Risk Analysis, Vol. 34, No. 3, 2014 DOI: 10.1111/risa.12127

Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models

Mark Steyvers,[1,*] Thomas S. Wallsten,[2] Edgar C. Merkle,[3] and Brandon M. Turner[4]

We propose the use of signal detection theory (SDT) to evaluate the performance of both probabilistic forecasting systems and individual forecasters. The main advantage of SDT is that it provides a principled way to distinguish the response from system diagnosticity, which is defined as the ability to distinguish events that occur from those that do not. There are two challenges in applying SDT to probabilistic forecasts. First, the SDT model must handle judged probabilities rather than the conventional binary decisions. Second, the model must be able to operate in the presence of sparse data generated within the context of human forecasting systems. Our approach is to specify a model of how individual forecasts are generated from underlying representations and use Bayesian inference to estimate the underlying latent parameters. Given our estimate of the underlying representations, features of the classic SDT model, such as the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), follow immediately. We show how our approach allows ROC curves and AUCs to be applied to individuals within a group of forecasters, estimated as a function of time, and extended to measure differences in forecastability across different domains. Among the advantages of this method is that it depends only on the ordinal properties of the probabilistic forecasts. We conclude with a brief discussion of how this approach might facilitate decision making.

KEY WORDS: AUC; Bayesian methods; combining forecasts; evaluating forecasts; judgmental forecasting; probability forecasting; ROC analysis; signal detection theory

1. INTRODUCTION

A fundamental issue for forecasting binary events is to distinguish the occurrence of an event, E, from its nonoccurrence, $\neg E$. Signal detection theory (SDT)[1–3] offers a natural framework to analyze the performance of forecasting systems and individual human forecasters in these situations. The key idea is to characterize the distributions that explain the forecasts associated with E ("signal") separately from the forecasts associated with $\neg E$ ("noise"). Based on the underlying distributions, receiver operating characteristic (ROC) analysis can be used (as explained later) to derive measures of diagnosticity such as the area under the curve (AUC). The AUC measure assesses the ability to discriminate E from $\neg E$, unconfounded by response bias and base rates. In addition to characterizing forecasting performance, ROC analysis can be used to analyze performance under different decision thresholds and misclassification costs. SDT and ROC analyses have long been

[1]Department of Cognitive Sciences, University of California, Irvine, CA, USA.
[2]Department of Psychology, University of Maryland, College Park, MD, USA.
[3]Department of Psychological Sciences, University of Missouri, Columbia, MO, USA.
[4]Department of Psychology, Stanford University, Stanford, CA, USA.
*Address correspondence to Mark Steyvers, Department of Cognitive Sciences, 2316 Social & Behavioral Sciences Gateway Bldg., University of California, Irvine, CA 92697-5100, USA; mark.steyvers@uci.edu.

used to evaluate forecasting or diagnostic systems in a variety of areas (examples to follow), but have only recently been proposed in the prediction of uncertain world events.[4]

In this article, we will focus on the problem of evaluating individual human forecasters and aggregates of human forecasts related to uncertain world events such as the United Kingdom leaving the European Union, the winner of the 2012 South Korean presidential election, and group on declaring bankruptcy. Recent work in this area[4] assumes that the forecaster produces binary forecasts (e.g., "target event will occur"). However, forecasts for binary events are often uncertain and therefore accompanied by a confidence or probability estimate (e.g., there is a 70% chance that the event will occur within a specific time frame). One solution is to use rating-scale SDT models where a judgment is discretized into a number of ordered response bins.[3] We explore another solution in the form of a new type of SDT approach where we directly model the continuous judgments without assuming a discrete ordered scale.

Furthermore, in forecasts based on human judgment,[5] problems of data sparsity can arise that can complicate standard SDT analyses. For example, forecasters might choose to answer only a few forecasting problems or they might answer many. In addition, an individual who contributes probabilistic forecasts to only a few problems might also have an associated imbalance in the number of forecasts associated with E and $\neg E$. In an extreme situation, we might have a forecaster who contributes judgments to only problems that resolve as E or that resolve as $\neg E$. Data sparsity and imbalance can result in unstable parameter estimates, potentially leading one to draw inaccurate conclusions about the forecasters in question. Therefore, one should take care when estimating parameters and conducting ROC analyses in such cases.

Other challenges for the evaluation of human forecasts are that they might be dependent on the time of judgment relative to the resolution of the problem and/or on the forecasting domain. In the case of time, one might expect that forecast diagnosticity improves dynamically over time as the point of forecast resolution approaches. Similarly, one might expect that forecasting domains vary in inherent uncertainty, with some domains being more forecastable than others. Thus, a reasonable SDT analysis should account for differences in temporal dynamics as well as differences among forecasting domains.

To address these challenges, we introduce a new SDT modeling framework designed to handle judged probabilities rather than discrete decisions or confidence ratings. Our approach is to specify a model of how individual forecasting judgments are generated from underlying belief distributions, and use Bayesian inference to estimate the distributions. We will use the Beta distribution to model probabilistic judgments and refer to the overall approach as the Beta-SDT model. To address the data sparsity problem, we use a hierarchical Bayesian approach in which the belief distributions for individual forecasters are based on group-level belief distributions that characterize the commonalities across individuals. Hierarchical Bayesian SDT models have been used successfully to model individual differences as well as differences among items[6–8] and are useful in measuring individual performance in the context of sparse data.[9] We also show how the SDT approach can be extended to measure forecasting performance at different points in time relative to the closing date.[5] Finally, we show how to account for domain-specific forecasting differences. Generally, one attractive feature of our hierarchical Beta-SDT approach is that we can flexibly incorporate forecaster-specific and problem-specific effects.

After the Beta-SDT distributions are estimated, standard ROC analysis can be applied to the inferred distributions in order to evaluate performance. ROC analysis had its roots in statistical decision theory, psychophysics, and radar engineering,[10,11] but is now used in many areas to evaluate predictive models including artificial intelligence and machine learning,[12–14] medicine[15–17] and meteorology,[18–21] and other domains.[22] We will focus on the AUC as a summary statistic of the ROC curve. The AUC is a widely used measure for ranking and classification performance.[23] We use the AUC as a way to assess the ability of individual forecasters and forecasting systems to discriminate between E and $\neg E$. The advantages of the AUC as an index of diagnostic performance are outlined in Section 2.1.

An advantage of the Bayesian approach to analyzing probabilistic forecasts is that it naturally leads to distributions over the AUC value for a

[5]The closing date is the date at which the forecasting problem will resolve and is closely related to the forecasting horizon, i.e., the length of time into the future for which forecasts are to be prepared. In the data sets that we focus on, each forecasting problem is associated with a fixed closing date but individual forecasters can choose to forecast on any date before the closing date, effectively leading to different forecasting horizons across forecasters.

particular forecaster or forecasting method. These distributions can be expressed as credible intervals—the Bayesian version of a confidence interval—which is important when interpreting the performance of any predictive method.[24] For example, we might want to know which individual forecasters are performing better than chance or whether one forecaster is performing better (in terms of AUC) than another forecaster.

In the rest of the article, we first give a brief introduction to empirical ROC curves that can be used to calculate the AUC without using an SDT model. We discuss the potential problems of estimating the AUC through empirical ROC curves. We then introduce the Beta-SDT model for judged probabilities that gives a parametric model for ROC analysis. We discuss a number of Beta-SDT modeling variants that allow us to measure differences among individuals as well as forecasting domains, and we also specify a temporal SDT model that can measure changes in discrimination ability over time. We apply the models to data from a recent forecasting study,[5] comparing the inferred AUC values at the level of individual forecasters as well as aggregates of forecasting judgments. Finally, we compare the AUC to conventional measures of forecasting performance (i.e., the Brier score and its decompositions) and argue that the AUC is preferable to these measures if the goal is to measure diagnosticity.

2. EMPIRICAL ROC ANALYSIS

ROC analysis is used across many disciplines as a way to evaluate the diagnosticity of human judgments and artificial systems.[10–13,16,17,22,25] We will first provide a basic review of *empirical* ROC analysis in the context of forecasting systems, in which no assumptions are made about underlying belief distributions, and therefore no SDT assumptions are needed for the analysis. For a more in-depth tutorial on ROC analysis, we refer the reader to Refs. 23, 26, and 27.

In our forecasting context, individual forecasters and forecasting systems produce a probability judgment on a continuous scale about E. We can obtain binary classifications by comparing the probability judgments against a decision threshold. For all cases where the judgment exceeds or equals the threshold, we predict E, and $\neg E$ otherwise. ROC analysis provides a graphical means to explore the tradeoff between hit rates and false alarm rates at various decision thresholds. The hit rate (also known as sensitivity) measures the proportion of those cases in which E occurs (or is true) that the decisionmaker (DM) (or system) forecasted E. The false alarm rate (also known as the complement of specificity, or 1-specificity) measures the proportion of those cases in which E does not occur (i.e., $\neg E$ is true) that the DM (or system) forecasted E. The ROC curve specifically traces the hit rates against the false alarm rates under decision thresholds that vary from the minimum to the maximum value (see Fig. 1). High decision thresholds lead to relatively few forecasts of E and therefore to low hit and false alarm rates; and conversely, low decision thresholds lead to many forecasts of E and correspondingly high hit and false alarm rates.

