Evaluation of probabilistic disease forecasts

Hughes, G; Burnett, FJ

Published in:
Phytopathology

DOI:
10.1094/PHYTO-01-17-0023-FI

First published: 14/07/2017

Document Version
Peer reviewed version

Evaluation of Probabilistic Disease Forecasts

Gareth Hughes and Fiona J. Burnett

Crop and Soil Systems Research Group, SRUC, Edinburgh EH9 3JG, UK

Corresponding author: G. Hughes; E-mail address: gareth.hughes@sruc.ac.uk

ABSTRACT

The statistical evaluation of probabilistic disease forecasts often involves calculation of metrics defined conditionally on disease status, such as sensitivity and specificity. However, for the purpose of disease management decision making, metrics defined conditionally on the result of the forecast – predictive values – are also important, although less frequently reported. In this context, the application of scoring rules in the evaluation of probabilistic disease forecasts is discussed. An index of separation with application in the evaluation of probabilistic disease forecasts, described in the clinical literature, is also considered and its relation to scoring rules illustrated. Scoring rules provide a principled basis for the evaluation of probabilistic forecasts used in plant disease management. In particular, the decomposition of scoring rules into interpretable components is an advantageous feature of their application in the evaluation of disease forecasts.

Additional keywords: Brier score, divergence score, resolution, reliability, uncertainty, PSEP, expected mutual information, G^2 test, McFadden's R^2.

The evaluation of a predictive system in disease management is not a single procedure (Gent et al. 2011, 2013). Initially, during the development of such a system, evaluation is largely based on the calculation of metrics that characterize the accuracy of predictions. Then, during implementation, evaluation of a system includes assessment of its uptake by users, and of its application to provide predictions that contribute to a current disease management decision process. Such direct application in decision making often decreases over time, giving way to indirect application as users gain and deploy their enhanced understanding of disease management in the pathosystem of concern. Assessment of this contribution to disease management decision making via user education may also be counted as part of the evaluation process for a system. And for developers, an awareness of the attributes of predictive systems regarded by users as successful – in terms both of uptake and application and of contribution to an enhanced understanding of disease management – may help to guide progress towards the next generation of systems.

These facets of evaluation are not independent. If a system produces predictions that are insufficiently accurate for use in a decision process, its uptake and application will be low and its impact on understanding of disease management in the pathosystem of concern will be negligible. Thus the foundation of a successful system is the accuracy of its predictions. It is this aspect of forecast evaluation that is the focus of the analysis presented here. In particular, we are concerned with predictions that take the form of probability forecasts, and methods used for evaluation of the accuracy of such forecasts (Broeker 2012).

Predictive systems in disease management are often based on the provision of probability forecasts, although in practice such forecasts are not typically issued in probabilistic terms. The same is true in clinical diagnosis (Graf et al. 1999). In both cases, operational classification of

subjects (i.e., crops or patients) is often based on the assessment of risk relative to a threshold, and the resulting forecast issued as 'intervention required' (i.e., risk threshold exceeded) or as 'intervention not required' (risk threshold not exceeded). Thus it is left implicit that predictive systems are imperfect and that the forecasts issued in these terms are probabilistic and should be interpreted in the context of the system's previously-characterized accuracy metrics. In practice, it is of course hard to tell whether interpretation is always nuanced in this way. Note that we must rely on previously-characterized metrics because the classification of subjects in a disease management process may lead to an intervention made with the aim of changing the (predicted) outcome; therefore it is difficult to evaluate performance when a predictive system is operational (Hughes et al. 2017).

For meteorological applications, in contrast, probability forecasts are often communicated in explicitly probabilistic terms (e.g., "70% chance of rain tomorrow"); and while users may choose to take mitigating action on the basis of such a forecast, the available actions do not include interventions that can change the outcome in terms of the actual weather that occurs. Either it rains, or it does not rain, regardless. Thus for meteorological probability forecasts, it is possible to undertake evaluation on the basis of comparison of the forecast weather to the corresponding actual weather. An important methodology used by meteorologists for the evaluation of probability forecasts in this way is the calculation of a *scoring rule*. It is convenient to think of the use of a scoring rule as a way of attaching a score to probability forecasts in order to provide a quantitative assessment of the success of the predictive system (Broeker 2012).

The analysis presented here provides a phytopathological perspective on the application of scoring rules, in particular the Brier score (Brier 1950) and the divergence score (Weijs et al. 2010), for evaluation of probabilistic disease forecasts. An index of separation proposed in the

clinical literature for the evaluation of probabilistic disease forecasts (Altman and Royston 2000) is also considered in this context. Thus we are concerned here with the evaluation of probabilistic forecasts on the basis of predictive values (probabilities defined conditionally on the result of the forecast) rather than the calculation of sensitivity and specificity (probabilities defined conditionally on the disease status). The decomposition of scoring rules into interpretable components (uncertainty, resolution, reliability) is discussed. The analysis is supported by numerical examples based on phytopathological data sets from the literature.

ANALYSIS

The phytopathological setting. It is not our purpose here to give an account of the experimental and analytical work that underpins development of the evidential basis for predictive systems providing probability forecasts for crop disease management. Detailed explanatory descriptions of such work (for two-forecast-category systems), including identification of risk factors, statistical modelling of disease risk, construction of a receiver operating characteristic (ROC) curve, choice of an appropriate risk threshold, and determination of the corresponding accuracy metrics defined conditionally on disease status (i.e., sensitivity and specificity) can be found in, for example, Yuen et al. (1996) and Twengström et al. (1998) (from a study of Sclerotinia stem rot in Sweden) or De Wolf et al. (2003) and Madden (2006) (from a study of Fusarium head blight in the U.S.A.).

The context for the analysis to be described here is provided by Bayesian updating (e.g., Yuen and Hughes 2002, Madden 2006). The starting point is a prior probability, $\Pr(o_j)$, which is updated to a posterior probability, $\Pr(o_j|f_i)$, by use of a predictor that incorporates evidence related to risk factors (as in the examples referred to above). Thus the Bayesian posterior probabilities – also referred to as predictive values (see Table 1 in Madden 2006) – are metrics

defined conditionally on the result of the forecast. Here, we discuss the situation in which there are two outcome categories o_j for the actual status of a crop, with $j = 1$ denoting the *control* (no disease) category and $j = 2$ denoting the *case* (disease) category. The number of forecast categories is not limited to two by the analysis described, although in practice many predictive systems providing probability forecasts for crop disease management use two forecast categories f_i, with $i = 1$ denoting here the 'best' forecast (intervention not required) and $i = 2$ the 'worst' forecast (intervention required). This is not restrictive if the decision process in question presents only two alternative courses of action. In the equivalent clinical situation, it is not unusual to have up to four or five forecast categories (*diagnosis-related groups*, DRGs), in which case the category for the worst forecast would be (using the present notation) the f_i indexed by the largest i.

Now we have some notation, we can write the ROC-based metrics (for a two-forecast-category system) sensitivity and specificity as, respectively, $\Pr(f_2|o_2)$ and $\Pr(f_1|o_1)$. Sensitivity is the proportion of cases with an 'intervention required' forecast (the true positive proportion, TPP), and specificity is the proportion of controls with an 'intervention not required' forecast (the true negative proportion, TNP). These accuracy metrics, respectively characterizing the proportion of actual epidemics correctly predicted and the proportion of actual non-epidemics correctly predicted, are widely cited in the evaluation of probabilistic disease forecasts with two forecast categories. In essence, they summarize the evidence related to disease risk factors as provided by a predictive system, independent of the prior probability.

