of CSS is described in section 2, together with the basic ideas behind the cost-loss relation, since not everyone may be familiar with it. The concept of effective cost loss ratio (ECLR) is also described. It is the ratio of the average cost to the average loss. Examples of possible usage of CSS are described in section 3. A short summary including some remarks is in section 4.

2. Derivation of the CSS

a. The basic assumptions behind the score

It is natural to start with the concept of the cost-loss relation. It is based on the assumption that the user of the forecasts has two alternatives. The first one is to protect which gives a cost C , regardless of if AW occurs or not. The second one is to not protect. Following early papers (e.g., Murphy 1977) this leads to a loss L_t in the case of AW. This loss is assumed to totally vanish in the case of protection. But in reality, the protection may rather be seen as a mitigation strategy, minimizing the impact, rather than to completely avoid the loss caused by AW. Therefore, the cost-loss concept proposed by Richardson (2000) is used here instead of the original one. If the mitigation effect of taking action is L , the net loss in the case of AW and action becomes $L_t - L$.

Acting or not acting together with AW or with no AW gives four different outcomes, which are shown in Table 1.

Consider that a probability forecast p is issued for the AW. The two expected economical outcomes are as follows: (The real outcome may be different than the expected one.)

TABLE 1. Table of the economical outcome of AW and non-AW. If a user acts, AW leads to a protection cost C regardless of if AW occurs or not. It also leads to a "residual" loss of $L_t - L$ in the case of AW. The term L is the mitigation effect of the action taken. No action leads to no cost if the forecast is correct, but leads to a loss L_t in the case of AW.

	AW happens	Non-AW happens
User acts	$C + L_t - L$	C
User does not act	L_t	0

- Protection/mitigation: $C + p(L_t - L)$.
- No protection/mitigation: $pL_t + (1 - p) \times 0 = pL_t$.

Protection/mitigation is expected to be the best alternative when it is cheaper than no action. This means that

$$C + p(L_t - L) < pL_t, \quad (1)$$

or

$$p > C/L. \quad (2)$$

The ratio C/L is the cost-loss ratio X . Users with $X < p$ are more sensitive to AW and should protect and thus receive a protection cost C , whereas the others should not, since the protection cost is too high. In the case of AW they instead get the loss L . In this case the loss becomes the missing mitigation effect, and this meaning is kept for the rest of this paper.

There are at least two issues regarding the construction of a verification score based on the cost-loss concept:

- 1) How do we know the different cost-loss relations of the users? We must know both how much money is at stake for every user, as well as the number of users for all different cost-loss relations.
- 2) If succeeding in this, how do we construct a suitable verification score based on this information?

The first issue is the hardest one and is briefly discussed in section 2d. Once the relative importance of the user's cost-loss relations is known, or just guessed, it is possible to create a verification score using a continuous loss function representing how important different cost-loss relations are.

b. What is a density function of the user's loss?

In the examples discussed below a rational behavior is assumed in order to minimize the expected expenses.

First consider a simple case with a discontinuous function of the user's loss. It is a single user with a loss of \$1,000 (U.S. dollars) and the cost-loss relation of 0.5. Thus, the protection cost is \$500. (It is possible to

create a score based on this type of discontinuous function, but it is not a topic of this paper.)

Then imagine a group of users with the same loss, \$1,000 all together, and a uniformly spread cost–loss ranging from 0.4 to 0.6. This gives uniform density function of loss of $\$1,000/(0.6-0.4) = 5000$ with the unit $\$/(C/L)$, where C/L is the cost–loss ratio. The cost–loss ratio is dimensionless. The density function of the cost increases linearly from 2000 at $X = 0.4$ to 3000 at $X = 0.6$.

The cost for the group is still \$500 since the mean density of the cost is 2500 and $2500 \times 0.2 = 500$.

c. The difference between a density function and a probability function of the user's loss

Instead of a group of users with uniform density function of loss one may consider a single user with a uniformly likelihood of having a cost–loss relation between 0.4 and 0.6. A function of her/his loss with a probability density of loss of $\$5,000/(C/L)$ leads to an expected mean loss of $\$5,000 \times (0.6-0.4) = \$1,000$. In the same way, the probability density of cost increases linearly from 2000 at $X = 0.4$ to 3000 at $X = 0.6$.

So although the meaning of the values is different, it all ends up with the same result. Which of the two concepts—a density function or a probability function of the user's loss (or cost)—is the most appropriate is related to our knowledge of this matter.

d. How to determine the density/probability function of the user's loss?

As mentioned earlier, this is not straightforward. In the paper by [Roebber and Bosart \(1996\)](#), different procedures for determining the cost–loss relations for the forecast users are described. They all require a lot of interviews with users of the forecasts about their protection costs and the loss in case of AW. Typical values found are between 0.35 and 0.5. A study by [Liljas and Murphy \(1994\)](#) suggests that ECLR is likely to be close to 0.3 in the case of a warning for storm, which is defined as mean wind exceeding 24.5 m s^{-1} .

One example where it seems less hard to determine the cost–loss relations is users of “significant” weather forecasts for aviation. Significant weather means fog, freezing rain, thunderstorms, or any other AW. A common format for forecasts for aviation purposes is the terminal area forecast (TAF) in which a limited number of probabilities for significant weather is allowed. When AW or significant weather is forecast, the route may be prolonged in order to avoid it. This costs extra fuel (=cost). In this case the route is to the nearest airport without a forecast of AW. But if AW occurs without being properly forecast, the route will be prolonged even more, because of a forced redirection to a

safer destination. The extra fuel cost for the detour is the loss.

Another way to get a hint of the costs and losses of importance for TAF is just to examine the rules for TAF. If the risk of AW is less than 30% it is normally ignored, and if it is 50% or more a categorical forecast for AW is used. The only probabilities used explicitly are 30% and 40%. From this it follows that the main density is in the range of 0.2–0.5.

e. The derivation of the score

Contrary to the papers mentioned earlier, the score derived here is based on a loss distribution that does not necessarily cover the whole range of cost–loss ratios from zero to one. It is assumed to begin at a cost–loss ratio A and end at a cost–loss ratio B , where both values are in the interval from zero to one. This also includes the possibility for discontinuous endpoints of the distributions. Probabilities outside the interval $(A:B)$ are truncated to be inside (p_i) , where $p_i = \max[A, \min(p, B)]$ since the integrals outside are zero. The index “ i ” will be explained below.

Users for which $X \leq p$ receive a protection cost. It is the integral from A to the forecast probability p_i of the density function of the user's cost $[\text{CTOT}(A, p_i)]$. If the density function of the user's loss is $F(X)$, the density function of the user's cost becomes $F(X)X$, so $\text{CTOT}(A, p_i)$ becomes

$$\text{CTOT}(A, p_i) = \int_A^{p_i} F(X)X dX. \quad (3)$$

The relation $C = XL$ is utilized in Eq. (3). In case of AW, all other users with $X > p$ are assumed not to protect. The integral of those user's loss $[\text{LTOT}(p_i)]$ goes from p_i to B :

$$\text{LTOT}(p_i) = \int_{p_i}^B F(X) dX. \quad (4)$$

A verification dataset often consists of a limited number M of different probability values (p_i) , for example, 0.0, 0.1, 0.2, $\dots$, 1.0. The economical outcome of the forecasts $[\text{Tot}(p_i, o_i)]$ becomes

$$\begin{aligned} \text{Tot}(p_i, o_i) &= o_i \text{LTOT}(p_i) + \text{CTOT}(A, p_i) \\ &= o_i \int_{p_i}^B F(X) dX + \int_A^{p_i} F(X)X dX. \end{aligned} \quad (5)$$

Here o_i is the observed frequency of AW when the probability p_i is issued. Index “ i ” may also denote the number of the particular case in the statistical sample. Then o_i is 1 in the case of AW, else 0. It may also be the relative frequency of AW for all probability forecasts of p_i . This mean value is often labeled “the sample climatology” in literature.

