Electronic Journal of Statistics
Vol. 18 (2024) 4992–5044
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2316

Decompositions of the mean continuous ranked probability score

Sebastian Arnold[*1], **Eva-Maria Walz**[*2,3], **Johanna Ziegel**[1] **and Tilmann Gneiting**[3,2]

[1]*Seminar for Statistics, ETH Zurich,*
e-mail: sebastian.arnold@stat.math.ethz.ch; ziegel@stat.math.ethz.ch

[2]*Institute for Stochastics, Karlsruhe Institute of Technology (KIT),*
e-mail: eva-maria.walz@kit.edu

[3]*Computational Statistics (CST) group, Heidelberg Institute for Theoretical Studies,*
e-mail: tilmann.gneiting@h-its.org

Abstract: The continuous ranked probability score (crps) is the most commonly used scoring rule in the evaluation of probabilistic forecasts for real-valued outcomes. To assess and rank forecasting methods, researchers compute the mean crps over given sets of forecast situations, based on the respective predictive distributions and outcomes. We propose a new, isotonicity-based decomposition of the mean crps into interpretable components that quantify miscalibration (MCB), discrimination ability (DSC), and uncertainty (UNC), respectively. In a detailed theoretical analysis, we compare the new approach to empirical decompositions proposed earlier, generalize to population versions, analyse their properties and relationships, and relate to a hierarchy of notions of calibration. The isotonicity-based decomposition guarantees the nonnegativity of the components and quantifies calibration in a sense that is stronger than for other types of decompositions, subject to the nondegeneracy of empirical decompositions. We illustrate the usage of the isotonicity-based decomposition and miscalibration–discrimination (MCB–DSC) plots in case studies from weather prediction and machine learning.

MSC2020 subject classifications: Primary 62G08.
Keywords and phrases: Continuous ranked probability score, score decomposition, isotonic distributional regression.

Received December 2023.

Contents

arXiv: 2311.14122
*These two authors contributed equally to this work.

1. Introduction

Probabilistic predictions are forecasts in the form of predictive probability distributions, which ought to be as sharp as possible subject to calibration (Gneiting, Balabdaoui and Raftery, 2007). Informally, predictive distributions are calibrated if they provide a statistically coherent explanation of the outcomes. Sharpness, on the other hand, quantifies how well one can discriminate different scenarios for future events according to the forecast and is a property of the forecast only. For the comparative evaluation of probabilistic forecasts, proper scoring rules should be employed (Gneiting and Raftery, 2007). A proper scoring rule assigns a numerical score to a probabilistic forecast with corresponding observed realization, and addresses calibration and sharpness simultaneously. If we compare two competing forecasts according to their scores, it is natural to

ask in which aspect one forecast is superior to the other. This motivates the decomposition of average realized scores into more interpretable terms measuring calibration, discrimination ability, and uncertainty, respectively.

Historically, the first score decomposition was introduced by Murphy (1973), who proposed a decomposition of the mean Brier score (BS). For a sequence of forecast–observation pairs $(p_1, y_1), \ldots, (p_n, y_n)$, consisting of predictive probabilities $p_i \in [0, 1]$ and corresponding binary outcomes $y_i \in \{0, 1\}$, the empirical average Brier score

$$\overline{\mathrm{BS}} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2$$

quantifies the overall performance of the assessed forecasts based on the actual observations. Murphy (1973) motivates a decomposition of $\overline{\mathrm{BS}}$ into interpretable components: a term measuring miscalibration (MCB) or reliability, a term measuring discrimination ability (DSC) or resolution, and a term quantifying the overall uncertainty (UNC) of the outcome. Originally derived as a vector partition by Murphy (1973), Siegert (2017) gives a persuasive interpretation of the Murphy decomposition: Suppose that the forecasts $p_1, \ldots, p_n$ attain a small number m of values only ($m < n$). For $i = 1, \ldots, n$, consider the conditional event probability q_i, i.e., the proportion of realized binary events when the forecast value was p_i. Denote by $\overline{\mathrm{BS}}_c$ the empirical Brier score of the calibrated forecasts $q_1, \ldots, q_n$, and by $\overline{\mathrm{BS}}_r$ the empirical Brier score with respect to the static reference forecast $r = (1/n) \sum_{i=1}^{n} y_i$, namely,

$$\overline{\mathrm{BS}}_c = \frac{1}{n} \sum_{i=1}^{n} (q_i - y_i)^2 \quad \text{and} \quad \overline{\mathrm{BS}}_r = \frac{1}{n} \sum_{i=1}^{n} (r - y_i)^2 . \tag{1}$$

Siegert (2017) shows that the Murphy decomposition reads as

$$\overline{\mathrm{BS}} = \underbrace{\left(\overline{\mathrm{BS}} - \overline{\mathrm{BS}}_c\right)}_{\overline{\mathrm{MCB}}} - \underbrace{\left(\overline{\mathrm{BS}}_r - \overline{\mathrm{BS}}_c\right)}_{\overline{\mathrm{DSC}}} + \underbrace{\overline{\mathrm{BS}}_r}_{\overline{\mathrm{UNC}}} . \tag{2}$$

The terms $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$ of this exact decomposition reveal deeper insight into the performance of the assessed forecasts: The predictive probabilities are calibrated if they are close to their conditional event probabilities, and hence, low values of $\overline{\mathrm{MCB}}$ indicate a good performance in terms of calibration. A perfectly calibrated forecast sequence can be constructed by issuing the marginal probability r over all instances. Even though perfectly calibrated, such a sequence would not be informative, since the same predictive probability is issued throughout. For such a sequence, we would obtain $\overline{\mathrm{DSC}} = 0$, which has a negative effect on the score, whereas larger values of $\overline{\mathrm{DSC}}$ are obtained if the calibrated forecasts can discriminate different scenarios better than the reference forecast. Finally, the $\overline{\mathrm{UNC}}$ component informs about the inherent difficulty of the prediction problem and is independent of the forecasts.

The rationale behind the decomposition in (2) can be summarized as the following recipe: Having available a calibration method that transforms the original

forecasts $p_1, \ldots, p_n$ into calibrated forecasts $q_1, \ldots, q_n$, and a reference forecast that depends on the empirical distribution of the outcomes only, one can measure miscalibration as the difference in the mean score of the original forecasts to the calibrated ones, resulting in the $\overline{\text{MCB}}$ term, discrimination ability as the difference in the mean score of the calibrated forecasts to the reference forecast, resulting in the $\overline{\text{DSC}}$ term, and uncertainty as the mean score with respect to the reference forecast, resulting in the $\overline{\text{UNC}}$ term. The CORP (Consistent, Optimally binned, Reproducible, and PAV algorithm based) score decomposition suggested by Dimitriadis, Gneiting and Jordan (2021) uses this general recipe, with the specific innovation that the calibrated forecasts $q_1, \ldots, q_n$ are computed by applying nonparametric isotonic regression on the vector $(y_1, \ldots, y_n)$ with respect to the order induced by $(p_1, \ldots, p_n)$. The CORP approach enforces isotonicity between the original forecasts and the calibrated ones, which "is natural, as decreasing estimates are counterintuitive, routinely being dismissed as artifacts by practitioners" (Dimitriadis, Gneiting and Jordan, 2021, p. 4). If we consider, e.g., the calibrated probability over all events where we predicted a positive outcome with probability 0.5, then we should expect this value to be smaller than the calibrated probability over all events where we predicted a positive outcome with probability 0.6. As hinted at by Bentzien and Friederichs (2014), Siegert (2017), Leutbecher and Haiden (2021), and Gneiting, Lerch and Schulz (2023), and discussed in detail by Gneiting and Resin (2023), the ideas of the Murphy decomposition and enforced isotonicity extend to scores other than the Brier score and general types of statistical functionals.