2.1. The AUC

The area under the ROC curve traced by the hit and false alarm rates as the decision threshold varies serves as a useful summary statistic of performance. A perfect AUC of 1 is achieved when all probability judgments associated with E are higher than judgments associated with $\neg E$. This perfect performance yields a single point at $(0, 1)$, the resulting ROC consists of a line from $(0, 0)$ to $(0, 1)$ and another from $(0, 1)$ to $(1, 1)$, and as a consequence AUC $= 1$. A random guessing strategy (e.g., generating random probabilities between 0 and 1) leads to an AUC of 0.5.[6] An AUC smaller than 0.5 is indicative of a system that is diagnostic but where better performance can be obtained by reporting the complement of the probability judgments. In many practical situations, the AUC lies between 0.5 and 1. A chance-corrected AUC is also known as the Gini coefficient,[28] G (i.e., $G = 2AUC - 1$, the area above the diagonal).

The AUC has a natural interpretation as a ranking measure: it is equivalent to the probability that a forecaster assigns a higher probability judgment to a randomly chosen E relative to a randomly chosen $\neg E$. Therefore, what matters are the ordinal relationships among the judged probabilities and not the numerical values of the probabilities *per se*. For example, if all forecasts associated with E and $\neg E$ are judged with values of 0.51 and 0.49, respectively, this would lead to a perfect AUC value of 1 even though the forecasts are hardly separated

[6]Note that while a random guessing strategy always leads to an AUC of 0.5, the converse is not necessarily true. An AUC of 0.5 and an ROC curve close to the diagonal is required to indicate a random guessing strategy.

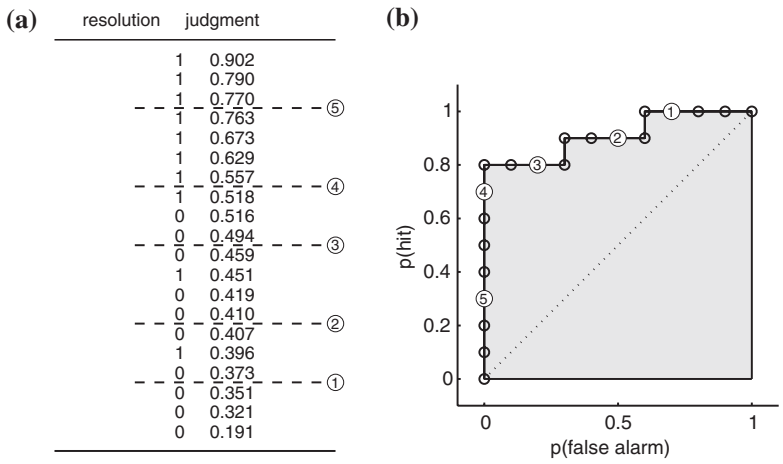

Fig. 1. An example of constructing an empirical ROC curve from probabilistic forecasting judgments. The probabilistic judgments from 20 problems are arranged in order of magnitude along with the associated resolution (1=E is true, 0=$\neg E$ is true). The ROC plot shows the hit rate against the false alarm rate under a variety of decision thresholds varying from 0 to 1. One practical method considers only $K+1$ thresholds when there are K unique values observed in the forecasts. In this example, this method leads to 21 possible thresholds. Five example decision thresholds are shown in (a) with corresponding ROC points in (b). One way to derive the AUC value is by calculating the area under the curve shown in gray. For this example, the AUC is 0.91.

quantitatively. In this example, what matters is that the forecasts associated with E are *consistently* higher than the forecasts associated with $\neg E$. Any AUC value higher than 0.5 indicates that the forecaster is performing above chance in discriminating between E and $\neg E$.

One attractive feature of the AUC is that it is independent of any threshold required for making decisions (i.e., deciding whether a forecast should lead to an E or $\neg E$ decision based on the judged probability). This feature is useful in deployment contexts where the costs of misclassifications (e.g., incorrectly predicting the occurrence of an event or missing the actual occurrence of an event) cannot be determined in advance, and a measure is needed that averages over a wide range of misclassification costs and associated decision thresholds. From another perspective, given a well-established ROC, the DM can estimate hit and false alarm probabilities associated with any forecast, combine them with her cost-benefit matrix, and make a decision accordingly. Another attractive feature of the AUC is that it is independent of the extent to which forecasts are calibrated. As we indicated above, any strictly monotonic transformation of the probabilistic forecasts will yield the same AUC. What matters are the *ordinal* relationships among the probabilities assigned to E and $\neg E$. This is important because statistical techniques can be used to calibrate individual forecasters,[29] but the fundamental measure of interest is the discrimination ability of individual forecasters. Indeed, because it is only the ordinal relationships among the judgments that are needed for the AUC, judged probabilities are not required. A discrete ordered scale also would do. For example, a seven-category rating scale might be used with ratings such as *very unlikely*, *unlikely*, *somewhat unlikely*, *50-50*, *somewhat likely*, *likely*, and *very likely*. Therefore, the same measure of diagnosticity can be applied to both probabilistic forecasts based on probability judgments and judgments based on a discrete ordered scale.

3. BETA-SDT MODELS FOR PROBABILISTIC FORECASTS

In the basic SDT approach to forecasting of dichotomous outcomes, all resolved forecasting problems are separated into two classes: problems where the event of interest occurred (E, or "signal" trials) and problems where the event did not occur ($\neg E$, or "noise" trials). Prior applications of SDT to forecasting[4,18–20,30] have assumed that the forecasting system produces discrete judgments in the form of warnings—either a warning is produced about an upcoming event or it is not. Models for this case have often taken the standard Gaussian SDT approach where internal evidence values are sampled from latent Gaussian distributions. These values are then compared against a decision threshold. For cases where the evidence exceeds the threshold, a warning is issued. It is important to note that although the vast majority of SDT models make Gaussian assumptions, the underlying theory of signal detection does not depend on the form of the distributions and other types of distributions are possible.[31] In principle, the standard SDT approach can be extended to probability judgments on a rating scale by assuming that the judgments fall into a

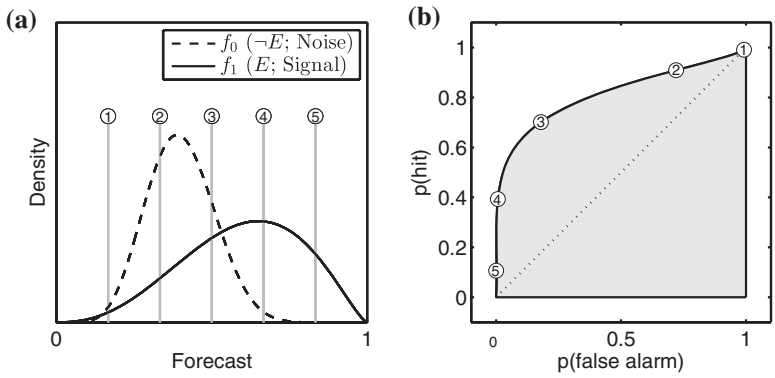

Fig. 2. Illustration of the Beta-SDT model and ROC analysis. In the SDT analysis, all resolved forecasting problems are separated into two classes: problems where the event of interest occurred (E; "signal" trials) and problems where the event did not occur ($\neg E$; "noise" trials). The probabilistic forecasts in these two classes are modeled by two separate Beta distributions, f_1 and f_0 (a). The ROC plot in (b) traces the hit and false alarm rates for all possible decision thresholds. Five points are marked corresponding to the thresholds shown in (a). The AUC value corresponds to the area under the curve shown in gray. In this example, the AUC is 0.82.

predetermined number of ordered categories (e.g., 0–0.1, 0.1–0.2, etc.), which requires multiple internal response criteria to be estimated.

The key difference in our approach is that we model the judgments as arising from a simple family of distributions based on the Beta distribution, as opposed to the normal. In addition, our approach does not require any response criteria to be estimated. Fig. 2 illustrates the basic approach. The probabilistic forecasts for the two classes of problems are modeled by two separate probability distributions. We assume that $f_1(y)$ and $f_0(y)$ describe the density functions for probabilistic forecasts y associated with signal and noise trials, respectively. Just as in standard SDT, the parameters that govern the shape and location of the signal and noise distribution are latent and need to be estimated on the basis of the observed data. The main difference is that in standard SDT, the observed data consist of discrete judgments or rating scales whereas in our approach, the observed data consist of continuous probability judgments. These judgments are assumed to be the direct result of a sampling process from the underlying signal and noise distribution. In standard SDT, the internal evidence values sampled from the signal and noise distributions are first compared to a decision or rating scale threshold in order to generate a discrete judgment.

The ROC plot, then, is simply a plot of the decumulative distribution (i.e., one minus the cumulative distribution) under $f_1(y)$ as a function of the decumulative under $f_0(y)$, calculated for each point from 0 to 1 on the probability judgment axis. The former is the hit rate, the probability of calling a signal a signal and the latter is the false alarm rate, the probability of calling a nonsignal (noise) a signal . Note that this process is conceptually identical to that used to create ROC plots in standard SDT, in which the hit and false alarms rates are the areas to the right of the threshold under the signal and noise distributions, respectively, for any given threshold location.

In the remainder of this section, we describe how we can model the underlying belief distribution f_k with a Beta distribution with separate parameters for the signal ($k = 1$) and the noise ($k = 0$) distributions. The Beta distribution[7] is a natural distribution for modeling probabilistic forecasts[31] because it inherently produces values bounded between 0 and 1.