So far, the method described for generating a general score is similar to earlier papers. However, the relation

$$\int F(\psi)\psi d\psi = F_I(\psi)\psi - F_{II}(\psi), \quad (6)$$

is utilized here to make the final score a little bit easier to use. Here F_I is the antiderivative (or “primitive function”) to F and F_{II} is the antiderivative to F_I . The relation in Eq. (6) is derived using integration by parts. The integration now yields

$$\begin{aligned} \text{Tot}(p_i, o_i) = & o_i \underbrace{[F_I(B) - F_I(p_i)]}_{\text{loss} = \text{LTOT}(p_i), p_i \leq X} \\ & + \underbrace{p_i F_I(p_i) - AF_I(A) - F_{II}(p_i) + F_{II}(A)}_{\text{cost} = \text{CTOT}(A, p_i), p_i > X}. \end{aligned} \quad (7) \quad \text{or}$$

A perfect forecast, which means predicting the correct event with a probability of 100%, results in a protection cost of $\text{CTOT}(A, B)$ when AW occurs. To get a verification score that is zero for perfect forecasts, $\text{CTOT}(A, B) = BF_I(B) - AF_I(A) - F_{II}(B) + F_{II}(A)$ is omitted in Eq. (7) in the case of AW. Second, a division with $\text{CTOT}(A, B)$ makes the score more handy, since 100% probability for AW when AW does not occur gives the score one. (There are exceptions when this division should be avoided due to numerical reasons.) The result of those modifications is the following score $\text{CSS}(p_i, o_i, A, B)$:

$$\begin{aligned} \text{CSS}(p_i, o_i, A, B) \\ = \frac{o_i \text{LTOT}(p_i) + \text{CTOT}(A, p_i) - o_i \text{CTOT}(A, B)}{\text{CTOT}(A, B)}, \end{aligned} \quad (8)$$

$$\frac{(p_i - o_i)F_I(p_i) - F_{II}(p_i) + F_{II}(A) - AF_I(A) + o_i[F_{II}(B) - F_{II}(A) + F_I(B)(1 - B) + AF_I(A)]}{BF_I(B) - AF_I(A) - F_{II}(B) + F_{II}(A)}. \quad (9)$$

Here, $A = 0$ and $B = 1$ yields

$$\text{CSS}(p_i, o_i, 0, 1) = \frac{(p_i - o_i)F_I(p_i) - F_{II}(p_i) + F_{II}(0) + o_i[F_{II}(1) - F_{II}(0)]}{F_I(1) - F_{II}(1) + F_{II}(0)}. \quad (10)$$

Failure of predicting AW $\text{CSS}(0, 1, A, B)$ leads to the following result:

$$\begin{aligned} \text{CSS}(0, 1, A, B) \\ = \frac{\text{LTOT}(A)}{\text{CTOT}(A, B)} - 1 \\ = \frac{F_I(B) - F_I(A)}{BF_I(B) - AF_I(A) - F_{II}(B) + F_{II}(A)} - 1. \end{aligned} \quad (11)$$

This is the inverted value of the ratio of the average cost to the average loss minus one. This ratio is furthermore referred to as the *effective cost-loss ratio*, ECLR in this paper. It can be written as

$$\begin{aligned} \text{ECLR} &= \frac{\int_A^B F(X)X dX}{\int_A^B F(X) dX} \\ &= \frac{BF_I(B) - AF_I(A) - F_{II}(B) + F_{II}(A)}{F_I(B) - F_I(A)}. \end{aligned} \quad (12)$$

ECLR becomes lower when low cost-loss ratios have larger loss density functions than higher ones and vice versa. Thus, ECLR measures how important low cost-loss ratios are compared to higher ones. The whole

group of users with a particular ECLR also responds in the same way as one single user having a cost-loss ratio of ECLR in the case of categorical forecasts (but not if the forecasts are probabilistic).

CSS can be regarded as the total relative expense of a group of users with the sensitivity of AW represented by the loss-density function F . An alternative interpretation is that CSS is the expected relative expense of a single user with a probability function of F for the various loss values for all possible cost-loss relations.

Assuming a forecast sample with a finite number of probability values p_i and taking the sum of all forecasts, the mean of the score becomes

$$\overline{\text{CSS}(p_i, o_i, A, B)} = \sum_{i=1}^M \frac{n_i}{N} \text{CSS}(p_i, o_i, A, B), \quad (13)$$

where N is the total number of forecasts and M is the number of different probabilities in the verification dataset. The number of cases for each subset is denoted n_i .

Similar to Roulston and Smith (2002) and Richardson (2001), the function is based on the loss for different

cost–loss ratios X . In [Stäel von Holstein \(1970\)](#) and [Murphy \(1969\)](#), the utility is used instead. It is basically a linear transformation in which every value α_{ij} in the 2×2 matrix in [Table 1](#) is recalculated by the transformation $(L - \alpha_{ij})/L$. This makes the score positively oriented but is in other respects similar. Here, the scores are negatively oriented, which means that lower values are better than higher ones.

f. Proof that CSS is a strictly proper scoring rule

A strictly proper scoring rule does not support “hedging,” which means setting a probability value different from one’s true belief in order to benefit from a better expected score ([Stäel von Holstein 1970](#); [Murphy 1969,1973,1977](#)). Following [Murphy \(1993\)](#), a strictly proper score encourages the forecaster to issue “consistent” forecasts.

Using [Eq. \(9\)](#), but for simplicity omitting division with the constant positive term

$$BF_I(B) - AF_I(A) - F_{II}(B) + F_{II}(A), \quad (14)$$

and the derivation with respect to p_i becomes

$$\begin{aligned} d[\text{CSS}(p_i, o_i, A, B)]/d(p_i) \\ = (p_i - o_i)F(p_i) + F_I(p_i) - F_{II}(p_i) = (p_i - o_i)F(p_i). \end{aligned} \quad (15)$$

The extreme value is found when $(p_i - o_i)F(p_i) = 0$, which means that $p_i = o_i$. The second derivation of $\text{CSS}(p_i)$ is $(p_i - o_i)F_{-1}(p_i) + F(p_i)$, where $F_{-1}(p_i)$ is the derivative to $F(p_i)$. For $p_i = o_i$ the result is just $F(p_i)$. If $F(p_i) > 0$ and $p_i = o_i$, it is a minimum value and thus $\text{CSS}(p_i, o_i)$ is by definition a strictly proper scoring rule. But this is of course only valid in the range $(A:B)$. Outside this range $F(p_i)$ is zero, and the score is only proper since hedging outside this range does not affect the user’s decision-making nor the value of score.

g. Redefinition of the resolution and the reliability terms

In the early 1970s, Allan Murphy showed that the Brier score could be decomposed into three different terms: reliability, resolution, and a term that represents the result of a climatological forecast ([Murphy 1973](#)). Reliability measures the degree of correspondence between the forecast and the observed relative frequency. Resolution measures the ability of the forecasts to separate the low risk situations of AW from the ones with high risk. This decomposition is also possible with CSS and is easiest by adopting the technique proposed by [Siegert \(2017\)](#).

Following [Siegert \(2017\)](#), [Eq. \(13\)](#) is transformed to two differences and a remaining term. Those three terms become the reliability, the resolution, and the climatological forecast:

$$\begin{aligned} \sum_{i=1}^M \frac{n_i}{N} \left\{ \underbrace{\text{CSS}(p_i, o_i, A, B) - \text{CSS}(o_{ip}, o_i, A, B)}_{\text{REL}} \right. \\ \left. - \underbrace{\text{CSS}(c, o_i, A, B) + \text{CSS}(o_{ip}, o_i, A, B)}_{-\text{RES}} \right. \\ \left. + \underbrace{\text{CSS}(c, o_i, A, B)}_{\text{CLIM}} \right\}. \end{aligned} \quad (16)$$

Here, c is the “true” climate or the “sample” climatology. REL is the reliability term, RES is the resolution term, and CLIM is the result of a climatological forecast. Notice that o_{ip} is a forecast so $o_{ip} = \max[A, \min(o_i, B)]$.

3. Examples of possible usage of CSS

a. Some general remarks

This score may be used in at least four different ways:

- Examining what kind of cost or loss density functions correspond to the existing probability scores.
- Creating new probability scores.
- Examining how the behavior of the score is dependent on the loss density functions.
- Using the new scores for real-time verification.

A few examples of those different usages are presented in the following subsections.

b. The loss density functions consistent with existing probability scores

The Brier score, the logarithmic score, and the spherical score may all be derived from different assumptions of the loss density function.