While it is beyond doubt that sensitivity and specificity are useful metrics, they do not represent a complete evaluation of a predictive system. This can be seen from, for example, Table 2 of Madden (2006). For a predictor with TPP = 0.833 and TNP = 0.844, with prior

probabilities (or *disease prevalence*) Pr(o_2) = 0.05, 0.36, and 0.85, the corresponding posterior probabilities are Pr($o_2|f_2$) = 0.22, 0.75 and 0.97, respectively. In the first example, where disease prevalence is 5%, consider a crop of unknown status for which there is an 'intervention required' forecast. High sensitivity and specificity values notwithstanding, there is still only a <25% chance that the crop actually does require intervention, so the forecast contributes little to the decision process. In the second example, disease prevalence is <50% but when the evidence related to risk factors results in an 'intervention required' forecast for a crop of unknown status, there is a >50% chance that the crop actually does require intervention. Thus this example illustrates the most useful kind of result supporting disease management decision making, in that the predictive system produces a posterior probability that might plausibly result in a different management decision to one that was based on the prior probability alone, made without recourse to evidence related to risk factors. In the third example, an 'intervention required' forecast is effectively redundant in relation to the decision process, since a crop of unknown status has an 85% chance of requiring intervention on the basis of disease prevalence alone, without need for any further evidence. Increasing this to a 97% chance is inconsequential in terms of the decision on whether or not to intervene.

Thus there are aspects of the performance of a predictive system in relation to disease management decision making that are characterized by prior and posterior probabilities. The probability of requirement for intervention given the forecast result depends both on the evidence related to disease risk factors as provided by a predictive system and on the disease prevalence. Scoring rules provide a basis for evaluating the performance of probability forecasts in this respect, as discussed below. The phytopathological data sets used here for the purpose of numerical illustration of the application of scoring rules are given in Table 1.

A brief introduction to scoring rules for probability forecasts. Two scoring rules are discussed here; the Brier score (Brier 1950) and the divergence score (Weijs et al. 2010), both first described in the meteorological literature. Both meet the criteria for strictly proper scoring rules (Gneiting and Raftery 2007). Both are penalty scores; that is, a less accurate forecast incurs a higher score. In meteorological application, the long term average frequency of a weather event is termed the *climatological probability*. Predictive values obtained by updating a climatological probability to a forecast probability for a weather event are not necessarily Bayesian posteriors. For example, based on an assessment of current atmospheric conditions, a probability forecast for the weather event of interest is usually issued in one of a number of pre-specified forecast categories (*allowed probabilities*). By the standards of disease forecasting, the number of categories used by meteorologists may be large; Table 8.2 of Wilks (2011), for example, shows a predictive system with 12 allowed forecast probabilities.

From a phytopathological perspective, it is desirable to place application of the Brier score and the divergence score explicitly in the context of Bayesian updating. That is to say, starting from prior probability $\Pr(o_2)$, a forecast updates this (in the two-forecast-category case) to either $\Pr(o_2|f_1)$ or $\Pr(o_2|f_2)$. Subsequently, the true status – either control ($o_j = 0$) or case ($o_j = 1$) becomes known. If the true status is control, then f_1 was the correct forecast and f_2 incorrect. If the true status is case, then f_2 was the correct forecast and f_1 incorrect. The Brier scores for individual forecasts are given by $(o_j - f_i)^2$, where observation $o_j \in [0,1]$ and (in the two-forecast-category case) forecast $f_i \in [\Pr(o_2|f_1), \Pr(o_2|f_2)]$. Similarly, the divergence scores for individual forecasts are given by the Kullback-Leibler divergences:

$$D_{KL}(o_j \| f_i) = o_j \cdot \ln\left(\frac{o_j}{f_i}\right) + (1 - o_j) \cdot \ln\left(\frac{1 - o_j}{1 - f_i}\right)$$

with $D_{KL}\left(o_j \| f_i\right) \geq 0$, and for calculation purposes (here and throughout) we take $0 \cdot \ln(0) = 0$, recalling $\lim_{x \to 0}\left(0 \cdot \ln\left(0\right)\right) = 0$. Both scoring rules attach a score to a forecast according to the distance (or *divergence*) between the forecast value and the true value. Smaller distances represent better forecasts, so individual scores increase with increasing inaccuracy. Usually, the frequency-weighted average score over a set of forecasts is presented. Thus for the Brier score (BS) we have:

$$\mathrm{BS} = \frac{1}{N} \cdot \sum_{ij} n_{ij} \cdot \left(o_j - f_i\right)^2 \tag{1}$$

and for the divergence score (DS):

$$\mathrm{DS} = \frac{1}{N} \cdot \sum_{ij} n_{ij} \cdot D_{KL}\left(o_j \| f_i\right) \tag{2}$$

where n_{ij} denotes the number of subjects in forecast category i and outcome category j, such that the total number of subjects is $N = \sum_{ij} n_{ij}$. As outlined in the introductory section, the true status of some subjects (specifically those with an 'intervention required' forecast that was then actioned) cannot be retrieved from an operational predictive system in disease management, so scoring rules are calculated from the same data sets for untreated subjects from which sensitivity and specificity values are calculated; that is to say, from data where both the forecast category and the actual status are known.

Both the Brier score and the divergence score are examples of Bregman divergences (Bregman 1967, Hughes and Topp 2015). In this format, the scores for individual forecasts are given by:

$$D_B\left(o_j \| f_i\right) = g\left(o_j\right) - g\left(f_i\right) - \left(o_j - f_i\right) \cdot g'\left(f_i\right) \tag{3}$$

with $D_B(o_j \| f_i) \geq 0$, and where $g(\bullet)$ is a convex function chosen to match the particular score to be calculated. For the Brier score, $g(\bullet) = (\bullet)^2$ (see Figure 1); for the divergence score, $g(\bullet) = -\mathrm{H}(\bullet) = \sum \Pr(\bullet) \cdot \ln(\Pr(\bullet))$ (i.e., the negative of the binary Shannon entropy function, see Figures 2, 3, and 4). The notation $g'(\bullet)$ denotes the slope of a tangent to the curve $g(\bullet)$. The frequency-weighted average score over a set of forecasts is then:

$$\frac{1}{N} \cdot \sum_{ij} n_{ij} \cdot D_B(o_j \| f_i) \tag{4}$$

For numerical calculations based on $g(\bullet) = -\mathrm{H}(\bullet)$, Bregman divergences are denominated in units depending on the choice of logarithmic base; since natural logarithms are used here the appropriate unit is the *nit* (Theil 1967).

An index of separation, PSEP. Altman and Royston's (2000) paper "What do we mean by validating a prognostic model?" relates to the evaluation of probabilistic disease forecasts, and remains influential in the clinical literature (see, for example, Collins and Altman 2013, Sharples and Nashef 2013). A simple index of separation, PSEP, is proposed for evaluation of the performance of predictive models:

$$\mathrm{PSEP} = \Pr(o_2 | f_{\text{worst } i}) - \Pr(o_2 | f_{\text{best } i}) \tag{5}$$

For the two-forecast-category case, this is $\mathrm{PSEP} = \Pr(o_2 | f_2) - \Pr(o_2 | f_1)$, in which case PSEP may be written in terms of sensitivity, specificity and prior probability, via Bayes' rule. We have $0 \leq \mathrm{PSEP} \leq 1$ (i.e., PSEP is measured on a probability scale; within which larger values are more desirable).

