ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluating probabilistic classifiers: The triptych[☆]

Timo Dimitriadis[a,b], Tilmann Gneiting[b,c,*], Alexander I. Jordan[b], Peter Vogel[d]

[a] Alfred-Weber-Institute for Economics, Heidelberg University, Germany
[b] Computational Statistics (CST) group, Heidelberg Institute for Theoretical Studies, Germany
[c] Institute for Stochastics, Karlsruhe Institute of Technology, Germany
[d] CSL Behring Innovation GmbH, Germany

ARTICLE INFO

Keywords:
Calibration error
Economic utility
Logarithmic score
MCB–DSC plot
Misclassification loss
Proper scoring rule
Score decomposition
Sharpness principle

ABSTRACT

Probability forecasts for binary outcomes, often referred to as probabilistic classifiers or confidence scores, are ubiquitous in science and society, and methods for evaluating and comparing them are in great demand. We propose and study a triptych of diagnostic graphics focusing on distinct and complementary aspects of forecast performance: Reliability curves address calibration, receiver operating characteristic (ROC) curves diagnose discrimination ability, and Murphy curves visualize overall predictive performance and value. A Murphy curve shows a forecast's mean elementary scores, including the widely used misclassification rate, and the area under a Murphy curve equals the mean Brier score. For a calibrated forecast, the reliability curve lies on the diagonal, and for competing calibrated forecasts, the ROC and Murphy curves share the same number of crossing points. We invoke the recently developed CORP (Consistent, Optimally binned, Reproducible, and Pool-Adjacent-Violators (PAV) algorithm-based) approach to craft reliability curves and decompose a mean score into miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components. Plots of the DSC measure of discrimination ability versus the calibration metric MCB visualize classifier performance across multiple competitors. The proposed tools are illustrated in empirical examples from astrophysics, economics, and social science.

1. Introduction

Across science and society, probability forecasts for the occurrence of a binary outcome, also referred to as probabilistic classifiers or confidence scores, are widely used. Prominent examples include a patient's recovery or survival, weather events, solar flares, the designation of email as spam, credit approval, and recidivism of criminal defendants, to name but a few applications. Our ability to develop and improve probability forecasts depends on the availability of diagnostic tools for assessing and comparing predictive power.

While some applications call for a single numerical performance measure, with forecast contests and leader boards being prime examples, the condensation of forecast quality into a single number prevents detailed analyses. As Janssens (2020) notes,

> "Some prediction researchers prefer one metric or graph that captures the overall performance of prediction models. Others prefer one for each different aspect of performance, such as calibration, discrimination, predictive value (risks), and utility".

[☆] The results presented in this paper were reproduced by the Editor-in-Chief on 7 October 2023. The complete reproducibility package, including data and code, is available at https://github.com/TimoDimi/replication_triptych.

* Corresponding author at: Computational Statistics (CST) group, Heidelberg Institute for Theoretical Studies, Germany.

E-mail addresses: timo.dimitriadis@awi.uni-heidelberg.de (T. Dimitriadis), tilmann.gneiting@h-its.org (T. Gneiting), alexander.jordan@h-its.org (A.I. Jordan), petervogel1991@googlemail.com (P. Vogel).

https://doi.org/10.1016/j.ijforecast.2023.09.007

Table 1
Probability forecasts for class C1.0+ solar flares at a prediction horizon of a day ahead from a joint test set within calendar years 2016 and 2017 (Leka & Park, 2019; Leka et al., 2019): Acronym, source, mean Brier score, mean logarithmic (Log) score, and misclassification rate (MR). Details of the data example are discussed in Section 6.1.

Probability forecast		Mean score		
Acronym	Source	Brier	Log	MR
NOAA	National Oceanic and Atmospheric Administration	0.144	0.449	0.205
SIDC	Royal Observatorium Belgium	0.172	0.515	0.263
ASSA	Korean Space Weather Agency	0.184	∞	0.273
MCSTAT	Trinity College Dublin	0.193	0.587	0.275

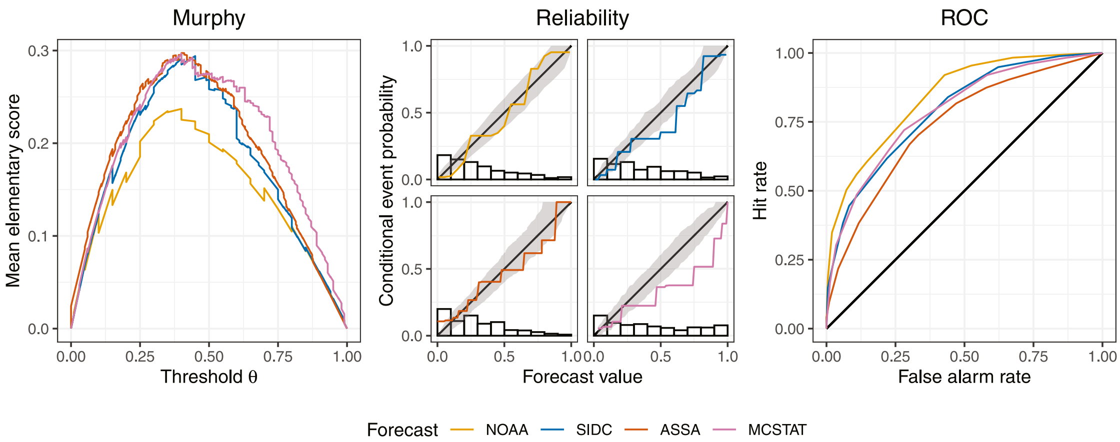

Fig. 1. Triptych of diagnostic graphics for evaluating and comparing the probability forecasts of class C1.0+ solar flares from Table 1: Murphy curves (lower is better), reliability curves (close to diagonal is preferred) with 90% consistency bands, and ROC curves (higher is better).

Not surprisingly, numerous types of diagnostic graphics for evaluating probability forecasts exist (Filho et al., 2023; Murphy & Winkler, 1992; Prati et al., 2011), and practitioners may wonder which ones are preferred.

In this article, we propose using a triptych of diagnostic graphics and provide theoretical support for our choices. The triptych consists of reliability curves in the recently proposed CORP (Consistent, Optimally binned, Reproducible, and Pool-Adjacent-Violators (PAV) algorithm-based) form to assess calibration (Dimitriadis et al., 2021), receiver operating characteristic (ROC) curves to judge discrimination ability (Fawcett, 2006; Swets, 1973), and Murphy curves for the assessment of overall predictive performance and utility (Ehm et al., 2016). Fig. 1 illustrates the triptych for probabilistic classifiers from an astrophysical forecast challenge (Leka & Park, 2019; Leka et al., 2019) as introduced in Table 1 and discussed in detail in Section 6.1.

From the left, Murphy curves assess overall predictive performance in terms of proper scoring rules (Ehm et al., 2016). To provide background, a scoring rule assigns a score $\mathsf{S}(x, y)$ to each pair of a probability forecast $x \in [0, 1]$ and a binary outcome $y \in \{0, 1\}$, where 1 stands for an event and 0 for a non-event. A scoring rule is proper if a forecaster minimizes the expected score by issuing a probability forecast that corresponds to her true belief, with the Brier score $\mathsf{S}(x, y) = (x - y)^2$ and the logarithmic (Log) score $\mathsf{S}(x, y) = -y \log x - (1 - y) \log(1 - x)$ being prominent examples (Gneiting & Raftery, 2007). Scores then are averaged over a test set, and the forecast with the smallest mean score is considered best. The widely used misclassification rate (MR) arises as a special case, namely, by assigning a score of 1 if the probability forecast is less than $\frac{1}{2}$ and the event realizes, or the forecast is greater than $\frac{1}{2}$ and the event does not realize, and assigning a score of 0 otherwise. Distinct proper scoring rules may yield distinct forecast rankings, so practitioners may wonder which one to use, and guidance is essential. In the case of a binary outcome, proper scoring rules can be represented as mixtures over so-called elementary scoring rules. Consequently, we can reconstruct a forecast's score under any given proper rule if we know its scores under the elementary rules. Fortunately, the family of the elementary scoring rules is linearly parameterized by a threshold or cost-loss parameter θ. A Murphy curve depicts the mean elementary score as a function of the threshold θ, with lower scores being preferable. The height of the Murphy curve at $\theta = \frac{1}{2}$ equals the misclassification rate, and the area under the Murphy curve equals the mean Brier score. If a forecast has a Murphy curve below a competitor's, it is superior in terms of any proper scoring rule and has superior economic utility to any decision maker. For example, we see from the Murphy curves in Fig. 1 that the NOAA forecast dominates the ASSA forecast, regardless of intended use.

A probability forecast is calibrated if, conditional on any forecast value p, the event realizes in $100 \cdot p$ percent of the instances considered. Reliability curves visualize calibration by plotting an estimate of the conditional event probability (CEP) as a function of the forecast value. While reliability curves close to the diagonal are compatible with calibration assumptions, notable departures from the diagonal suggest miscalibration and can be interpreted diagnostically. We adopt the recently proposed CORP approach of Dimitriadis et al. (2021) for the estimation of CEPs by nonparametric isotonic regression, as illustrated in Fig. 1, where the SIDC and MCSTAT forecasts exhibit overprediction, with estimated CEPs below the diagonal.

Receiver operating characteristic (ROC) curves visualize the discrimination ability of the forecasts — that is, they judge to what extent the forecast values distinguish situations with lower or higher true event probabilities. Specifically, as one issues hard classifiers based on successively higher forecast thresholds, a ROC curve plots the hit rate (HR) on the ordinate against the false alarm rate (FAR) on the abscissa. As the ROC curve is invariant under strictly increasing transformations of the forecast values, it diagnoses discrimination ability only, while ignoring calibration issues. Hit rates close to 1 and false alarm rates close to 0 are desirable, so ROC curves at the upper left indicate superior discrimination ability. In the ROC curves in Fig. 1, the NOAA forecast shows the highest and the ASSA forecast the lowest discrimination ability. In conclusion, the NOAA forecast performs best regarding scoring rules and economic utility, featuring excellent calibration and superior discrimination ability.

The choice of the triptych graphics reflects theoretically supported desirable properties. Reliability curves exclusively diagnose calibration, ROC curves assess discrimination ability only, and Murphy curves quantify overall predictive performance. Moreover, the novel Fact B and Theorem 2 below demonstrate that, under perfect calibration, Murphy curves and ROC curves yield congruent insights, as they share the same number of crossing points.

Following the pioneering work of Murphy (1973), researchers have sought decompositions of mean scores into intuitively appealing components that reflect calibration and discrimination, respectively. We utilize the CORP decomposition of Dimitriadis et al. (2021), which decomposes a mean score

$$\bar{\mathsf{S}} = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC}$$

into readily interpretable components that represent miscalibration (MCB), discrimination (DSC), and uncertainty (UNC), respectively. In contrast to earlier approaches, CORP reliability curves and score components do not depend on user choices or tuning parameters, and they show appealing finite and large sample optimality properties. The mean score $\bar{\mathsf{S}}$ equals a weighted area under the Murphy curve and serves as a summary measure of predictive performance. The MCB component quantifies deviations of the CORP reliability curve from the diagonal and can be used as a calibration metric. The DSC component is an appealing alternative to the widely used Area Under the ROC Curve (AUC) measure of discrimination ability.

If many competing forecasting methods are to be compared, the triptych graphics yield crowded displays. With such settings in mind, we propose a simple alternative, namely, MCB–DSC plots, that show, for each competitor involved, the DSC measure plotted against the MCB component, augmented by parallel contour lines that indicate an equal mean score. Due to their simplicity and the joint assessment of overall predictive ability, calibration, and discrimination, MCB–DSC plots visualize strengths and weaknesses of forecasting methods and facilitate the identification of methods of interest that can be analyzed in more detail via the triptych graphics. In Fig. 2, we show Brier score MCB–DSC plots for probability forecasts from solar flare and social science forecast contests.

While there is a rich literature on the evaluation of probabilistic classifiers and associated graphical displays, as reviewed by Murphy and Winkler (1992), Prati et al. (2011), Richardson (2012), Alba et al. (2017), Filho et al. (2023), and Xenopoulos et al. (2023), and variants of the joint triptych display feature in the extant literature, as in Figure 1 of Flach (2017) and graphics in Taillardat and Mestre (2020), the original contributions of our work include the presentation of the triptych as an argumentatively complete set of displays, connections to the CORP approach of Dimitriadis et al. (2021), the development of MCB–DSC plots, and novel theoretical results proved in the Appendix.

The remainder of the article is organized as follows. Sections 2, 3, and 4 look at the individual triptych displays by discussing proper scoring rules and Murphy curves, CORP reliability curves and score decompositions, and ROC curves, respectively. Section 5 argues for the simultaneous use of the triptych displays, studies the connections between the individual displays, and discusses MCB–DSC plots. In particular, we show that for two competing forecasts that are calibrated, Murphy and ROC curves yield congruent insights, as they share the same number of crossing points. In Section 6, we apply the proposed methods in case studies from astrophysics, economics, and social science. The paper closes in Section 7. A software implementation of the proposed tools and material for replicating the results in the article (Dimitriadis & Jordan, 2023a, 2023b) is available for R (R Core Team, 2022).

2. Scoring rules assess overall predictive performance

Comparative assessments of overall forecast quality rely on proper scoring rules that encourage honest and careful forecasting (Brier, 1950; Gneiting & Raftery, 2007). A scoring rule is a function $\mathsf{S}(x, y)$ that assigns a numerical score or penalty based on the probability forecast $x \in [0, 1]$ and the binary outcome $y \in \{0, 1\}$, where 1 stands for an event and 0 for a non-event. Infinite penalties are permitted only if an outcome was declared to have probability zero and, thus, be impossible. Throughout the paper, we assume that scoring rules are negatively oriented, so smaller scores are preferable.