It is straightforward to start with the simplest possible assumption: a uniform distribution as of users with different cost–loss ratios, $F = 1$.

Omitting the constants in the integration since they do not change the final result, the functions become

- $F_I = X$, and
- $F_{II} = (1/2)X^2$.

With [Eqs. \(10\)](#) and [\(13\)](#) together with $A = 0$ and $B = 1$, CSS becomes

TABLE 2. Table of some parameters related to the scores derived by different choices of loss density functions. Note that both the logarithmic score and the spherical score are positively oriented, but CSS is negatively oriented. This means that the result becomes $-(\text{logarithmic score})$ and $1 - (\text{spherical score})$, respectively. The division with $\text{CTOT}(0, 1)$ is omitted for the logarithmic score for numerical reasons, which also leads to $\text{CSS}(1, 0)$ becoming different from one. $\text{CSS}(\bar{o}_i, \bar{o}_i)$ is sometimes referred to as the “uncertainty.”

Parameter	Brier	Logarithmic	Spherical
F	1	$1/X + 1/(1 - X)$	$[X^2 + (1 - X)^2]^{-3/2}$
F_I	X	$\ln X - \ln(1 - X)$	$(2X - 1)[X^2 + (1 - X)^2]^{-1/2}$
F_{II}	$\frac{1}{2}X^2$	$X \ln X - (1 - X) \ln(1 - X) - 2X$	$[X^2 + (1 - X)^2]^{1/2}$
$\text{CSS}(p_i, o_i)$	$(p_i - o_i)^2 + o_i(1 - o_i)$	$-o_i \ln(p_i) - (1 - o_i) \ln(1 - p_i)$	$1 - [o_i p_i + (1 - o_i)(1 - p_i)][p_i^2 + (1 - p_i)^2]^{-1/2}$
$\text{CSS}(\bar{o}_i, \bar{o}_i)$	$\bar{o}_i(1 - \bar{o}_i)$	$-\bar{o}_i \ln(\bar{o}_i) - (1 - \bar{o}_i) \ln(1 - \bar{o}_i)$	$1 - [\bar{o}_i^2 + (1 - \bar{o}_i)^2]^{1/2}$
$\text{CSS}(0, 1)$	1	∞	1
$\text{CSS}(1, 0)$	1	∞	1
$\text{CSS}(0, 0)$	0	0	0
$\text{CSS}(1, 1)$	0	0	0

$$\begin{aligned}
 & \sum_{i=1}^M \frac{n_i}{N} \frac{p_i p_i - \frac{1}{2} p_i^2 + 0 + o_i \left(\frac{1}{2} - 0 - p_i \right)}{1 - \frac{1}{2} + 0} \\
 &= \sum_{i=1}^M \frac{n_i}{N} [(p_i - o_i)^2 + o_i(1 - o_i)]. \quad (17)
 \end{aligned}$$

For $o_i = 0$, Eq. (17) becomes $(p_i - 0)^2$ and for $o_i = 1$, it is $(p_i - 1)^2$. Both are identical to the Brier score.

The functions for the logarithmic score and the spherical score are handled in a similar way. The loss functions for all three scores are in Table 2.

The decomposition of CSS into three terms is exemplified by $F = 1$ (Brier score). The reliability is

$$\begin{aligned}
 & \sum_{i=1}^M \frac{n_i}{N} [\text{CSS}(p_i, o_i) - \text{CSS}(o_i, o_i)] \\
 &= \sum_{i=1}^M \frac{n_i}{N} (p_i - o_i)^2 = \sum_{i=1}^M \frac{n_i}{N} \text{RELB}(p_i, o_i). \quad (18)
 \end{aligned}$$

Here, $o_{ip} = o_i$ since $A = 0$ and $B = 1$.

$\text{RELB}(c, o_i)$ is the reliability term for the Brier score. Similarly, the resolution is also identical to that of the Brier score:

$$\begin{aligned}
 & \sum_{i=1}^M \frac{n_i}{N} [\text{CSS}(c, o_i) - \text{CSS}(o_i, o_i)] \\
 &= \sum_{i=1}^M \frac{n_i}{N} (c - o_i)^2 = \sum_{i=1}^M \frac{n_i}{N} \text{RESB}(c, o_i). \quad (19)
 \end{aligned}$$

The “climate” term becomes $(c - \bar{o}_i)^2 + (1 - \bar{o}_i)\bar{o}_i$, where $\bar{o}_i$ is the sample climatology. The climate term is easily derived by using Eqs. (9) or (10) together with Eq. (13) but using $(c, \bar{o}_i)$ instead of (p_i, o_i) . If $c = \bar{o}_i$, the result is the “uncertainty” term (UNSB) $(1 - \bar{o}_i)\bar{o}_i$,

which measures the spread or uncertainty of the observations.

c. Some general comments about choosing a suitable density function

There are many different types of functions that may be considered for describing the user’s loss, or the uncertainty of the loss for a particular user. Based on the studies mentioned in section 2d it is tentative to propose the following for choosing a density function:

- 1) Because of the uncertainty of this matter, it is less likely that highly nonlinear functions such as exponential functions or gamma functions should fit well with our present knowledge. Smoothed functions should be considered in the first place.
- 2) For practical reasons one should use functions that are easy for making an antiderivative. Avoid complicated functions if possible.

In this study, only simple functions will be considered. Bimodal or multimodal functions may be of

TABLE 3. Table of some parameters for the AS.

Parameter	Asymmetric score
F	$1 - X$
F_I	$-\frac{1}{2}(1 - X)^2$
F_{II}	$\frac{1}{6}(1 - X)^3$
$\text{CSS}(p_i, o_i)$	$(p_i - o_i)^2(3 - 2p_i - o_i) + o_i(1 - o_i)(2 - o_i)$
$\text{CSS}(\bar{o}_i, \bar{o}_i)$	$\bar{o}_i(1 - \bar{o}_i)(2 - \bar{o}_i)$
$\text{CSS}(0, 1)$	2
$\text{CSS}(1, 0)$	1
$\text{CSS}(0, 0)$	0
$\text{CSS}(1, 1)$	0