Altman and Royston's (2000) account of PSEP is concerned mainly with its perceived advantages; particularly its low computational load and its interpretability as a measure of separation between DRGs (forecast categories). Here our interest is in the analytical properties of PSEP (as compared to scoring rules), but at the outset it is worth considering why separation between forecast categories is important in the forecast evaluation process. Here we offer a simple heuristic view. Before the forecast, the best evidence-based decision we can make is based on the prior probability $\Pr(o_2)$. The forecast, incorporating evidence related to risk factors, then allows us to update this to a posterior probability $\Pr(o_2|f_i)$, the f_i representing the available forecast categories. In assigning subjects to appropriate forecast categories based on posterior probabilities rather than to a single category based on a prior probability, we are in essence modelling observed variation in a manner analogous to the analysis of a simple treatment-comparison experiment in which we anticipate that the treatment means will provide a better description of variation than the overall mean alone.

Altman and Royston (2000) consider the Brier score, as follows. "The Brier score has several pleasant mathematical properties, but it has the drawback that it lacks an obvious interpretation other than in general terms − the bigger the score, the worse the quality of the prediction. A cruder but more interpretable statistic is the difference between observed and predicted probabilities at the group level (PSEP), though of course more than one measure may be used." For the Brier score, taking the average score for a data set (equation 1) provides a value in the range $0 \leq \mathrm{BS} \leq 1$ (Wilks 2011), the same as the range for PSEP. For PSEP, however, the bigger the score, the better the quality of the prediction. The use of both PSEP and BS in the course of a forecast evaluation would require that the two measures were independent, but as we shall now see, this is not the case.

The example used here is Scenario A (Table 1). Incidentally, Scenario A was originally chosen in order to provide an example from a pathosystem where probability forecasts are usually made in more than two categories (Nutter et al. 2002, Esker et al. 2006), but the published external validation data supported only a two-forecast-category calculation. Analytically, the link between PSEP and the Brier scores for individual forecasts as Bregman divergences (Figure 1) and between PSEP and the divergence scores for individual forecasts as Bregman divergences (Figure 2) is provided by the forecast probabilities for f_1 and f_2, which define both PSEP (equation 5) and the gradients of the tangents to the convex function $g(f)$ (equation 3).

The goal here is not to establish any quantitative equivalence between PSEP and the scoring rules BS and DS; numerical results are provided for the convenience of readers who wish to use the analysis as a template for calculations. For Scenario A, PSEP = 0.454 (equation 5). This is shown diagrammatically in both Figure 1 (for the purpose of illustrating the link with the Brier score) and Figure 2 (for the purpose of illustrating the link with the divergence score). Figure 1 illustrates the calculation of Brier scores for individual forecasts as Bregman divergences (equation 3). Figure 2 illustrates the calculation of divergence scores for individual forecasts as Bregman divergences (equation 3). The frequency-weighted average Brier score over the set of forecasts for Scenario A is then BS = 0.230 (equation 4), identical to the value calculated via equation 1. The frequency-weighted average divergence score over the set of forecasts for Scenario A is then DS = 0.650 nits (equation 4), identical to the value calculated via equation 2.

Resolution, RES. Having characterized the non-independence of PSEP and the BS and DS scoring rules, such that PSEP is (qualitatively) an inverse of BS and of DS, it would be useful at this stage to characterize a probability forecast evaluation measure for which PSEP is a direct

analogue. To do so, we take advantage of the analysis by means of which both BS and DS can be decomposed into terms denoted uncertainty (UNC), resolution (RES) and reliability (REL) (Murphy 1973, Weijs et al. 2010), such that:

$$\left.\begin{matrix}\text{BS}\\ \\ \text{DS}\end{matrix}\right\} = \text{UNC} - \text{RES} + \text{REL} \qquad (6)$$

This decomposition has the advantage that it supplies a useful interpretation of the Brier score and of the divergence score in very specific terms. UNC quantifies our state of knowledge based only on the prior probability Pr(o_2), RES refers to the extent to which forecasts separate subjects into different groups, and REL refers to the extent of agreement between forecast probabilities and observed frequencies. UNC, RES and REL are all ≥ 0 (see, e.g., Hughes and Topp 2015). For an hypothetical perfect forecaster, RES = UNC and REL = 0, so the scoring rule (BS or DS) = 0 (equation 6). For a typical (imperfect) forecaster, RES < UNC and REL > 0, so the scoring rule (BS or DS) > 0 (equation 6). Smaller BS or DS scores indicate better forecaster performance; thus for RES, larger values (≥0) are more desirable; while for REL, smaller values (≥0) are more desirable.

The notation used for equations 7-9 below identifies the context in which data are used in analyses based on the decomposition of a scoring rule. Consider Scenario A (Table 1), where there are 12 observed cases out of 14 'intervention required' forecasts; then Pr($o_2|f_2$) = 12/14 = 0.857. In Bayesian disease forecasting as described thus far, the probability forecast and the observed frequency are identical. The adopted notation is required for when this is not so. Thus, f_i denotes the categories for forecast probabilities, and d_i the categories for the corresponding observed frequencies. The prior probability Pr(o_2) is calculated as the overall observed frequency of cases and denoted $\bar{d}$. Note that these notational issues do not arise in non-Bayesian weather

forecasting, where the observed frequencies usually differ to some extent from the corresponding probability forecasts (see, for example, Table 8.2 in Wilks 2011, Table 1 in Hughes and Topp 2015). Now, writing the analysis in terms of Bregman divergences provides a common format for the decomposition of both the BS and DS scoring rules (Hughes and Topp 2015):

$$\left.\begin{aligned} \mathrm{UNC} &= u\left(\bar{d}\right) \\ \mathrm{RES}_i &= D_B\left(d_i \| \bar{d}\right) = g\left(d_i\right) - g\left(\bar{d}\right) - \left(d_i - \bar{d}\right)\cdot g'\left(\bar{d}\right) \\ \mathrm{REL}_i &= D_B\left(d_i \| f_i\right) = g\left(d_i\right) - g\left(f_i\right) - \left(d_i - f_i\right)\cdot g'\left(f_i\right) \end{aligned}\right\} \tag{7}$$

Given the appropriate convex function for calculation of the Bregman divergences (as described above) and an appropriate choice of uncertainty function $u\left(\bar{d}\right)$, equation 7 applies equally to both the Brier score and the divergence score (Hughes and Topp 2015). Because of this we only need show one such analysis. We adopt the divergence score for the purpose of illustration because it allows some useful information theoretic interpretations. In equation 7, $u\left(\bar{d}\right) = -\sum_j \Pr\left(o_j\right)\cdot \ln\left(\Pr\left(o_j\right)\right)$ (the binary Shannon entropy of the prior distribution of observations) is the uncertainty function for the decomposition of the divergence score. RES_i and REL_i represent, respectively, resolution and reliability components for group *i*. The corresponding overall resolution and reliability components are, respectively:

$$\mathrm{RES} = \frac{1}{N}\cdot \sum_i n_i \cdot D_B\left(d_i \| \bar{d}\right) \tag{8}$$

$$\mathrm{REL} = \frac{1}{N}\cdot \sum_i n_i \cdot D_B\left(d_i \| f_i\right) \tag{9}$$

Here, RES is the probability forecast evaluation measure of particular interest, because (like PSEP) RES is a measure of separation between groups (Wilks 2011). The link between PSEP and the RES component of the divergence score decomposition is illustrated diagrammatically in

Figure 3. It is apparent that PSEP, measured on a probability scale, is an analogue of RES, measured on an information scale. Further, Weijs et al. (2010) show that the RES component of the divergence score decomposition is the *expected mutual information* between forecasts and observations. Applications of expected mutual information in the evaluation of clinical diagnostics go back at least as far as Metz et al. (1973), while Benish (2003) provides a useful overview. More recent phytopathological perspectives on expected mutual information can be found in Hughes (2012) and Hughes and McRoberts (2014).

The numerical results for Scenario A (Table 1) show PSEP = 0.454 (equation 5) as before. For a tangent to $g(d)$ drawn at $\bar{d}$, the observed frequencies for d_1 and d_2 that define PSEP also define the Bregman divergences for the required RES components (Figure 3). The frequency-weighted average divergence over the set of forecasts for Scenario A is then RES = 0.037 nits (equation 8).