2.1. Proper and strictly proper scoring rules

A scoring rule is proper if, given a Bernoulli random variable Y with success probability p,

$$\mathbb{E}\,[\mathsf{S}(p, Y)] \le \mathbb{E}\,[\mathsf{S}(x, Y)] \tag{2.1}$$

for all forecast values x. It is strictly proper if, additionally, equality in (2.1) implies that $x = p$ so that the true success

Fig. 2. Brier score MCB–DSC plots for competitors in forecast contests for (a) class C1.0+ solar flares (Leka et al., 2019), and (b) job training in the Fragile Families Challenge (Salganik et al., 2020a). The colours in panel (a) align with Fig. 1; in panel (b), benchmark forecasts are represented in green. The green square at the origin represents the ex post best constant forecast, that is, the unconditional event frequency, and the thick green line separates forecasts that are better (above the line) and worse (below the line) than this baseline. Details of the data examples from astrophysics and social science are discussed in Sections 6.1 and 6.3, respectively. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

probability is the unique minimizer of the expected score. The key benefit of propriety is the implicit enforcement of honest and careful forecasts: If a forecaster believes that an event has success probability p, then p is her best forecast in terms of the expected score or penalty. In practice, for a given record

$$(x_1, y_1), \ldots, (x_n, y_n) \tag{2.2}$$

of probability forecasts $x_1, \ldots, x_n$ and associated binary outcomes $y_1, \ldots, y_n$, the mean score

$$\bar{\mathsf{S}} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{S}(x_i, y_i) \tag{2.3}$$

is used to rank competing forecasts. The expression on the right-hand side of (2.3) corresponds to the expectation $\mathbb{E}\,[\mathsf{S}(X, Y)]$ when the tuple (X, Y) of random quantities follows the joint empirical distribution of the record (2.2).

The most popular examples of strictly proper scoring rules are the Brier score (BS) and the Logarithmic score (LogS), defined by

$$\mathsf{BS}(x, y) = (x - y)^2 \tag{2.4}$$

and

$$\mathsf{LogS}(x, y) = -y \log(x) - (1 - y) \log(1 - x) \tag{2.5}$$

for $x \in [0, 1]$ and $y \in \{0, 1\}$. The zero–one loss or score

$$\mathsf{S}_{\frac{1}{2}}(x, y) = \mathbb{1}\left(x > \frac{1}{2}, y = 0\right) + \mathbb{1}\left(x < \frac{1}{2}, y = 1\right) + \frac{1}{2}\,\mathbb{1}\left(x = \frac{1}{2}\right) \tag{2.6}$$

is a prominent example of a scoring rule that is proper but not strictly proper. When averaged in the form of (2.3), it yields the widely reported misclassification rate. The zero–one loss arises as the special case $\theta = \frac{1}{2}$ of the general elementary scoring rule

$$\mathsf{S}_\theta(x, y) = 2\theta\, \mathbb{1}(x > \theta, y = 0) + 2(1 - \theta)\, \mathbb{1}(x < \theta, y = 1) + 2\theta(1 - \theta)\, \mathbb{1}(x = \theta) \tag{2.7}$$

with decision threshold or cost–loss parameter $\theta \in (0, 1)$, which is proper, but not strictly proper, as it only takes into account whether a predicted probability is smaller or larger than θ, so it cannot distinguish between forecasts that are on the same side of θ. From an economic perspective, S_θ specifies the loss of a rational decision maker when the ratio of the monetary cost of a false alarm versus the cost of a missed event equals $\theta/(1 - \theta)$; see Ehm et al. (2016) and references therein.[1] In turn, S_θ can be identified with the special case $t = c = \theta$ of the general, cost-weighted misclassification loss at decision threshold t and cost proportion c, as studied in the machine learning literature (Hand, 2009; Hernández-Orallo et al., 2011, 2012, 2013).

2.2. Representations of proper scoring rules

The special role of the elementary scoring functions S_θ from (2.7) is highlighted in a mixture representation studied by Schervish (1989). Subject to technical conditions that are immaterial in practice, every proper scoring rule

[1] The economic interpretation applies to the left-continuous version of S_θ in eq. (14) of Ehm et al. (2016). Here we use the symmetric version in (2.7), which assigns a fixed penalty of $2\theta(1-\theta)$ when $x = \theta$, independently of the binary outcome $y \in \{0, 1\}$. Both versions are proper, but not strictly proper.

admits a representation of the form

$$\mathsf{S}(x,y) = \int_0^1 \mathsf{S}_\theta(x,y)\,\mathrm{d}H(\theta) \tag{2.8}$$

for forecast values $x \in [0,1]$ and outcomes $y \in \{0,1\}$, where H is a measure that assigns non-negative weight to cost–loss parameters $\theta \in (0,1)$. The corresponding score is strictly proper if the assigned weight is positive almost everywhere. The elementary score S_η arises when H is a point measure that assigns mass one to $\eta \in (0,1)$ and no mass elsewhere, the Brier score emerges when the mixing measure H is uniform, and the logarithmic score arises when H has density proportional to $(\theta(1-\theta))^{-1}$. Hence, the logarithmic score assigns infinite mass to the integrand in (2.8) at the very boundaries of the unit interval, discouraging predictions with forecast probabilities at or near 0 or 1. It may render a single (and, hence, a mean) score infinite, as for the ASSA forecast from Table 1, which is a value judgment that some authors consider "unacceptable" (Selten, 1998, p. 51).

The mixture representation (2.8) is also a powerful tool for constructing proper scoring rules. For instance, Buja et al. (2005) introduce the flexible Beta family that arises when the mixing measure H has a density proportional to a Beta density. The members of the Beta family include the Brier score, the logarithmic score, and the H-measure of Hand (2009).

An alternative, essentially equivalent characterization is due to Savage (1971), who showed that subject to technical conditions, any proper scoring rule allows a representation of the form

$$\mathsf{S}(x,y) = \phi(y) - \phi(x) - \phi'(x)(y-x), \tag{2.9}$$

where the function ϕ is convex with subgradient ϕ'. A subgradient is a generalized version of the classical derivative; whenever the latter exists, the subgradient equals the derivative. In relation to the mixture representation from (2.8) it holds that $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta) = \phi''(\theta)\,\mathrm{d}\theta$, with slight technical adaptations when ϕ is convex, but not strictly convex (Gneiting & Raftery, 2007). Under the Savage representation (2.9) the Brier score arises when $\phi(t) = t^2$, the logarithmic score emerges when $\phi(t) = t\log(t) + (1-t)\log(1-t)$, and the elementary scoring rule S_θ arises under the convex, but not strictly convex, function $\phi(t) = 2\max(\theta t, (1-\theta)(1-t))$.

In practice, it is not uncommon that different proper scoring rules yield distinct forecast rankings. For example, in Table 1, the Brier score and the logarithmic score disagree in ranking the ASSA and MCSTAT forecasts. As there is no apparent reason for a specific strictly proper scoring rule to be preferred over any other, a natural question is which one to choose (Merkle & Steyvers, 2013).

2.3. Murphy curves

The mixture representation (2.8) allows for a compelling resolution of the challenge for guidance in choosing proper scoring rules. As the representation shows, any strictly proper scoring rule arises as a mixture over the family of the elementary scoring functions S_θ from (2.7). Thus, it suffices to consider the family of mean elementary scores,

$$\bar{\mathsf{S}}_\theta = \frac{1}{n}\sum_{i=1}^{n} \mathsf{S}_\theta(x_i,y_i), \tag{2.10}$$

where $\theta \in (0,1)$. Ehm et al. (2016) proposed plots of the Murphy curve, that is, the graph of $\bar{\mathsf{S}}_\theta$ as a function of the decision threshold or cost–loss parameter $\theta \in (0,1)$, which allows users to assess forecast performance with respect to all scoring rules simultaneously. In particular, the height of a Murphy curve at $\theta = \frac{1}{2}$ equals the misclassification rate, but note that sole focus on the misclassification rate as a general measure of predictive performance is problematic because any single S_θ represents a particular economic scenario as mentioned in Section 2.1 and fails to be strictly proper. For a general measure, looking at the area under a Murphy curve, which equals the mean Brier score, is better.

Hernández-Orallo et al. (2011) had proposed the same tool under the name of Brier curve, and related displays had been studied by Murphy and Winkler (1992) and Drummond and Holte (2006), among other authors. The term Murphy diagram is also used, particularly when multiple Murphy curves for competing forecasts are shown within a single display. If a forecast exhibits a lower mean score than another under all S_θ, then it dominates the competitor in terms of any proper scoring rule S, as lower values under S_θ carry over to S through integration in the mixture representation (2.8).

Fig. 3 looks at the Murphy curves for the class C1.0+ solar flare forecasts. The panel at left compares the leading contenders from Table 1, the NOAA, and the SIDC forecasts. We note that NOAA has smaller $\bar{\mathsf{S}}_\theta$ from (2.10) nearly throughout, so integration over $\bar{\mathsf{S}}_\theta$ with respect to typically used mixing measures yields smaller mean scores. The panel at right compares the MCSTAT and ASSA forecasts, and we see that, while for decision thresholds θ up to about 0.30, the former has lower $\bar{\mathsf{S}}_\theta$, the situation is reversed for θ greater than 0.50. Hence, it depends on the mixing measure H in (2.8) whether a scoring rule prefers the MCSTAT or the ASSA forecast.

To summarize, Murphy curves assess overall predictive performance, with the evaluation being complete in terms of proper scoring rules and economic utility. Still, a more detailed assessment of the merits and deficiencies of competing forecasts is often desirable. For example, a forecast might be deficient by systematically over or underpredicting or by an inability to distinguish between instances of higher and lower true CEPs. In a nutshell, these two deficiencies correspond to a lack of calibration and discrimination, respectively. While Murphy curves rank forecasts with such deficiencies lower, they cannot diagnose the form and extent of these issues. Reliability and ROC curves, to which we turn in the sequel, serve this purpose.

3. Reliability curves assess calibration

A crucial, desirable property of a probabilistic classifier is that, when looking back at a collection of forecasts

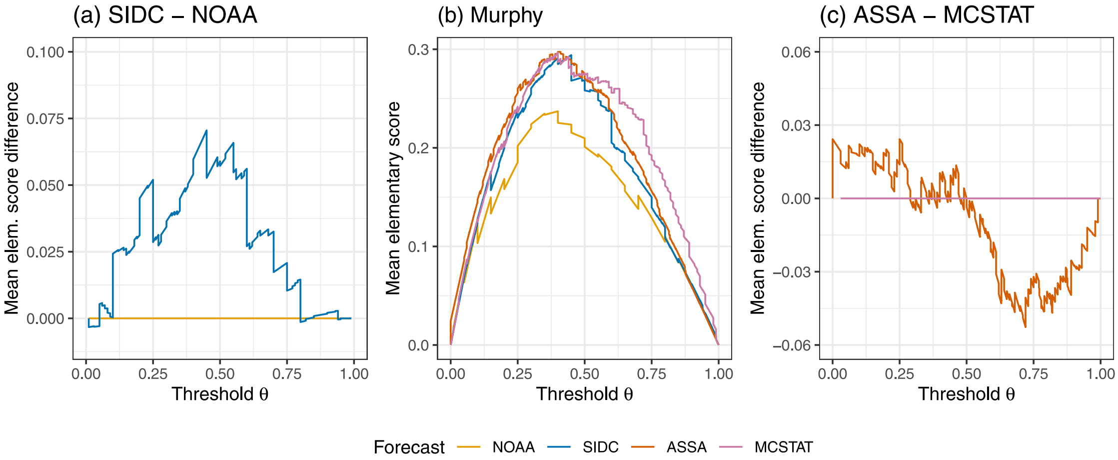

Fig. 3. Murphy curves for the probability forecasts of class C1.0+ solar flares from Table 1. Panel (b) shows the mean elementary score from (2.10); panels (a) and (c) show the difference between the mean elementary score for SIDC and NOAA, and for ASSA and MCSTAT, respectively.

and associated binary outcomes, whenever the forecast value x was issued, the outcome ought to occur in about $100 \cdot x$ percent of the respective instances. To formalize this property, it is useful to think of the forecast and the outcome as random variables X and Y, respectively, with joint distribution $\mathbb{Q}$. Then the probability forecast X is calibrated (Bröcker, 2012b; Lindley, 1982; Schervish, 1989) if the conditional event probability,

$$\mathrm{CEP}(x) = \mathbb{Q}(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x], \tag{3.1}$$

agrees with the forecast value x for all relevant $x \in [0, 1]$. By Theorem 2.11 of Gneiting and Ranjan (2013), the condition in (3.1) serves as the unified notion of calibration for binary outcomes.

3.1. Reliability curves

Calibration is typically assessed graphically via reliability curves (Bröcker & Smith, 2007; Murphy & Winkler, 1977, 1992) that plot an estimated version of the conditional event probability CEP(x) against the forecast value x, with deviations from the diagonal suggesting lack of calibration. Classical approaches to estimating CEP(x) rely on binning and counting and have been hampered by ad hoc implementation decisions and instability under unavoidable choices regarding binning (Arrieta-Ibarra et al., 2022; Roelofs et al., 2022). To resolve these issues, Dimitriadis et al. (2021) introduced the CORP (Consistent, Optimally binned, Reproducible, and Pool-Adjacent-Violators (PAV) algorithm-based) reliability curve that plots an estimate of CEP(x) obtained through nonparametric isotonic regression, subject to the regularizing constraint of isotonicity in x, as implemented via the Pool-Adjacent-Violators (PAV) algorithm (Ayer et al., 1955; De Leeuw et al., 2009).