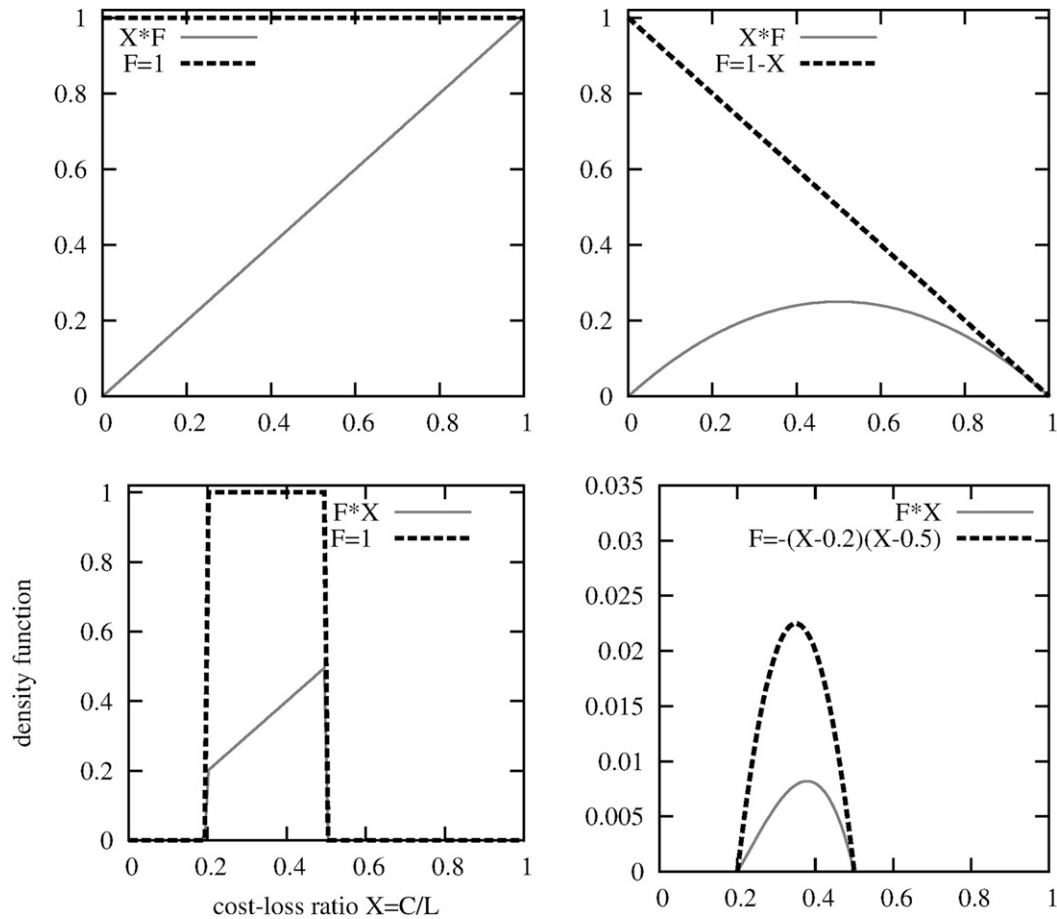


FIG. 1. Diagrams showing the density functions for the loss with dashed lines and corresponding cost functions with thinner solid lines. Density is on the vertical axis, and the cost-loss relation is on the horizontal axis. (top left) Brier score (top right) AS, (bottom left) LIN, and (bottom right) PAR.

interest for special considerations but are not a topic of this paper.

d. Creating a new probability score: An asymmetric score

The Brier score is the result of CSS with a uniform distribution of the loss density function. This leads to an ECLR of 0.5. This may be reasonable in many cases, especially when the knowledge of the user's cost-loss ratio is poor. If the score is assumed to be closer to the expected user's cost and loss distributions, the score measures the "type 3 goodness" or "value" in a more precise way (Murphy 1993; Richardson 2001). This is of particular interest when verifying severe weather events. The studies mentioned earlier support that most of the users have cost-loss ratios below 0.5. If the goal is to create a score that is closer to the findings discussed here, but without being too specific, one may consider the asymmetric score (AS) described below.

It is based on a distribution in which low cost-loss ratios are more important than higher ones, but all densities are still in the whole range from zero to (almost) one. The distribution chosen is $1 - X$, which leads

TABLE 4. Table of probability values, corresponding number of AW [>35 mm (12 h) $^{-1}$], and number of forecasts for each forecast probability.

Probability	Observations > 35 mm	Cases
0.0	54	153 582
0.1	16	284
0.2	11	79
0.3	5	23
0.4	12	20
0.5	9	16
0.6	3	7
0.7	4	7
0.8	7	9
0.9	6	8
1.0	4	5
Sum	131	154 040

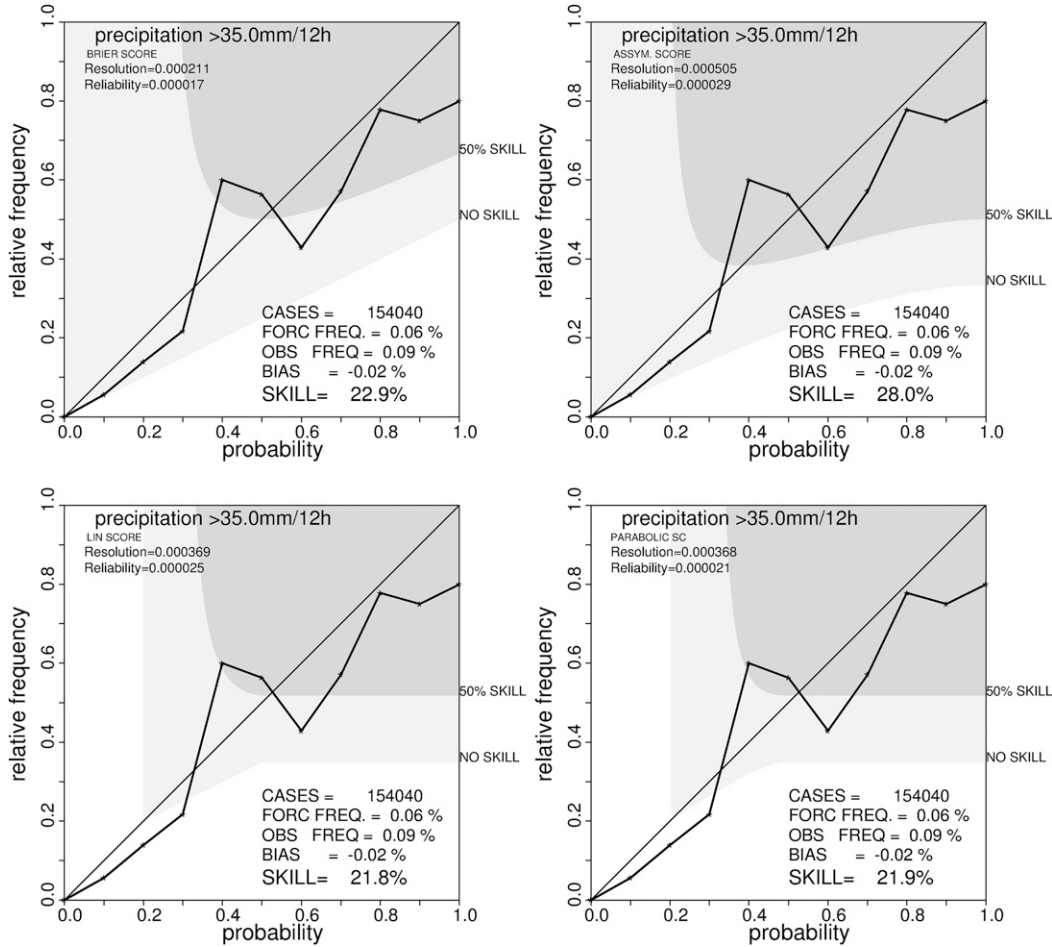


FIG. 2. Reliability diagram for forecasts of $35 \text{ mm (12 h)}^{-1}$ or more for four different scores. The number of cases for each probability value is given in Table 4. The shading in the diagram is based on calculation of the skill score for a large number of pairs (p_i, o_i) covering the area in the figure. The skill scores are derived as in Eq. (23), which means that the “sample” climatology is used as reference. The border between light gray and white is the “no skill line,” and darker gray shadings mean more than 50% skill. Ideally, the probabilities and the observed frequencies should match, and this is shown as a 45° black line. (top left) Brier score, (top right) AS, (bottom left) linear score, and (bottom right) parabolic score.

to an ECLR of 1/3. The features of the new score are seen in Table 3, and the density function of the cost and loss is in Fig. 1 (top-right panel). The corresponding density functions for the Brier score are in the top-left panel. (The density functions in the bottom panels are discussed later.)

The AS has so far only been used occasionally at the Swedish Meteorological and Hydrological Institute (SMHI) and has not been publicized in any scientific paper.

The reliability for the AS (RELAS) becomes

$$\begin{aligned} \text{RELAS} &= \sum_{i=1}^M \frac{n_i}{N} (p_i - o_i)^2 (3 - 2p_i - o_i) \\ &= \sum_{i=1}^M \frac{n_i}{N} \text{RELB}(p_i, o_i) (3 - 2p_i - o_i). \end{aligned} \quad (20)$$

In RELAS the term $\text{RELB}(p_i, o_i)$ is multiplied by an “asymmetrical factor” $3 - 2p_i - o_i$, which is close to zero when both p_i and o_i are close to one, reflecting that the loss density function is close to zero for X being near one. The opposite is found when p_i and o_i are close to zero. Then this factor is near to 3, which is a result of a high density function when X is close to zero. This implies the enhanced importance of users with low cost-loss ratios. The reason for p_i being multiplied with a factor of 2, whereas o_i is not, is that variation of o_i only leads to a (linear) increase/decrease of LTOT [see e.g., Eq. (5)] whereas varying p_i means varying the density function itself as well as varying the part of it being used.