In this paper, we focus on the continuous ranked probability score (crps; Matheson and Winkler, 1976). The crps is one of the most prominent scoring rules for the evaluation of probabilistic forecasts for real-valued outcomes and is popular across application areas and methodological communities; see, e.g., Gneiting et al. (2005), Hothorn, Kneib and Bühlmann (2014), Pappenberger et al. (2015), Rasp and Lerch (2018), and Gasthaus et al. (2019). The crps is defined in terms of any cumulative distribution function (cdf) F on $\mathbb{R}$ and $y \in \mathbb{R}$, and given by

$$\text{crps}(F, y) = \int_{\mathbb{R}} \big(F(z) - \mathbb{1}\{y \le z\}\big)^2 \, \mathrm{d}z.$$

For a sequence of forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, comprising a predictive distribution F_i and a corresponding real-valued outcome y_i, the mean crps,

$$\overline{\text{CRPS}} = \frac{1}{n} \sum_{i=1}^{n} \text{crps}(F_i, y_i) \tag{3}$$

serves to quantify the overall performance of the forecasts. Possible decompositions of the mean score at (3) have been discussed in the literature, with the most prominent approaches being introduced by Hersbach (2000) and Candille and Talagrand (2005). These methods offer promising solutions but come with severe limitations, which we discuss in detail in Section 2. In a nutshell, the

Hersbach decomposition lacks a theoretical background and the desirable property that the components of the decomposition are nonnegative, whereas the decomposition of Candille and Talagrand (2005) is not practically feasible, as acknowledged in their article. Another approach for decomposing the mean crps is by exploiting its representation as an integral over Brier scores, compare (6), and then integrating existing decompositions of $\overline{\text{BS}}$. Similarly, the crps can be expressed as an integral over quantile scores, see (7), and existing decompositions for quantile scores can be leveraged to decompose the mean score at (3). However, these approaches have the drawback that miscalibration and discrimination ability are not measured with respect to the full probabilistic forecasts but only with respect to individual threshold or quantile levels.

In this article, we propose a new decomposition of the mean crps based on Isotonic Distributional Regression (IDR; Henzi, Ziegel and Gneiting, 2021). In the case of binary outcomes, Dimitriadis, Gneiting and Jordan (2021) argue that isotonicity between the original and the calibrated probabilities is a natural constraint, as it implies the preservation of the order structure in the transition from the original to the calibrated forecasts. This argument generalizes to the real-valued setting, since it is natural to assume that the conditional law of the outcome, given the forecast, should tend to be small (large) if the predictive distribution is small (large), where notions of small and large are understood with respect to the usual stochastic order. IDR is a nonparametric distributional regression technique that honors the shape constraint of isotonicity. Applying IDR to the data $(F_1, y_1), \ldots, (F_n, y_n)$ yields calibrated forecasts, whereas the marginal distribution of the outcomes $y_1, \ldots, y_n$ serves as static reference forecast. The general recipe of the Murphy decomposition at (1) and (2) then yields mean scores for the calibrated forecast and the reference forecast, respectively, and a corresponding exact decomposition,

$$\overline{\text{CRPS}} = \overline{\text{MCB}}_{\text{ISO}} - \overline{\text{DSC}}_{\text{ISO}} + \overline{\text{UNC}}_0,$$

of the mean crps at (3), to which we refer as the isotonicity-based decomposition. The isotonicity-based approach guarantees the nonnegativity of the three components, and the miscalibration term admits a persuasive interpretation in terms of calibration. Importantly, our method yields meaningful decompositions if sufficiently many pairs of the raw forecasts $F_1, \ldots, F_n$ are comparable under the stochastic (partial) order. This condition may be restrictive, and we develop and discuss remedies for this limitation.

While for binary events there is a universal notion of calibration (Gneiting and Ranjan, 2013, Theorem 2.11), for real-valued random outcomes distinct notions of calibration are found in the literature (Dawid, 1984; Diebold, Gunther and Tay, 1998; Strähl and Ziegel, 2017; Arnold, Henzi and Ziegel, 2023), as reviewed by Gneiting and Resin (2023). The strongest notion is auto-calibration and, ideally, one would like to measure miscalibration as deviation from auto-calibration, as targeted by the decomposition of Candille and Talagrand (2005). However, the Candille–Talagrand approach yields degenerate empirical decompositions. Therefore, we quantify miscalibration as the deviation from isotonic

calibration, as introduced by Arnold and Ziegel (2024) in a study of the population version of IDR. Isotonic calibration is closer to auto-calibration than the notions of calibration targeted by the Hersbach decomposition, or by the aforementioned decompositions based on Brier or quantile scores.

We would like to clarify that in context of forecast evaluation, the term 'calibration' is used in at least two distinct, though related, senses. In a first sense the term 'calibration' refers to the statistical consistency between forecasts and outcomes as in the previous paragraph. In a second sense, the very same term 'calibration' – or, perhaps more adequately, 're-calibration' – refers to the process of adjusting (improving) a forecast in such a way that the adjusted (improved) forecast is 'calibrated' in the first sense.

The remainder of the paper is organized as follows. Section 2 reviews the previously proposed decompositions and their properties. In Section 3, we develop the empirical version of the new isotonicity-based decomposition, followed by a thorough study of the population versions of the various types of decomposition and their properties in Section 4, with particular emphasis on calibration. In Section 5, we apply the proposed isotonicity-based decomposition in case studies from meteorology and machine learning, and we propose a succinct graphical way of comparing competing forecast methods in miscalibration–discrimination (MCB–DSC) plots. The main part of the paper closes with a discussion in Section 6. Proofs, technical comments, and a series of detailed analytic examples in population settings are available in Appendices A through F.

Replication materials and code in R (R Core Team, 2023) for the computation of the isotonicity-based decomposition and MCB–DSC plots are publicly available at repositories (Walz, 2022, 2023).

2. Previously proposed empirical decompositions

2.1. Preliminaries

Throughout the article, we denote by $\mathcal{P}(\mathbb{R})$ the class of all probability distributions on $\mathbb{R}$ with finite first moment. We treat its elements interchangeably as probability measures or cumulative distribution functions (cdfs).

Single-valued forecasts for functionals of an unknown quantity should be compared using consistent scoring functions (Gneiting, 2011). For example, the *quadratic score* $\mathrm{s}(x, y) = (x - y)^2$, and the piecewise linear *quantile score*

$$\mathrm{qs}_\alpha(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)\,(x - y), \tag{4}$$

where $x, y \in \mathbb{R}$, are consistent scoring functions for the mean functional, and for the quantile at level $\alpha \in (0, 1)$, respectively. In other words, $\int (x - y)^2 \,\mathrm{d}F(y)$ is minimal when x is the mean of $F \in \mathcal{P}(\mathbb{R})$, and $\int \mathrm{qs}_\alpha(x, y) \,\mathrm{d}F(y)$ is minimal when x is a quantile of F at level $\alpha \in (0, 1)$.

Probabilistic forecasts specify a probability measure over all possible values of the outcome, and predictive performance ought to be be compared and evaluated using proper scoring rules (Gneiting and Raftery, 2007). A popular proper

scoring rule for probability forecasts of a binary outcome is the *Brier score*

$$\mathrm{s_B}(p, y) = (p - y)^2, \tag{5}$$

where $p \in [0, 1]$ and $1 - p$ are the predicted probabilities of the outcomes $y = 1$ and $y = 0$, respectively. A key example of a proper scoring rule for predictive distributions over $\mathbb{R}$ is the *continuous ranked probability score* (crps), defined for all $F \in \mathcal{P}(\mathbb{R})$ and $y \in \mathbb{R}$, and given equivalently by

$$\mathrm{crps}(F, y) = \int \mathrm{s_B}(F(z), \mathbb{1}\{y \le z\})\, \mathrm{d}z \tag{6}$$

$$= \int_0^1 \mathrm{qs}_\alpha(F^{-1}(\alpha), y)\, \mathrm{d}\alpha, \tag{7}$$

where $\mathrm{s_B}$ and qs_α are defined at (5) and (4), respectively, and where F^{-1} denotes the quantile function defined as $F^{-1}(\alpha) = \inf\{z \in \mathbb{R} \mid F(z) \ge \alpha\}$ for $\alpha \in (0, 1)$. The representation at (7) is due to Laio and Tamea (2007).

We consider a collection

$$(F_1, y_1), \ldots, (F_n, y_n) \tag{8}$$

of tuples that comprise a forecast $F_i \in \mathcal{P}(\mathbb{R})$ in the form of a cdf and the respective outcome $y_i \in \mathbb{R}$, where $i = 1, \ldots, n$. Our aim is to decompose the empirical mean score,

$$\overline{\mathrm{CRPS}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{crps}(F_i, y_i), \tag{9}$$

of the forecast–observation pairs at (8) into three distinct components, namely, miscalibration ($\overline{\mathrm{MCB}}$), discrimination ($\overline{\mathrm{DSC}}$), and uncertainty ($\overline{\mathrm{UNC}}$). The following desirable properties are relevant.