For a given record of the form (2.2), suppose without loss of generality that $x_1 \leq \cdots \leq x_n$. As specified in Algorithm 1, where we specialize descriptions in Gneiting and Resin (2023), the PAV algorithm generates a sequence

Algorithm 1: PAV algorithm based on data of the form (2.2)

Input: $(x_1, y_1), \ldots, (x_n, y_n) \in [0, 1] \times \{0, 1\}$ where $x_1 \leq \cdots \leq x_n$
Output: calibrated values $\hat{x}_1, \ldots, \hat{x}_n$
partition into groups $G_{1:1}, \ldots, G_{n:n}$ and let $\hat{x}_i = y_i$ for $i = 1, \ldots, n$
while *there are groups* $G_{k:i}$ *and* $G_{(i+1):l}$ *such that* $\hat{x}_1 \leq \cdots \leq \hat{x}_i$ *and* $\hat{x}_i > \hat{x}_{i+1}$ **do**
  merge $G_{k:i}$ and $G_{(i+1):l}$ into $G_{k:l}$ and let $\hat{x}_i = \frac{1}{l-k+1} \sum_{j=k}^{l} y_j$ for $i = k, \ldots, l$
end

$$\hat{x}_1 \leq \cdots \leq \hat{x}_n \tag{3.2}$$

of recalibrated values, such that the empirical measure of $(\hat{x}_1, y_1), \ldots, (\hat{x}_n, y_n)$ satisfies (3.1). In a nutshell, the algorithm partitions the index set $\{1, \ldots, n\}$ into groups $G_{k:l} = \{k, \ldots, l\}$ of consecutive integers. Successive groups get pooled iteratively if the conditional event probability (CEP) in the preceding group exceeds the CEP in the subsequent group.

While isotonic regression and the PAV algorithm have been well known as tools for re-calibration (Zadrozny & Elkan, 2002), their usage in constructing reliability curves is a recent development. The CORP reliability curve shows the piecewise linear curve that connects the points $(x_1, \hat{x}_1), \ldots, (x_n, \hat{x}_n)$. Horizontal segments in this piecewise linear graph correspond to indices i with $x_i < x_{i+1}$ and $\hat{x}_i = \hat{x}_{i+1}$; diagonal segments originate from indices i where $x_i < x_{i+1}$ and $\hat{x}_i < \hat{x}_{i+1}$. If the original forecast is calibrated, then $x_1 = \hat{x}_1, \ldots, x_n = \hat{x}_n$, and the reliability curve lies on the diagonal. Otherwise, systematic deviations from the diagonal suggest a lack of calibration.

In contrast to classical approaches that estimate a reliability curve by assigning forecast values to bins and counting events per bin, which mandates user choices, the

CORP approach does not require any tuning parameters. It benefits from the regularizing constraint of isotonicity and has appealing finite sample optimality and asymptotic consistency properties (Dimitriadis et al., 2021). If desired, CORP reliability curves allow for an interpretation in terms of binning and counting by identifying any horizontal segment with a bin and interpreting the CEP as the corresponding empirical event frequency. Generally, we supplement the reliability curve with a histogram depicting the forecast values' unconditional distribution.

Returning to the solar flare forecasts from Table 1, the CORP reliability curves in Fig. 1 are supplemented by consistency bands (Bröcker & Smith, 2007), which we generate as described by Dimitriadis et al. (2021, Section S3). Segments of a reliability curve that lie considerably outside the consistency band, which represents 90 percent of the reliability curves that arise under the assumption of a calibrated forecast, suggest that the lack of calibration should not be attributed to estimation noise alone. The NOAA and ASSA forecasts are well-calibrated, with reliability curves primarily within the consistency bands. In contrast, the SIDC and MCSTAT forecasts show substantial underprediction.

3.2. Empirical score decomposition: Miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components

For several decades, researchers have sought decompositions of the mean score $\bar{\mathsf{S}}$ from (2.3) into nonnegative components that allow for persuasive interpretation (Blattenberger & Lad, 1985; Bröcker, 2009; DeGroot & Fienberg, 1983; Ferro & Fricker, 2012; Murphy, 1973; Siegert, 2017; Yates, 1982). Typically, a decomposition involves a reliability or miscalibration (MCB) term that indicates how much the predicted probabilities differ from the conditional event frequencies, a resolution or discrimination (DSC) term that measures a forecast's ability to distinguish between events and non-events, and an uncertainty (UNC) component that quantifies the inherent difficulty of the prediction problem but does not depend on the forecast under consideration. While extant approaches lack stability under mandatory user decisions, particularly about binning, and may fail to provide an exact decomposition or might yield components that fail to be nonnegative, the CORP approach yields a new type of decomposition that resolves these issues.

As before, for a given record (2.2) suppose without loss of generality that $x_1 \leq \cdots \leq x_n$, and let $\hat{x}_1 \leq \cdots \leq \hat{x}_n$ denote the PAV re-calibrated values from (3.2), as plotted in the CORP reliability curve. Furthermore, let $r = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the realized unconditional event frequency. With S being any proper scoring rule, let

$$\bar{\mathsf{S}}_{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{S}(\hat{x}_i, y_i) \quad \text{and} \quad \bar{\mathsf{S}}_{\mathrm{R}} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{S}(r, y_i) \tag{3.3}$$

denote the mean score for the (re)Calibrated probabilities and the constant Reference forecast r, respectively. Using (3.3), the mean score $\bar{\mathsf{S}}$ from (2.3) decomposes as

$$\bar{\mathsf{S}} = \underbrace{\left(\bar{\mathsf{S}} - \bar{\mathsf{S}}_{\mathrm{C}}\right)}_{\mathrm{MCB}} - \underbrace{\left(\bar{\mathsf{S}}_{\mathrm{R}} - \bar{\mathsf{S}}_{\mathrm{C}}\right)}_{\mathrm{DSC}} + \underbrace{\bar{\mathsf{S}}_{\mathrm{R}}}_{\mathrm{UNC}}. \tag{3.4}$$

The miscalibration term MCB $= \bar{\mathsf{S}} - \bar{\mathsf{S}}_{\mathrm{C}}$ equals the difference in the mean score of the original versus the (re)calibrated forecast. It expresses deviations of the CORP reliability curve from the diagonal in terms of the score under consideration. The discrimination component DSC $= \bar{\mathsf{S}}_{\mathrm{R}} - \bar{\mathsf{S}}_{\mathrm{C}}$ quantifies how much the (re)calibrated forecast improves upon the reference score $\bar{\mathsf{S}}_{\mathrm{R}}$ that is based on a calibrated but constant forecast, and we note that, by construction, DSC is invariant under strictly increasing transformations of the forecast values. While small values of MCB are preferable, so are large values of the DSC component. The uncertainty term UNC $= \bar{\mathsf{S}}_{\mathrm{R}}$ is independent of the forecast at hand and provides a natural benchmark, as it equals the score of the (ex-post) best constant forecast. In contrast to earlier types of decomposition, the CORP decomposition from (3.4) is exact and guarantees that MCB ≥ 0 with equality if the original forecast is calibrated, and DSC ≥ 0 with equality if the (re)calibrated forecast is constant (Dimitriadis et al., 2021, Theorem 1).

The CORP decomposition applies under any proper scoring rule S. When S is the Brier score, it agrees with the classical decomposition of Murphy (1973) in the special case where the bins reduce to unique forecast values with an associated nondecreasing sequence of conditional event frequencies (Dimitriadis et al., 2021, Theorem 2). Under the misclassification rate that arises from (2.3) under the zero–one score in (2.6), the components admit appealing interpretations in terms of the original, the (re)calibrated, and the constant reference forecast being on the same or distinct side(s) of $\frac{1}{2}$. Table 2 shows the CORP decomposition of the mean Brier score, the mean logarithmic score, and the misclassification rate for the solar flare forecasts from Table 1. The MCB components confirm the visual appearance of the CORP reliability curves in Fig. 1. The NOAA forecast exhibits the least and MCSTAT the most pronounced lack of calibration. As discussed, the mean score $\bar{\mathsf{S}}$ and the MCB component under the logarithmic score are infinite for the ASSA forecast. We defer consideration of the DSC components to Section 4, where we focus on ROC curves.

3.3. Calibration metrics and the Brier score MCB component

Recently, there has been a surge of interest in calibration metrics in the machine learning literature. The widely used metric of the Expected (or Estimated) Calibration Error (ECE: Guo et al., 2017; Naeini et al., 2015) depends on binning and counting and thus is subject to the aforementioned types of instabilities (Dimitriadis et al., 2021; Roelofs et al., 2022) and biases (Bröcker, 2012a; Ferro & Fricker, 2012). In a recent review, Arrieta-Ibarra et al. (2022, p. 3) summarize that

> "the classical empirical calibration errors based on binning vary significantly based on the choice of bins. The choice of bins is fairly arbitrary and enables the analyst to fudge results (whether purposefully or unintentionally)".

To address these issues, Roelofs et al. (2022) use equal-mass bins and select the number of bins as large as possible while preserving isotonicity in the calibration curve. Arrieta-Ibarra et al. (2022) and Bröcker (2022) recommend graphical displays, calibration metrics, and tests

Table 2
CORP decomposition of the mean score in (2.3) for the probability forecasts of class C1.0+ solar flares from Table 1, under the Brier score (UNC = 0.211), the logarithmic score (UNC = 0.614), and misclassification rate (UNC = 0.303).

Forecast	Brier score			Logarithmic score			Misclassification rate		
	$\bar{\mathrm{S}}$	MCB	DSC	$\bar{\mathrm{S}}$	MCB	DSC	$\bar{\mathrm{S}}$	MCB	DSC
NOAA	0.144	0.006	0.073	0.449	0.027	0.191	0.205	0.004	0.102
SIDC	0.172	0.014	0.053	0.515	0.036	0.135	0.263	0.038	0.078
ASSA	0.184	0.007	0.035	∞	∞	0.085	0.273	0.006	0.036
MCSTAT	0.193	0.034	0.052	0.587	0.101	0.128	0.275	0.042	0.071

based on cumulative differences between predicted and observed event probabilities. While approaches of this type have appealing mathematical properties, cumulative quantities lack intuition and ease of interpretation. Similar to the method proposed by Roelofs et al. (2022), the CORP approach to reliability curves enforces isotonicity but uses the PAV algorithm to select the number and the arrangement of the bins in fully automated, optimal ways (Dimitriadis et al., 2021). The CORP decomposition from (3.3) and (3.4) is based on the CORP reliability curve. When S is the Brier score, it yields an MCB component that reduces to a classical calibration metric under modest conditions (Dimitriadis et al., 2021, Theorem 2). We propose using the Brier score MCB component as a calibration metric.

4. Receiver operating characteristic (ROC) curves visualize discrimination ability

While calibration is an important quality of probabilistic classifiers, a calibrated forecast is not necessarily powerful, as it may lack the ability to discriminate between events of low and high event probability. ROC curves are key in assessing this ability (Egan, 1975; Fawcett, 2006; Swets, 1973). In a nutshell, ROC curves visualize potential predictive power, detached from calibration considerations.

4.1. ROC curves

To introduce ROC curves, suppose that we use the threshold t to construct a hard classifier from the probability forecast x in the usual way by predicting an event ($y = 1$) if $x > t$ and predicting a non-event ($y = 0$) if $x \leq t$. For a record of the form (2.2), the resulting False Alarm Rate (FAR) and Hit Rate (HR) are given by

$$\mathrm{HR}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 1, x_i > t)}{\sum_{i=1}^{n} \mathbb{1}(y_i = 1)} \tag{4.1}$$

and

$$\mathrm{FAR}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 0, x_i > t)}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)}, \tag{4.2}$$

respectively. The ROC curve is the piecewise linear curve that connects the at most $n+1$ unique points of the form $(\mathrm{FAR}(t), \mathrm{HR}(t))$ that arise as the threshold t decreases.[2] Informally, the threshold t parameterizes the ROC curve, with the points (0,0) and (1,1) corresponding to $t \geq 1$ and $t < 0$, respectively.

ROC curves assess the discrimination ability of forecasts and can be interpreted diagnostically (Marzban, 2004). If the empirical conditional distributions for data (2.2) given an event ($y = 1$) and a non-event ($y = 0$) coincide, then the forecast is unable to distinguish between events and non-events, and its ROC curve lies on the diagonal. The larger the separation between these conditional distributions, the higher the discriminatory power of the forecast and the further to the upper left the ROC curve. For a perfectly discriminating probabilistic classifier, there is a threshold value t such that $y_i = 0$ if $x_i \leq t$ and $y_i = 1$ if $x_i > t$, and hence it exhibits an ideal ROC curve along the left and upper edges of the unit square. By construction, ROC curves are invariant under strictly increasing transformations of the forecast values.

An often neglected but important consideration concerns the concavity of ROC curves.[3] In practice, the original ROC curves constructed from empirical data (2.2) almost inevitably fail to be concave, as illustrated on the forecasts from Table 1 in the left-hand panel of Fig. 4. This observation is explained by Theorems 3 and 4 of Gneiting and Vogel (2022), according to which a ROC curve is concave if, and only if, the conditional event probability is nondecreasing with the forecast value x, which for empirical data is hardly ever the case. ROC curves assess discrimination ability — that is, potential predictive ability — and while potential predictive ability is ignorant of calibration, it can only be assessed under the assumption of larger forecast values implying higher event probabilities. In this light, displays of nonconcave ROC curves have been harshly criticized, with researchers positing that they "must be considered irrational" and "unethical when applied to medical decisions" (Pesce et al., 2010).

[2] Some researchers talk of ROC as relative operating characteristic (Swets, 1973), and the hit rate is also referred to as probability of detection, recall, sensitivity, or true positive rate. The false alarm rate is also known as the probability of false detection, fall-out, or false positive rate. It equals one minus the specificity, selectivity, or true negative rate. See https://en.wikipedia.org/wiki/Precision_and_recall#Definition_(classification_context), accessed 27 November 2022. Moreover, some researchers define the ROC curve as a display that connects the points $(1 - \mathrm{HR}(t), 1 - \mathrm{FAR}(t))$ for $t \in [0, 1]$, resulting in a curve that is mirrored at the anti-diagonal of the unit square, which maintains the interpretation that ROC curves at upper left are desirable (Hernández-Orallo et al., 2013).

[3] The machine learning literature uses the terms convex and convexity, following the seminal work of Fawcett (2006). Conventions in the mathematical literature suggest that concave and concavity are preferred terms.