We will describe three versions of Beta-SDT models for the evaluation of probabilistic forecasts. The first model applies to situations where a number of individual forecasters are forecasting on overlapping sets of problems and the diagnosticity of each individual forecaster needs to be assessed. In this approach, we use a hierarchical Bayesian model to address data sparsity problems, in which some individual forecasters produce judgments on a small number of problems and some problems have only a small number of forecasts. The second model extends the hierarchical approach to model differences among forecasting domains and the time of judgment relative to the forecasting horizon. In this approach, the signal and noise distributions are made functionally dependent on the time of judgment as well as the forecasting domain, allowing us to estimate how fast diagnosticity improves as forecasters come closer to the forecasting horizon and more accurate information becomes available. Finally, the third model, achieved by removing the hierarchical component from the first model, applies to situations where only

[7]The Beta distribution has the density function $f_k(y|a_k, b_k) = \frac{1}{\mathrm{B}(a_k,b_k)} y^{a_k-1}(1-y)^{b_k-1}$, where $\mathrm{B}(a, b)$ is the Beta function, $\mathrm{B}(a, b) = \int_0^1 y^{a-1}(1-y)^{b-1}\mathrm{d}y$.

a single forecasting system or individual forecaster is evaluated and there are no issues of data sparsity.

3.1. Basic Model

The hierarchical model handles data sparsity by associating judges with their own signal and noise distributions, pooling statistical strength across judges, and estimating the distributions at the individual judge level.

To introduce notation, let N be the number of resolved events and M be the number of judges. Let $y_{i,j}$ represent the probabilistic forecast of the ith judge for the jth event where $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N\}$. Let x_j represent the coded outcome for Event j, setting $x_j = 0$ if the jth event did not occur (corresponding to a noise trial) and $x_j = 1$ if the jth event did occur (corresponding to a signal trial).

Each forecast $y_{i,j}$ is assumed to arise from one of two Beta distributions (corresponding to the signal and noise distributions) depending on the outcome x_j:

$$y_{i,j} \sim \begin{array}{l} \text{Beta}\,(a_{i,0}, b_{i,0}))\,, \text{ if } x_j = 0, \\ \text{Beta}\,(a_{i,1}, b_{i,1})\,, \quad \text{if } x_j = 1. \end{array} \tag{1}$$

The parameters $a_{i,k}$ and $b_{i,k}$, where $k \in \{0, 1\}$, determine the shapes of the Beta distributions for the signal and noise conditions. In the model, we determine these parameters in terms of mean and precision parameters:

$$\begin{aligned} a_{i,k} &= \mu_{i,k} \exp(\xi_{i,k}) \\ b_{i,k} &= (1 - \mu_{i,k}) \exp(\xi_{i,k}). \end{aligned} \tag{2}$$

The first parameter μ, $0 < \mu < 1$, is the mean of the distribution. The second parameter ξ is a precision parameter that determines how concentrated the distribution is around the mean. At the next level, we assume that the means of the individual judge distributions are sampled from a logit-normal distribution:

$$\text{logit}(\mu_{i,k}) \sim \text{N}(\mu_k^*, \phi_k^*), \tag{3}$$

where μ_k^* and ϕ_k^* are the mean and precision of this group-level distribution.

In Equation (2), it is convenient to use an exponential transformation on ξ. This allows us to model the precision parameters $\xi_{i,k}$ by a normal distribution:

$$\xi_{i,k} \sim \text{N}(\delta_k^*, \tau_k^*). \tag{4}$$

To complete the model, we need to specify prior probability distributions on the parameters μ_k^*, ϕ_k^*, δ_k^*, and τ_k^*. One approach would be to inform our priors by previous experimental research or theory to find plausible ranges for these parameters. However, we are unaware of prior research (or theory). Therefore, we set the priors with the goal to be as uninformative as possible about the means and precisions of the probabilistic forecasts consistent with computational simplicity. Specifically, we use normal priors on the means and inverse Gamma (IG) priors on the precisions:

$$\mu_k^* \sim \text{N}(0, 0.35), \tag{5}$$

$$\phi_k^* \sim \text{IG}(1, 1), \tag{6}$$

$$\delta_k^* \sim \text{N}(2, 0.01), \tag{7}$$

$$\tau_k^* \sim \text{IG}(1, 1). \tag{8}$$

The N(0, 0.35) prior on μ_k^* in Equation (5) was chosen such that the means $\mu_{i,k}$ in the probability scale (i.e., after the inverse logit transform in Equation (3)) are approximately uniform. The IG(1,1) priors in Equations (5) and (7) are noninformative priors for precision parameters. The prior on δ_k^* places most density on low precisions (high variances). This is useful because previous research has found the Beta distribution's precision parameter to generally be overestimated.[32] However, this distribution also places enough density on high precisions to counteract the density on low precisions, resulting in a minimally informative prior. For similar priors in other Beta-distributed models, see Ref. 33.Taken together, these priors only provide minimal information about $\mu_{i,k}$ and $\xi_{i,k}$.

3.2. Incorporating Temporal and Domain Effects

We can extend the hierarchical model by incorporating the effects of the time of judgment and forecasting domain. With respect to forecasting domain, we assume that it is easier to provide accurate forecasts in some domains than in others and so they will be associated with higher diagnosticity, and therefore greater area under the ROC. With respect to the role of time, we expect that forecasts made closer to the forecasting horizon will be more accurate than those made earlier in time.

We incorporate time effects by allowing the mean μ of the signal and noise distributions in Equation (3) to vary as a function of time. We consider growth functions of the form $\text{logit}(\mu) = b + \exp(-\gamma t)(a - b)$, where t is the time at which the judge forecasts the event, expressed in days before resolution, a is the intercept when $t = 0$, b is the asymptotic value as t approaches infinity, and γ is a scaling parameter. There are many ways to parameterize a, b, and γ on the basis of item, person, and outcome differences.

To introduce notation, let K be the number of forecasting domains ($K < N$), and $c_j \in \{1, \ldots, K\}$ represent the forecasting domain for the jth event. In the hierarchical model, we replace Equation (3) with:

$$\begin{aligned} \text{logit}(\mu_{i,j}) &= b_{1,i,x_j} + b_{3,c_j,x_j} + \exp(-\gamma_{c_j} t_{i,j}) \\ &\quad (b_{0,i,x_j} + b_{2,c_j,x_j} - b_{1,i,x_j} - b_{3,c_j,x_j}) \\ b_{0,i,x_j} &\sim \text{N}(\mu_{0,x_j}, \phi_{0,x_j}) \\ b_{1,i,x_j} &\sim \text{N}(\mu_{1,x_j}, \phi_{1,x_j}) \\ b_{2,c_j,x_j} &\sim \text{N}(\mu_{2,x_j}, \phi_{2,x_j}) \\ b_{3,c_j,x_j} &\sim \text{N}(\mu_{3,x_j}, \phi_{3,x_j}), \end{aligned} \tag{9}$$

where t_{ij} is the time at which Judge i forecasted Event j, expressed in days before resolution. The b_1 and b_3 parameters, respectively, account for forecaster and domain differences in the intercept of the growth function, while the b_0 and b_2 parameters, respectively, account for forecaster and domain differences in the asymptote of the growth function. Note that in this parameterization, the scaling variable γ_k in the growth function depends on the forecasting domain k, allowing for differences in the temporal dynamics of diagnosticity across domains.

To complete the model, we place uninformative priors on the means and precisions of the b parameters:

$$\begin{aligned} &\mu_{0,x_j} \sim \text{N}(0, 0.35),\ \mu_{1,x_j} \sim \text{N}(0, 0.35), \\ &\mu_{2,x_j} \sim \text{N}(0, 0.35),\ \mu_{3,x_j} \sim \text{N}(0, 0.35), \\ &\phi_{0,x_j} \sim \text{IG}(1, 1),\ \phi_{1,x_j} \sim \text{IG}(1, 1), \\ &\phi_{2,x_j} \sim \text{IG}(1, 1),\ \phi_{3,x_j} \sim \text{IG}(1, 1). \end{aligned} \tag{10}$$

Note that these priors are equivalent to the priors used in Equations (5) and (6) and lead to an approximately uniform distribution (*a priori*) over the b parameters.

We also place a uniform prior on the temporal scaling variable γ_k:

$$\gamma_k \sim \text{Uniform}(0.0001, 0.75). \tag{11}$$

This prior allows for a flexible range in the temporal scaling of the growth function.

3.3. Nonhierarchical Model

Finally, we can also evaluate diagnosticity of a single forecaster in situations where data sparsity is not an issue. In this case, we can create a nonhierarchical model variant of the basic model. To simplify notation, we can omit the index for judges and let y_j represent the probabilistic forecast for event j where $j \in \{1, \ldots, N\}$. Each forecast y_j is sampled from either the signal and noise distribution depending on the outcome x_j:

$$y_j \sim \begin{array}{l} \text{Beta}\,(\mu_0 \exp(\xi_0), (1 - \mu_0)\exp(\xi_0)) \text{ if } x_j = 0, \\ \text{Beta}\,(\mu_1 \exp(\xi_1), (1 - \mu_1)\exp(\xi_1)) \text{ if } x_j = 1. \end{array} \tag{12}$$

Instead of specifying group-level distributions as the basic model, we can complete the model by placing priors on the means and precisions μ_k and ξ_k, respectively. As before, we place a normal prior on the logit of μ_k and a normal prior on ξ_k:

$$\text{logit}(\mu_k) \sim \text{N}(0, 0.35),\ \xi_k \sim \text{N}(2, 0.01). \tag{13}$$

3.4. ROC and AUC Analysis with Beta-SDT Models

Once the signal and noise densities (f_1 and f_0, respectively) are estimated in any of three modeling approaches we described, the procedure for calculating the hit (F_1) and false alarm rates (F_0) involves integrating over the signal and noise densities f_1 and f_0:

$$F_1 = \int_c^1 f_1(y)\text{d}y, \text{ and} \tag{14}$$

$$F_0 = \int_c^1 f_0(y)\text{d}y, \tag{15}$$

where c is a point on the probability forecast axis. Based on the signal and noise densities or the hit and false alarm rates, the AUC can then be computed[24]

by calculating:

$$AUC = \int\limits_{y_1 > y_2} \int f_1(y_1)\mathrm{d}y_1\, f_0(y_2)\mathrm{d}y_2 = \int_0^1 F_1(y)\mathrm{d}F_0(y). \tag{16}$$

For the Beta distribution, the AUC has an analytic expression if the parameters of the Beta distribution have integer values.[31] This does not apply in our situation, and we have to use numerical integration to calculate Equations (14)–(16).