(E_1) The decomposition is exact, i.e.,

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}} - \overline{\mathrm{DSC}} + \overline{\mathrm{UNC}}.$$

(E_2) The components $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$ are nonnegative.
(E_3) The decomposition is not degenerate. Here, a decomposition is *degenerate* if $\overline{\mathrm{MCB}} = \overline{\mathrm{CRPS}}$ whenever $F_1, \ldots, F_n$ are pairwise distinct.
(E_4) The $\overline{\mathrm{DSC}}$ component vanishes if $F_1 = \cdots = F_n$.
(E_5) The $\overline{\mathrm{UNC}}$ component can be expressed in terms of the empirical distribution of the outcomes $y_1, \ldots, y_n$.

These conditions do not depend on the use of any specific scoring rule; they are desirable for decompositions of mean scores in general. An exact decomposition (E_1) is desirable, since it allows us to fully decompose the mean score. A degenerate decomposition is undesirable, as in typical practice, such as in the case studies in Section 5, the issued forecast distributions are pairwise distinct,

and then the method is useless. A static forecast, i.e., $F_1 = \cdots = F_n$, has no discrimination ability, hence E_4 is desirable. Requirement E_5 is natural since intrinsic uncertainty does not depend on the activities of forecasters.

Finally, we argue that there ought to be a population version of the decomposition that applies to any admissible joint distribution $\mathbb{P}$ of tuples (F, Y). Furthermore, the population version ought to reduce to the empirical version if $\mathbb{P}$ is the empirical measure for the data at (8). We study decompositions at the population level in Section 4.

In what follows, let δ_y denote the Dirac or point measure in $y \in \mathbb{R}$, and let the marginal law $\hat{F}_{\text{mg}} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ denote the empirical distribution of the outcomes $y_1, \ldots, y_n$. All decompositions studied in the following quantify uncertainty via the mean score of the static forecast $\hat{F}_{\text{mg}}$, and for ease of reference we define

$$\overline{\text{UNC}}_0 = \overline{\text{CRPS}}_{\text{mg}} = \frac{1}{n}\sum_{i=1}^n \text{crps}(\hat{F}_{\text{mg}}, y_i). \tag{10}$$

2.2. Candille–Talagrand decomposition

Candille and Talagrand (2005) naturally extend the idea of the Murphy decomposition at (2). To describe their approach, let $\hat{F}_i$ be the auto-calibrated version of the forecast F_i in (8), i.e., let $\hat{F}_i$ be the normalized version of $\sum_{j=1}^n \mathbb{1}\{F_j = F_i\}\,\delta_{y_j}$ for $i = 1, \ldots, n$, and let

$$\overline{\text{CRPS}}_{\text{ac}} = \frac{1}{n}\sum_{i=1}^n \text{crps}(\hat{F}_i, y_i)$$

be the mean score of the auto-calibrated forecast. Candille and Talagrand (2005) define miscalibration and discrimination components as

$$\overline{\text{MCB}}_{\text{CT}} = \overline{\text{CRPS}} - \overline{\text{CRPS}}_{\text{ac}} \quad \text{and} \quad \overline{\text{DSC}}_{\text{CT}} = \overline{\text{CRPS}}_{\text{mg}} - \overline{\text{CRPS}}_{\text{ac}}, \tag{11}$$

respectively, to yield the *Candille–Talagrand* (CT) *decomposition*

$$\overline{\text{CRPS}} = \overline{\text{MCB}}_{\text{CT}} - \overline{\text{DSC}}_{\text{CT}} + \overline{\text{UNC}}_0. \tag{12}$$

The Candille–Talagrand decomposition tackles the core idea of auto-calibration and satisfies properties E_1, E_2, E_4, and E_5, but fails to satisfy the nondegeneracy condition E_3, which prohibits its practical use.

To avoid a degenerate decomposition, one might partition the forecasts into equivalence classes of cdfs that are considered identical when calibrating (Candille and Talagrand, 2005, p. 2147). However, the choice of such a partition is challenging and the decomposition depends on its effects, akin to the effects of binning on the Murphy decomposition and the classical reliability diagram for probability forecasts of a binary event as described by Dimitriadis, Gneiting and Jordan (2021) and references therein.

2.3. Brier score based decomposition

The Brier score based representation of individual crps values at (6) implies that

$$\overline{\mathrm{CRPS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{crps}(F_i, y_i) = \int_{-\infty}^{\infty} \overline{\mathrm{BS}}_z \, \mathrm{d}z, \tag{13}$$

where

$$\overline{\mathrm{BS}}_z = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_{\mathrm{B}}(F_i(z), \mathbb{1}\{y_i \le z\}).$$

In this light, a natural way of decomposing $\overline{\mathrm{CRPS}}$ lies in integrating a given decomposition of the mean Brier score, as proposed and implemented by Ferro and Fricker (2012), Tödter and Ahrens (2012), and Lauret, David and Pinson (2019), among other authors.

Specifically, suppose that, for each $z \in \mathbb{R}$, there is a decomposition $\overline{\mathrm{BS}}_z = \overline{\mathrm{MCB}}_{\mathrm{BS},z} - \overline{\mathrm{DSC}}_{\mathrm{BS},z} + \overline{\mathrm{UNC}}_{\mathrm{BS},z}$ of the mean Brier score. Then we can define

$$\overline{\mathrm{MCB}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{MCB}}_{\mathrm{BS},z} \, \mathrm{d}z, \quad \overline{\mathrm{DSC}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{DSC}}_{\mathrm{BS},z} \, \mathrm{d}z, \tag{14}$$

and

$$\overline{\mathrm{UNC}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{UNC}}_{\mathrm{BS},z} \, \mathrm{d}z. \tag{15}$$

The CORP approach of Dimitriadis, Gneiting and Jordan (2021) yields a compelling decomposition of the mean Brier score, which does neither require tuning, nor binning of the assessed predictive probabilities, and enforces the natural shape constraint of isotonicity between the predictive probabilities and the calibrated forecasts. Throughout this article, we decompose the mean Brier score by the CORP approach and refer to the induced decomposition, namely,

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{DSC}}_{\mathrm{BS}} + \overline{\mathrm{UNC}}_{\mathrm{BS}}, \tag{16}$$

as the *Brier score based* (BS) decomposition of $\overline{\mathrm{CRPS}}$. Details of this approach are reviewed in Appendix A.1, where we prove the following result.

Proposition 2.1. *For the Brier score based decomposition at* (16) *it holds that* $\overline{\mathrm{UNC}}_{\mathrm{BS}} = \overline{\mathrm{UNC}}_0$*, and the decomposition satisfies properties* E_1*,* E_2*,* E_3*,* E_4*, and* E_5*.*

Despite these favorable properties, the Brier score based decomposition is subject to shortcomings and inconsistencies, due to the isolated treatment of probability forecasts at fixed thresholds. For discussion, we refer the reader to Section 2.6 and Appendix A.

2.4. Quantile score based decomposition

In view of the quantile score representation of the crps at (7), a natural approach to decomposing the mean score $\overline{\text{CRPS}}$ leverages decompositions of the mean quantile score at (4). Specifically, the quantile score representation implies that

$$\overline{\text{CRPS}} = \frac{1}{n}\sum_{i=1}^{n} \text{crps}(F_i, y_i) = \int_0^1 \overline{\text{QS}}_\alpha \, \text{d}\alpha,$$

where

$$\overline{\text{QS}}_\alpha = \frac{1}{n}\sum_{i=1}^{n} \text{qs}_\alpha(F_i^{-1}(\alpha), y_i).$$

Suppose that for each $\alpha \in (0,1)$, there is a decomposition $\overline{\text{QS}}_\alpha = \overline{\text{MCB}}_{\text{QS},\alpha} - \overline{\text{DSC}}_{\text{QS},\alpha} + \overline{\text{UNC}}_{\text{QS},\alpha}$ of the mean quantile score, and define $\overline{\text{MCB}}_{\text{QS}}$ as the integral of $\overline{\text{MCB}}_{\text{QS},\alpha}$ over $\alpha \in (0,1)$, and similarly for the discrimination and uncertainty components. The CORP score decomposition of Dimitriadis, Gneiting and Jordan (2021) and its core idea of isotonicity as a shape constraint between issued and calibrated forecasts extend naturally to quantiles, as discussed by Gneiting and Resin (2023, Section 3.3) and Gneiting et al. (2023, Section 3.3). Throughout the article, we decompose the mean quantile score by the CORP approach and refer to the resulting decomposition, namely,

$$\overline{\text{CRPS}} = \overline{\text{MCB}}_{\text{QS}} - \overline{\text{DSC}}_{\text{QS}} + \overline{\text{UNC}}_{\text{QS}}, \tag{17}$$

as the *quantile score based* (QS) decomposition of $\overline{\text{CRPS}}$. For details, we refer the reader to Appendix A.2 where we prove the following result.