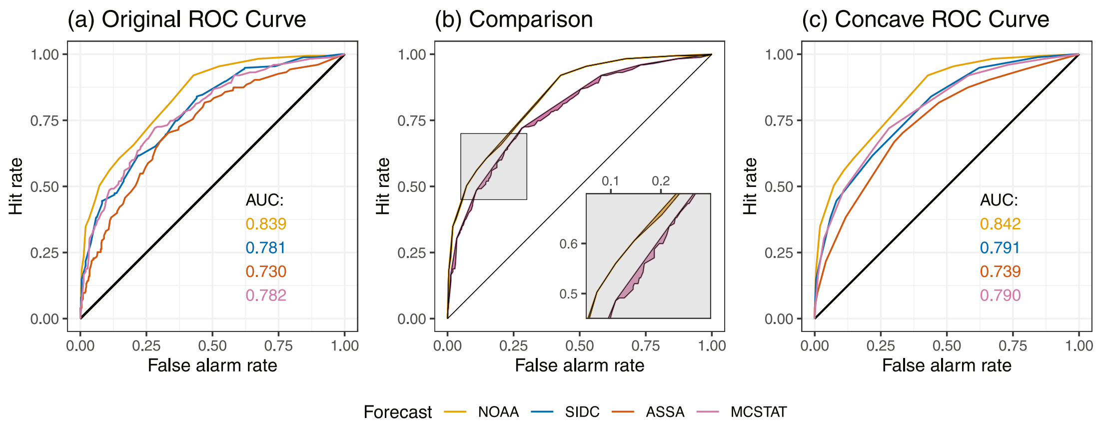

Fig. 4. ROC curves for the probability forecasts of class C1.0+ solar flares from Table 1. Panel (a) shows the original ROC curves for the forecasts $x_1, \ldots, x_n$ from (2.2), panel (c) the concave ROC curves for the PAV re-calibrated forecasts $\hat{x}_1, \ldots, \hat{x}_n$ from (3.2). In panel (b), the original and the concave ROC curves are shown for the NOAA and SIDC forecasts. The magnified details demonstrate that the original ROC curve morphs into its concave hull.

Fortunately, there is a straightforward remedy. If one computes the ROC curve from the PAV transformed forecast values (3.2) in lieu of the original forecasts from (2.2), the ROC curve morphs into its concave hull, that is, the smallest concave curve that lies to its upper left (Fawcett & Niculescu-Mizil, 2007). The corrected, concave version of the ROC curve exclusively compares discrimination ability, as differences in calibration get eliminated through re-calibration. In contrast, while original ROC curves focus on discrimination ability, conditional event frequencies that fail to be monotone generate confounding effects.[4]

We strongly recommend using concave ROC curves computed from PAV-transformed forecast values in triptych graphics for empirical data. In Fig. 4, the left-hand panel illustrates the original versions of the ROC curves for the solar flare forecasts, the right-hand panel displays the concave ROC curves, as in the triptych graphics in Fig. 1, and the middle panel illustrates the transition from the original curve to the concave hull. The NOAA forecast discriminates the most, and the ASSA forecast the least. The MCSTAT and SIDC forecasts exhibit roughly equal discrimination ability, with ROC curves that are nested in between the curves for the NOAA and ASSA forecasts.

4.2. The area under the curve (AUC) measure and the Brier score discrimination (DSC) component

Myriads of scientific papers have employed the Area Under the ROC Curve (AUC: Bradley, 1997; DeLong et al., 1988; Hanley & McNeil, 1982; Marzban, 2004) measure to compare the predictive performance of probabilistic classifiers. AUC admits an appealing interpretation as the probability of a value drawn at random from the empirical distribution of forecast values for an event being higher than a value drawn from the distribution for a non-event. An AUC value of 1 signifies perfect discrimination ability; a value of $\frac{1}{2}$ indicates no discrimination, corresponding to the trivial ROC curve on the diagonal. A value smaller than $\frac{1}{2}$ implies that interchanging the predictions for 0 and 1 would improve forecast accuracy. As the ROC curve is invariant under strictly increasing transformations, so is AUC, and we note that AUC exclusively concerns discrimination ability while ignoring (mis)calibration.

Consequently, AUC has severe limitations as an overall performance metric. To summarize arguments of Hand (2009) informally, AUC can be interpreted in the form of (2.8) with a mixing measure H that depends on the forecast values in complex ways. As argued by Hand (2009), Hand and Anagnostopoulos (2023), such a dependence is "absurd" and may entail "seriously misleading results". Indeed, Example 3 of Byrne (2016) demonstrates that deliberately misspecified classifiers may yield higher expected AUC than the true underlying probabilities. While these perspectives are qualified by an alternative representation of AUC studied by Flach et al. (2011), and by Theorem 6 of Byrne (2016) according to which the true probabilities yield the highest expected AUC when participants in forecast contests are informed a priori of the numbers of events and non-events, the limitations mentioned above remain, and AUC generally is not suitable as an omnibus performance measure.

Instead, both AUC and the DSC measure from the CORP score decomposition in (3.4) are measures of discrimination ability. Under calibration, the MCB component in (3.4) vanishes, and as the UNC component is independent of the forecast, $\bar{S}$ equals the DSC component, except for an additive constant. Furthermore, if S is the Brier score, then DSC reduces to a classical component, subject to conditions (Dimitriadis et al., 2021, Theorem

[4] In general, the transformation from the original probabilities $x_1 \leq \cdots \leq x_n$ to the PAV transformed, re-calibrated probabilities $\hat{x}_1 \leq \cdots \leq \hat{x}_n$ is monotonic, but not strictly monotonic, so a change in the ROC curve does not contradict the aforementioned invariance under strictly increasing transformations.

2). These relationships suggest using the Brier score DSC component as an attractive alternative to AUC if a measure of discrimination ability is sought. While both AUC and DSC are invariant under strictly increasing classifier transformations, DSC $= \bar{\mathrm{S}}_{\mathrm{R}} - \bar{\mathrm{S}}_{\mathrm{C}}$ admits an appealing interpretation in terms of the mean Brier score $\bar{\mathrm{S}}_{\mathrm{C}}$ for the PAV (re)calibrated forecast, up to the constant $\bar{\mathrm{S}}_{\mathrm{R}}$. This interpretation is retained when S is the logarithmic score or any other proper scoring rule.

5. Putting it together: The triptych graphics and MCB–DSC plots

Considering the class C1.0+ solar flares example from the previous sections and Figs. 1–4, the NOAA forecast is superior in all facets. However, rankings of the other forecasts from Table 1 depend on the criteria used: The ASSA forecast is well calibrated but exhibits poor discrimination ability. The MCSTAT and SIDC forecasts show discrimination ability between the NOAA and ASSA forecasts but lack calibration. Murphy curves provide an overall assessment of predictive performance, covering both calibration and discrimination ability, and favor the NOAA forecast, followed by the SIDC forecast. In contrast, the MCSTAT and ASSA forecast rankings depend on the scoring rule used.

5.1. Theoretical guarantees

In the triptych graphics, information about overall predictive ability in the Murphy curves is disentangled into facets of calibration, as displayed in CORP reliability curves, and facets of discrimination ability, as visualized by ROC curves. We now summarize existing and new theoretical results that support this intuition and provide new insights about links between the displays, with technical details being available in the Appendix. The findings are illustrated in idealized settings, where we consider the joint distribution of a pair (X, Y) of random variables, with X representing the probability forecast and Y the binary outcome. The triptych graphics in these ideal settings derive from the population quantities and can be interpreted as the triptych graphics that arise when a record of the form in (2.2) is generated by ever larger samples from the joint distribution of (X, Y). As the populations involved show nondecreasing CEPs, original and re-calibrated probabilities coincide and yield the same concave ROC curve.

We begin with a discussion of the role of calibration. As noted, any forecast can be re-calibrated ex-post by applying the PAV algorithm that converts the original forecast probabilities $x_1 \le \cdots \le x_n$ from (2.2) into the calibrated probabilities $\hat{x}_1 \le \cdots \le \hat{x}_n$ from (3.2). The following stylized fact summarizes findings from Schervish (1989, Theorem 6.3) and Holzmann and Eulert (2014, Corollary 2).

Fact A. *If a probability forecast fails to be calibrated, its re-calibrated version is superior in terms of Murphy curves.*

To illustrate Fact A, we consider the idealized Scenario A, where the binary outcome Y has event probability X_0, uniformly distributed on the unit interval. We compare to the probability forecast $X_1 = \frac{3}{8} + \frac{1}{4}X_0$, which is a strictly increasing transformation of X_0. Part (a) of Fig. 5 shows idealized triptych plots, where density plots for the unconditional distribution of the forecast values augment the reliability curves. While X_0 and X_1 have the same discrimination ability and identical ROC curves, X_1 fails to be calibrated, whereas X_0 is calibrated. In fact, X_0 is the re-calibrated version of X_1 and thus is superior in terms of Murphy curves.

The following fact summarizes a crucial, novel finding. We provide a rigorous version as Theorem 2 in the Appendix, which also contains its proof.

Fact B. *For two competing probability forecasts that are both calibrated, the number of crossing points of the ROC curves equals the number of crossing points of the Murphy curves.*

For illustration, we tend to Scenario B, where the forecasts X_1 and X_2 are calibrated. Let X_0 be uniformly distributed and the outcome Y have true event probability X_0. We consider the probability forecasts

$$X_1 = \begin{cases} X_0 & \text{if } X_0 < \frac{1}{4}, \\ \frac{1}{2} & \text{if } \frac{1}{4} \le X_0 \le \frac{3}{4}, \\ X_0 & \text{if } X_0 > \frac{3}{4}, \end{cases} \tag{5.1}$$

and

$$X_2 = \begin{cases} \frac{1}{8} & \text{if } X_0 < \frac{1}{4}, \\ X_0 & \text{if } \frac{1}{4} \le X_0 \le \frac{3}{4}, \\ \frac{7}{8} & \text{if } X_0 > \frac{3}{4}, \end{cases} \tag{5.2}$$

respectively. The triptych plots in part (b) of Fig. 5 illustrate that the ROC and Murphy curves share the same number, namely two, of crossing points.

In particular, Fact B implies that if two competing probability forecasts are calibrated, then there is a superiority relation in terms of ROC curves if, and only if, there is a superiority relation in terms of Murphy curves. The following fact sharpens this statement and relates to considerations of sharpness (Gneiting et al., 2007). Informally, a probability forecast is sharper than another if its forecast values are closer to the most confident values of 0 and 1, respectively. In the Appendix, we state and prove a rigorous version of the subsequent Fact C in Theorem 3.

Fact C. *Suppose two competing probability forecasts are calibrated, and one is sharper than the other. In that case, the sharper one is superior in terms of both ROC curves and Murphy curves, and vice versa.*

For an illustration in terms of nested information sets, which imply Murphy dominance as proved by Holzmann and Eulert (2014, Corollary 4) and Krüger and Ziegel (2021, Proposition 3.1), we consider Scenario C. Specifically, let the binary outcome Y have true event probability $X_0 = \Phi(\sum_{i=1}^{4} a_i)$, where a_1, a_2, a_3, and a_4, respectively, are independent standard normal variates, and Φ is the cumulative distribution function of the standard normal

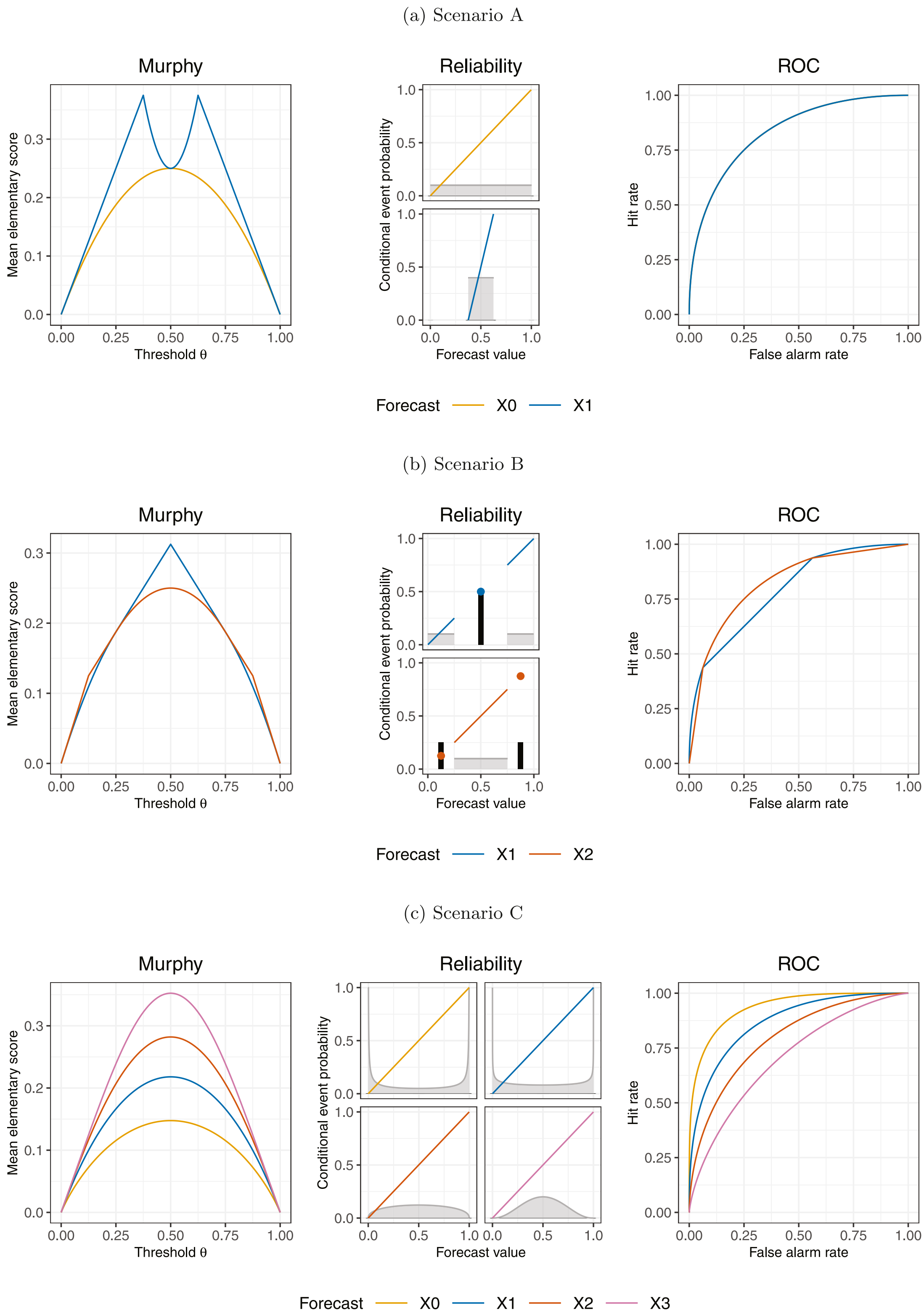

Fig. 5. Triptych displays in idealized Scenarios A, B, and C.

distribution. We consider the probability forecasts

$$X_j = \Phi\left(\frac{1}{(j+1)^{1/2}}\sum_{i=1}^{4-j} a_i\right) \tag{5.3}$$

for $j = 0, 1, 2,$ and 3. So, there are four independent sources of information, represented by $a_1, a_2, a_3,$ and a_4, and the forecast X_j provides the correct specification of the event probability conditional on $4 - j$ sources being available. Thus, the information sets are nested, the forecasts are calibrated, and they exhibit an increase in sharpness as j decreases. The triptych graphs in part (c) of Fig. 5 illustrate the increase in sharpness and the associated gain in terms of both ROC and Murphy curves. Pairwise comparisons between the forecasts illustrate the relationships guaranteed by Fact C.