4. DATA AND MODEL ESTIMATION

To evaluate the signal detection modeling approach for probabilistic forecasting judgments, we use data from a total of 1,309 participants collected by the aggregative contingent estimation system (ACES), a large-scale project for collecting and combining forecasts of many widely dispersed individuals (http://www.forecastingace.com/aces). A preliminary description of the data-collection procedure can be found in Ref. 5.Volunteer participants were asked to estimate the probability of various future events' occurrences, such as the outcome of presidential elections in Taiwan and the potential of a downgrade of Greek sovereign debt. Participants were free to log on to the website at their convenience and forecast any items of interest. A median of 52 forecasters contributed to each forecasting problem. For this article, we focused on a subset of 176 resolved binary forecasting problems. The forecasting problems were categorized into *a priori* $K = 5$ forecasting domains, including politics and policy ($N = 82$), business and economy ($N = 39$), science and technology ($N = 16$), military and security ($N = 23$), and sports, health, and social ($N = 16$). All forecasting problems involved a standard way of framing the event and were presented in the form: Will event A happen by date B? This last constraint excluded a small number of events from the current analysis where the event was framed in terms of a deviation from status quo (e.g., will A remain true by date B?). In this data set, 39 of the 176 events happened before the closing date ($x_j = 1$), such that the base rate of event occurrence, $\bar{x} = 0.22$.

4.1. Parameter Inference

We used JAGS[34] to estimate the joint posterior distribution of each set of model parameters in the Beta-SDT models. For each model, we obtained 1,000 samples from the joint posterior after a burn-in period of 1,000 samples, and we also collapsed across seven chains.

4.2. Performance Measures

We use a number of measures to evaluate forecasting performance, including AUC, Brier scores, and measures of (mis)calibration, as explained below. For each of these measures, we evaluate individual forecasters as well as aggregates of probabilistic forecasts. To simplify notation, we omit the indexing over individuals and the resulting value of the performance statistic refers to a specific forecaster or aggregation method.

4.2.1. AUC

One of the advantages of the posterior sampling approach in the Beta-SDT models is that we can infer distributions of the AUC value. We obtain these distributions by calculating the estimated signal and noise densities f_0 and f_1 for each posterior parameter sample. We can then calculate Equations (14)–(16) for each posterior sample in order to get distributions over the AUC value. From these distributions, we will derive the 95% credible intervals. We will also calculate the empirical AUC value as explained in Section 2 when sufficient data are available. This allows us to compare the results from the Beta-SDT procedure with the results obtained through standard empirical ROC analysis.

4.2.2. Global Calibration Offset (GCO)

A previous study using the same data found that many individual forecasters are poorly calibrated and systematically overestimate the likelihood of future events.[29] We will evaluate the calibration of forecasters with a single measure that assesses the degree to which forecasters overestimate the likelihood of future events. Our measure, which we call GCO, measures the log-difference between the mean (expected) probabilistic forecasts, derived from the Beta-SDT model, and the base rate of events. Specifically, the GCO is based on the contrast between the expectation of probabilistic forecasts $\hat{y}$ derived from the Beta-SDT model and the base rate of events $\bar{x}$:

$$\begin{aligned} \mathrm{GCO} &= \log(\hat{y}/\bar{x}) \\ &= \log\big((\bar{x}\hat{y}_1 + (1-\bar{x})\hat{y}_0)/\bar{x}\big). \end{aligned} \tag{17}$$

This definition of GCO assumes that $\bar{x}$ can never be zero and that any processes that lead to over- and underestimation have multiplicative effects on the judged probabilities.[8] Note also that in the second equation, the expectation of probabilistic forecast is based on the empirical base rates of event occurrence—it is simply the average of the (inferred) means of the signal and noise distributions weighted by the empirical base rates of event occurrence. Overall, a GCO value is greater (less) than zero if the expected probabilistic forecasts are greater (smaller) than the empirical base rate. A GCO value of zero indicates that there are neither overestimation nor underestimation errors.

4.2.3. Brier Scores and Mean Predictive Error (MPE)

Finally, we will also evaluate models through use of the Brier score.[35–37] Because all events involved only two outcomes, the Brier score for the jth event can be expressed as:

$$B_j = (x_j - y_j)^2, \tag{18}$$

where y_j is the probabilistic forecast for Event j and x_j is the resolution of the jth event. Thus, in this definition of the Brier score, the best score B_j is zero, and the worst possible score is one. After the Brier scores are obtained for each event, we compute the MPE by averaging the Brier scores across the number of events N:

$$MPE = \frac{1}{N}\sum_{j=1}^{N} B_j. \tag{19}$$

4.3. Forecast Aggregation Methods

There exists a large body of work focused on the use of statistical models for combining individual subjective probability judgments into a single probability estimate.[38–42] A simple form of aggregation, namely, the unweighted linear average, has proven to be effective in many situations.[35] The goal of this article is not to propose novel aggregation methods for probabilistic forecasts. Instead, we investigate a number of simple aggregation procedures[29] that allow us to highlight the effects of aggregation on diagnosticity and GCO.

[8] An alternative definition of GCO could be based on the difference between $\hat{y}$ and $\bar{x}$.

- ULinOP. The unweighted linear opinion pool (ULinOP) is simply the unweighted average of probabilistic forecasts across judges. Thus, predictions λ_j are obtained by evaluating $\lambda_j = \frac{1}{n_j}(\sum_{i=1}^{n_j} y_{i,j})$, where n_j is the number of responses obtained on event j.
- Calibrated ULinOP. The ULinOP is not necessarily calibrated[29,42] and can be associated with systematic forecasting errors. A simple aggregation method is to recalibrate the unweighted average using a monotonic transformation function f such that $\lambda_j = f(\frac{1}{n_j}(\sum_{i=1}^{n_j} y_{i,j}))$. In this procedure, we chose the linear in log-odds transformation function[43] to recalibrate the unweighted average, where $f(p) = \delta p^{\gamma}/(\delta p^{\gamma} + (1-p)^{\gamma})$, and γ and δ are parameters. We followed the procedures of Ref. 29 to estimate these parameters for our data set.
- Calibrated Time-Weighted Average. Another procedure is to take the time of judgment relative to the forecasting horizon into account. The idea is to upweight forecasts closer to the forecasting horizon as these are expected, on average, to be more accurate. We implemented this in a weighted averaging scheme $\lambda_j = f((\sum_{i=1}^{n} w_{i,j}\, y_{i,j})/(\sum_{i=1}^{n} w_{i,j})))$, where the weight for each judgment is based on an exponential decay function, $w_{i,j} \propto \exp(-t_{i,j}\, c)$, $t_{i,j}$ is the time of judgment expressed in number of days before the forecasting horizon, and c is a scaling constant.
- Guess Baseline. Our final forecasting procedure does not involve aggregating the forecasts at all, but instead relies on the base rate of event occurrence such that $\lambda_j = \bar{x}$. Therefore, using this method, a constant probabilistic forecast is used across all forecasting problems. This method results in zero GCO because the average forecasted probability exactly matches the base rate. Even though this particular method is not psychologically interesting (as it does not rely on human judgment) and computationally simplistic (there is no procedure for estimating the base rate in an online fashion), it is helpful for illustration purposes because it is associated, by definition, with zero diagnosticity (i.e., an AUC value of 0.5) and has zero GCO.

5. RESULTS

In the sections below, we evaluate forecasting performance of individual forecasters as well as

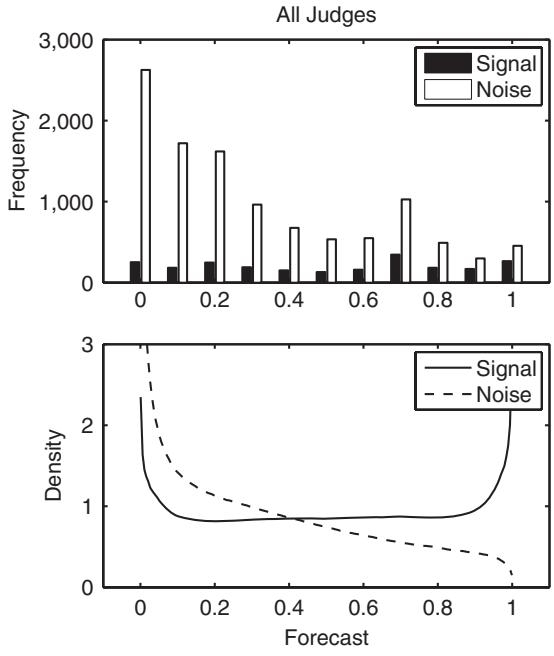

Fig. 3. Empirical results and posterior predictive distributions of the hierarchical Beta-SDT model. The top panel shows the frequency counts of the probabilistic forecasts across all judges separated into problems where the event did or did not occur (signal and noise trials, respectively). The bottom panel shows the posterior predictive distributions when sampling new judges from the population-level distributions in the model.

forecast aggregates using various Beta-SDT models. We first describe the results of the basic hierarchical model and show individual forecaster differences in diagnosticity and calibration. We then show how we can use the nonhierarchical model to evaluate forecast aggregates. By combining the performance measures for individual forecasters and aggregates into one visualization, we show how forecast aggregates improve on the diagnosticity of the majority of individual forecasters. Finally, we show how the hierarchical model with a temporal component allows us to evaluate the temporal changes in forecasting performance as well as differences in those dynamics across forecasting domains.