Expected mutual information, $\mathbf{I_M}(o,f)$. The analysis of Scenario A established the non-independence of PSEP and expected mutual information. This was achieved by using Bregman divergences to calculate the RES component of the decomposition of the divergence score. However, this is not necessarily the way that expected mutual information would be routinely calculated in order to characterize the relationship between forecasts and observations for a single data set. The example used here is Scenario B (Table 1). Scenario B was selected in order to provide an example from a study where the validation data for a risk prediction model were presented as a 2×2 prediction-realization table. In order to calculate some reference values, the numerical data for Scenario B are first normalized. Hughes et al. (2015) show a normalized 2×2 prediction-realization table in both notational and data formats, and provide background related to calculations based on equations 10-15 below.

Expected mutual information between forecasts and observations may be calculated directly from the normalized prediction-realization table as:

$$\mathrm{I_M}(o,f) = \sum_i \sum_j \Pr(o_j \cap f_i) \cdot \ln\left[\frac{\Pr(o_j \cap f_i)}{\Pr(o_j)\cdot\Pr(f_i)}\right] \tag{10}$$

from which we obtain $I_M(o,f) = 0.340$ nits for Scenario B. Proceeding instead step-by-step, the entropy based on the prior probability is:

$$\mathrm{H}(o) = -\sum_j \Pr(o_j)\cdot\ln(\Pr(o_j)) \tag{11}$$

and for Scenario B, $H(o) = 0.641$ nits. The entropy $H(o)$ can be thought of as characterizing information or uncertainty. Either $H(o)$ characterizes the amount of uncertainty before use of the predictor or, alternatively, $H(o)$ characterizes the amount of information needed to completely resolve that uncertainty. The entropies based on the posterior probabilities are:

$$\mathrm{H}(o|f_i) = -\sum_j \Pr(o_j|f_i)\cdot\ln(\Pr(o_j|f_i)) \tag{12}$$

and then:

$$\mathrm{H}(o|f) = -\sum_i \Pr(f_i)\mathrm{H}(o|f_i) \tag{13}$$

and for Scenario B, $H(o|f) = 0.301$ nits. The conditional entropy $H(o|f)$ is the remaining uncertainty, on average, after use of the imperfect binary predictor, or, alternatively, the amount of information still needed to resolve that remaining uncertainty. Then we note:

$$\mathrm{I_M}(o,f) = \mathrm{H}(o) - \mathrm{H}(o|f) \tag{14}$$

and for Scenario B, $I_M(o,f) = 0.641 - 0.301 = 0.340$, as previously. We can see from equation 14 that expected mutual information $I_M(o,f)$ is a measure the average reduction in entropy $H(o)$

resulting from use of the imperfect binary predictor, or, alternatively, the average amount of information supplied by the predictor. Normalized expected mutual information is:

$$\textit{normalized } \mathrm{I_M}(o,f) = \frac{\mathrm{H}(o) - \mathrm{H}(o|f)}{\mathrm{H}(o)} \tag{15}$$

which for Scenario B is equal to 0.530.

In their model validation, Harikrishnan and del Río (2008) use chi-squared statistics and the R^2 value from a linear regression analysis of predicted frequency on observed frequency, of the kind often used in the validation of disease simulation studies (see, for example, Dias et al. 2014). Here, we consider first the likelihood-ratio chi-squared statistic, denoted G^2 (Agresti 2012). Of interest here is that there is a relationship between G^2 and expected mutual information: $G^2 = 2 \cdot N \cdot \mathrm{I_M}(o,f)$ (Attneave 1959). For example, for Scenario B, referring to the original 2×2 prediction-realization table from Harikrishnan and del Río (2008) and following Agresti (2012), we calculate G^2 = 67.931; then note that 67.931/(2·100) = 0.340, identical to $\mathrm{I_M}(o,f)$ from equation 10 or equation 14. From an historical perspective, note that the Pearson chi-squared was originally described in order to meet the need for an approximate but more conveniently calculable form of G^2.

Now consider a binary logistic regression of the 2×2 prediction-realization table for Scenario B. This analysis yields estimates of the posterior log odds as:

$$\left.\begin{aligned} \mathrm{logit}(o_2|f_1) &= -2.213 + 4.816 \cdot (0) \\ \mathrm{logit}(o_2|f_2) &= -2.213 + 4.816 \cdot (1) \end{aligned}\right\}$$

from which the corresponding estimates of $\Pr(o_2|f_1)$ and $\Pr(o_2|f_2)$ are respectively 0.099 and 0.931, exactly as in Table 1. Of interest here is that goodness-of-fit as measured by McFadden's

(pseudo-) R^2 (McFadden 1974) is identical to the normalized expected mutual information for the 2×2 prediction-realization table (Hauser 1978). For Scenario B, McFadden's $R^2 = 0.530$, as given in the model summary provided by the statistical software used to calculate the logistic regression analysis, and identical to *normalized* $I_M(o,f)$ from equation 15. Application of binary logistic regression is a satisfying approach to the analysis of explained variation for disease forecasts as represented by a 2×2 table, because it maintains the classification of subjects into forecast categories as the basis for the calculation.

Reliability, REL. The analysis of Scenario B established the role of expected mutual information – via the G^2 test and McFadden's R^2 – in characterizing the relationship between forecasts and observations on the basis of a single data set. More important, perhaps, is evaluation of probabilistic disease forecasts using independent data. This may happen when a predictive system is developed using data collected over a period of time, then tested using data collected over a subsequent period (e.g., Esker et al. 2006, Bondalapati et al. 2012), or when a system developed in one location is applied in another (e.g., De Wolf et al. 2003, Duttweiler et al. 2008).

The examples used here are Scenarios C1 and C2 (Table 1). These scenarios were selected in order to provide an example from a study where both training data and validation data for a risk prediction model were presented. For this example, the validation data set provides data that meet the original study's requirements in terms of sensitivity and specificity. If we calculate expected mutual information (equation 10) for Scenario C1 (training data set) and C2 (validation data set) we obtain $I_M(o,f) = 0.177$ and 0.172, in nits (equation 10), respectively, so resolution is consistent between C1 and C2. Note also that from Table 1, we obtain PSEP = 0.55 (C1) and 0.57 (C2) (equation 5), which would satisfy Altman and Royston's (2000) validation criterion.

What is of particular interest here is characterization of the reliability (REL) component of the decomposition of the divergence score (equation 9). For Scenario C1, the (Bayesian) probability forecasts and the observed frequencies are identical. Thus $f_i = d_i$, as a result of which $\mathrm{REL}_i = 0\ \forall i$ (equation 7) and REL = 0 (equation 9). The decomposition of the divergence score (equation 6) becomes DS = UNC – RES, which is the same as $\mathrm{H}(o|f) = \mathrm{H}(o) - \mathrm{I}_{\mathrm{M}}(o, f)$ (rearranging equation 14). Theil (1967) discusses $\mathrm{H}(o|f)$, the equivalent of DS when REL = 0, as *information inaccuracy*. Consider the amount of information that would be required from a forecast to take us from the prior probability to the correct identification of the actual status: if the forecaster in use is imperfect, it can only supply enough information to take us part of the way, from the prior to the posterior probability. Thus there is a deficit, the amount of information still required to take us from the posterior probability to the actual status. This, taken on average over all forecast-observation combinations, is the information inaccuracy. For Scenario C1, $\mathrm{H}(o|f)$ = 0.358 nits (equation 13).