5.2. *Visualizing classifier performance for many competitors simultaneously: MCB–DSC plots*

It is not uncommon that a multitude of competing forecasts are to be compared, with forecast contests being prime examples of such settings (Leka et al., 2019; Salganik et al., 2020a). Considering all competing forecasts in the triptych graphics results in overcrowded displays. However, the components of the CORP decomposition of a mean score $\bar{\mathrm{S}}$ from (3.4) can serve as numerical summaries. For a succinct comparison that considers multiple facets of forecast performance we propose a simple tool, which we call an MCB–DSC plot, namely, a scatter plot of the miscalibration (MCB) versus the discrimination (DSC) component of the CORP decomposition, augmented with a set of parallel contour lines that according to (3.4) correspond to an equal mean score. Notably, the joint consideration of the MCB and DSC components enables a comparison in terms of the mean score $\bar{\mathrm{S}}$ as well, contrary to the joint use of the (traditional) Brier score reliability component and AUC in the extant literature, as exemplified in Hewson and Pillosu (2021, Figure 2).

MCB–DSC plots admit appealing interpretations that apply under any choice of the underlying proper scoring rule S, as summarized now.

- For any forecast method considered, the mean score $\bar{\mathrm{S}}$ and the associated MCB and DSC components from (3.4) can be read off immediately. The UNC component depends on the outcomes only, is shared by all methods considered, and equals the label attached to the diagonal (that is, the thick green line with a unit slope that originates from the lower left corner where MCB $= 0$ and DSC $= 0$).
- The origin of the coordinate system in an MCB–DSC plot, where MCB $=$ DSC $= 0$, corresponds to the best constant forecast, namely, the unconditional event frequency in the test set. As noted, the diagonal corresponds to its mean score, namely, $\bar{\mathrm{S}}_{\mathrm{R}} =$ UNC. Forecasts that appear above the diagonal perform better than this reference; forecasts below the diagonal perform worse.
- The mean score $\bar{\mathrm{S}}_{\mathrm{C}} =$ UNC $-$ DSC of the PAV-(re)calibrated forecast corresponds to the DSC component, up to a constant and sign. This illustrates that the forecast with the largest DSC component has the greatest potential, provided (re)calibration is an option.

Figs. 2 and 7 show MCB–DSC plots for competing solar flare forecasts (Leka & Park, 2019; Leka et al., 2019), and for a considerably larger number of forecasts for a binary outcome from the Fragile Families Challenge (Salganik et al., 2020a), at which we take a closer look in Section 6.3. We focus on the Brier score decomposition, which is of particular appeal, as all terms involved are guaranteed to be finite, the mean Brier score $\bar{\mathrm{S}}$ equals the area under the Murphy curve, and under modest conditions, the Brier score MCB component reduces to a classical measure of deviations from the diagonal in a reliability curve (Dimitriadis et al., 2021). Under the logarithmic score, the mean score $\bar{\mathrm{S}}$ equals a weighted area under the Murphy curve, and both $\bar{\mathrm{S}}$ and the MCB component may become infinite. While MCB–DSC plots might mask details of the predictive performance, they are well suited to select subsets of interesting forecasts that can be analyzed further by plotting triptych graphics.

5.3. *Uncertainty quantification via confidence bands*

Our choices of details in the triptych graphics are tailored to typical desiderata in forecast evaluation. Murphy and ROC curves assess economic utility and discrimination ability, respectively, which should be compared between forecasts. This is facilitated by using a single panel to allow the curves to be compared against one other. For a reliability curve, the degree of deviation from the diagonal is relevant, and consistency bands address the natural question of whether (or not) observed differences between a CORP reliability curve and the diagonal can reasonably be attributed to chance alone despite the forecast probabilities being perfectly calibrated. One panel per forecast is the preferred visualization format for the joint display of a reliability curve and the associated consistency band.

While consistency bands only bear relevance to the reliability curve, confidence bands aim to quantify sampling variability conditional on the data record at hand. So, they apply to Murphy, reliability, and ROC curves. Here, the focus lies on the uncertainty attached to the respective estimate, and we change the visualization format to one panel per forecast. As a baseline method for uncertainty quantification, we use bootstrap case resampling to generate confidence bands; that is, we draw samples (with replacement, and of the same size n) from the collection (2.2) of prediction-observation tuples, find the respective (Murphy, reliability, or ROC) curve for each bootstrap sample, and construct confidence bands based on pointwise quantiles. In Fig. 6, we return to the C1.0+ solar flare data and show a modified version of the triptych, where confidence bands are attached to the individual curve estimates, that represent 90 percent of the curves that arise under the bootstrap scheme.[5]

[5] For Murphy curves and reliability curves, the confidence bands show the collection of pointwise confidence intervals given the parameter on the horizontal axis. For ROC curves, confidence intervals are

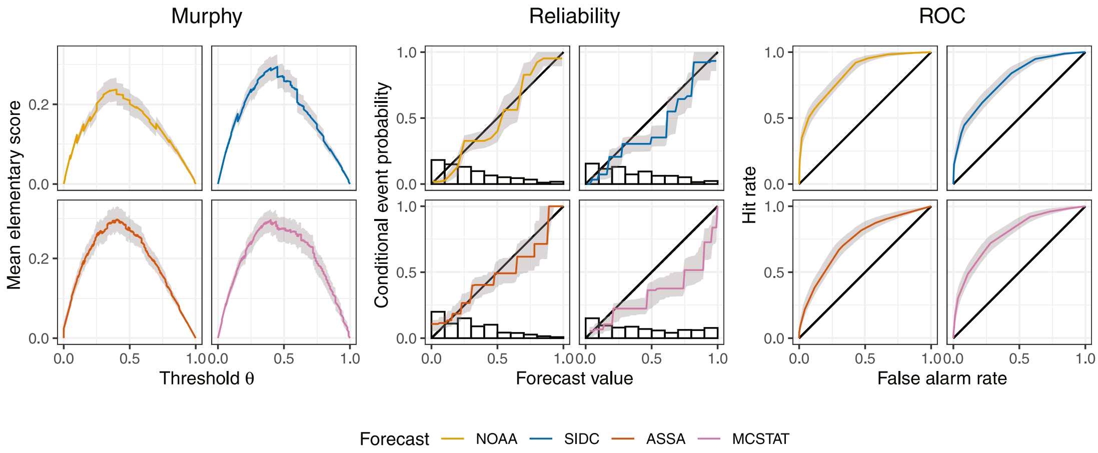

Fig. 6. Modified triptych graphics for probability forecasts of class C1.0+ solar flares from Table 1 and Figs. 1–4: Murphy curves (left panel), reliability curves (middle panel), and ROC curves (right panel), with uncertainty quantification via 90% confidence bands.

While case resampling may not always be valid, it is a reasonable starting point, and alternatives such as block-bootstrapping for time-series data can be readily employed. Other alternatives that are straightforward to implement include the resampling of observations from a (non)parametric estimate of the conditional event probability; for example, Dimitriadis et al. (2021) describe the use of the isotonic estimate for the resampling of reliability curves. Care must be taken when bootstrapping the MCB component in MCB–DSC plots due to its finite sample bias for forecasts calibrated on the population level. A more in-depth analysis of these issues is deferred to future work.

6. Empirical examples

We illustrate the use of the triptych displays and MCB–DSC plots for probabilistic classifiers from the academic literature in astrophysics, economics, and the social sciences.

6.1. Solar flares

Solar flares are energetic phenomena with potentially disastrous effects on modern terrestrial communications systems. Numerous forecasting systems for solar flares have been developed due to the increased availability of astrophysical data in real time. In a series of workshops, a data repository for comparative evaluation has been created (Barnes et al., 2016; Leka & Park, 2019; Leka et al., 2019).

We consider operational probability forecasts at a prediction horizon of a day ahead for solar flares of class C1.0+ and M1.0+, in exceedance of 10^{-6} and 10^{-5} Watts per square meter, respectively, as issued in calendar years 2016 and 2017. While Leka and Park (2019), Leka et al. (2019) describe 11 and 19 competing forecasts for C1.0+ and M1.0+ flares, respectively, there are substantial amounts of missing data in the records. As fair comparisons require evaluation on a joint set of forecast situations, we restrict our analysis to test sets of nine forecasts for C1.0+ flares on 577 days, as analyzed in Figs. 1 and 2, and 17 forecasts for M1.0+ flares on 431 days. On these test sets, records are complete, and flares have unconditional event frequencies of 30.3 and 3.5 percent, respectively.

Turning to M1.0+ flares, Fig. 7 shows MCB–DSC plots under the Brier score and the logarithmic score. Notably, under the logarithmic score, most forecasts are outperformed by the best constant forecast. For the triptych graphics in Fig. 8, we select the NOAA forecast, which performs well under either scoring rule, and the NICT forecast, which is by far the best-performing method in terms of the Brier score. Furthermore, we consider the MCSTAT forecast, which is poorly calibrated, and the ASSA forecast as a technique that lacks discrimination ability. Due to the low unconditional event probability, most forecast values are small. The NOAA forecast is of high quality in every regard. The NICT forecast is a hard classifier; it only issues forecast probabilities of 0 and 1. It performs best under most thresholds θ in the Murphy curves, except at very high values. Not surprisingly, it is penalized by an infinite mean logarithmic score. The MCSTAT forecast exhibits good discrimination ability but overpredicts; the conditional event frequency is persistently lower than the forecast value, resulting in poor overall performance. These issues can be addressed by (re)calibration, as opposed to the lack of discrimination ability of the ASSA forecast, which cannot be remedied.

6.2. Survey of professional forecasters (SPF) probability forecasts of economic recessions

We study probability forecasts for US GDP recessions: quarters with a negative real GDP growth rate. The database of the Survey of Professional Forecasters (SPF: Croushore & Stark, 2019) includes probability forecasts for a GDP decline in the current and the following four

created along parallel lines that connect points with the same index (for technical details, cf. Fig. 12).

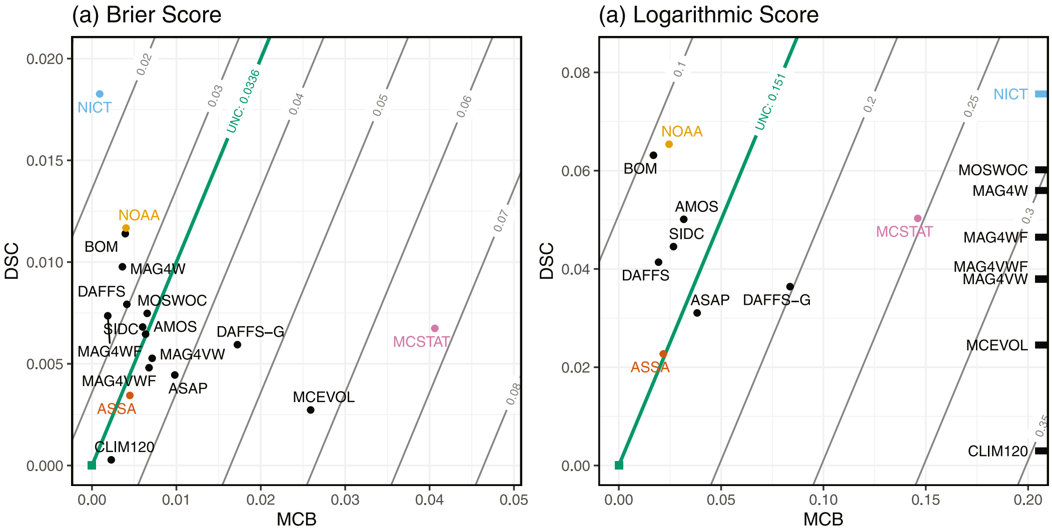

Fig. 7. MCB–DSC plots for probability forecasts of class M1.0+ solar flares under (a) the Brier score and (b) the logarithmic score. Colours align with Fig. 8. The green square at the origin represents the ex post best constant forecast, that is, the unconditional event frequency, and the thick green line separates forecasts that are better (above the line) and worse (below the line) than this baseline. Forecasts in panel (b) along the right margin have an infinite mean logarithmic score. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