The resolution (RESAS) is similar to RELAS:

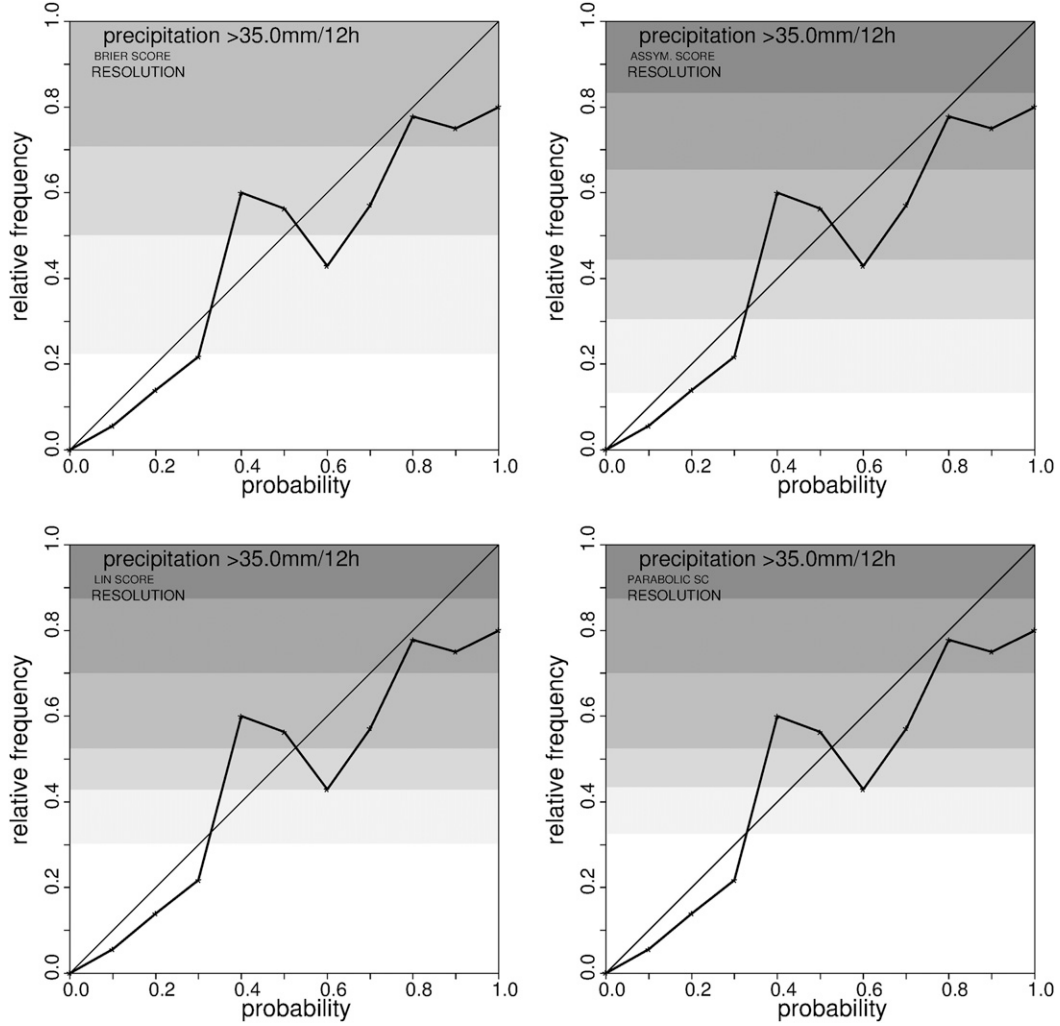


FIG. 3. Diagrams showing the resolution for different combination of forecasts and observed relative frequencies. The relative frequencies for the 35-mm precipitation forecasts are included. Darker shading means higher resolution values (higher resolution). The shadings are following the thresholds: 0.01, 0.25, 0.5, 1.0, and above 1.5 (=darkest) No shading for resolution values below 0.01. (top left) Brier score, (top right) AS, (bottom left) linear score, and (bottom right) parabolic score.

$$\text{RESAS} = \sum_{i=1}^M \frac{n_i}{N} \text{RESB}(c, o_i)(3 - 2c - o_i). \quad (21)$$

The sample climatology or uncertainty for the AS (UNCAS) is

$$\text{UNCAS} = (2 - \bar{o}_i)(1 - \bar{o}_i)\bar{o}_i = \text{UNSB}(2 - \bar{o}_i) \quad (22)$$

and can be interpreted as a weighted uncertainty, in which variations near zero have higher weights than near unity.

e. A comparison between the Brier score and the AS

As mentioned in the beginning of this section, CSS makes it possible to examine how the result of a score is

altered when changing the loss density functions. In this case, the comparison is between a score based on a uniform distribution of loss, the Brier score, and a score based on a larger loss function for low cost-loss ratios, the AS.

The 12-h precipitation forecasts for autumn 2017 with the limited area model ensemble system used at Norway, Sweden, and Finland are used as example. The threshold selected is 35 mm. It is one of the thresholds for issuing a severe warning at SMHI. The outcome of the verification is seen in Table 4.

The result with the Brier score is seen in Fig. 2 in the top-left panel and the result with the AS in the top-right panel.

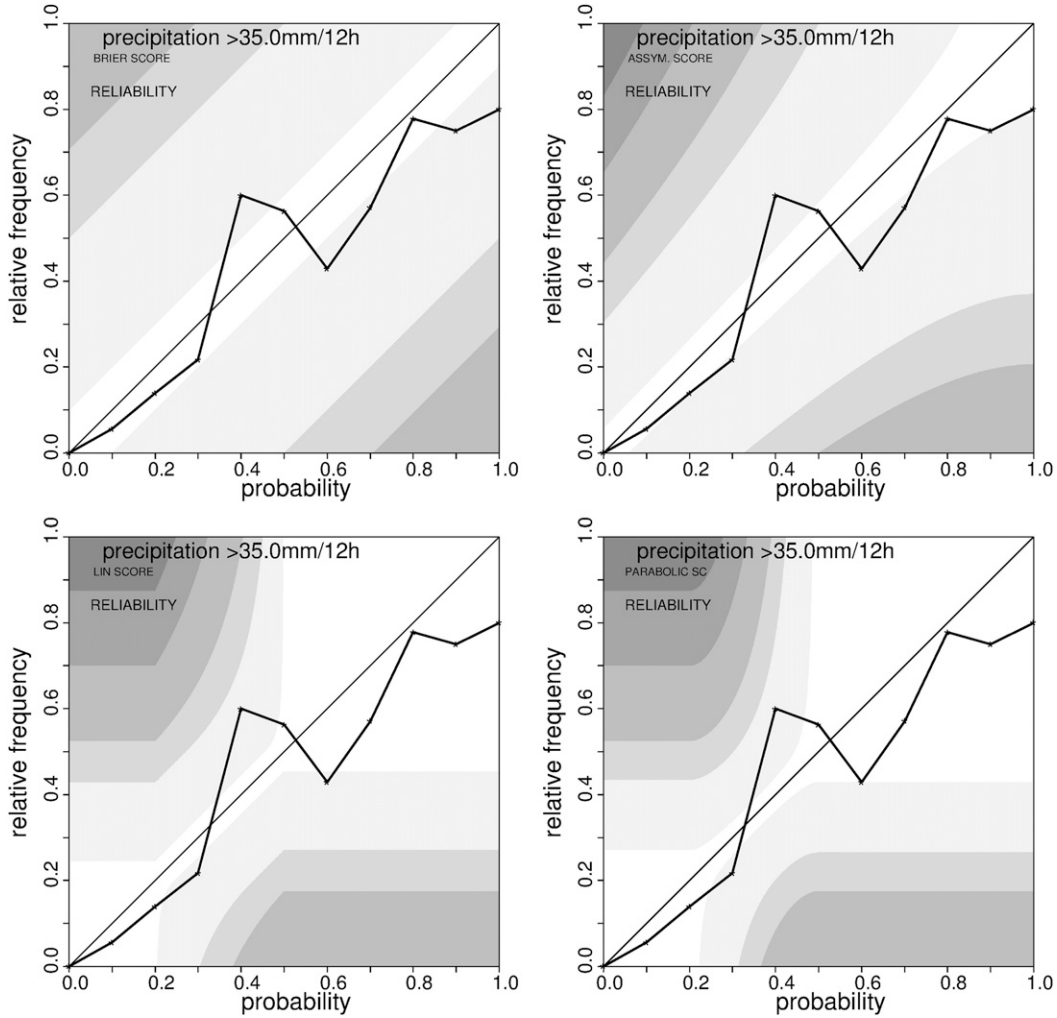


FIG. 4. Diagrams showing the reliability for different combination of forecasts and observed relative frequencies. The relative frequencies for the 35-mm precipitation forecasts are included. Darker shading means higher reliability values (less reliable). The shadings are following the thresholds: 0.05, 0.25, 0.5, 1.0, and above 1.5 (=darkest) No shading for reliability values below 0.05. (top left) Brier score, (top right) AS, (bottom left) linear score, and (bottom right) parabolic score.