Proposition 2.2. *For the quantile score based decomposition at* (17) *it holds that* $\overline{\text{UNC}}_{\text{QS}} = \overline{\text{UNC}}_0$*, and the decomposition satisfies properties* E_1*,* E_2*,* E_3*,* E_4*, and* E_5*.*

The quantile score based decomposition is subject to shortcomings in analogy to the issues with the Brier score based approach, due to the reliance on quantile forecasts at fixed levels; for further discussion see Section 2.6 and Appendix A.

2.5. Hersbach decomposition

The decomposition of Hersbach (2000) applies specifically to ensemble forecasts and operates under the implicit assumption of a continuous outcome. For the data at (8), assume that, for $i = 1, \ldots, n$, the forecast F_i is the empirical cdf of a fixed number m of values $x_1^i \le \cdots \le x_m^i$, where we allow for any real-valued outcome y_i, in contrast to Hersbach (2000), who assumes that $y_i \notin \{x_1^i, \ldots, x_m^i\}$. Figure 5 in Appendix B illustrates in detail how the case $y_i \in \{x_1^i, \ldots, x_m^i\}$ should be handled in the Hersbach decomposition.

The miscalibration component, which Hersbach (2000) refers to as reliability, is

$$\overline{\mathrm{MCB}}_{\mathrm{HBo}} = \sum_{\ell=0}^{m} \bar{g}_\ell \left(p_\ell - \bar{o}_\ell\right)^2,$$

where $p_\ell = \ell/m$ for $\ell = 0, \ldots, m$, and $\bar{g}_\ell$ is the average distance from x_ℓ^i to $x_{\ell+1}^i$, i.e.,

$$\bar{g}_\ell = \frac{1}{n} \sum_{i=1}^{n} (x_{\ell+1}^i - x_\ell^i) \tag{18}$$

for $\ell = 1, \ldots, m-1$. The term $\bar{o}_\ell$ can be interpreted as an approximate version of the average frequency of an outcome below the midpoint of the interval from x_ℓ^i to $x_{\ell+1}^i$; specifically, $\bar{o}_\ell = \bar{f}_\ell - \bar{m}_\ell$, where

$$\bar{f}_\ell = \frac{1}{n\bar{g}_\ell} \sum_{i=1}^{n} \mathbb{1}\{F_i(y_i) \le p_\ell\} \left(x_{\ell+1}^i - x_\ell^i\right) \tag{19}$$

and $\bar{m}_\ell = (1/(n\bar{g}_\ell)) \sum_{i=1}^{n} \mathbb{1}\{x_\ell^i < y_i < x_{\ell+1}^i\} \left(y_i - x_\ell^i\right)$ for $\ell = 1, \ldots, m-1$. To complete the specification, we let

$$\bar{o}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i < x_1^i\} \quad \text{and} \quad \bar{o}_m = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_m^i < y_i\},$$

and if these quantities are nonzero then we let

$$\bar{g}_0 = \frac{1}{n\bar{o}_0} \sum_{i=1}^{n} \mathbb{1}\{y_i < x_1^i\} \left(x_1^i - y_i\right) \quad \text{and} \quad \bar{g}_m = \frac{1}{n\bar{o}_m} \sum_{i=1}^{n} \mathbb{1}\{x_m^i < y_i\} \left(y_i - x_m^i\right),$$

respectively. The miscalibration component thus measures deviations from uniformity for the rank histogram of an ensemble forecast (Hamill, 2001; Gneiting, Balabdaoui and Raftery, 2007).

Hersbach (2000) defines the resolution (in our terminology, the discrimination) component $\overline{\mathrm{DSC}}_{\mathrm{HBo}} = \overline{\mathrm{MCB}}_{\mathrm{HBo}} + \overline{\mathrm{UNC}}_0 - \overline{\mathrm{CRPS}}$ as the remainder, to complete the *original Hersbach* (HBo) *decomposition*

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{HBo}} - \overline{\mathrm{DSC}}_{\mathrm{HBo}} + \overline{\mathrm{UNC}}_0. \tag{20}$$

For reasons that will become evident in Section 4.4, we introduce a slightly modified miscalibration component,

$$\overline{\mathrm{MCB}}_{\mathrm{HB}} = \sum_{\ell=1}^{m-1} \bar{g}_\ell \left(p_\ell - \bar{f}_\ell\right)^2, \tag{21}$$

and a respectively modified discrimination component, $\overline{\mathrm{DSC}}_{\mathrm{HB}} = \overline{\mathrm{MCB}}_{\mathrm{HB}} + \overline{\mathrm{UNC}}_0 - \overline{\mathrm{CRPS}}$, to yield the *modified Hersbach*, or simply *Hersbach* (HB) *decomposition*,

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{HB}} - \overline{\mathrm{DSC}}_{\mathrm{HB}} + \overline{\mathrm{UNC}}_0. \tag{22}$$

The interpretation of the miscalibration component remains unchanged, as the $\overline{\text{MCB}}_{\text{HB}}$ and $\overline{\text{MCB}}_{\text{HBo}}$ terms differ only slightly, with $\bar{f}_\ell$ in (21) being the approximate frequency (19) of an outcome y_i less than or equal to $x^i_{\ell+1}$. For a more detailed comparison and the proof of the following result, we refer the reader to Appendix B.

Proposition 2.3. *The original and modified Hersbach decompositions at* (20) *and* (22)*, respectively, satisfy properties* E_1*,* E_3*, and* E_5*, while properties* E_2 *and* E_4 *fail to hold.*

As discussed thus far, the Hersbach decomposition requires that the forecasts assume the form of an ensemble. Further shortcomings have been discussed in the literature (Siegert, 2017); in particular, it has been noted that the discrimination component $\overline{\text{DSC}}_{\text{HBo}}$ is defined "somewhat artificially" (Hersbach, 2000, p. 565) and that it can be negative, thus violating property E_2. The original Hersbach decomposition has been extended by Lalaurette so that it applies to forecasts with strictly increasing cdfs (Candille and Talagrand, 2005, Appendix A). We discuss and generalize Lalaurette's extension in Section 4.4, and our analysis demonstrates that the extensions can more naturally be interpreted as extensions of the modified Hersbach decomposition at (22). In Appendix D we describe empirical versions that apply in the general case of forecast distributions with finite support, and to mixed discrete-continuous distributions for nonnegative quantities, respectively.

2.6. Comparison and discussion

For an initial comparison of the different decompositions, we consider forecasts from the case studies in Section 5. The decompositions from Sections 2.2 through 2.5 all use the uncertainty component $\overline{\text{UNC}}_0$ at (10), and they specify the discrimination component as

$$\overline{\text{DSC}}_\bullet = \overline{\text{CRPS}} - \overline{\text{MCB}}_\bullet - \overline{\text{UNC}}_0,$$

where $\bullet$ indicates the type of decomposition, namely, the Candille–Talagrand (CT), the Brier score based (BS), the quantile score based (QS), or the modified Hersbach (HB) decomposition.

Table 1 displays the mean score $\overline{\text{CRPS}}$, the uncertainty component $\overline{\text{UNC}}_0$, and the various $\overline{\text{MCB}}_\bullet$ terms for the ENS forecast of precipitation accumulation at Frankfurt Airport, as studied in our Section 5.1 and Henzi, Ziegel and Gneiting (2021), and the EasyUQ forecasts for the Boston Housing and Wine data, as considered in our Section 5.2 and Walz et al. (2024). The ENS forecast is an ensemble forecast with $m = 52$ members and so the Hersbach decomposition at (20) applies; for the EasyUQ forecasts, we apply the generalized formula (46) from Appendix D. For the first two examples in the table, it holds that $\overline{\text{CRPS}} = \overline{\text{MCB}}_{\text{CT}} > \overline{\text{MCB}}_{\text{QS}} > \overline{\text{MCB}}_{\text{BS}} > \overline{\text{MCB}}_{\text{HB}}$, where the initial equality reflects the degeneracy of the Candille–Talagrand decomposition. In our experience, this chain of relations holds in many, though not all, empirical

TABLE 1
Candille–Talagrand (CT), quantile score based (QS), Brier score based (BS), and Hersbach (HB) decomposition of the mean score $\overline{\text{CRPS}}$*, as applied to the one-day ahead raw ensemble (ENS) forecast of precipitation accumulation at Frankfurt Airport (Section 5.1), and the EasyUQ forecast for the Boston and Wine data, respectively (Section 5.2).*

Forecast	$\overline{\text{CRPS}}$	$\overline{\text{MCB}}_{\text{CT}}$	$\overline{\text{MCB}}_{\text{QS}}$	$\overline{\text{MCB}}_{\text{BS}}$	$\overline{\text{MCB}}_{\text{HB}}$	$\overline{\text{UNC}}_0$
ENS	0.75	0.75	0.18	0.16	0.08	1.21
EasyUQ (Boston)	1.75	1.75	0.72	0.57	0.36	4.76
EasyUQ (Wine)	0.35	0.35	0.04	0.07	0.08	0.43

examples. However, as we state in further generality at (26) and in Corollary 4.6, it always holds that $\overline{\text{CRPS}} \geq \overline{\text{MCB}}_{\text{CT}} \geq \max\{\overline{\text{MCB}}_{\text{BS}}, \overline{\text{MCB}}_{\text{QS}}\}$.