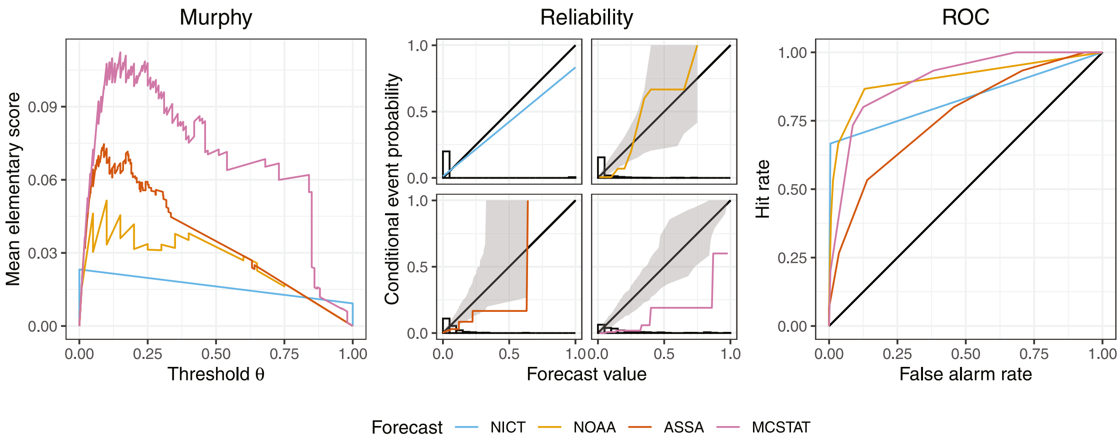

Fig. 8. Triptych graphics for probability forecasts of class M1.0+ solar flares. Reliability curves are shown on (the smallest contiguous interval containing) the support of the forecast distribution.

quarters from the fourth quarter of 1968 through the third quarter of 2019.[6] Following Lahiri and Wang (2013), we consider the mean of the individual SPF forecasts, which we denote SPF Consensus, and SPF forecaster #65, who reports the second most frequently among the survey participants. Lahiri and Wang (2013) study SPF probability forecasts through the first quarter of 2011 by evaluating calibration, assessing potential predictive ability through ROC curves, and reporting mean Brier and mean logarithmic scores. While their analysis is in the spirit of the triptych approach, it differs by necessity, as the methods proposed here depend on recent methodological advances (Dimitriadis et al., 2021; Ehm et al., 2016) that were not available to Lahiri and Wang (2013).

The triptych graphics in Fig. 9 compare SPF Consensus forecasts at prediction horizons ranging from current quarter nowcasts to four quarters ahead. The test set comprises 195 quarters between the second quarter of 1971 and the first quarter of 2019, with an unconditional event frequency of 0.16. We note the unsurprising yet drastic effects of the forecast horizon on the quality of the

[6] SPF forecasts are available under https://www.philadelphiafed.org/surveys-and-data/recess and binary outcomes under https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/routput. Data for the third quarter of 1975 are missing.

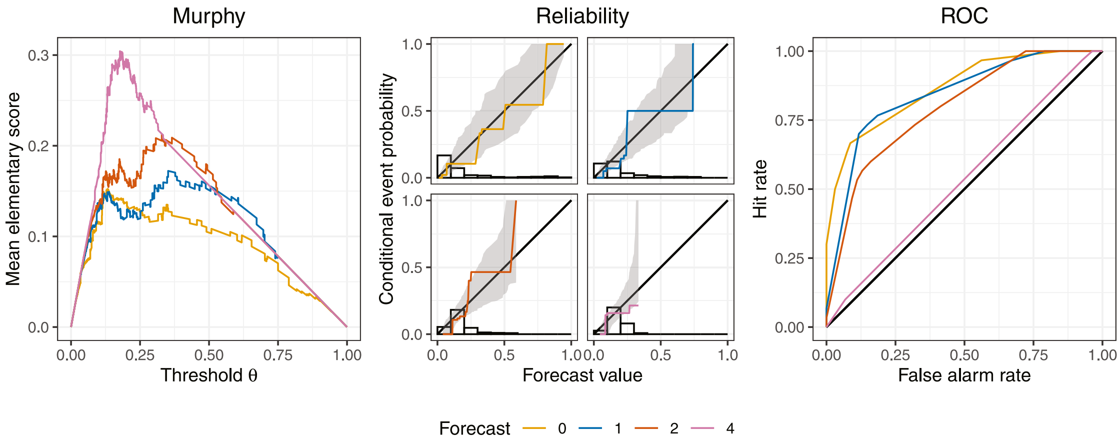

Fig. 9. Triptych graphics for SPF Consensus forecasts of US recessions at different prediction horizons.

Table 3
CORP decomposition of mean Brier score for probability forecasts of US recessions from SPF Consensus and SPF #65. The UNC component at 0.177 equals the mean Brier score for the best constant forecast, namely, the unconditional event frequency in the test set.

Forecast	$h=1$			$h=2$			$h=4$		
	$\bar{S}$	MCB	DSC	$\bar{S}$	MCB	DSC	$\bar{S}$	MCB	DSC
SPF Consensus	0.118	0.045	0.104	0.144	0.043	0.075	0.177	0.018	0.018
SPF #65	0.143	0.019	0.053	0.207	0.043	0.013	0.212	0.036	0.001

SPF Consensus forecast. While forecasts are reasonably well calibrated at all prediction horizons,[7] we see a diminishing range of the forecast values and the associated consistency bands. Accordingly, as shown by the ROC curves, the discrimination ability and, as visualized by the Murphy curves, the overall predictive ability deteriorate dramatically with the prediction horizon. Forecasts four quarters ahead have virtually no discrimination ability. Table 3 concerns a test set of 61 quarters between 1972 and 2006, for which SPF forecaster #65 predictions are available, with unconditional event frequency 0.23. The SPF Consensus forecast outperforms SPF forecaster #65 at all prediction horizons considered. The predictive performance of the individual forecaster falls markedly below that of the unconditional reference forecast at a prediction horizon of two quarters already. The SPF Consensus forecast maintains superior discrimination ability and overall performance at a prediction horizon of two quarters. It performs on par with the unconditional reference forecast at a prediction horizon of four quarters ahead.

6.3. Fragile families challenge

The Fragile Families Challenge (Salganik et al., 2020a, 2021, 2020b) is a scientific mass collaboration where teams supplied predictions for six (three binary and three real-valued) variables about life trajectories of children and families based on a rich data set from the Fragile Families and Child Wellbeing Study (Reichman et al., 2001). Salganik et al. (2020a) posit in their abstract that

> "despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model".

We use the triptych methodology to shed detailed light on this claim for one of the binary outcomes in the study, namely, eviction (from a family's home or apartment for not paying rent or mortgage). For eviction and also for the binary outcome job training (specifically, primary caregiver participation in job training) analyzed in panel (b) of Fig. 2, probability forecasts were sought for a holdout set of 1103 families, with 160 teams providing valid contributions. In addition, the Challenge organizers supplied nine benchmark forecasts based on commonly used statistical and machine learning techniques. The unconditional event frequency in the holdout set is 0.059 for eviction and 0.246 for job training.

The substantial numbers of up to 169 forecasts to be compared discourage the immediate use of the triptych graphics. To enable the selection of methods of interest, Fig. 10 shows MCB–DSC plots for the eviction data set under the Brier score and the logarithmic score, respectively. Benchmark forecasts are represented in green and cluster near the origin; they are well-calibrated but lack discrimination. Interestingly, a substantial number of competitors

[7] At prediction horizons of two and four quarters ahead, the consistency bands are not formally valid, as they depend on the assumption of independence of the instances, and thus, they provide qualitative guidance only. We encourage future work on the construction of consistency bands under dependencies.

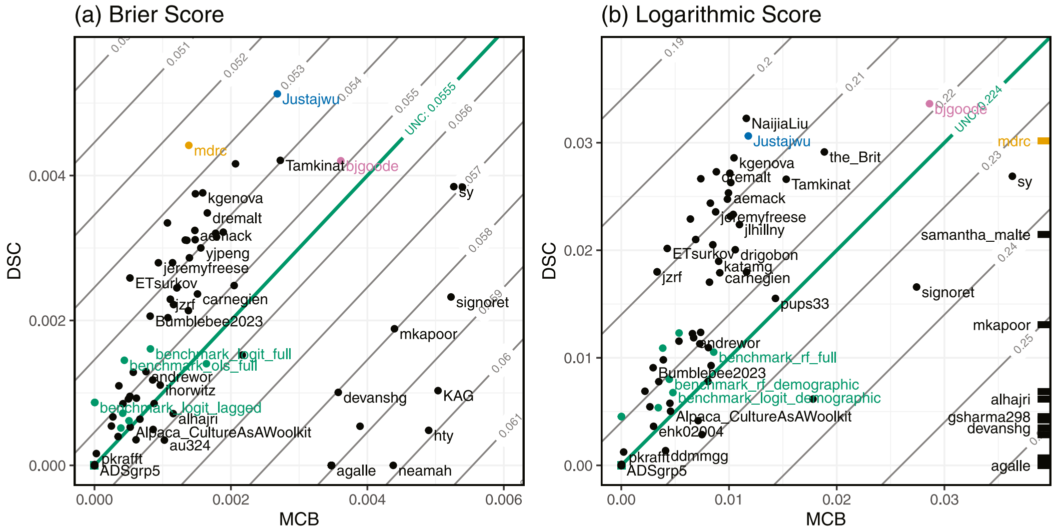

Fig. 10. MCB–DSC plots for probability forecasts of eviction from the Fragile Families Challenge under (a) the Brier score and (b) the logarithmic score. Benchmark forecasts are marked in green; the other colours align with Fig. 11. The green square at the origin represents the ex post best constant forecast, that is, the unconditional event frequency, and the thick green line separates forecasts that are better (above the line) and worse (below the line) than this baseline. Forecasts in panel (b) along the right margin have infinite mean logarithmic scores. Various forecasts are not represented in the displays due to trivial submissions (Salganik et al., 2020a, Table S5), overlap in symbols or labels, or a particularly poor (but finite) mean score. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

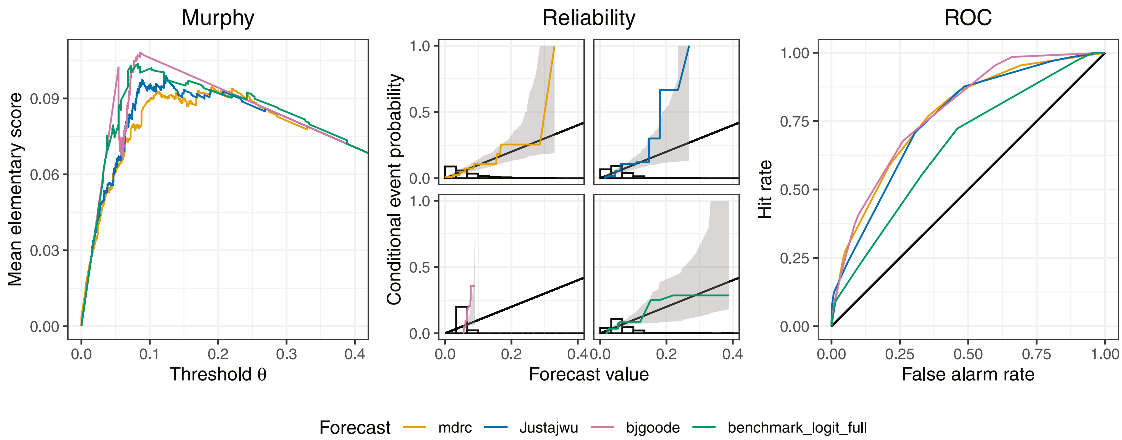

Fig. 11. Triptych graphics for probability forecasts of eviction from the Fragile Families Challenge.

outperform the benchmarks of Salganik et al. (2020a) and the best constant forecast in terms of both scores, though the improvement is small. However, for the job training data set in panel (b) of Fig. 2, none of the teams show predictive ability superior to the benchmark forecasts.

For the triptych graphics in Fig. 11, we select mrdc as the best forecast in terms of the Brier score, bjgoode, and Justajwu as the best-discriminating forecasts with respect to the Brier score and the logarithmic score, respectively, and the baseline technique benchmark_logit_full of Salganik et al. (2020a). The mrdc, bjgoode, and Justajwu forecasts outperform the baseline model in terms of discrimination ability, as depicted by the ROC curve. Due to the low unconditional event frequency, the forecasts take on values below 0.40 only, and we restrict the reliability curves, the respective consistency bands, and the Murphy curves in the triptych displays to suitable ranges. The baseline model is particularly well calibrated, making it competitive in overall predictive performance, as demonstrated in the Murphy curves. However, (re)calibrated versions of the mrdc, bjgoode, and Justajwu forecasts are bound to outperform the benchmarks by notable margins.

7. Discussion

In this paper, we have proposed the joint use of a triptych of diagnostic graphics in the evaluation of probability forecasts, including reliability curves in the recently proposed CORP form to assess calibration, the concave variant of receiver operating characteristic (ROC) curves to elucidate discrimination ability, and Murphy curves for the overall assessment of predictive performance and economic utility. For a succinct overview of the performance of multiple forecasts, we have introduced MCB–DSC plots that leverage the CORP decomposition of a mean proper score into miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components. A software implementation of the proposed tools and material for replicating the results in the article (Dimitriadis & Jordan, 2023a, 2023b) is available for R (R Core Team, 2022).

Our work builds on and supplements, and in a sense completes, extant software for the evaluation of probabilistic classifiers, or probabilistic forecasts in general, including but not limited to the `ROCR` (Sing et al., 2005), `pROC` (Robin et al., 2011), `verification` (NCAR - Research Applications Laboratory, 2015), and `reliabilitydiag` (Dimitriadis et al., 2021) packages in R, and the `PyCalib` package (Perello-Nieto et al., 2021) in Python. Arguably, closest in spirit are the `classifierplots` (Defazio & Campbell, 2020) package, which generates a "grid of diagnostic plots" that includes reliability curves and ROC curves, and the interactive `Calibrate` approach of Xenopoulos et al. (2023). However, these packages do not use the CORP approach of Dimitriadis et al. (2021) to generate reliability curves and score decompositions, nor do they implement Murphy curves.