5.1. Evaluating Individual Forecasters

The top panel of Fig. 3 shows the frequency distribution of probabilistic forecasts across all judges. Because the majority of events did not occur, there are more probabilistic forecasts associated with noise trials than signal trials. The results show that the distribution of forecasts on noise trials is skewed toward the lower probabilities. In contrast, the distribution of forecasts on signal trials is fairly uniform with no indication that higher probabilistic forecasts are favored for events that eventually will occur.

The bottom panel of Fig. 3 shows the posterior predictive distribution of forecasts for the basic hierarchical model. These correspond to the forecast distributions for a new (simulated) judge sampled from the group-level distributions using the distributions in Equations (3) and (4). These posterior predictive distributions describe what the forecasts of an "average" judge looks like according to the model. Note that the empirical distributions and predictions of the model match reasonably well with the one exception that the model favors a U-shaped Beta distribution for the signal trials in contrast to the relatively flat empirical distribution.

Fig. 4 illustrates the results of the hierarchical model for three individual judges. The top panels show the empirical distributions of forecasts. Note that judge #133 performs reasonably well and tends to separate forecasts for signal and noise trials. Judge #935 shows poor performance and does not seem to distinguish forecasts based on eventual outcome. Judge #1198 highlights a common issue when analyzing sparse forecasts from individual users. This judge has provided only a single forecast for a problem where the event did not occur. The middle panels show the inferred distributions for these individual judges. Note that for judge #1198, the model is able to infer a signal distribution even though the judge never contributed any forecasts for this condition. The hierarchical model in this case simply uses the parameter estimates at the group level to infer a distribution. The bottom panels show the inferred ROC curves. Importantly, the ROC curves shown are the mean curves. The model actually infers a distribution over ROC curves from which a distribution over AUC values can be calculated. The 95% credible intervals of AUC values are shown in the figure. Note that the model infers the AUC for judge #935 around 0.5, consistent with the empirical distributions of forecasts not discriminating between signal and noise trials. For judge #1198, the range of AUC values is much larger, indicating large uncertainty about the diagnosticity of this judge. The range includes AUC = 0.5, suggesting the judge might be completely undiagnostic, but most of the range exceeds 0.5, which implies that it is likely that the judge performs better than chance. This inference is based on only a single forecast (and, of course, on the assumptions in the model). However, this one forecast

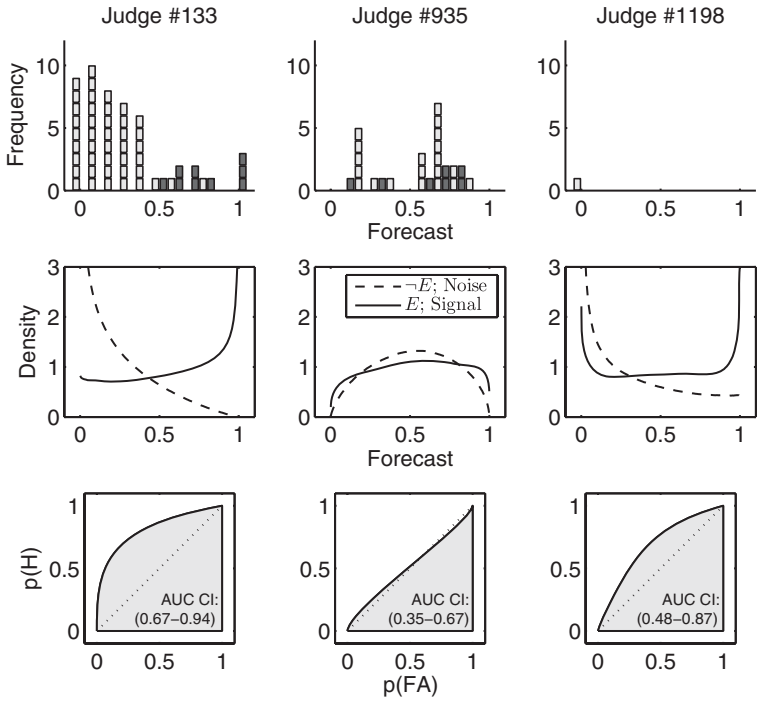

Fig. 4. Results of the hierarchical Beta-SDT model for three individual judges. The top row shows the frequency counts of the judged probabilities separated into problems where the event did or did not occur (filled and open squares, respectively). The middle row shows the estimated Beta distributions. The bottom row shows the ROC curves with the 95% credible interval of the AUC values. Note that the left and middle columns show the results of judges with relatively many judgments, one associated with a high AUC (left column) and the other with a chance-level AUC (middle column). The right column shows a judge with a single probability judgment associated with an event that did not occur. This judge is associated with an expected AUC better than chance but there is a great uncertainty about the exact AUC value.

was quite accurate, which makes it more likely than not (but by no means guaranteed) that this judge will perform well on other problems.

Fig. 5 shows the inferred AUC distributions for individual forecasters. The results show that the majority of individual forecasters have an AUC value around 0.65, but that there are significant individual differences—some forecasters appear to be performing at or near chance whereas others are much better than average.

5.2. Evaluating Forecast Aggregates

Up to this point, we have observed the diagnosticity of individual forecasters. We can also investigate how forecasting performance, as measured by AUC, MPE, and GCO, changes when aggregating over individual forecasts. Table I and Fig. 6 show the forecasting performance for the four aggregation procedures. The results show the AUC values derived from empirical ROC analysis as well as the nonhierarchical Beta-SDT model. The values are quite similar to each other, showing that the two procedures result in the same AUC (as long as the data set is not sparse). The results also show that the effect of (model) calibration (row 2 of Table I) has no effect on the AUC value, but it does remove the overestimation bias (GCO > 0). This is not surprising as the calibration is designed to remove the systematic error (leading to zero GCO) but cannot improve diagnosticity—the calibration procedure involves a strictly monotonic transformation of the probabilistic forecast and the AUC is not sensitive to such transformations. The results also show that taking a weighted average leads to higher diagnosticity relative to an unweighted average (compare the second and third rows of Table I). As a reminder, in the weighted average procedure, recent forecasts are upweighted relative to older forecasts. The difference in diagnosticity demonstrates that forecasters are better able to discriminate between signal and noise as time approaches the forecasting horizon. Finally, the results show that the guessing strategy involving a constant baseline probability is associated with zero diagnosticity (i.e., an AUC around 0.5) and no overestimation (i.e., a GCO value of 0).

Interestingly, the MPE based on Brier scores shows significant changes across all four methods. This is because the Brier score does not separate between diagnosticity and bias components of performance. This could potentially lead to the wrong conclusions. The guessing strategy, for example, shows an MPE that is similar to the MPE of the ULinOP. Based on these results, one might question whether

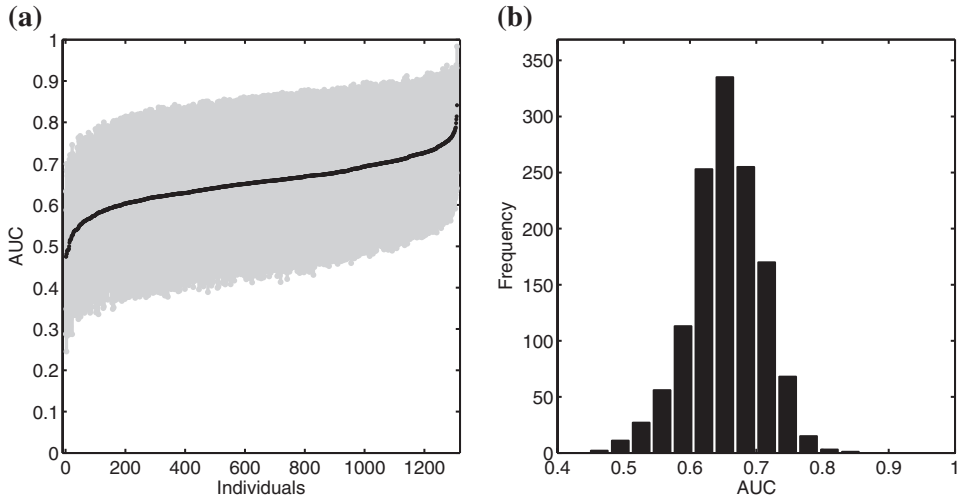

Fig. 5. Estimated AUC values of individual forecasters using the hierarchical Beta-SDT model. Panel (a) shows the mean AUC values for each forecaster ordered from worst to best along with the 95% credible interval shown in the gray area. Panel (b) shows the distribution of AUC values across forecasters.