How we then treat the analysis of reliability when it comes to Scenario C2 depends on our view of the evaluation process. If we regard Scenario C2 as supplying new probability forecasts, then REL = 0 again and calculations yield $\mathrm{H}(o|f)$ = 0.506 nits. The increase in information inaccuracy for Scenario C2 over that of Scenario C1 arises because H(o) for Scenario C2 (= 0.678) is larger than that of Scenario C1 (= 0.535) (equation 11), while the $\mathrm{I}_{\mathrm{M}}(o,f)$ values are similar (= 0.177 nits (C1) and 0.172 nits (C2), as above). The difference between H(o) values reflects the change in prior probability Pr(o_2) between the two scenarios (see Table 1).

If, instead, we regard Scenario C1 as having established probability forecasts for f_1 (= 0.058) and f_2 (= 0.609) that are applicable to Scenario C2, we note that the observed frequencies for d_1

(=3/12) and d_2 (= 14/17) from C2 now differ from the corresponding forecasts. It is these differences between probability forecasts and observed frequencies that are measured by REL. The reliability components REL_i are calculated as Bregman divergences as in equation 7. For Scenario C2, this calculation is illustrated in Figure 4. The frequency-weighted average reliability over the set of forecasts for Scenario C2 is then REL = 0.144 nits (equation 9).

Essentially, we now have two versions of equation 6 for the decomposition of the divergence score relating to Scenario C2. They illustrate different perspectives on the evaluation process. Recall that for the overall score (DS) and for reliability (REL), smaller values are more desirable; while for resolution (RES), larger values are more desirable (all components are ≥ 0). Either:

$$\text{DS } (= 0.506) = \text{UNC } (= 0.678) - \text{RES } (= 0.172)$$

(in which REL is implicitly taken to be equal to zero by use of the observed frequencies for Scenario C2 as the forecast probabilities) or:

$$\text{DS } (= 0.650) = \text{UNC } (= 0.678) - \text{RES } (=0.172) + \text{REL } (= 0.144)$$

(in which REL explicitly accounts for discrepancies between the observed frequencies from Scenario C2, the validation data, and the forecast probabilities from Scenario C1, the training data) (all quantities in nits). The components of the decomposition are independent (they measure different aspects of forecaster performance), so the calculation of reliability (rather than the implicit assumption that it is equal to zero) does not affect the calculation of the uncertainty and resolution components. The difference between the two versions simply reflects different perceptions of the need (or otherwise) to account for differences between forecast probabilities from the training data set (Scenario C1) and the observed frequencies from the validation data set (Scenario C2). Here, DS increases (by an amount equal to REL) when the lack of agreement

between the observed frequencies (from C2) and the forecast probabilities (from C1) is taken into account. REL characterizes a difference between training data and validation data by applying the forecast probabilities from the former to the calculation of the scoring rule for the latter. In evaluations where REL is implicitly taken to be equal to zero, it would be good practice to make a clear statement to that effect.

DISCUSSION

Altman and Royston (2000) introduced PSEP in the clinical literature as a simple index of prognostic information with application in the validation of probabilistic disease risk prediction models. Of particular interest was the performance of predictive models applied to subjects other than those whose data had been used for model derivation. Specifically, the idea of greater or lesser separation between the observed frequencies for the 'worst' and 'best' forecast categories as a measure of prognostic information was considered attractive, being both interpretable and pragmatic. The Brier score (Brier 1950), a strictly proper scoring rule for use in the evaluation of probability forecasts, was deemed to be lacking in interpretability. Altman and Royston's (2000) misgivings notwithstanding, we note that the Brier score has subsequently been discussed in the context of performance evaluation for clinical risk prediction models by, for example, Gerds et al. (2008) and Steyerberg et al. (2010).

Here, analysis of the Brier score and also the divergence score, another strictly proper scoring rule for use in the evaluation of probability forecasts (Weijs et al. 2010) shows how PSEP is related to both these scoring rules. In particular, PSEP is an analogue of the resolution (RES) component of the scoring rule decomposition, a measure of separation between observed frequencies for forecast categories (Wilks 2011). Thus PSEP offers no more interpretability than either the Brier score or the divergence score. In the specific case of the information-theoretic

divergence score decomposition, RES is identical to the expected mutual information between forecasts and observations.

PSEP may be simple to calculate (equation 5), but this simplicity is not as straightforward as it may seem, specifically for models with more than two forecast categories. In such cases, only the data from the 'best' and 'worst' forecast categories are used in the PSEP calculation. So, for example, in Table III of Altman and Royston (2000) data from only 49% (Hong Kong) and 47% (Guangzhou and Shanghai) of subjects in the validation samples are used in the PSEP calculation – but the unused data must still be collected to enable that calculation. Once resources have been allocated to the collection of validation data, it would seem desirable to use all those data in the model evaluation process, increased computational load notwithstanding. This is achieved by the adoption of a scoring rule approach.

The decomposition of scoring rules into uncertainty, resolution and reliability components offers interpretability beyond assessment of separation between forecast categories. For both the overall Brier score and the overall divergence score, smaller values are more desirable. For the information theoretic divergence score (DS), where the resolution component RES is equal to expected mutual information $I_M(o,f)$, we have interpretations of resolution in terms of the likelihood-ratio chi-squared statistic G^2 (Agresti 2012) and McFadden's (1974) R^2 measure of explained variation for logistic regression (Hauser 1978). Larger values of RES are indicative of a greater reduction of the uncertainty (UNC) component of DS because the observed frequencies for the different forecast categories really are separate from each other. This contributes to a smaller overall score. Larger values of the reliability (REL) component indicate greater discrepancy between observed frequencies and the corresponding forecast probabilities, which contributes to a larger overall score. Essentially, for probability forecasts with application in

disease management decision making, resolution is a measure of separation between the observed frequencies for the adopted forecast categories and reliability is then a measure of the mismatch between the observed frequencies and the probability forecasts for those forecast categories.

The majority of evaluations of probabilistic disease forecasts with clinical applications are based on measures defined conditionally on disease status (i.e., sensitivity and specificity) (Shiu and Gastonis 2008). The same appears to be true for forecasts with phytopathological applications. However, also important in disease management decision making are measures defined conditionally on the result of the forecast (i.e., the predictive values), although these are more are difficult to evaluate. Shiu and Gastonis (2008) provide an overview of the problem and some possible solutions. The application of scoring rules – and in particular the information theoretic divergence score of Weijs et al. (2010) and its decomposition – is a useful addition to the available methodology for evaluation of the accuracy of probabilistic disease forecasts deployed in phytopathological applications.

ACKNOWLEDGMENT

SRUC receives grant-in-aid from the Scottish Government.

LITERATURE CITED

Agresti, A. 2012. Categorical Data Analysis, 3rd ed. Wiley: Chichester, UK.

Altman, D. G., and Royston, P. 2000. What do we mean by validating a prognostic model? Stat. Med. 19:453-473.

Attneave, F. 1959. Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results. Holt, Rinehart and Winston: New York.

Benish, W. A. 2003. Mutual information as an index of diagnostic test performance. Method. Inform. Med. 42:260-264.

Bondalapati, K. D., Stein, J. M., Neate, S. M., Halley, S. H., Osborne, L. E., and Hollingsworth, C. R. 2012. Development of weather-based predictive models for Fusarium head blight and deoxynivalenol accumulation for spring malting barley. Plant Dis. 96:673-680.

Bregman, L. M. 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. 7:200-217.

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

Broecker, J. 2012. Probability forecasts. Pages 119-139 in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd ed. I. T. Jolliffe, and D. B. Stephenson, eds. Wiley: Chichester, UK.

Collins, G., and Altman, D. 2013. Design flaws in EuroSCORE II. Eur. J. Cardiothorac. Surg. 43:871.