Given the general theory of calibration and score decompositions developed by Gneiting and Resin (2023), the triptych approach to the diagnostic evaluation of probability forecasts might serve as a blueprint for evaluation strategies in similar settings, including but not limited to ordinary least squares regression, forecasts in the form of the expected value of a general real-valued outcome, quantile regression, and quantile forecasts. The case of quantiles has been studied by Gneiting et al. (2023), whose toolbox includes variants of Murphy curves, CORP reliability curves, and the CORP score decomposition. The recently developed universal ROC (UROC) curve of Gneiting and Walz (2022) generalizes the ROC curve from the classical case of a binary outcome to a general real-valued outcome, and the UROC curve might join CORP reliability curves and Murphy curves to form triptych graphics in the above types of settings. While currently available implementations of the triptych graphics and MCB–DSC plots involve static graphics only, ever-increasing numbers of competitors in forecast contests (Makridakis et al., 2022; Salganik et al., 2020a) may warrant the development of interactive versions, where users can select competitors of interest in an MCB–DSC plot and generate the respective triptych graphics on the fly.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Timo Dimitriadis, Tilmann Gneiting, and Alexander Jordan gratefully acknowledge support from the Klaus Tschira Foundation. Timo Dimitriadis gratefully acknowledges financial support from the German Research Foundation (DFG) through grant number 502572912. DFG funded the work of Peter Vogel through grant number 257899354. We thank three anonymous referees, Kajal Lahiri, Johannes Resin, Stefan Trautmann, and participants at the 2019 and 2022 International Symposium on Forecasting in Thessaloniki and Oxford, respectively, COMPSTAT 2022 in Bologna, and CFE 2022 in London for comments and advice. Special thanks to Davide Hailer for providing and reviewing code.

Appendix. Theoretical guarantees: Rigorous statements and proofs

We work within the prediction space setting of Gneiting and Ranjan (2013) and Gneiting and Vogel (2022), where we represent the probability forecast and the binary outcome as random variables X and Y, respectively, with joint distribution $\mathbb{Q}$, where $Y = 1$ represents an event and $Y = 0$ a non-event, with both types of outcomes having strictly positive probability. The symbol F denotes the marginal cumulative distribution function (CDF) of the forecast X. When comparing competing probability forecasts X_1 and X_2 for the binary outcome Y, we also denote their joint distribution by $\mathbb{Q}$.

The probability forecast X is *calibrated* if $\mathbb{Q}(Y = 1 \mid X) = X$ almost surely. We define the conditional CDF $F_{|1}(t) = \mathbb{Q}(X \le t \mid Y = 1)$ and $F_{|0}(t) = \mathbb{Q}(X \le t \mid Y = 0)$ such that, for any threshold value $t \in [0, 1]$, the population versions of *hit rate* (HR) and *false alarm rate* (FAR) are given by

$$\mathrm{HR}(t) = 1 - F_{|1}(t) = \mathbb{Q}(X > t \mid Y = 1)$$

and

$$\mathrm{FAR}(t) = 1 - F_{|0}(t) = \mathbb{Q}(X > t \mid Y = 0),$$

respectively. If $F_{|0}$ and $F_{|1}$ are continuous and strictly increasing, the ROC curve can be identified with a function $R : [0, 1] \to [0, 1]$, where $R(0) = 0$, $R(1) = 1$, and $R(p) = 1 - F_{|1}(F_{|0}^{-1}(1 - p))$ for $p \in (0, 1)$. In the general setting, including but not limited to the case of empirical distributions for data of the form (2.2), the *raw ROC diagnostic* is the set-theoretic union of the points of the form $(\mathrm{FAR}(t), \mathrm{HR}(t))'$ within the unit square, from which the *ROC curve* is obtained by linear interpolation (Gneiting & Vogel, 2022). On the support of F, we can index the ROC curve in terms of $c = F(t)$, with a natural extension to $c \in [0, 1]$ via linear interpolation.

Except for assuming that the competing probability forecasts X_1 and X_2 are calibrated and requiring that $\mathbb{Q}(Y = 0) \in (0, 1)$, we do not impose any regularity conditions on the distribution $\mathbb{Q}$. Our proofs rely on the novel

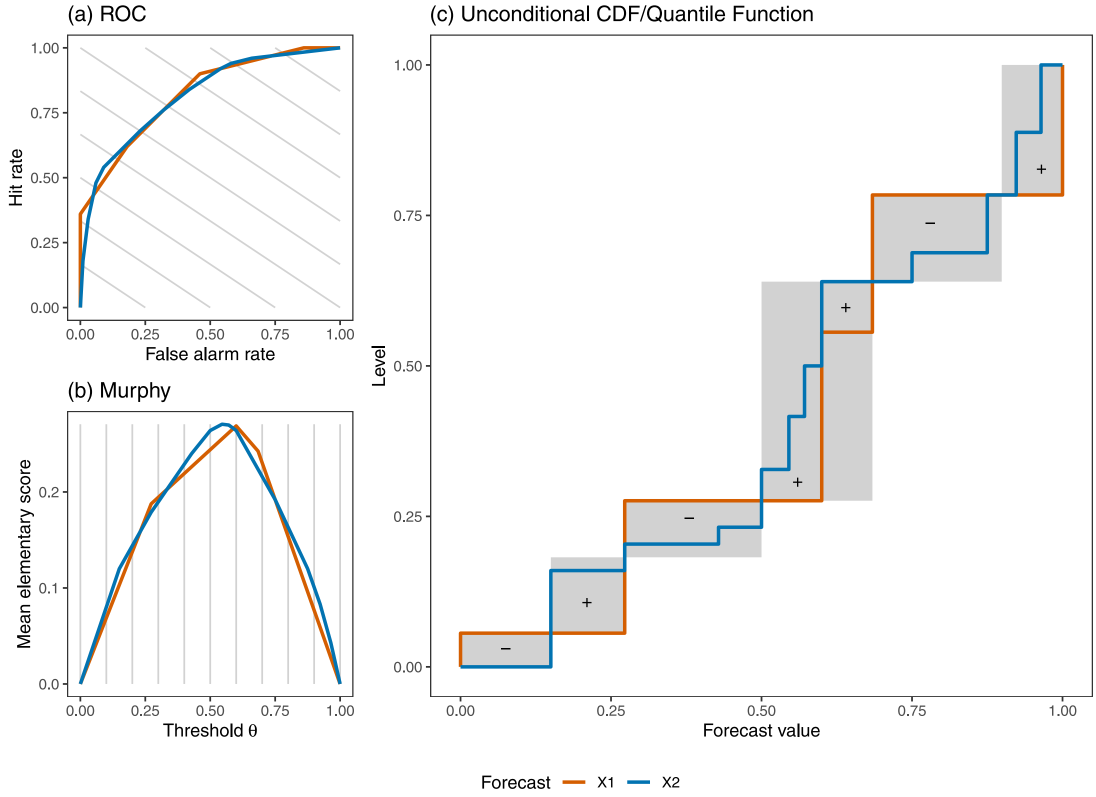

Fig. 12. Schematic illustration of the technical concepts in the statements and proofs of Lemma 1 and Theorem 2. Given calibrated classifiers X_1 and X_2 for the binary outcome Y, where $0 < \pi_0 = \mathbb{Q}(Y = 0) < 1$, the panel at right shows the unconditional CDFs, F_1 and F_2, and generalized inverses or quantile functions, Q_1 and Q_2, respectively. At lower left, the Murphy curve difference $\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta)$ arises as the vertical difference between the two Murphy curves at the cost–loss parameter θ. The panel at upper left concerns the ROC curve difference $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c)$. The parallel lines with slope $-\pi_0/(1-\pi_0)$ connect points on the two ROC curves with the same index, c.

Lemma 1 that expresses the difference of the Murphy curves,

$$2\,\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta) = \mathbb{E}_{\mathbb{Q}}\,\mathrm{S}_\theta(X_1, Y) - \mathbb{E}_{\mathbb{Q}}\,\mathrm{S}_\theta(X_2, Y)$$

at the cost–loss parameter $\theta \in (0, 1)$, as sketched in the lower left display of Fig. 12, and the difference of the ROC curves, $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c)$, at $c = F(t)$ in terms of the unconditional CDFs, F_1 and F_2, and the associated quantile functions, Q_1 and Q_2, of the classifiers X_1 and X_2. As the proof of Theorem 2 demonstrates, and the upper left panel of Fig. 12 illustrates, the respective points on the ROC curves lie on parallel lines with slope $-\pi_0/(1-\pi_0)$, and the ROC curve difference $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c)$ is taken in this direction. We note the close relation to the idea that underlies the *Kendall curve* (Hernández-Orallo et al., 2013), which considers the difference between a ROC curve and its optimal version, that is, the left and upper boundary of the unit square, in exactly this direction. The Kendall curve then presents the information contained in the ROC curve in cost space, which coincides with the abscissa of the Murphy curve.

Lemma 1. *Suppose that X_1 and X_2 are calibrated probability forecasts for the binary outcome Y, where $\mathbb{Q}(Y = 0) \in (0, 1)$. Then, the Murphy curve difference and the ROC curve difference are*

$$\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta) = \int_{[0,\theta]} (F_2(x) - F_1(x))\,\mathrm{d}x$$

and

$$\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c) = \int_{[0,c]} (Q_1(\alpha) - Q_2(\alpha))\,\mathrm{d}\alpha,$$

where F_1 and F_2 are the unconditional CDFs for X_1 and X_2, and Q_1 and Q_2 are the left-continuous generalized inverses to F_1 and F_2, respectively.

Proof. To simplify notation, let X be a probability forecast for Y with unconditional CDF F and generalized inverse Q. Let $F_{|i} = \mathbb{Q}(X \le \cdot \mid Y = i)$ and $\pi_i = \mathbb{Q}(Y = i)$ for $i \in \{0, 1\}$.

For the Murphy curve difference, consider the elementary scoring function S_θ from (2.7). For $\theta \in (0, 1)$ the expected elementary score of X is

$$\mathbb{E}_{\mathbb{Q}}\mathrm{S}_\theta(X, Y) = 2\theta\,\mathbb{Q}(X > \theta, Y = 0) \\ + 2(1-\theta)\,(\mathbb{Q}(X < \theta, Y = 1) + \theta\,\mathbb{Q}(X = \theta))\,.$$

The Murphy curve is the graph of the map $\mathrm{MC}\colon \theta \mapsto \mathbb{E}_{\mathbb{Q}}\mathrm{S}_\theta(X, Y)$. Under calibration of X, we have $x = \mathbb{Q}(Y =$

$1 \mid X = x)$ and $\theta\, \mathbb{Q}(X = \theta) = \mathbb{Q}(X = \theta, Y = 1)$, which yields

$$\begin{aligned}\mathrm{MC}(\theta) &= 2\theta\, \mathbb{Q}(Y = 0) - 2\theta\, \mathbb{Q}(X \le \theta) + 2\mathbb{Q}(X \le \theta, Y = 1) \\ &= 2\theta\pi_0 - 2\theta F(\theta) + 2\int_{[0,\theta]} x \,\mathrm{d}F(x) \\ &= 2\theta\pi_0 - 2\int_{[0,\theta]} F(x)\,\mathrm{d}x\end{aligned}$$

using integration by parts for Lebesgue–Stieltjes integrals. Therefore, a version of the Murphy curve difference of two forecasts X_1 and X_2 is

$$\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta) = \tfrac{1}{2}\mathrm{MC}_{X_1}(\theta) - \tfrac{1}{2}\mathrm{MC}_{X_2}(\theta) = \int_{[0,\theta]} (F_2(x) - F_1(x))\,\mathrm{d}x$$

for $\theta \in [0, 1]$, as claimed.

Turning to the ROC curve distance, the raw ROC diagnostic of the probability forecast X is the set-theoretic union of the points of the form $(\mathrm{FAR}(t), \mathrm{HR}(t))' = (1 - F_{|0}(t), 1 - F_{|1}(t))'$ for all threshold values t. Under calibration of X and using substitution for Lebesgue–Stieltjes integrals, we have

$$\pi_1 F_{|1}(t) = \int_{[0,t]} x\,\mathrm{d}F(x) = \int_{[0,F(t)]} Q(\alpha)\,\mathrm{d}\alpha.$$

Furthermore, $(F \circ Q \circ F)(\theta) = F(\theta) = \pi_0 F_{|0}(\theta) + \pi_1 F_{|1}(\theta)$, for Q is a generalized inverse to F. We use the idea that underlies the construction of the rate-driven cost curve and the Kendall curve (Hernández-Orallo et al., 2013), namely, to substitute $Q(c)$ for t, where $c = F(t)$, in concert with the above facts, to write the points in the raw ROC diagnostic as

$$\left(1 - \frac{1}{\pi_0}\left(c - \int_{[0,c]} Q(\alpha)\,\mathrm{d}\alpha\right), 1 - \frac{1}{\pi_1}\int_{[0,c]} Q(\alpha)\,\mathrm{d}\alpha\right)',$$

where $c \in \mathrm{Im}(F) = \{F(t) : t \in \mathbb{R}\}$. These expressions interpolate linearly when $c \in [0, 1] \setminus \mathrm{Im}(F)$, since $Q(\alpha) = Q(\min(\mathrm{Im}(F) \cap [\alpha, 1]))$. Therefore, the ROC curve, as a linear interpolation of the raw ROC diagnostic, is the graph of the map

$$\mathrm{ROC}^{\mathrm{curve}}_X : [0, 1] \to [0, 1]^2,$$

$$c \mapsto \left(1 - \frac{1}{\pi_0}\left(c - \int_{[0,c]} Q(\alpha)\,\mathrm{d}\alpha\right), 1 - \frac{1}{\pi_1}\int_{[0,c]} Q(\alpha)\,\mathrm{d}\alpha\right)'.$$

Given competing probability forecasts X_1 and X_2, the vector-valued difference between the ROC curves at $c \in [0, 1]$ is

$$\begin{aligned}&\mathrm{ROC}^{\mathrm{curve}}_{X_1}(c) - \mathrm{ROC}^{\mathrm{curve}}_{X_2}(c) \\ &\quad= \left(\frac{1}{\pi_0}, -\frac{1}{\pi_1}\right)' \left(\int_{[0,c]} (Q_1(\alpha) - Q_2(\alpha))\,\mathrm{d}\alpha\right),\end{aligned}$$

which demonstrates that pointwise differences between ROC curves are to be measured along lines with slope $-\pi_0/\pi_1$ in the ROC curve plot, as illustrated in the upper left panel of Fig. 12. The factor at right is chosen as the pointwise distance between the ROC curves for X_1 and X_2 at index $c \in [0, 1]$, that is, $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c) = \int_{[0,c]} (Q_1(\alpha) - Q_2(\alpha))\,\mathrm{d}\alpha$, as claimed. □

A function on the unit interval has *n sign changes* if there exists a partition of the unit interval with $n + 1$ members, such that the function is nonnegative (nonpositive) with at least one nonzero value on the first and nonpositive (nonnegative) with at least one nonzero value on the second of any two consecutive members of the partition.