While the Candille–Talagrand decomposition seems attractive and preferable from theoretical perspectives, the degeneracy prohibits its practical use. The Hersbach decomposition has been popular in the specific setting of ensemble forecasts, but has serious shortcomings including but not limited to the possibility of a negative discrimination component. The Brier score and quantile score based decompositions have desirable properties, but they define the components of the decomposition in terms of isolated functionals (probabilities and quantiles, respectively) rather than the entire predictive distributions, which is "unsatisfactory" (Ferro and Fricker, 2012, p. 1958) and entails the artifacts described in Remarks A.1 and A.2, respectively. Furthermore, it is not obvious whether the Brier score based or the quantile score based decomposition ought to be preferred. In this light, there remains the need for a decomposition that is both practically feasible and theoretically justifiable and appealing.

3. Empirical isotonicity-based decomposition

We propose a method that builds on the idea of the Candille–Talagrand decomposition, but replaces auto-calibration with a slightly weaker notion of calibration, namely, isotonic calibration. The resulting isotonicity-based decomposition, which we develop in this section, can be interpreted as a nondegenerate approximation to the Candille–Talagrand decomposition.

3.1. Empirical isotonicity-based decomposition

Recall that we denote by $\mathcal{P}(\mathbb{R})$ the class of the probability distributions on $\mathbb{R}$ with finite first moment. For cdfs F, G, F is stochastically smaller than or equal to G, for short $F \leq_{\text{st}} G$, if $F(x) \geq G(x)$ for all $x \in \mathbb{R}$. The stochastic order defines a partial order on $\mathcal{P}(\mathbb{R})$ and we refer to Müller and Stoyan (2002) and Shaked and Shanthikumar (2007) for comprehensive studies.

In the spirit of the Candille–Talagrand decomposition, a calibration tool ought to be applied to the assessed forecasts $F_1, \ldots, F_n$ from (8), and we pro-

pose that this tool be isotonic distributional regression (IDR; Henzi, Ziegel and Gneiting, 2021). IDR is a nonparametric distributional regression method under the shape constraint of isotonicity between covariates and responses: For training data consisting of covariates $x_1, \dots, x_n$ in a partially ordered set $(\mathcal{X}, \preceq)$ and real-valued responses $y_1, \dots, y_n$, Henzi, Ziegel and Gneiting (2021) prove that there exists a unique minimizer of the criterion

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{crps}(P_i, y_i) \tag{23}$$

in the space of the n-tuples $(P_1, \dots, P_n)$ of cdfs that satisfy $P_i \leq_{\mathrm{st}} P_j$ if $x_i \preceq x_j$ and $P_i = P_j$ if $x_i = x_j$ for $i, j = 1, \dots, n$, and we refer to this minimizer as the IDR solution.

The constraint of isotonicity between the assessed and the calibrated forecasts is natural, and hence, we apply IDR to the data $(F_1, y_1), \dots, (F_n, y_n)$ at (8), where the covariates are $F_1, \dots, F_n$ and the partial order on the covariate space is the stochastic order. Let $\check{F}_1, \dots, \check{F}_n$ denote the calibrated forecasts that are obtained by using IDR, let

$$\overline{\mathrm{CRPS}}_{\mathrm{ISO}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{crps}(\check{F}_i, y_i) \tag{24}$$

denote the mean score of the calibrated forecasts, let the marginal forecast $\hat{F}_{\mathrm{mg}}$ and its mean score $\overline{\mathrm{UNC}}_0 = \overline{\mathrm{CRPS}}_{\mathrm{mg}}$ be defined as at (10), and let

$$\overline{\mathrm{MCB}}_{\mathrm{ISO}} = \overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}_{\mathrm{ISO}} \quad \text{and} \quad \overline{\mathrm{DSC}}_{\mathrm{ISO}} = \overline{\mathrm{CRPS}}_{\mathrm{mg}} - \overline{\mathrm{CRPS}}_{\mathrm{ISO}}.$$

Then the *isotonicity-based* (ISO) decomposition

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{ISO}} - \overline{\mathrm{DSC}}_{\mathrm{ISO}} + \overline{\mathrm{UNC}}_0 \tag{25}$$

differs from the Candille–Talagrand decomposition at (11) by the choice of the calibration method only, as it draws on the slightly weaker notion of isotonic calibration in lieu of auto-calibration. The isotonicity-based decomposition has desirable and appealing properties, as follows.

Proposition 3.1. *The isotonicity-based decomposition at* (25) *satisfies* E_1, E_2, E_3, E_4, *and* E_5. *Furthermore,* $\overline{\mathrm{MCB}}_{\mathrm{ISO}} = 0$ *if, and only if,* $F_i = \check{F}_i$ *for* $i = 1, \dots, n$, *and* $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$ *if, and only if,* $\check{F}_i = \hat{F}_{\mathrm{mg}}$ *for* $i = 1, \dots, n$.

Proof. By definition, the isotonicity-based decomposition satisfies properties E_1 and E_5. The IDR solution is the unique minimizer of the criterion (23) over all n-tuples of distributions $(P_1, \dots, P_n)$ that are stochastically ordered with the same order relations as the covariates. Therefore, $(F_1, \dots, F_n)$ is an admissible n-tuple of distributions in the minimization problem, whence $\overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq 0$. A further admissible n-tuple in the minimization problem is the constant n-tuple with entries $\hat{F}_{\mathrm{mg}}$, whence $\overline{\mathrm{DSC}}_{\mathrm{ISO}} \geq 0$, so E_2 is satisfied. The examples in the case study in Section 5 imply that the isotonicity-based decomposition satisfies

E_3. Assume now that $F_1 = \cdots = F_n$. Then we obtain $\hat{F}_{\mathrm{mg}}$ as the IDR solution, whence $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$, so E_4 is satisfied. Finally, if $\overline{\mathrm{MCB}}_{\mathrm{ISO}} = 0$ then $F_i = \check{F}_i$, since IDR is the unique minimizer of the criterion at (23), and analogously, if $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$ then $\check{F}_i = \hat{F}_{\mathrm{mg}}$ for $i = 1, \ldots, n$. □

In the pure form presented thus far, the isotonicity-based decomposition is fully automated in the sense that it does not involve any tuning parameter. For the examples in Table 1, $\overline{\mathrm{MCB}}_{\mathrm{ISO}}$ equals 0.34, 0.80, and 0.072, respectively, and so $\overline{\mathrm{MCB}}_{\mathrm{ISO}}$ is larger than $\overline{\mathrm{MCB}}_{\mathrm{BS}}$ (which equals 0.068 in the third example) and $\overline{\mathrm{MCB}}_{\mathrm{QS}}$ and smaller than the essentially useless $\overline{\mathrm{MCB}}_{\mathrm{CT}} = \overline{\mathrm{CRPS}}$ term. As we demonstrate in Section 4.5, it is always true that

$$\overline{\mathrm{CRPS}} \geq \overline{\mathrm{MCB}}_{\mathrm{CT}} \geq \overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq \max\{\overline{\mathrm{MCB}}_{\mathrm{BS}}, \overline{\mathrm{MCB}}_{\mathrm{QS}}\}. \tag{26}$$

In view of these theoretical guarantees in concert with its non-degeneracy and generality, we contend that the isotonicity-based method is more compelling than the Brier score or quantile score based decompositions.

3.2. Computational implementation

When the predictive distributions are empirical distributions, stochastic order relations can be found by comparing the cdfs at a finite number of real numbers, namely, the respective jump points. If the predictive distributions are parametric, analytical results in terms of the parameters may be available; see, e.g., Shaked and Shanthikumar (2007) and the proof of Proposition 1 in Gneiting and Vogel (2022).