Table I. AUC, Global Calibration Offset (GCO), and MPE (Brier) Scores for Four Aggregation Methods

Model	Empirical AUC	AUC	GCO	MPE (Brier)
ULinOp	0.834	0.835 (0.753–0.901)	0.605 (0.552–0.656)	0.153 (0.135–0.171)
Calibr. ULinOp	0.828	0.822 (0.732–0.892)	0.042 (−0.077–0.150)	0.120 (0.090–0.152)
Calibr. Weighted Average	0.931	0.932 (0.873–0.968)	0.129 (0.012–0.238)	0.072 (0.052–0.098)
Guess Baseline	0.492	0.488 (0.381–0.592)	−0.009 (−0.016 to −0.001)	0.166 (0.131–0.202)

Notes: The empirical AUC is the area under the curve derived from empirical ROC analysis. AUC is the area under the curve estimated by the nonhierarchical Beta-STD model. The ranges provide the 95% confidence interval.

human forecasters exceed the performance of simple guessing strategies that do not rely on any human judgment. However, based on the AUC, it is clear that aggregates of human probabilistic forecasts carry important diagnostic value that is not present in random guessing strategies.

Overall, the results for AUC and GCO show that some aggregation procedures perform well on diagnosticity but not calibration, and vice versa. Therefore, the AUC can be used to identify aggregation procedures that perform well on diagnosticity, which is arguably the most important goal when developing aggregation methods.

5.3. Comparing Individual Forecasters and Aggregates

We can also investigate how the performance of forecast aggregates compares with the performance of individual forecasters. Fig. 7, left panel, shows the AUC plotted against GCO for individual forecasters as well as the four aggregation methods. The results show that all aggregation procedures except the guessing heuristic are associated with a much higher diagnosticity than the majority of individual forecasters. However, just taking the unweighted average (ULinOP) does not reduce the tendency to overestimate event probabilities (measured by GCO) relative to the individual forecasters. This result is not surprising because averaging is not expected to remove any systematic bias. Visualizing the aggregation procedures and individual forecasters in the AUC versus GCO space clarifies which components of forecasting performance are improved under the different aggregation procedures.

Fig. 7, right panel, shows the Brier scores for individual forecasters in the AUC versus GCO space. Note that many individual forecasters have similar Brier scores but different AUC and GCO values. Fig. 8 plots the Brier score as a function of AUC with a point for each forecaster and points with 95% credible intervals for the various aggregation methods.

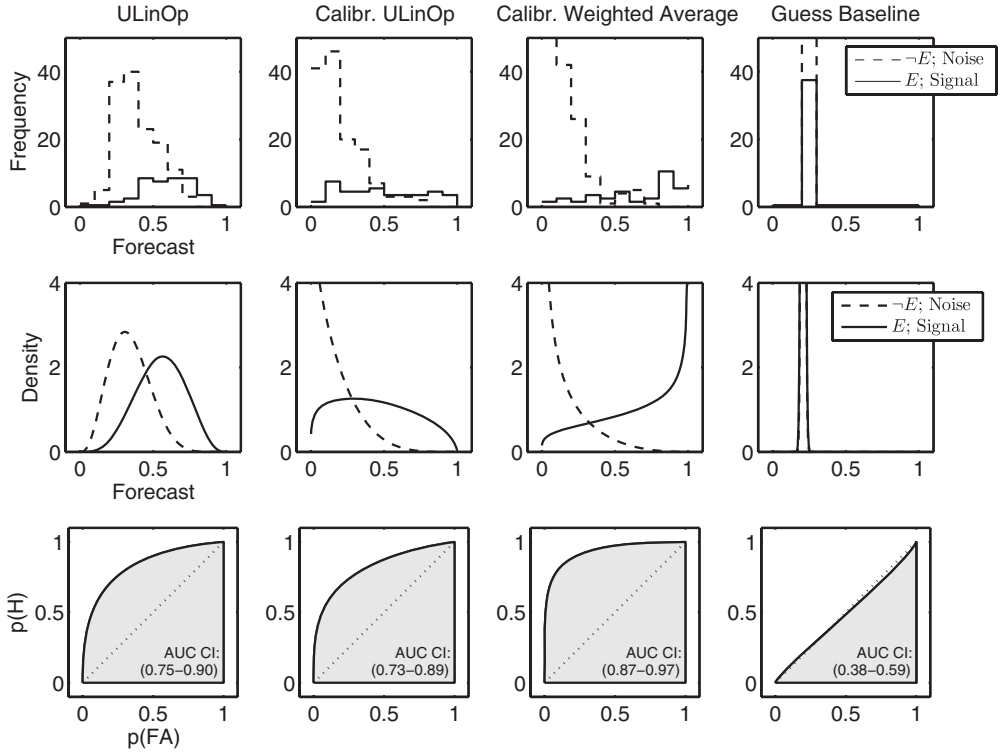

Fig. 6. Results of the basic Beta-SDT model applied to four aggregation models. The top row shows the frequency counts of the aggregated judgments across forecasting problems separated into problems where the event did or did not occur (solid and dashed lines, respectively). The middle row shows the estimated Beta distributions. The bottom row shows the ROC curves with the 95% credible interval for the AUC values.

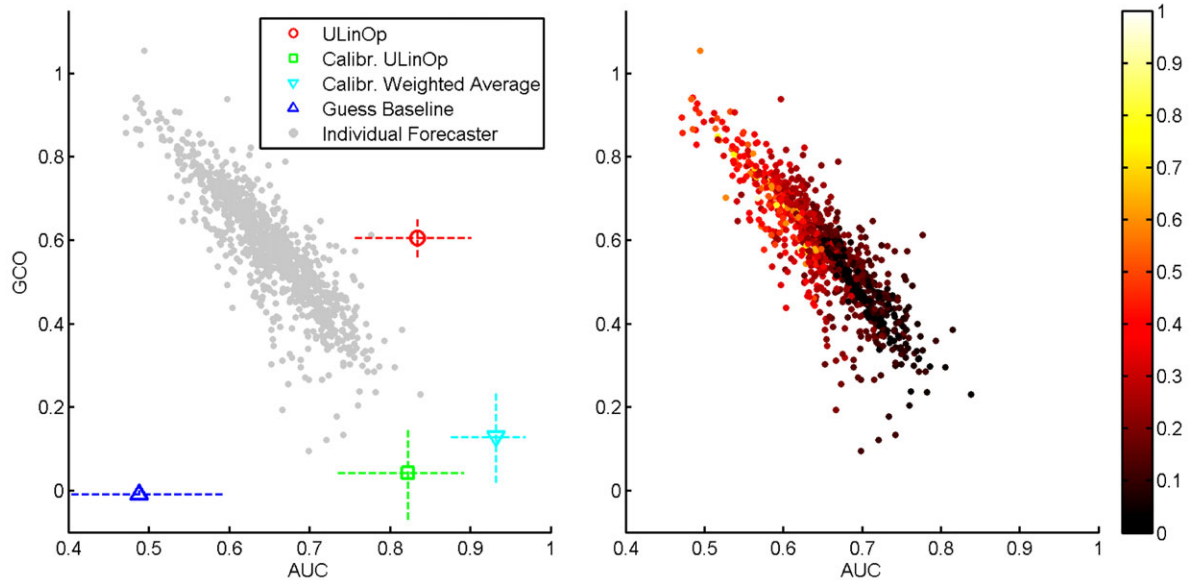

Fig. 7. Estimated global calibration offset (GCO) plotted against estimated AUC values. The left panel shows individual forecasters as well as aggregation models. The right panel indicates the Brier scores for individual forecasters with the colormap as shown on the right side. For individual forecasters, only the mean performance numbers are visualized.

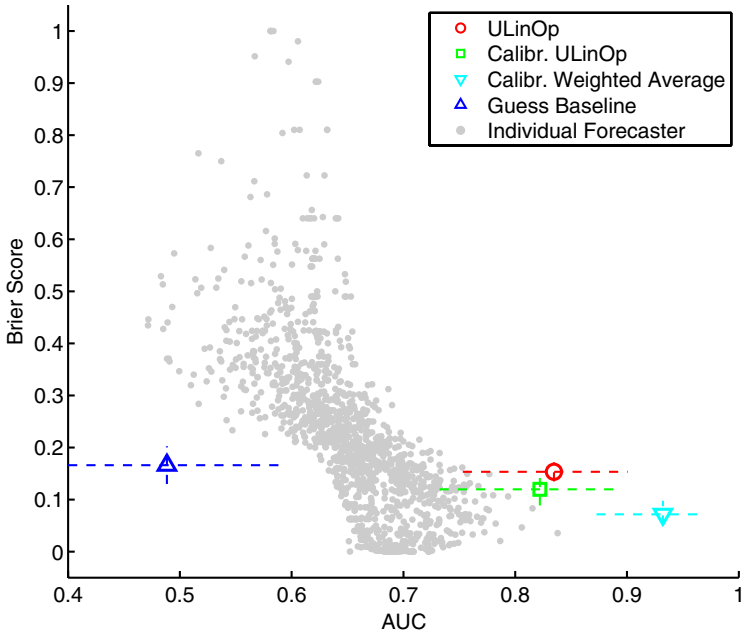

Fig. 8. Brier scores plotted against estimated AUC values for individual forecasters and aggregation models.

This visualization highlights the differences between the two indices. Note the severe nonmonotonicity in the relationship. This nonmonotonicity occurs because, as Yates (1982) has demonstrated, the Brier score reflects three components: diagnosticity, calibration, and problem difficulty (base rate). For any given base rate, procedures or methods that decrease Brier scores could be either improving calibration or diagnosticity. Because calibration can always be improved via monotonic transformation, the more fundamental issue is to assess improvement in diagnosticity. AUC provides this measure uncontaminated by other factors.