The shading of the areas with different skill in figures in this paper follows some of the suggestions for an “attributes diagram” in Hsu and Murphy (1986) and Mason (2004).

Although the results are rather similar, one clear difference is that the skill is higher with the AS. Recalling that any skill score (SS) with the reference

forecast being the sample climatology (UNC) can be derived as

$$SS = \frac{(\text{REL} - \text{RES} + \text{UNC}) - \text{UNC}}{0 - \text{UNC}} = \frac{\text{RES} - \text{REL}}{\text{UNC}}, \quad (23)$$

and the asymmetric skill score becomes

$$\text{ASS} = \sum_{i=1}^M n_i \frac{[\text{RESB}(\bar{o}_i, o_i)(3 - 2\bar{o}_i - o_i) - \text{RELB}(p_i, o_i)(3 - 2p_i - o_i)]}{N \times \text{UNSB}(2 - \bar{o}_i)}. \quad (24)$$

Equations (20)–(22) have been utilized here. Equation (24) gives a hint of what causes the difference between

the Brier skill score and that of the AS. The main reason is that the resolution term is higher with the asymmetric

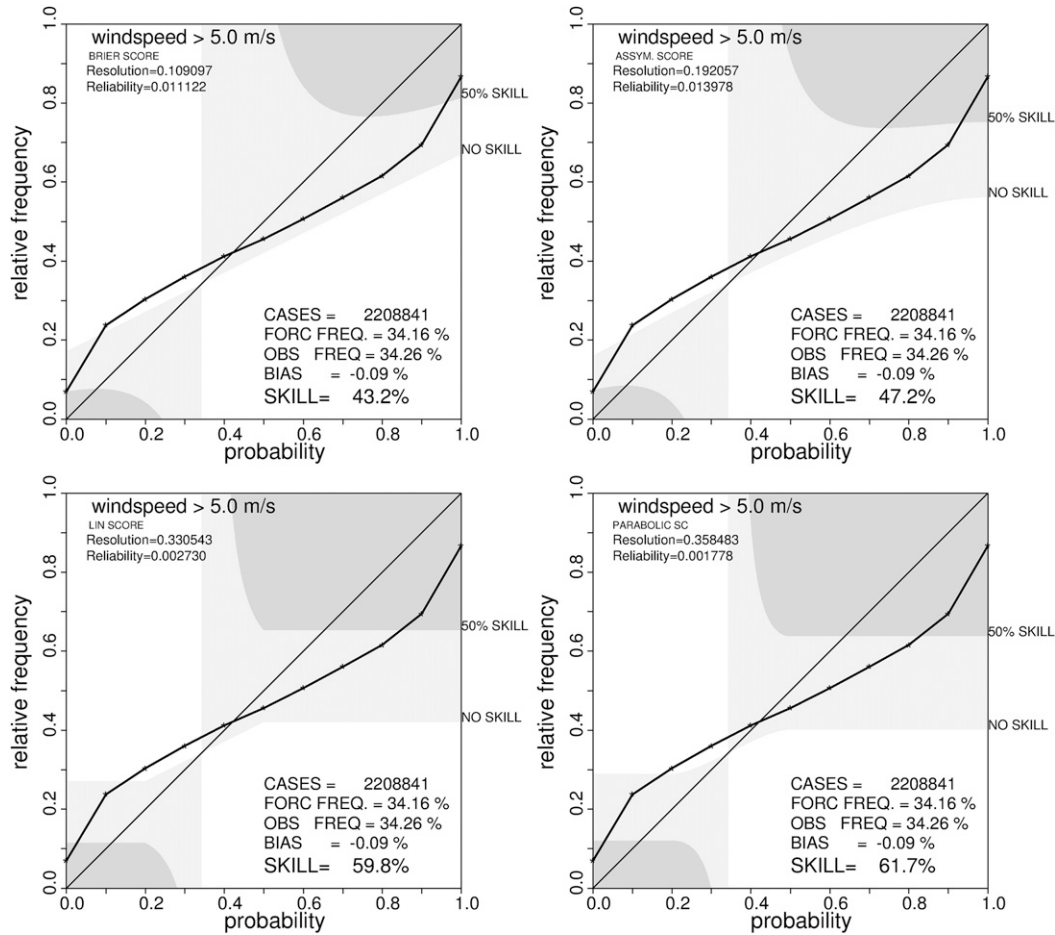


FIG. 5. Reliability diagram for forecasts of 5 m s^{-1} wind speed or more. See Fig. 2 for details.

skill score. This is pronounced when the probabilities are of the order of 0.5 and the forecasts are by and large reliable. Since RELB is small here, the term $3 - 2\bar{o}_i - o_i$ is around 2.5, whereas $(2 - \bar{o}_i)$ is about 1.5. In this case the skill increases with a factor of about $2.5/1.5 = 5/3$. This is also visible in the top panels of Fig. 3, in which the resolution for the two scores is shown. Darker shading means higher resolution, and it is higher for the AS when the relative frequency is high. This also reflects a possible higher value of the forecasts in this particular verification dataset if users with lower cost-loss relations are more important.

In the top panels of Fig. 4, the reliability for BS and AS is seen. The top left is the Brier score and to the right is the AS. The AS has a broader area of low values of reliability for high relative frequencies because of the asymmetric term in RELAS. It is consistent with the idea that lack of reliability is less harmful when it affects users with low loss densities.

In the top panels of Fig. 5, the Brier score and AS are compared, but for ensemble forecasts of 10-m wind

speed exceeding 5 m s^{-1} . The verification period is December 2018–February 2019. This example is selected because 5 m s^{-1} is an ordinary wind speed so the sample climatology is higher and the probability values are more evenly spread, (Table 5). The forecasts are

TABLE 5. Table of probability values, corresponding number of AW ($> 5 \text{ m s}^{-1}$), and number of forecasts for each forecast probability.

Probability	Observations $> 5 \text{ m s}^{-1}$	Cases
0.0	80 217	1 141 337
0.1	28 113	118 096
0.2	22 295	73 361
0.3	20 514	56 906
0.4	20 221	49 032
0.5	20 941	45 889
0.6	24 016	47 357
0.7	28 971	51 634
0.8	38 024	61 725
0.9	61 898	89 066
1.0	411 522	474 438
Sum	756 732	2 208 841