De Wolf, E. D., Madden, L. V., and Lipps, P. E. 2003. Risk assessment models for wheat Fusarium head blight epidemics based on within-season weather data. Phytopathology 93:428-435.

Dias, A. P. S., Li, X., and Yang, X. B. 2014. Modeling the effects of cloudy weather on regional epidemics of soybean rust. Plant Dis. 98:811-816.

Duttweiler, K. B., Gleason, M. L., Dixon, P. M., Sutton, T. B., McManus, P. S., and Monteiro, J. E. B. A. 2008. Adaptation of an apple sooty blotch and flyspeck warning system for the Upper Midwest United States. Plant Dis. 92:1215-1222.

Esker, P. D., Harri, J., Dixon, P. M., and Nutter, F. W., Jr. 2006. Comparison of models for forecasting of Stewart's disease of corn in Iowa. Plant Dis. 90:1353-1357.

Gent, D. H., De Wolf, E. D., and Pethybridge, S. J. 2011. Perceptions of risk, risk aversion, and barriers to adoption of decision support systems and integrated pest management: an introduction. Phytopathology 101:640-643.

Gent, D. H., Mahaffee, W. F., McRoberts, N., and Pfender, W. F. 2013. The use and role of predictive systems in disease management. Annu. Rev. Phytopathol. 51:267-289.

Gerds, T. A., Cai, T., and Schumacher, M. 2008. The performance of risk prediction models. Biom. J. 50:457-479.

Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102:359-378.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. 1999. Assessment and comparison of prognostic classification schemes for survival data. Stat. Med. 18:2529-2545.

Harikrishnan, R., and del Río, L. E. 2008. A logistic regression model for predicting risk of white mold incidence on dry bean in North Dakota. Plant Dis. 92:42-46.

Hauser, J. R. 1978. Testing the accuracy, usefulness, and significance of probabilistic choice models: an information-theoretic approach. Oper. Res. 26:406-421.

Hughes, G. 2012. Applications of Information Theory to Epidemiology. The American Phytopathological Society: St. Paul, MN.

Hughes, G., and McRoberts, N. 2014. The structure of diagnostic information. Australas. Plant Pathol. 43:267-286.

Hughes, G., McRoberts, N., and Burnett, F. J. 2015. Information graphs for binary predictors. Phytopathology 105:9-17.

Hughes, G., McRoberts, N., and Burnett, F. J. 2017. Resolution of probabilistic weather forecasts with application in disease management. Phytopathology 107:158-162.

Hughes, G., and Topp, C. F. E. 2015. Probabilistic forecasts: scoring rules and their decomposition and diagrammatic representation via Bregman divergences. Entropy 17:5450-5471.

Madden, L. V. 2006. Botanical epidemiology: some key advances and its continuing role in disease management. Eur. J. Plant Pathol. 115:3-23.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. Pages 105-142 in: Frontiers in Econometrics. P. Zarembka, ed. Academic Press: New York.

Metz, C. E., Goodenough, D. J., and Rossmann, K. 1973. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. Radiology 109:297-303.

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

Nutter, F. W., Jr., Rubsam, R. R., Taylor, S. E., Harri, J. A., and Esker, P. D. 2002. Use of geospatially-referenced disease and weather data to improve site-specific forecasts for Stewart's disease of corn in the U.S. corn belt. Comput. Electron. Agric. 37:7-14.

Sharples, L. D., and Nashef, S. A. M. 2013. Reply to Collins and Altman. Eur. J. Cardiothorac. Surg. 43:872.

Shiu, S.-Y., and Gatsonis, C. 2008. The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. Phil. Trans. R. Soc. A 366:2313-2333.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattane, M. W. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21:128-138.

Theil, H. 1967. Economics and Information Theory. North-Holland: Amsterdam.

Twengstöm, E., Sigvald, R., Svensson, C., and Yuen, J. 1998. Forecasting Sclerotinia stem rot in spring sown oilseed rape. Crop Prot. 17:405-411.

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

Wilks, D. S. 2011. Statistical Methods in the Atmospheric Sciences, 3rd ed. Academic Press, Oxford, UK.

Yuen, J. E., and Hughes, G. 2002. Bayesian analysis of plant disease prediction. Plant Pathol. 51:407-412.

Yuen, J., Twengström, E., and Sigvald, R. 1996. Calibration and verification of risk algorithms using logistic regression. Eur. J. Plant Pathol. 102:847-854.

TABLE 1. Data (correct to 3 decimal places (d.p.)) [a,b]

	Scenario A	Scenario B	Scenario C1	Scenario C2
$Pr(o_2\|f_1)$	0.403 (56/139)	0.099 (7/71)	0.058 (6/104)	0.250 (3/12)
$Pr(o_2\|f_2)$	0.857 (12/14)	0.931 (27/29)	0.609 (28/46)	0.824 (14/17)
$Pr(o_2)$	0.444 (68/153)	0.340 (34/100)	0.227 (34/150)	0.586 (17/29)

[a] Source: Scenario A, see Table 5 (Stevens Model) in Esker et al. (2006); Scenario B, see Table 2 in Harikrishnan and del Río (2008); Scenarios C1 and C2, see Table 4 (Model #3) in Bondalapati et al. (2012).

[b] Notation: $Pr(o_2)$, prior probability of disease or need for a control intervention (the complement is $Pr(o_1)$, prior probability of no disease or no need for a control intervention); $Pr(o_2|f_2)$, posterior probability of disease or need for a control intervention given a forecast of disease or need for a control intervention (the complement is $Pr(o_1|f_2)$, posterior probability of no disease or no need for a control intervention given a forecast of disease or need for a control intervention); $Pr(o_2|f_1)$, posterior probability of disease or need for a control intervention given a forecast of no disease or no need for a control intervention (the complement is $Pr(o_1|f_1)$, posterior probability of no disease or no need for a control intervention given a forecast of no disease or no need for a control intervention).

Fig. 1. Brier score and index of separation (PSEP), Scenario A (see Table 1). For probability forecasts *f*, the curve $g(f) = f^2$ (solid line) is the basis for the scoring rule known as the Brier score. Here, tangents to the curve (long-dashed lines) are drawn at probability forecasts $f = \Pr(o_2|f_2) = 0.857$ and $f = \Pr(o_2|f_1) = 0.403$, both points marked ● on the curve and the horizontal axis. Short-dashed lines show the projections from the points marked ● on the curve to the horizontal axis. PSEP is calculated as the horizontal difference between these projections, $\Pr(o_2|f_2) - \Pr(o_2|f_1) = 0.454$. The tangent at $f = 0.857$ has slope $g'(f) = 1.714$, and intersects the vertical axis where $f = 0$ at $g(f) = -0.734$ (□) and the vertical axis at $f = 1$ at $g(f) = 0.980$ (■). The tangent at $f = 0.403$ has slope $g'(f) = 0.806$, and intersects the vertical axis where $f = 0$ at $g(f) = -0.162$ (Δ) and the vertical axis at $f = 1$ at $g(f) = 0.644$ (▲). The vertical distances between the curve and the intersections of the tangents at the vertical axis where $f = 0$ (0.162 and 0.734) and at the vertical axis where $f = 1$ (0.020 and 0.356) are Brier scores for individual forecasts calculated as Bregman divergences (equation 3). The frequency-weighted average Bregman divergence is then the Brier score (BS = 0.230, equation 4). All calculations correct to 3 d.p.