In relevant applications, the stochastic order in its pure form might be too strong for our purposes, since it does not allow for crossings of the forecast cdfs. For example, for Gaussian forecasts $F = \mathcal{N}(\mu, \sigma^2)$ and $G = \mathcal{N}(\nu, \tau^2)$, F and G only order with respect to the stochastic order in case of $\sigma = \tau$, a condition which is rarely satisfied if parameters are estimated from data.[1] Generally, if F and G are members of a location-scale family, they are stochastically ordered if, and only if, they have equal scale parameter, subject to minimal conditions. If only very few forecasts in the dataset are comparable with respect to the stochastic order, applying IDR results in calibrated forecasts that are close to Dirac measures in the corresponding observations. Hence, in principle, the isotonicity-based decomposition faces the same problem as the Candille–Talagrand decomposition in this setting. However, we argue that there is a convincing remedy to the issue.

Consider settings where only few of the predictive distributions F_i in the collection at (8) are comparable with respect to the stochastic order. Frequently,

[1]For an explicit example, consider the Laplace method in Section 5.2, where the predictive distributions are Gaussian with estimated mean and estimated variance. Therefore, there are no pairs of comparable cdfs in the pure form of the isotonicity-based decomposition. However, if the approximate implementation proposed in this section is applied, the fraction of comparable cdfs rises to values of about 0.90 and higher, depending on the dataset.

predictive distributions fail to order due to crossings of the cdfs in a far tail. Recent work by Brehmer and Strokorb (2019) and Taillardat et al. (2023) casts doubt on the ability of the average crps to distinguish tail behaviour of the forecast distribution, which provides support for the evaluation of the forecasts on a bounded interval only. Motivated by these findings, instead of decomposing the original mean score $\overline{\text{CRPS}}$ as given in (9), we decompose

$$\overline{\text{CRPS}}^{(a,b)} = \frac{1}{n}\sum_{i=1}^{n} \text{crps}(\tilde{F}_i^{(a,b)}, y_i), \tag{27}$$

where for lower and upper threshold values $a \leq \min\{y_1, \ldots, y_n\}$ and $b \geq \max\{y_1, \ldots, y_n\}$, respectively,

$$F_i^{(a,b)}(x) = \begin{cases} 0, & x < a, \\ F_i(x), & x \in [a, b), \\ 1, & x \geq b, \end{cases} \tag{28}$$

for $i = 1, \ldots, n$. Given an error tolerance $\epsilon > 0$, we determine the thresholds a and b such that the condition

$$\left| \overline{\text{CRPS}} - \overline{\text{CRPS}}^{(a,b)} \right| = \overline{\text{CRPS}} - \overline{\text{CRPS}}^{(a,b)} < \epsilon \tag{29}$$

is satisfied, where the equality holds since $\overline{\text{CRPS}} \geq \overline{\text{CRPS}}^{(a,b)}$. Condition (29) is equivalent to

$$I(a,b) = \frac{1}{n}\sum_{i=1}^{n} \left(\int_{-\infty}^{a} F_i(x)^2 \, \mathrm{d}x + \int_{b}^{\infty} (1 - F_i(x))^2 \, \mathrm{d}x \right) < \epsilon.$$

A simple method for determining the thresholds a and b to be used in (28) is described in Algorithm 1. If the support of the predictive distributions is bounded from above or below (e.g., in the case of precipitation accumulations, which are necessarily nonnegative), it is natural to set a or b equal to the respective bound (e.g., $a = 0$ for precipitation accumulations).

Algorithm 1 Determination of thresholds a, b from data $(F_1, y_1), \ldots, (F_n, y_n)$.

$\epsilon = \overline{\text{CRPS}}/1000$
$a = \min\{y_1, \ldots, y_n\}$ and $b = \max\{y_1, \ldots, y_n\}$
if $I(a,b) \geq \epsilon$ **then**
    $\delta = (b - a)/100$
    while $I(a,b) \geq \epsilon$ **do**
        $a = a - \delta$ and $b = b + \delta$
    end while
end if
return a, b

The computation of this modified isotonicity-based decomposition remains of complexity $\mathcal{O}(n^2)$. Furthermore, the following result shows that, even with the approximation, theoretical guarantees from (26) continue to hold.

Proposition 3.2. *Let* $\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{ISO}} - \overline{\mathrm{DSC}}_{\mathrm{ISO}} + \overline{\mathrm{UNC}}_0 = \overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{DSC}}_{\mathrm{BS}} + \overline{\mathrm{UNC}}_0$ *denote decompositions for data* $(F_1, y_1), \ldots, (F_n, y_n)$, *and let*

$$\overline{\mathrm{CRPS}}^{(a,b)} = \overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)} - \overline{\mathrm{DSC}}_{\mathrm{ISO}}^{(a,b)} + \overline{\mathrm{UNC}}_0 = \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)} - \overline{\mathrm{DSC}}_{\mathrm{BS}}^{(a,b)} + \overline{\mathrm{UNC}}_0$$

denote the respective decompositions for modified data $(F_1^{(a,b)}, y_1), \ldots, (F_n^{(a,b)}, y_n)$, *where* $F_1^{(a,b)}, \ldots, F_n^{(a,b)}$ *derive from* $F_1, \ldots, F_n$ *as in* (28). *Then* $I(a,b) = \overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}^{(a,b)} < \epsilon$ *implies that*

$$\overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq \overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)} \geq \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)} > \overline{\mathrm{MCB}}_{\mathrm{BS}} - \epsilon. \tag{30}$$

Proof. The properties of the IDR solution imply $\overline{\mathrm{CRPS}}_{\mathrm{ISO}} \leq \overline{\mathrm{CRPS}}_{\mathrm{ISO}}^{(a,b)} \leq \overline{\mathrm{CRPS}}^{(a,b)} \leq \overline{\mathrm{CRPS}}$, and we conclude that

$$\overline{\mathrm{MCB}}_{\mathrm{ISO}} = \overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}_{\mathrm{ISO}} \geq \overline{\mathrm{CRPS}}^{(a,b)} - \overline{\mathrm{CRPS}}_{\mathrm{ISO}}^{(a,b)} = \overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)}.$$

To complete the proof, we apply the inequality (26) to the modified data to yield $\overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)} \geq \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)}$, and we note that $a \leq \min\{y_1, \ldots, y_n\}$ and $b \geq \max\{y_1, \ldots, y_n\}$, whence $\overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)} = I(a,b) < \epsilon$. □

Assume that the predictive cdfs belong to a location-scale family with full support, i.e., there exists a distribution $F_0 \in \mathcal{P}(\mathbb{R})$ with full support on $\mathbb{R}$ such that for $i = 1, \ldots, n$ and $x \in \mathbb{R}$, $F_i(x) = F_0((x-\mu_i)/\sigma_i)$ for some location $\mu_i \in \mathbb{R}$ and scale $\sigma_i > 0$. Then for any $i, j = 1, \ldots, n$, the stochastic order relations between the modified distributions can be obtained based on the parameters (Gneiting and Vogel, 2022, proof of Proposition 1), in that

$$F_i^{(a,b)} \leq_{\mathrm{st}} F_j^{(a,b)}$$

if, and only if, $\mu_i \leq \mu_j$ and either $\sigma_i = \sigma_j$ or $(\mu_i\sigma_j - \mu_j\sigma_i)/(\sigma_j - \sigma_i) \notin [a,b]$. In more complex but not uncommon situations, e.g., when the predictive distributions are mixtures of Gaussians, it may be hard to decide analytically whether or not there is a stochastic dominance relation between any two such distributions. A remedy is then to numerically evaluate and compare the cdfs on a suitably chosen grid of threshold values. As a default we suggest and use an equidistant grid from a to b of size 5000. As long as the grid is sufficiently dense, order relations hardly ever change with the size of the grid, as experimental experience demonstrates.

In order to increase the number of comparable pairs amongst $F_1, \ldots, F_n$, it may also appear natural to exchange the stochastic order with a weaker partial order on $\mathcal{P}(\mathbb{R})$, rather than restricting the support of the predictive distributions to a bounded interval $[a,b] \subseteq \mathbb{R}$. However, we show in Appendix C that isotonic calibration is generally only compatible with the stochastic order. Therefore, the stochastic order is the only valid choice of a partial order if IDR is applied to generate a calibrated forecast for an isotonicity-based approach in the spirit of the Candille–Talagrand decomposition.

TABLE 2

Properties of the empirical Candille–Talagrand (CT), isotonicity-based (ISO), Brier score based (BS), quantile score based (QS), and Hersbach (HB) decomposition of the mean crps. For properties $E_1, \ldots, E_5$ see Section 2.1. Computational complexity is quantified via a lower bound to the number of floating point operations in terms of the sample size n at (8).