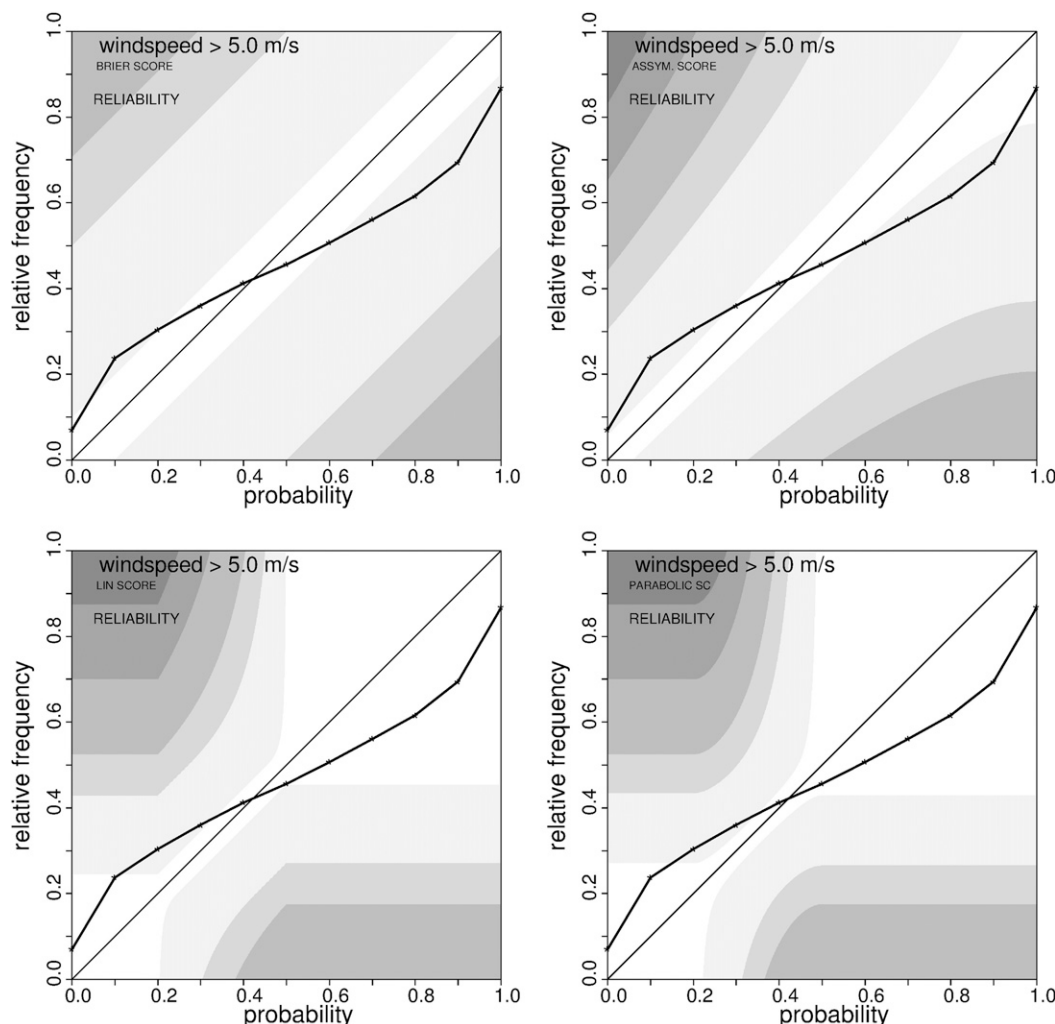


FIG. 6. Diagrams showing the reliability for different combination of forecasts and observed relative frequencies. The relative frequencies for the 5 m s^{-1} wind speed forecasts are included. In other respects the diagrams are similar to Fig. 4. (The shading is actually identical, since reliability is independent of the value of the climatological forecast.)

somewhat overconfident since the observed frequency is generally too high for low probabilities and vice versa. Also here, the skill is higher with the AS, but the relative difference is not great. One reason for the difference is that the lack of reliability is somewhat larger for higher probabilities, which has less impact with the AS (Fig. 6, top-right panel).

There are 10 members only in this ensemble system. In the papers by Ferro et al. (2008) and Ferro (2014) it is shown that the scores should be corrected, accounting for the low number of ensembles. This has not been done here but is kept in mind for further development of the verification system.

f. The result when using truncated density functions

In section 2d different studies indicate that the density of the user's loss (and /or cost) is likely to be below 0.5. The AS is consistent with this, but may be regarded as an

alternative when the knowledge of the user's cost end loss is still rather uncertain, since the whole range of cost-loss ratios from zero to (almost) one is assumed. This score also corresponds to a situation when one knows that the distribution is likely to be that broad.

To analyze the effect of using truncated density functions, two alternatives will be compared with each other, as well as with the Brier score and the AS. Based from the studies discussed above the values $A = 0.2$ and $B = 0.5$ are selected, so that below 0.2 and above 0.5 the density function is zero. Now consider two alternatives:

- 1) A uniform loss distribution between A and B as for the Brier score, called the linear score (LIN).
- 2) A parabolic distribution that becomes zero at A and B , called the parabolic score (PAR). The distribution is

TABLE 6. Table of some parameters for LIN and PAR.

Parameter	LIN	PAR
F	1	$-(X - A)(X - B)$
F_I	X	$-\frac{1}{6}(X - A)^2(2X - 3B + A)$
F_{II}	$\frac{1}{2}X^2$	$-\frac{1}{12}(X - A)^3(X - 2B + A)$
$CSS(p_i, o_i)$	$\frac{100}{21}[(p_i - o_i)^2 + o_i(1 - o_i)] - o_i - \frac{4}{21}$	$\frac{10000}{189}[-(p_i - 0.2)^2(4P_i - 2.6)(p_i - o_i) + (p_i - 0.2)^3(p_i - 0.8)] - \frac{13}{7}o_i$
$CSS(\bar{o}_i, \bar{o}_i)$	$\frac{100}{21}[(\bar{o}_i - \bar{o}_i)^2 + \bar{o}_i(1 - \bar{o}_i)] - \bar{o}_i - \frac{4}{21}$	$\frac{10000}{189}[-(\bar{o}_i - 0.2)^2(4\bar{o}_i - 2.6)(\bar{o}_i - \bar{o}_i) + (\bar{o}_i - 0.2)^3(\bar{o}_i - 0.8)] - \frac{13}{7}\bar{o}_i$
$CSS(0, 1)$	$\frac{13}{7}$	$\frac{13}{7}$
$CSS(1, 0)$	1	1
$CSS(0, 0)$	0	0
$CSS(1, 1)$	0	0

$-(X - A)(X - B)$, so the maximum is at $(A + B)/2 = 0.35$ with the value 0.0225.

None of those scores have been published before and are used here as examples of what is possible to create with the CSS. The distributions of cost and loss are seen in the bottom panels of Fig. 1. The LIN is to the left and the PAR is to the right. In Table 6 the features of the two scores are summarized. The PAR is a rather complex score, but the LIN is simpler and has resemblances with the Brier score, which is expected since both are based on a uniform loss distribution.

The result for the 35-mm precipitation forecasts is seen in the bottom panels of Fig. 2. The skill is a little lower compared to that of the Brier score and the AS. The reason is that the focus for the LIN and the PAR is the cost–loss range (0.2:0.5), and the reliability is a little lower here. Another reason is lower resolution for probabilities of 10%, 20%, and 30% (Fig. 3, lower panel) compared to the Brier score and the AS. Numerically, the lower resolution for the LIN compared to the Brier score is explained by examining the resolution for the LIN (RESLIN):

$$\begin{aligned} \text{RESLIN} &= \sum_{i=1}^M \frac{n_i}{N} \frac{100}{21} [(c - o_i)^2 - (o_{ip} - o_i)^2] \\ &= \sum_{i=1}^M \frac{n_i}{N} \frac{100}{21} [\text{RESB}(c, o_i) - (o_{ip} - o_i)^2]. \end{aligned} \quad (25)$$

Below 0.2 ($=A$) both c and o_{ip} are truncated to 0.2 so RESLIN becomes zero, but RESB has no such compensating term, so RESB is nonzero. A similar compensating term is also present for the resolution with the PAR (not shown).

The result for wind speed exceeding 5 m s^{-1} is quite different, Fig. 5. The skill is considerable higher with LIN and with the PAR even more compared to, for

example, the Brier score. The reason is high reliability in the range (0.2:0.5) and especially at 0.4 that is close to the maximum of the cost and loss density functions for the PAR (Fig. 1). The less good reliability for probabilities near 10% and above 60% (Fig. 6) reduce the Brier skill score and the asymmetric skill score, whereas the effect for the reliability with the LIN and the PAR is less since it is outside the cost–loss range (0.2:0.5). The resolution for the different scores is seen in Fig. 7.

4. A short summary and some remarks

A new verification score for probability forecasts, the continuous specific score (CSS) is presented. It is based on a continuous density function of users with different cost–loss relations. The benefit with CSS is that this score makes it possible to examine how different assumptions about the density function of the loss (or cost) affect probability scores. Another benefit is that it is fairly easy to construct new verification scores, based on the knowledge of the user's loss density functions. One example of this is an asymmetric score, which is a possible alternative to the Brier score for users with presumably low cost–loss ratios. Since the score is more user oriented, CSS has better reliability and resolution properties than the Brier score.

The scores LIN and PAR are also possible outcomes from CSS. Here, all cost–loss relations are assumed to be in the range (0.2:0.5), which is consistent with many studies and with the rules for TAF. The distribution function for the PAR is continuous at the zero points, which may be appealing, but the nonlinearity makes the score numerically complicated. The LIN is simpler and resembles the Brier score.