Fig. 2. Divergence score and index of separation (PSEP), Scenario A (see Table 1). For probability forecasts *f*, the curve $g(f) = -\mathrm{H}(f)$ (solid line) is the basis for the scoring rule known as the divergence score. Here, tangents to the curve (long-dashed lines) are drawn at probability forecasts $f = \Pr(o_2|f_2) = 0.857$ and $f = \Pr(o_2|f_1) = 0.403$, both points marked ● on the curve and the horizontal axis. Short-dashed lines show the projections from the points marked ● on the curve to the horizontal axis. PSEP is calculated as the horizontal difference between these projections, $\Pr(o_2|f_2) - \Pr(o_2|f_1) = 0.454$. The tangent at $f = 0.857$ has slope $g'(f) = 1.791$, and intersects the vertical axis where $f = 0$ at $g(f) = -1.945$ (□) and the vertical axis at $f = 1$ at $g(f) = -0.154$ (■). The tangent at $f = 0.403$ has slope $g'(f) = -0.393$ and intersects the vertical axis where $f = 0$ at $g(f) = -0.516$ (Δ) and the vertical axis at $f = 1$ at $g(f) = -0.909$ (▲). The vertical distances between the curve and the intersections of the tangents at the vertical axis where $f = 0$ (= 0.516 and 1.945) and at the vertical axis where $f = 1$ (= 0.154 and 0.909) are divergence scores for individual forecasts (in nits) calculated as Bregman divergences (equation 3). The frequency-weighted average Bregman divergence is then the divergence score (DS = 0.650 nits, equation 4). All calculations correct to 3 d.p.

Fig. 3. Resolution and index of separation (PSEP), Scenario A (see Table 1). The solid line is the curve $g(d) = -\mathrm{H}(d)$ for observed frequency of case status *d*. Here, a tangent to the curve (long-dashed line) is drawn at $\bar{d} = \Pr(o_2) = 0.444$, marked ● on the curve. The tangent has slope $g'(\bar{d})$ = −0.223. Short-dashed lines show the projections of the observed frequencies $d = 0.857$ and $d = 0.403$ from the curve onto the horizontal axis. PSEP is calculated as the horizontal difference between these projections (= 0.454). The vertical distances between the curve and the tangent (i.e., between points marked ▼ and ▲) at *d* = 0.857 (= 0.369) and at *d* = 0.403 (= 0.004) are Bregman divergences (in nits) (equation 7). The frequency-weighted average Bregman divergence is then resolution (RES = 0.037 nits, equation 8). All calculations correct to 3 d.p.

Fig. 4. Reliability, Scenario C1, C2 (see Table 1). The solid line is the curve $g(f) = -\mathrm{H}(f)$ for probability forecasts *f*. Here, tangents to the curve (long-dashed lines) are drawn at probability forecasts based on Scenario C1, $f = \Pr(o_2|f_2) = 0.609$ and $f = \Pr(o_2|f_1) = 0.058$, both points marked ● on the curve. The tangent at $f = 0.609$ has slope $g'(f) = 0.442$, the tangent at $f = 0.058$ has slope $g'(f) = -2.793$. Short-dashed lines show the projections of the observed frequencies based on Scenario C2, $d = 0.824$ and $d = 0.250$, from the curve onto the horizontal axis. The vertical distances between the curve and the tangent (i.e., between points marked ▼ and ▲) at $d = 0.824$ (= 0.108) and at $d = 0.250$ (= 0.195) are Bregman divergences (in nits) (equation 7). The frequency-weighted average Bregman divergence is then reliability (REL = 0.144 nits, equation 9). All calculations correct to 3 d.p.

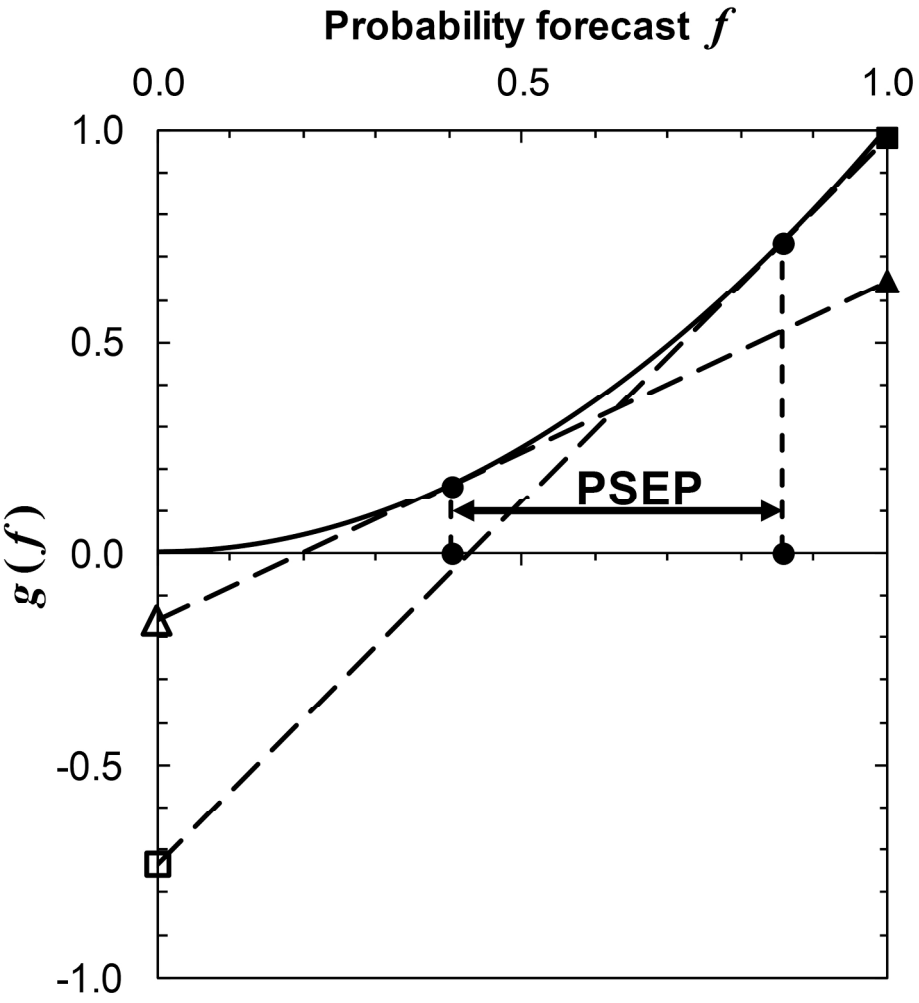

Figure 1. Caption in main document.

111x115mm (600 x 600 DPI)