We state and prove a rigorous version of Fact B in Section 5.1.

Theorem 2. *Suppose that X_1 and X_2 are calibrated probability forecasts for the binary outcome Y, where $\mathbb{Q}(Y = 0) \in (0, 1)$. Let F_1 and F_2 denote the unconditional CDFs of X_1 and X_2, and suppose that $F_1 - F_2$ has a finite number $n \ge 1$ of sign changes. Then the following hold:*

(a) The ROC curve difference $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}$ has at most $n - 1$ sign changes.
(b) The Murphy curve difference $\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}$ and the ROC curve difference $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}$ have the same number of sign changes.

Proof. For part (a), note that the integrand $Q_1 - Q_2$ of $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}$ has the same number of sign changes as $F_1 - F_2$ and that no additional sign change can be introduced by integration. Since $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c)$ evaluates to 0 at both $c = 0$ (by definition) and $c = 1$ (due to calibration, which implies that $\mathbb{E}X_1 = \mathbb{E}X_2$), the number of sign changes of the integral must be smaller than the number of sign changes of the integrand.

For part (b), consider the two partitions of the unit interval generated by sign changes of the integrands $F_2 - F_1$ and $Q_1 - Q_2$, respectively. As both partitions have the same number of elements, the elements can be matched pairwise to create blocks, as illustrated in the right panel of Fig. 12. At the beginning and end of every block (the bottom left or top right corner, respectively), we have equality of the differences $\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta)$ and $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c)$, and within a block either both differences are nonincreasing or both are nondecreasing. Therefore, in a single block, both differences experience a single sign change, or neither does. □

The assumption of finitely many sign changes in $F_1 - F_2$ is not particularly restrictive, as it is satisfied whenever either X_1 or X_2 has finite support (covering all empirical cases) or when both have a finite number of (potentially interval-valued) modes. Otherwise, the statement in part (b) continues to hold whenever the number of sign changes in either $\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}$ or $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}$ is finite. Furthermore, the assumption of calibration guarantees the existence of at least one sign change in $F_1 - F_2$ whenever $F_1 \ne F_2$, since otherwise F_1 and F_2 are stochastically ordered, which implies $\mathbb{E}X_1 \ne \mathbb{E}X_2$ and contradicts the assumption of calibration.

Informally, a probability forecast for a binary outcome is sharper than another if its forecast values are closer to the most confident values of 0 and 1, respectively (Gneiting et al., 2008). To formalize the notion of sharpness we follow Krüger and Ziegel (2021) and define X_1 to be *sharper* than X_2 if it is greater in convex order, that is, if $\mathbb{E}\,\phi(X_1) \ge \mathbb{E}\,\phi(X_2)$ for all convex functions ϕ on the unit interval, with strict inequality for some ϕ. As Krüger and

Ziegel (2021, p. 974) note, given the assumption that forecasts are calibrated, being larger in convex order implies that the forecast values are more spread out toward the most confident probabilities of 0 and 1.

In Fact C in Section 5.1, we express the idea that if competing probability forecasts are calibrated, and one of them is sharper than the other, then the sharper one is superior in terms of both ROC curves and Murphy curves, and vice versa. To state a rigorous version of this fact, we say that X_1 *dominates* X_2 *in the ROC sense* if $\mathrm{D}^{\mathrm{ROC}}_{X_1,X_2}(c) \leq 0$ for all $c \in [0, 1]$ with strict inequality at some c. Similarly, X_1 *dominates* X_2 *in the Murphy sense* if $\mathrm{D}^{\mathrm{MC}}_{X_1,X_2}(\theta) \leq 0$ for all $\theta \in [0, 1]$ with strict inequality at some θ.

Theorem 3. *Suppose that X_1 and X_2 are calibrated probability forecasts for the binary outcome Y, where $\mathbb{Q}(Y = 0) \in (0, 1)$. Then the following relations are equivalent:*

(i) X_1 is sharper than X_2.
(ii) X_1 dominates X_2 in the ROC sense.
(iii) X_1 dominates X_2 in the Murphy sense.

Proof. The equivalence of statements (i) and (iii) is a special case of Theorem 3.1 in Krüger and Ziegel (2021), and the equivalence of statements (ii) and (iii) follows from the arguments in the proof of part (b) of Theorem 2. □

Theorem 3 demonstrates that if probability forecasts are calibrated, then comparisons in terms of sharpness, discrimination ability, and proper scoring rules yield congruent insights, as illustrated in Scenario C and part (c) of Fig. 5 in Section 5.1. Related results have been discussed by Krzysztofowicz and Long (1990, p. 670), Wilks (2019, p. 416), and references therein.

References

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., McGinn, T., & Guyatt, G. (2017). Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *Journal of the American Medical Association*, *318*, 1377–1384.

Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., & Xu, C. (2022). Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, *23*, 1–54.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silvermann, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, *26*, 641–647.

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., Yuan, Y., Zhang, J., McAteer, R. T. J., Bloomfield, D. S., Higgins, P. A., Gallagher, P. T., Falconer, D. A., Georgoulis, M. K., Wheatland, M. S., Balch, C., Dunn, T., & Wagner, E. L. (2016). A comparison of flare forecasting methods. I. Results from the "all-clear" workshop. *Astrophysical Journal*, *829*, 89.

Blattenberger, G., & Lad, F. (1985). Separating the brier score into calibration and refinement components: A graphical exposition. *American Statistician*, *39*, 26–32.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, *135*, 1512–1519.

Bröcker, J. (2012a). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, *39*, 655–667.

Bröcker, J. (2012b). Probability forecasts. In I. T. Jolliffe, & D. B. Stephenson (Eds.), *Forecast verification: a practitioner's guide in atmospheric science* (2nd ed.). (pp. 119–184). Wiley.

Bröcker, J. (2022). Uniform calibration tests for forecasting systems with small lead time. *Statistics and Computing*, *32*, 1–13.

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, *22*, 651–661.

Buja, A., Stuetzle, W., & Shen, Y. (2005). *Loss functions for binary class probability estimation and classification: Structure and application*: *Working paper*, available at http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf.

Byrne, S. (2016). A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electronic Journal of Statistics*, *10*, 380–393.

Croushore, D., & Stark, T. (2019). Fifty years of the survey of professional forecasters. *Economic Insights*, *4*, 1–11.

De Leeuw, J., Hornik, K., & Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, *32*(5), 1–24.

Defazio, A., & Campbell, H. (2020). Classifierplots: Generates a visualization of classifier performance as a grid of diagnostic plots. R package version 1.4.0, available at https://cran.r-project.org/package=classifierplots.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Statistician*, *32*, 12–22.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, *118*, Article e2016191118.

Dimitriadis, T., & Jordan, A. I. (2023a). Triptych: Diagnostic graphics to evaluate forecast performance. R package version 0.1.2, available at https://cran.r-project.org/package=triptych.

Dimitriadis, T., & Jordan, A. I. (2023b). Replication material for "evaluating probabilistic classifiers: The triptych". Available at https://github.com/TimoDimi/replication_triptych.

Drummond, C., & Holte, R. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, *65*, 95–130.

Egan, J. P. (1975). *Signal dectection theory and ROC analysis*. New York: Academic Press.

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B*, *78*, 505–562.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Fawcett, T., & Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, *68*, 97–106.

Ferro, C. A. T., & Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, *138*, 1954–1960.

Filho, T. S., Song, H., Perelló-Nieto, M., Santos-Rodríguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*, 3211–3260.

Flach, P. A. (2017). Classifier calibration. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 210–217). New York: Springer.

Flach, P., Hernández-Orallo, J., & Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th international conference on machine learning*.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B*, *69*, 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Gneiting, T., & Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, in press, available at https://arxiv.org/abs/2108.03210.

Gneiting, T., Stanberry, L. A., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Rejoinder on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, *17*, 256–264.

Gneiting, T., & Vogel, P. (2022). Receiver operating characteristic (ROC) curves: Equivalences, beta model, and minimum distance estimation. *Machine Learning*, *111*, 2147–2159.

Gneiting, T., & Walz, E.-M. (2022). Receiver Operating Characteristic (ROC) movies, universal ROC (UROC) curves, and Coefficient of Predictive Ability (CPA). *Machine Learning*, *111*, 2769–2797.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., & Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, *10*, 597–621.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning*.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hand, D. J., & Anagnostopoulos, C. (2023). Notes on the H-measure of classifier performance. *Advances in Data Analysis and Classification*, *17*, 109–124.

Hanley, A., & McNeil, J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2011). Brier curves: A new cost-based visualisation of classifier performance. In *Proceedings of the 28th international conference on machine learning*.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, *13*, 2813–2869.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2013). ROC curves in cost space. *Machine Learning*, *93*, 71–91.

Hewson, T. D., & Pillosu, F. M. (2021). A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, *2*, 132.

Holzmann, H., & Eulert, M. (2014). The role of the information set for forecasting — with applications to risk management. *Annals of Applied Statistics*, *8*, 595–621.

Janssens, A. C. J. W. (2020). ROC curves for clinical prediction models part 4. Selection of the risk threshold — once chosen, always the same? *Journal of Clinical Epidemiology*, *126*, 224–225.

Krüger, F., & Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, *39*, 972–983.

Krzysztofowicz, R., & Long, D. (1990). Fursion of detection probabilities and comparison of multisensor systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *20*, 665–677.

Lahiri, K., & Wang, J. G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, *29*, 175–190.

Leka, K. D., & Park, S.-H. (2019). A comparison of flare forecasting methods II: Data and supporting code. http://dx.doi.org/10.7910/DVN/HYP74O, Harvard Dataverse, V1.

Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., Bloomfield, D. S., McCloskey, A. E., Delouille, V., Falconer, D., Gallagher, P. T., Georgoulis, M. K., Kubo, Y., Lee, K., Lee, S., Lobzin, V., Mun, J., Murray, S. A., Nageem, T. A. M. H., ... Terkildsen, M. (2019). A comparison of flare forecasting methods. II. Benchmarks, metrics, and performance results for operational solar flare forecasting systems. *Astrophysical Journal Supplement Series*, *243*, 36.

Lindley, D. V. (1982). The improvement of probability judgements. *Journal of the Royal Statistical Society Series A: General*, *145*, 117–126.

Makridakis, S., Petropoulos, F., & Spiliotis, E. (2022). Introduction to the M5 forecasting competition special issue. *International Journal of Forecasting*, *38*, 1279–1282.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting*, *19*, 1106–1114.

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*, 292–304.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *26*, 41–47.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, *7*, 435–455.

Naeini, P. N., Cooper, G. F., & Hausknecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 2901—2907).

NCAR - Research Applications Laboratory (2015). Verification: Weather forecast verification utilities. R package version 1.42, available at https://cran.r-project.org/package=verification.

Perello-Nieto, M., Song, H., Silva Filho, T., & Kängsepp, M. (2021). *PyCalib a library for classifier calibration*. Zenodo, http://dx.doi.org/10.5281/zenodo.5518877.

Pesce, L. L., Metz, C. E., & Berbaum, K. S. (2010). On the convexity of ROC curves estimated from radiological test results. *Academic Radiology*, *17*, 960–968.

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2011). A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, *23*, 1601–1618.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*, 303–326.

Richardson, D. S. (2012). Economic value and skill. In I. T. Jolliffe, & D. B. Stephenson (Eds.), *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). (pp. 167–184). Wiley.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 1–8.

Roelofs, R., Cain, N., Shlens, J., & Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *Proceedings of the 25th international conference on artificial intelligence and statistics*.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020a). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*, 8398–8403.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2021). Correction for Salganik et al., Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *118*, Article 2118703118.

Salganik, M., Lundberg, I., Kindel, A., & McLanahan, S. (2020b). Replication material for "Measuring the predictability of life outcomes using a scientific mass collaboration". http://dx.doi.org/10.7910/DVN/CXSECU, Harvard Dataverse, Harvard Dataverse, V3.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Schervish, M. J. (1989). A general method for comparing probability assessors. *The Annals of Statistics*, *17*, 1856–1879.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Empirical Economics*, *1*, 43–62.

Siegert, S. (2017). Simplifying and generalizing Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, *143*, 1178–1183.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*, 3940–3941.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, *182*, 990–1000.

Taillardat, M., & Mestre, O. (2020). From research to applications – Examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, *27*, 329–347.

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Amsterdam: Elsevier.

Xenopoulos, P., Rulff, J., Nonato, L. G., Barr, B., & Silva, C. (2023). Calibrate: Interactive analysis of probabilistic model output. *IEEE Transactions on Visualization and Computer Graphics*, *29*, 853–863.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132–156.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 694–699).