The tests discussed here show the importance of reliable probabilities within the range of cost–loss ratios

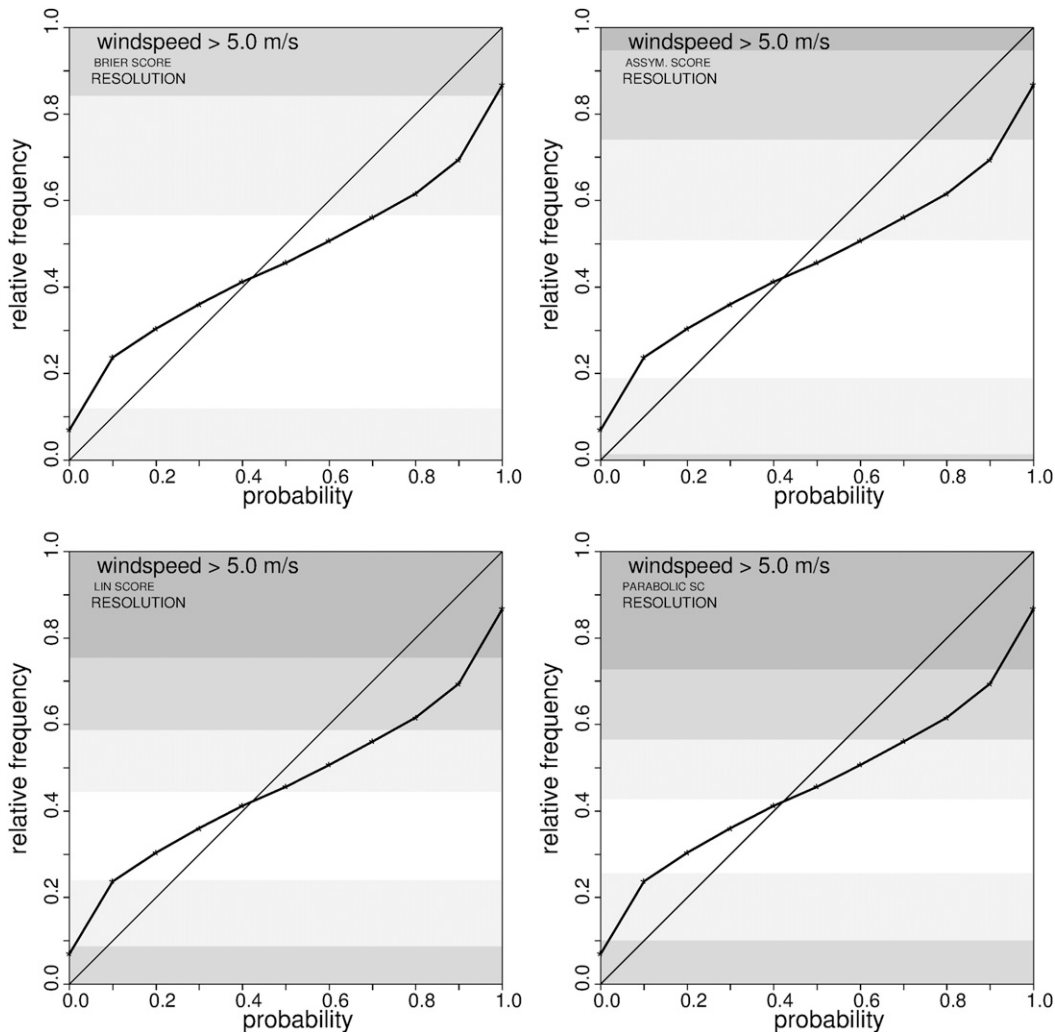


FIG. 7. Diagrams showing the resolution for different combination of forecasts and observed relative frequencies. In other respects the diagrams are similar to Fig. 3.

corresponding to the largest loss density functions. Another important aspect is that forecasts should have a resolution large enough to distinguish between the climatological probability and the risks that lies at least in the range of cost-loss ratios with the largest loss density functions.

This study is restricted to binary probability forecasts, but it is easy to extend the CSS to a larger number of probabilities by using different thresholds for the weather parameter of interest and then adding it all together. It is shown by Staël von Holstein (1970) that such a procedure makes the score sensitive to distance. Using this technique for CSS and $F = 1$ (Brier score) leads to the ranked probability score.

Acknowledgments. The author wishes to thank the three anonymous reviewers for their constructive suggestions for improving this manuscript.

REFERENCES

- Bickel, J. E., 2007: Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decis. Anal.*, **4**, 49–65, <https://doi.org/10.1287/deca.1070.0089>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Bröcker, J., 2009: Reliability, sufficiency, and the decomposition of proper scores. *Quart. J. Roy. Meteor. Soc.*, **135**, 1512–1519, <https://doi.org/10.1002/qj.456>.
- Ferro, C. A. T., 2014: Fair scores for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **140**, 1917–1923, <https://doi.org/10.1002/qj.2270>.
- , D. Richardson, and A. P. Viegel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability score. *Meteor. Appl.*, **15**, 19–24, <https://doi.org/10.1002/met.45>.
- Gneiting, T., and A. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc. B*, **14**, 107–114, <https://doi.org/10.1111/J.2517-6161.1952.TB00104.X>.

- Hsu, W. R., and A. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, [https://doi.org/10.1016/0169-2070\(86\)90048-8](https://doi.org/10.1016/0169-2070(86)90048-8).
- Liljas, E., and A. Murphy, 1994: Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bull. Amer. Meteor. Soc.*, **75**, 1227–1236, [https://doi.org/10.1175/1520-0477\(1994\)075<1227:AAHEPO>2.0.CO;2](https://doi.org/10.1175/1520-0477(1994)075<1227:AAHEPO>2.0.CO;2).
- Mason, S. J., 2004: On using “climatology” as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891–1895, [https://doi.org/10.1175/1520-0493\(2004\)132<1891:OUCAAR>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2).
- Murphy, A., 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. Appl. Meteor.*, **5**, 534–537, [https://doi.org/10.1175/1520-0450\(1966\)005<0534:ANOTUO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1966)005<0534:ANOTUO>2.0.CO;2).
- , 1969: Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which the knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteor.*, **8**, 863–873, [https://doi.org/10.1175/1520-0450\(1969\)008<0863:MOTUOP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1969)008<0863:MOTUOP>2.0.CO;2).
- , 1973: Hedging and skill scores for probability forecasts. *J. Appl. Meteor.*, **12**, 215–223, [https://doi.org/10.1175/1520-0450\(1973\)012<0215:HASSFP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0215:HASSFP>2.0.CO;2).
- , 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816, [https://doi.org/10.1175/1520-0493\(1977\)105<0803:TVOCCA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2).
- , 1993: What is it a good forecasting? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- , and R. W. Katz, Eds., 1985: *Probability, Statistics, and Decision Making in the Atmospheric Sciences*. Westview Press, 545 pp.
- Raiffa, H., 1969: Assessment of probabilities. Harvard University, School of Business, Cambridge, MA, 23 pp.
- Richardson, D., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, <https://doi.org/10.1002/qj.49712656313>.
- , 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of the ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, <https://doi.org/10.1002/qj.49712757715>.
- Roebber, P. J., and L. Bosart, 1996: The complex relationship between forecast skill and forecast value: A real world analysis. *Wea. Forecasting*, **11**, 544–559, [https://doi.org/10.1175/1520-0434\(1996\)011<0544:TCRBFS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1996)011<0544:TCRBFS>2.0.CO;2).
- Roulston, M., and L. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, [https://doi.org/10.1175/1520-0493\(2002\)130<1653:EPFUIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2).
- Siebert, S., 2017: Simplifying and generalising Murphy’s Brier score decomposition. *Quart. J. Roy. Meteor. Soc.*, **143**, 1178–1183, <https://doi.org/10.1002/qj.2985>.
- Staël von Holstein, C.S., 1970: A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteor.*, **9**, 360–364, [https://doi.org/10.1175/1520-0450\(1970\)009<0360:AFOSPS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1970)009<0360:AFOSPS>2.0.CO;2).