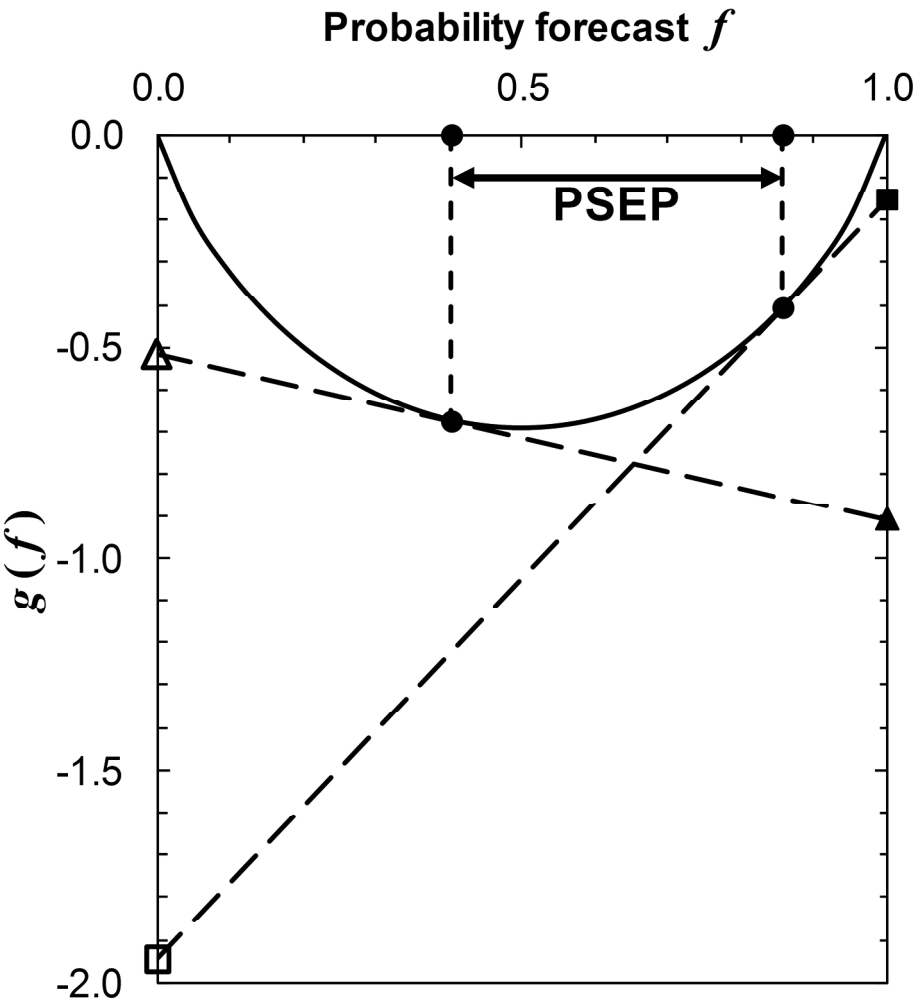

Figure 2. Caption in main document.

111x115mm (600 x 600 DPI)

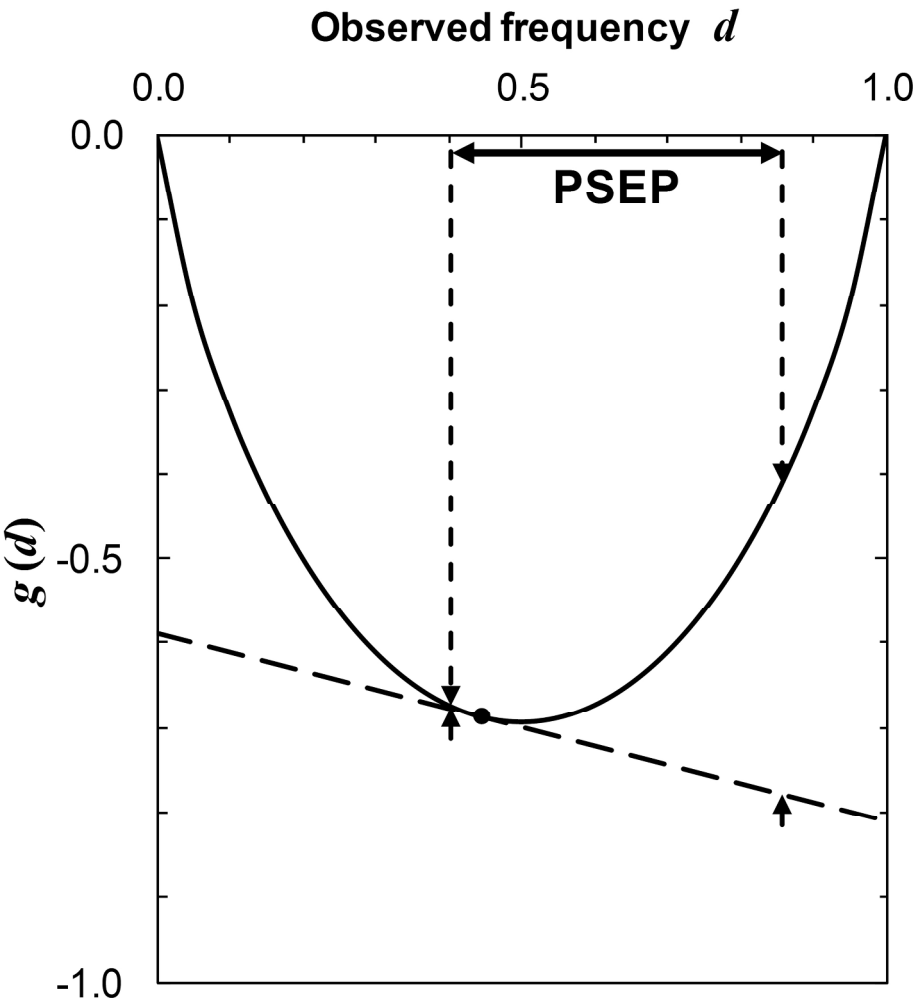

Figure 3. Caption in main document.

111x115mm (600 x 600 DPI)

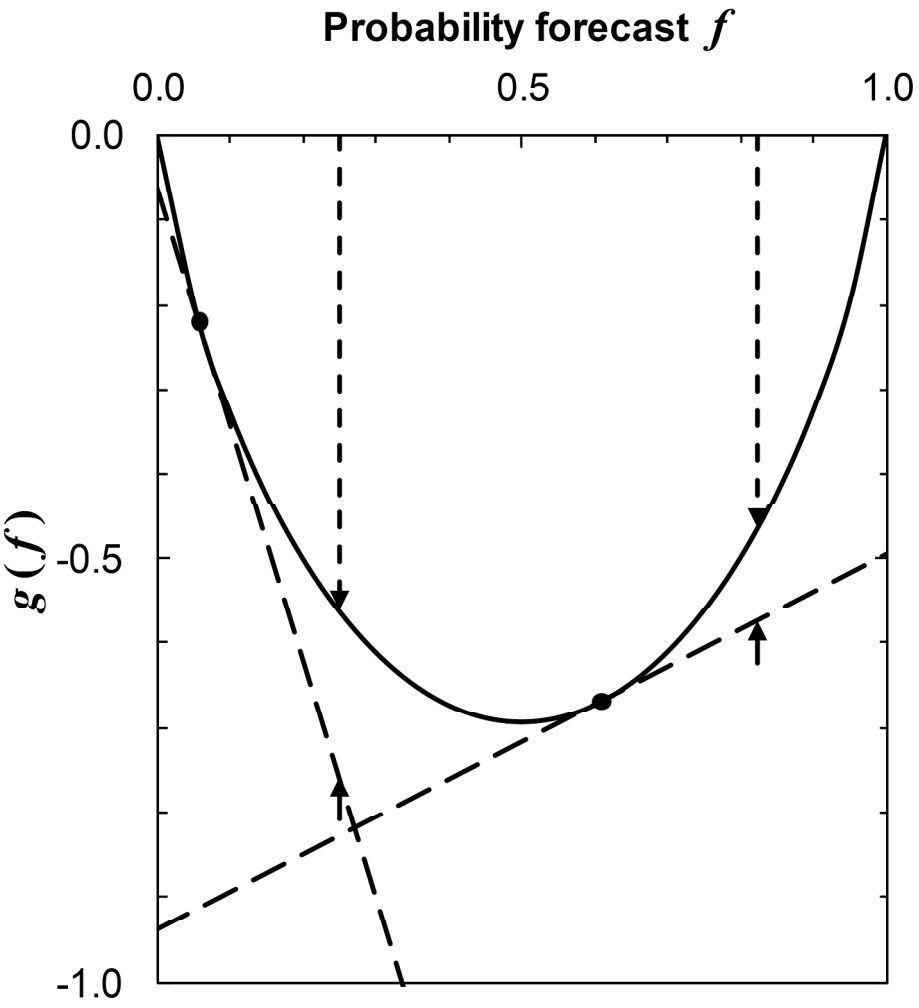

Figure 4. Caption in main document.

111x115mm (600 x 600 DPI)