Fig. 7 also shows a correlation between the estimated AUC and GCO values. Individual forecasters with higher diagnosticity tend to overestimate event probabilities to a lesser degree. This correlation stems from the fact that the base rate plays a role in both the GCO and AUC measures. In an analysis of the prior predictive distribution[44] of the model, we found that a small negative (positive) correlation can be expected, *a priori*, between AUC and GCO, when the base rate of events falls below 0.5 (above 0.5).

5.4. Evaluating the Effects of Time and Forecasting Domain

Finally, we will evaluate how the performance of individual forecasters changes over time and how the dynamics of this change varies among forecasting domains. Fig. 9 shows the empirical mean forecasts conditioned on event occurrence and nonoccurrence computed over a number of temporal ranges. The figure also shows the posterior predictive means (lines; collapsed across judges) from the hierarchical model that incorporates time effects. Note that there is a reasonable overlap between theory and data. For some forecasting domains, such as science and technology, there are only a few forecasting problems ($N = 16$), which makes the estimation of empirical means difficult. The majority of forecasting problems in the current data set fall in the politics and policy domain ($N = 82$), which leads to a clearer picture of the temporal dynamics. Overall, the results show that the signal and noise distributions separate when time approaches the forecasting horizon.

Fig. 10 shows the mean estimated AUC (across judges) as a function of time and forecasting domain. Note that these trends show the performance of the average individual forecaster. The figure also shows the 95% credible interval as dashed lines. The comparisons among forecasting domains reveal some interesting differences in temporal dynamics between domains. For example, for the politics and policy domain, the AUC quickly increases when fewer than 20 days remain in the forecasting period, suggesting that the type of problems in this domain can be resolved with recent information. In the domain of business

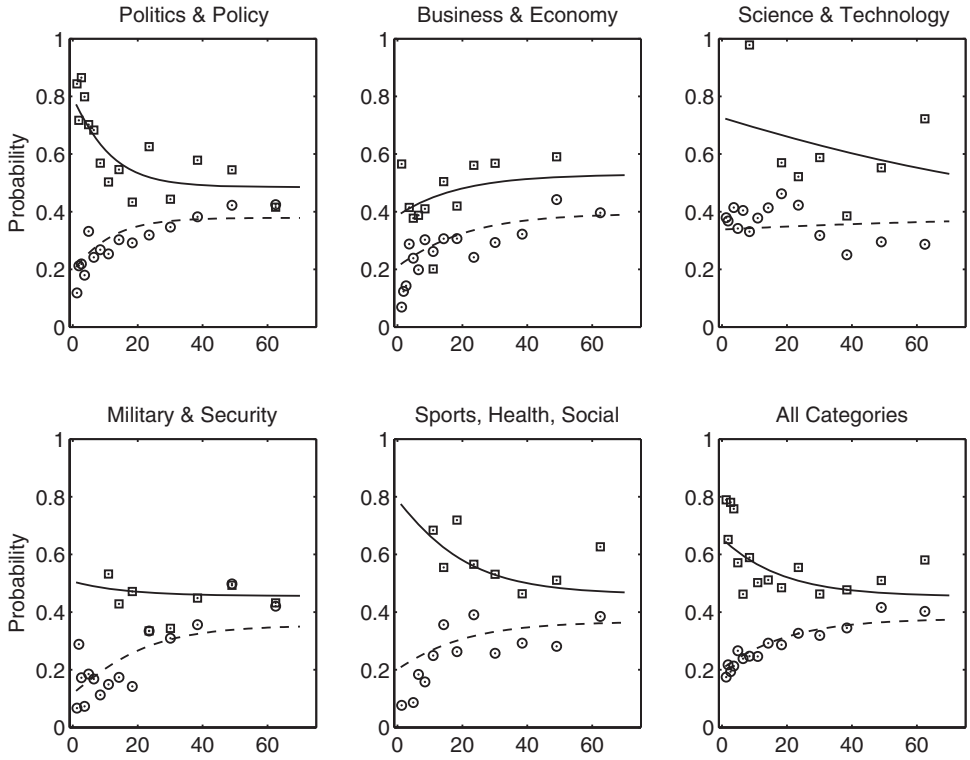

Fig. 9. Empirical and theoretical means of probabilistic forecasts given signal (event occurrence) and noise (event nonoccurrence), respectively, as a function of time (horizontal axis) and forecasting domain (panels). Note that these means are based on the average individual forecaster. Signal and noise distributions are represented by solid and dashed lines (model) and squares and circles (empirical data), respectively. Note that time is expressed as the number of days before the forecasting horizon. Thus, time from problem closure decreases to the left on the horizontal axis.

and economy on the other hand, the AUC changes slowly, indicating that accurate information available to forecasters does not accumulate quickly over time. The most important message in Fig. 10, however, is the general one that AUC can quantify diagnosticity, or forecasting difficulty, as a function of domain and horizon.

6. DISCUSSION

In this article, we introduced a novel Beta-SDT modeling approach to describing and evaluating probabilistic forecasts. We showed that the Beta-SDT model can be estimated on the basis of sparse data using Bayesian hierarchical modeling techniques. The hierarchical model allows us to estimate the underlying belief distributions for individual forecasters as well as for the group of forecasters as a whole. Furthermore, the model provides estimates of diagnosticity (AUC) along with credible intervals for individuals or groups, overall or for defined domains collapsed over time or as a function of time expressed as forecasting horizon.

Two features not emphasized earlier bear mentioning here. The first is that the GCO measure introduced above addresses a global aspect of calibration, which is forecasters' tendency to overestimate (or underestimate, if that turns out to be the case) the likelihood of events. By distinguishing AUC and GCO at the level of individual forecasters, we can compare the performance of a large number of forecasters with aggregates of the individual forecasts. Going beyond what we already know about the benefits of forecast aggregation,[29] Figs. 7 and 8 show that any of the aggregation methods improve diagnosticity (AUC) relative to the individual forecasts, some methods more than others, whereas recalibration is required to reduce the GCO. Aggregation via calibrated weighted averaging provides better AUC than does aggregation via the calibrated ULinOp model, but at the expense of GCO. We argue that this tradeoff is worthwhile as, within this framework, what is important are

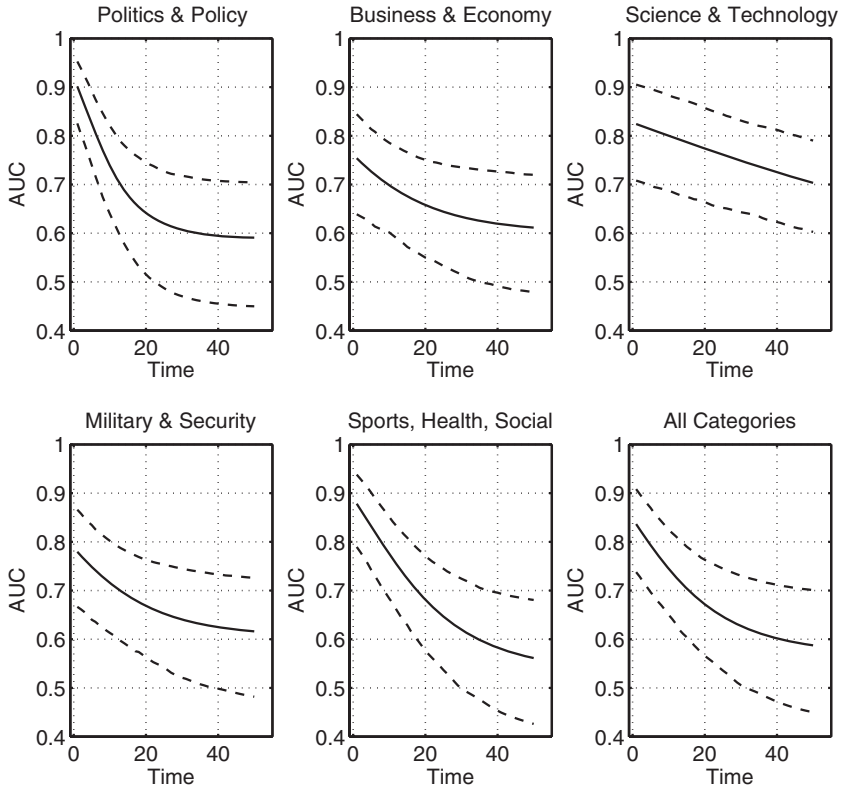

Fig. 10. AUC of the average individual forecaster as a function of time (horizontal axis) and forecasting domain (panels). The dashed lines show the 95% confidence interval across individual forecasters. Note that time is expressed as the number of days before the forecasting horizon.

the AUC and the hit and false alarm probabilities associated with any decision threshold, not the numerical values of the forecasts themselves.

6.1. Other Modeling Extensions

We have already discussed a number of variants of the Beta-SDT approach that allow us to investigate individual differences as well differences in the temporal dynamics and the effect of forecasting domain. One attractive feature of the Beta-SDT approach is that it can be extended in many other ways.

For example, the model can be extended to handle ordinal judgments even when expressed nonnumerically (e.g., *unlikely*, *perhaps*, *likely*) by including response thresholds for the samples from the underlying evidence distributions. In fact, the same latent belief distributions could be used to model the probabilistic forecasts as well as ordinal responses, allowing the mapping between different types of responses.

We have also focused in the current work on forecasting problems with only two possible outcomes. However, it is possible to extend the ROC analysis to multiple classes,[28,45] and these alternative ROC approaches can motivate different SDT models.