	E_1	E_2	E_3	E_4	E_5	Complexity
CT	✓	✓	✗	✓	✓	$\mathcal{O}(n)$
ISO	✓	✓	✓	✓	✓	$\mathcal{O}(n^2)$
BS	✓	✓	✓	✓	✓	$\mathcal{O}(n^2 \log n)$
QS	✓	✓	✓	✓	✓	$\mathcal{O}(n^2 \log n)$
HB	✓	✗	✓	✗	✓	$\mathcal{O}(n)$

We emphasize that we employ the approximations described in this section only when the predictive distributions have an absolutely continuous component. If the predictive distributions are discrete the pure form of the decomposition from Section 3.1 suffices. The pure form also suffices when it is known a priori that every single pair orders, e.g., when the predictive distributions are members of a location family with pairwise distinct centers. For illustration we refer the reader to our case studies and the penultimate paragraphs in Sections 5.1 and 5.2, respectively, where we indicate the use of the pure versus the approximate implementation.

Table 2 summarizes properties of the empirical isotonicity-based decomposition in a comparison to the types of decomposition from Section 2. For the statements of properties $E_1, \ldots, E_5$ we refer the reader to Section 2.1. The properties quoted for the isotonicity-based decomposition apply both to the pure and to the approximate implementation; for the latter, property E_1 is understood relative to the modified score at (27).

Computational complexity is quantified in floating point operations in terms of the number n of tuples (F_i, y_i) at (8). Concerning the isotonicity-based decomposition, the determination of the pairwise stochastic order relations between the distributions $F_1, \ldots, F_n$ requires $\mathcal{O}(n^2)$ operations. As IDR can be implemented in at most $\mathcal{O}(n^2)$ operations (Henzi, Ziegel and Gneiting, 2021; Henzi, Mösching and Dümbgen, 2022), the computation of the isotonicity-based decomposition is of complexity $\mathcal{O}(n^2)$. In contrast, the Brier score based and quantile score based decompositions require at least $\mathcal{O}(n)$ sortings of n real-valued quantities (cf. Appendices A.1 and A.2) and, hence, the implementation is of complexity at least $\mathcal{O}(n^2 \log n)$.

4. Population level analysis

In this section, we present population level versions of all decompositions which we have discussed so far, and we analyse their relations to notions of calibration.

The population quantity to be decomposed is the expected score

$$\mathbb{E}\,\mathrm{crps}(F, Y), \tag{31}$$

where the expectation is with respect to the joint law $\mathbb{P}$ of the random tuple (F, Y) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where F is a cdf-valued random quantity, which we interpret as the forecast, and the random variable Y is the real-valued outcome. Under simple conditions, which we state in Section 4.4, the expectation at (31) is finite. For subsequent use, we assume the existence of a standard uniform variable U on $(\Omega, \mathcal{F}, \mathbb{P})$, which is independent of (F, Y). Evidently, if $\mathbb{P}$ is the empirical distribution for the data at (8) the expectation at (31) reduces to the mean score $\overline{\mathrm{CRPS}}$ from (9).

In all types of decompositions the population version of the uncertainty component is the expected score

$$\mathrm{UNC}_0 = \mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}}, Y) \tag{32}$$

of the marginal law F_{mg} of Y. Again, the expectation is with respect to $\mathbb{P}$, and if $\mathbb{P}$ is the empirical distribution of the data at (8) then (32) reduces to (10). In this light, the decompositions at the population level read

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_\bullet - \mathrm{DSC}_\bullet + \mathrm{UNC}_0,$$

where $\bullet$ indicates the type, namely, CT, BS, QS, HB, or our new ISO. Therefore, it suffices to specify the miscalibration component $\mathrm{MCB}_\bullet$; the discrimination component is deduced as $\mathrm{DSC}_\bullet = \mathrm{MCB}_\bullet + \mathrm{UNC}_0 - \mathbb{E}\,\mathrm{crps}(F, Y)$.

4.1. Desiderata for decompositions at the population level

We adapt the desirable properties E_1 through E_5 for decompositions of a mean score from Section 2.1 to the population setting, as follows.

- (P_1) The decomposition is exact.
- (P_2) The components MCB, DSC, and UNC are nonnegative.
- (P_3) The MCB component vanishes if, and only if, the forecast is calibrated in a well defined sense.
- (P_4) The DSC component vanishes if the forecast is static, i.e., there is an $F_0 \in \mathcal{P}(\mathbb{R})$ such that $F = F_0$ almost surely.
- (P_5) The UNC component only depends on the unconditional distribution F_{mg} of the outcome.

The following property P_0 formalizes the natural reduction condition introduced and formulated at the end of Section 2.1.

- (P_0) When the joint law $\mathbb{P}$ is an empirical measure, MCB, DSC, and UNC reduce to the empirical components $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$, respectively.

Evidently, in concert with P_0 properties P_1 and P_2 are direct counterparts of properties E_1 and E_2, respectively. Concerning P_3, a notion of forecast calibration has to be specified. In the special case of a binary outcome, there is a unique,

clear-cut notion of calibration (Gneiting and Ranjan, 2013, Theorem 2.11). Here, we consider the case of a real-valued outcome, for which numerous notions of calibration exist (Gneiting and Resin, 2023), which we discuss in Section 4.3. Auto-calibration is the strongest such notion, but typically cannot be used in practice. Indeed, it turns out that E_3 and P_3 represent two sides of the same coin, in the sense that if a decomposition satisfies P_3 with respect to the strongest notion of auto-calibration, then E_3 is violated and the decomposition becomes degenerate. Conversely, if we wish for E_3 to hold, we ought to consider notions of calibration for P_3 that are weaker than auto-calibration. Requirement P_4 is natural, since a static forecast has no discrimination ability at all. Finally, property P_5 is motivated by the observation that intrinsic uncertainty does not depend on the forecast; evidently, the criterion is satisfied by UNC_0 at (32).

4.2. Isotonic conditional expectations and laws

The population versions of the isotonicity-based, Brier score based, and quantile score based decompositions rely on conditional expectations given σ-lattices and isotonic conditional laws. We give a short overview of the necessary concepts and refer to Arnold and Ziegel (2024) for further details. Readers not familiar with measure theory might skip the current section and intuitively think of the conditional expectation and the conditional law of a random variable Y given a σ-lattice $\mathcal{A}$, which we denote $\mathbb{E}(Y \mid \mathcal{A})$ and $P_{Y|\mathcal{A}}$, respectively, as classical conditional expectations and laws under the constraint of isotonicity.

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A subset $\mathcal{A} \subseteq \mathcal{F}$ is a *σ-lattice* if it is closed under countable unions and intersections and $\Omega, \emptyset \in \mathcal{A}$. Let $\mathcal{A} \subseteq \mathcal{F}$ be a σ-lattice and let X and Z be integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We call X *$\mathcal{A}$-measurable* if $\{X > x\} \in \mathcal{A}$ for all $x \in \mathbb{R}$ and define the *σ-lattice generated by X*, denoted by $\mathscr{L}(X)$, as the smallest σ-lattice which contains $\{X > x\}$ for all $x \in \mathbb{R}$. We call an $\mathcal{A}$-measurable random variable $\tilde{X}$ a *conditional expectation of X given $\mathcal{A}$*, for short $\mathbb{E}(X \mid \mathcal{A})$, if $\mathbb{E}(X\mathbb{1}_A) \leq \mathbb{E}(\tilde{X}\mathbb{1}_A)$ for all $A \in \mathcal{A}$ and $\mathbb{E}(X\mathbb{1}_B) = \mathbb{E}(\tilde{X}\mathbb{1}_B)$ for all $B \in \sigma(\tilde{X})$, where $\sigma(\tilde{X})$ denotes the σ-algebra generated by $\tilde{X}$. Brunk (1965) showed that $\mathbb{E}(X \mid \mathcal{A})$ is almost surely unique and coincides with the classical conditional expectations if $\mathcal{A}$ is a σ-algebra. Conditional expectations given σ-lattices are closely connected to isotonicity as illustrated in Arnold and Ziegel (2024). In particular, for any integrable random variable X and random variable Z, there exists an increasing Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(X \mid \mathscr{L}(Z)) = f(Z)$. This result is analogous to the well-known factorization result for classical conditional expectations given σ-algebras, with the difference that, additionally, f has to be increasing.