Finally, we have assumed that the resolution of the binary forecasting problems leads to an unambiguous assignment of E and $\neg E$. However, in some situations, the definition of E and $\neg E$ are arbitrary. For example, a temperature forecasting problem could be framed in terms of the temperature exceeding a given value or in terms of the temperature not exceeding a given value. In these cases, the SDT model can be simplified to enforce symmetry in the signal and noise distributions, i.e., $\mu_0 = 1 - \mu_1$, and $\xi_0 = \xi_1$.

6.2. Relationship to Brier Score and Related Measures

The relative performance of competing forecasting systems is often evaluated with Brier scores. We contend that the Brier score, or any strictly proper scoring rule, is not sufficient for evaluating forecasting systems, as these rules were developed to motivate honest reporting, not to compare systems. The

Brier score, as illustrated in our results, is sensitive to different components of forecasting performance, including diagnosticity and calibration. One possibility is to decompose the Brier scores into component scores to assess different dimensions of forecasting performance.[37,46,47] Although a detailed discussion of these decompositions is outside the scope of this article, we note a few differences with the Beta-SDT approach. First, it is not clear how to separate out different components of performance in situations where only sparse forecasting data are available. There is also no obvious way to take uncertainty about the measured values of the components into account. In addition, the AUC comes with a guarantee that it is insensitive to strictly monotonic transformations of probabilistic forecasts. In contrast, there are no corresponding results for components of the Brier score. Indeed, although AUC is a principled statistic based on underlying theory, the Brier score decomposition simply reflects a convenient variance partitioning. See Ref. 46 for further discussion of issues associated with interpreting partitioned components. Finally, one of the appealing properties of the Beta-SDT approach is that it can be viewed as a model that describes the generation of probabilistic forecasts, and can be extended in any number of ways to take additional covariates, different response types, and different types of forecasting problems into account. In contrast, the Brier score decomposition is based on a measurement approach that does not lend easily to extensions, primarily because it was not intended to provide an explanation of the process of forecasting.

7. CONCLUSIONS AND FINAL THOUGHTS

The developments in this article highlight two main points. (1) It is possible to distinguish in a principled way two important properties of forecasts in order to assess the effects of any methods for improving them or for quantifying differences in domain forecastability as a function of domain and/or time or overall. AUC is a principled, easy to understand, index of diagnosticity; and GCO, the log-difference between the mean forecast and the base rate, is an easy to understand index of bias. (2) The Beta-SDT model is a powerful tool for estimating AUC along with the precision of the estimate for individuals or groups overall or within domains, collapsed over time or as a function of time.

An additional point not emphasized until now, but that derives directly from SDT, is that if the ROC is sufficiently well specified for any given forecaster, group, or domain, it provides the DM with the probability estimates she needs in order to do expected utility analyses on potential decisions. That is, the probability forecasts, *per se*, are not sufficient for the DM to make a best decision. She also requires good estimates of the hit and false alarm rates, which are obtained from the ROC. Armed with estimates of these two rates as well as of the event probability, the DM can utilize cost or utility estimates of the two kinds of errors, a miss (complement of a hit) or a false alarm, and estimate the expected utility of acting as though the event will or will not occur.

Details on how such a decision policy might be implemented await further research, but we can address one potential criticism immediately. That criticism is that the utilities of outcomes of many decisions, e.g., public policy, national security, or personal health, cannot be numerically estimated and in that sense, expected utility is not the right decision criterion at all. To this point, we agree, but nevertheless it is the case that sensitivity analyses, accomplished by varying the ratios of the error costs, say from small to large, can be enormously helpful to the DM in arriving at a justifiable decision.

ACKNOWLEDGMENTS

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20059. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. government.

REFERENCES

1. Green DM, Swets JA. Signal Detection Theory and Psychophysics. New York: Wiley Press, 1966.
2. Egan JP. Recognition Memory and the Operating Characteristic. Bloomington, IN: Hearing and Communication Laboratory, Indiana University, Tech. Rep. AFCRC-TN-58-51, 1958.
3. Macmillan NA, Creelman CD. Detection Theory: A User's Guide. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

4. McClelland GH. Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. Pp. 83–100 in Fischhoff B, Chauvin C (eds). Intelligence Analysis: Behavioral and Social Scientific Foundations. Washington, DC: National Academies Press, 2011.
5. Warnaar D, Merkle E, Steyvers M, Wallsten T, Stone E, Budescu D, Yates J, Sieck W, Arkes H, Argenta C, Shin Y, Carter J. The aggregative contingent estimation system: Selecting, rewarding, and training experts in a wisdom of crowds approach to forecasting. Pp. 75–76 in Proceedings of the 2012 Association for the Advancement of Artificial Intelligence Spring Symposium Series. AAAI Technical Report SS-12-06, 2012.
6. DeCarlo LT. Signal detection theory with item effects. Journal of Mathematical Psychology, 2011; 55: 229–239.
7. Rouder JN, Lu J. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychonomic Bulletin and Review, 2005; 12: 573–604.
8. Rouder JN, Lu J, Sun D, Speckman PL, Morey RD, Naveh-Benjamin M. Signal detection models with random participant and item effects. Psychometrika, 2007; 72: 621–642.
9. Lee MD. Three case studies in the Bayesian analysis of cognitive models. Psychonomic Bulletin and Review, 2008; 15: 1–15.
10. Tanner WP, Swets JA. A decision-making theory of visual detection. Psychological Review, 1954; 61: 401–409.
11. Swets JA. The relative operating characteristic in psychology. Science, 1973; 182: 990–1000.
12. Marzban C. Performance measures and uncertainty. Pp. 49–75 in Haupt SE, Pasini A, Marzban C (eds). Artificial Intelligence Methods in the Environmental Sciences. Berlin Heidelberg: Springer-Verlag, 2009.
13. Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. Pp. 445–453 in Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997.
14. Spackman KA. Signal detection theory: Valuable tools for evaluating inductive learning. Pp. 160–163 in Proceedings of the Sixth International Workshop on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1989.
15. Lusted LB. Signal detectability and medical decision-making. Science, 1971; 171: 1217–1219.
16. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, UK: Oxford University Press, 2003.
17. Zhou XH, McClish DK, Obuchowski NA. Statistical Methods in Diagnositic Medicine. New York: John Wiley and Sons, 2002.
18. Harvey LO, Hammond KR, Lusk CM, Mross EF. The application of signal detection theory to weather forecasting behavior. Monthly Weather Review, 1992; 120: 863–883.
19. Levi K. A signal detection framework for the evaluation of probabilistic forecasts. Organizational Behavior and Human Decision Processes, 1985; 36: 143–166.
20. Mason SJ. A model for assessment of weather forecasts. Australian Metereological Magazine, 1979; 30: 291–303.
21. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and the relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Metereological Society, 2002; 128: 2145–2166.
22. Swets JA. Signal Detection Theory and ROC Analyses in Psychology and Diagnostics: Collected Papers. Mahwah, NJ: Erlbaum, 1996.
23. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006; 27: 861–874.
24. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Briefings in Bioinformatics, 2012; 13: 83–97.
25. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest, 2000; 1: 1–26.
26. Flach PA. ROC analysis. Pp. 869–875 in Sammut C, Webb GI (eds). Encyclopedia of Machine Learning, No. 19. Springer, 2010.
27. Wickens TD. Elementary Signal Detection Theory. New York: Oxford University Press, 2001.
28. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 2001; 45: 171–186.
29. Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS. Forecast aggregation via recalibration. Machine Learning, in press.
30. Kharin VV, Zwiers FW. On the ROC score of probability forecasts. Journal of Climate, 2003; 16: 4145–4150.
31. Marzban C. The ROC curve and the area under it as performance measures. Weather and Forecasting, 2004; 19: 1106–1114.
32. Kosmidis I, Firth D. A generic algorithm for reducing bias in parametric estimation. Electronic Journal of Statistics, 2010; 4: 1097–1112.
33. Smithson M, Merkle EC, Verkuilen J. Beta regression finite mixture models of polarization and priming. Journal of Educational and Behavioral Statistics, 2011; 36: 804–831.
34. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik K, Leisch F, Zeileis A (eds). Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003.
35. Armstrong JS. Principles of Forecasting. Norwell, MA: Kluwer Academic, 2001.
36. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 1950; 78: 1–3.
37. Murphy AH. A new vector partition of the probability score. Journal of Applied Meteorology, 1973; 12: 595–600.
38. Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G. The effects of averaging subjective probability estimates between and within judges. Journal of Experimental Psychology: Applied, 2000; 6: 130–147.
39. Clemen RT. Calibration and the aggregation of probabilities. Management Science, 1986; 32: 312–314.
40. Clemen RT, Winkler RL. Combining economic forecasts. Journal of Business and Economic Statistics, 1986; 4: 39–46.
41. Cooke RM. Experts in uncertainty: Opinion and subjective probability in science. New York: Oxford University Press, 1991.
42. Wallsten TS, Budescu DV, Erev I. Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 1997; 10: 243–268.
43. Gonzalez R, Wu G. On the shape of the probability weighting function. Cognitive Psychology, 1999; 38: 129–166.
44. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. New York: Chapman and Hall, 2004.
45. Wandashin MS, Mullen SJ. Multiclass ROC analysis. Weather and Forecasting, 2009; 24: 530–547.
46. Yaniv I, Yates JF, Smith JEK. Measures of discrimination skill in probabilistic judgment. Psychological Bulletin, 1991; 110: 611–617.
47. Yates JF. External correspondence: Decompositions of the mean probability score. Organizational Behavior and Human Performance, 1982; 30: 132–156.