Isotonic conditional laws can be defined in analogy to classical conditional laws. Specifically, the isotonic conditional law (ICL) of the random variable Y given $\mathcal{A}$, denoted $P_{Y|\mathcal{A}}$, is a Markov kernel from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\omega \mapsto P_{Y|\mathcal{A}}(\omega, (y, \infty))$ is a version of $\mathbb{P}(Y > y \mid \mathcal{A}) = \mathbb{E}(\mathbb{1}\{Y > y\} \mid \mathcal{A})$ for any $y \in \mathbb{R}$. Arnold and Ziegel (2024) show the existence and uniqueness of ICL. Equivalently, ICL emerges as the minimizer of an expected score, where

the scoring rule may be taken from a large class of proper scoring rules that includes the crps.

We are particularly interested in ICL with respect to the σ-lattice generated by the forecast F. We call $B \subseteq \mathcal{P}(\mathbb{R})$ an upper set if $P \in B$ and $P \leq_{\mathrm{st}} Q$ implies $Q \in B$ for $Q \in \mathcal{P}(\mathbb{R})$, and we denote by $\mathcal{U}$ the family of all upper sets in $\mathcal{P}(\mathbb{R})$. For the forecast F, we define the σ-lattice generated by F as the family of all preimages of measurable upper sets under F, i.e., $\mathscr{L}(F) = \{F^{-1}(B) \mid B \in \mathcal{B}(\mathcal{P}(\mathbb{R})) \cap \mathcal{U}\} \subseteq \mathcal{F}$, where $\mathcal{B}(\mathcal{P}(\mathbb{R}))$ denotes the σ-algebra on $\mathcal{P}(\mathbb{R})$ with respect to the weak topology. For details, we refer the reader to Definition 3.1 of Arnold and Ziegel (2024).

In a nutshell, $P_{Y|\mathscr{L}(F)}$ arises as the best available prediction for the distribution of Y, given all information in the forecast F, under the assumption that smaller (greater) values of F correspond to smaller (greater) values of the conditional law with respect to the stochastic order.

4.3. Calibration

A strong notion of calibration is auto-calibration, which formalizes the idea that the outcome is indistinguishable from a random draw from the posited distribution F. Specifically, the random forecast F is *auto-calibrated* (Tsyplakov, 2013) if $P_{Y|F} = F$, or equivalently

$$F(x) = \mathbb{P}(Y \leq x \mid F) \quad \text{almost surely for all } x \in \mathbb{R}. \tag{33}$$

For any threshold value $x \in \mathbb{R}$, we may condition on the random variable $F(x)$ instead of the random distribution F in (33), to obtain the weaker notion of threshold calibration. Specifically, the forecast F is called *threshold calibrated* (Henzi, Ziegel and Gneiting, 2021) if

$$F(x) = \mathbb{P}(Y \leq x \mid F(x)) \quad \text{almost surely for all } x \in \mathbb{R}.$$

Essentially, for a threshold calibrated forecast F, we can take $F(x)$ at face value for any $x \in \mathbb{R}$. In a slight adaptation of the definition in Gneiting and Resin (2023), we call the forecast F *quantile calibrated* if

$$F^{-1}(\alpha) = q_\alpha(Y \mid F^{-1}(\alpha)) \quad \text{almost surely for all } \alpha \in (0,1),$$

where for any $\alpha \in (0,1)$, $q_\alpha(Y \mid F^{-1}(\alpha))$ denotes the lower-α-quantile of the conditional law of Y given $F^{-1}(\alpha)$. Equivalently, one can think of $q_\alpha(Y \mid F^{-1}(\alpha))$ as a $\sigma(F^{-1}(\alpha))$-measurable random variable which minimizes $\mathbb{E}\,\mathrm{qs}_\alpha(G, Y)$ over all $\sigma(F^{-1}(\alpha))$-measurable random variables G; see Armerin (2014).

The forecast F is called *isotonically calibrated* if F is almost surely equal to the isotonic conditional law of Y given $\mathscr{L}(F)$, i.e., $F = P_{Y|\mathscr{L}(F)}$ almost surely. By Proposition 5.3 of Arnold and Ziegel (2024), auto-calibration implies isotonic calibration, and isotonic calibration implies threshold calibration and quantile calibration.

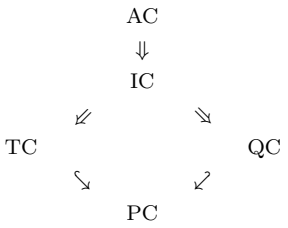

FIG 1. *Implications between auto-calibration (AC), isotonic calibration (IC), threshold calibration (TC), and quantile calibration (QC). Implications with respect to probabilistic calibration (PC) are indicated by hooked arrows and hold under Assumption 2.15 of Gneiting and Resin (2023).*

The probability integral transform (PIT) of the cdf-valued random quantity F is the random variable $Z_F = F(Y-) + U(F(Y) - F(Y-))$, where $F(y-) = \lim_{x \uparrow y} F(x)$ denotes the left-hand limit of F at $y \in \mathbb{R}$, with a random variable U that is standard uniform and independent of F and Y. The PIT of a continuous cdf F simplifies to $Z_F = F(Y)$. The forecast F is *probabilistically calibrated* if Z_F is uniformly distributed on the unit interval (Gneiting and Ranjan, 2013). Originally suggested by Dawid (1984), checks for probabilistic calibration, and for the uniformity of the closely related rank histogram, constitute a cornerstone of forecast evaluation (Diebold, Gunther and Tay, 1998; Hamill, 2001; Gneiting, Balabdaoui and Raftery, 2007). Under regularity conditions, a threshold calibrated or quantile calibrated forecast is probabilistically calibrated; details and a direct implication from isotonic calibration to a weak form of probabilistic calibration are available in Gneiting and Resin (2023, Section 3.3) and Arnold and Ziegel (2024, Appendix D), respectively. Figure 1 summarizes relationships between the notions of calibration discussed in this section.

4.4. Population level decompositions

We now give generalizations of the empirical decompositions discussed in Sections 2 and 3 that apply at the population level. Recall that we consider the joint law $\mathbb{P}$ of the random tuple (F, Y). As before, we let $\mathcal{P}(\mathbb{R})$ denote the class of the Borel probability measures on $\mathbb{R}$ that have a finite first moment. In the current and the subsequent subsection, we generally operate under the following regularity conditions. For proofs, we refer the reader to Appendix D.

Assumption 4.1. Let the marginal law F_{mg} of Y be such that $F_{\mathrm{mg}} \in \mathcal{P}(\mathbb{R})$, and suppose that

$$\mathbb{E} \int |x| \, \mathrm{d}F(x) = \mathbb{E}\, \mathbb{E}_F |X| < \infty. \tag{34}$$

In view of the kernel score representation of the crps (Gneiting and Raftery, 2007, eq. (21)), Assumption 4.1 implies that

$$\mathbb{E}\, \mathrm{crps}(F, Y) = \mathbb{E}\, \mathbb{E}(\mathrm{crps}(F, Y) \mid F)$$

$$= \mathbb{E}\left(\mathbb{E}_F(|X - Y| \mid F) - \frac{1}{2}\mathbb{E}_F(|X - X'| \mid F)\right)$$
$$\leq \mathbb{E}\,\mathbb{E}_F|X| + \mathbb{E}\,|Y| < \infty,$$

where X and X' are independent random variables with law F. Similarly, it follows that $\mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}}, Y) < \infty$. Furthermore, the properties of isotonic and standard conditional laws imply that $\mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y) \leq \mathbb{E}\,\mathrm{crps}(F, Y)$ and $\mathbb{E}\,\mathrm{crps}(P_{Y|F}, Y) \leq \mathbb{E}\,\mathrm{crps}(F, Y)$, respectively. In this light, Assumption 4.1 ensures that $\mathbb{E}\,\mathrm{crps}(F, Y)$, $\mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}}, Y)$, $\mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y)$, and $\mathbb{E}\,\mathrm{crps}(P_{Y|F}, Y)$ are finite.

The population version of the Candille–Talagrand decomposition at (12) is

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_{\mathrm{CT}} - \mathrm{DSC}_{\mathrm{CT}} + \mathrm{UNC}_0, \tag{35}$$

where UNC_0 is defined at (32), and

$$\mathrm{MCB}_{\mathrm{CT}} = \mathbb{E}\,\mathrm{crps}(F, Y) - \mathbb{E}\,\mathrm{crps}(P_{Y|F}, Y).$$

Similarly, the population version of the isotonicity-based decomposition at (25) is

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_{\mathrm{ISO}} - \mathrm{DSC}_{\mathrm{ISO}} + \mathrm{UNC}_0, \tag{36}$$

where

$$\mathrm{MCB}_{\mathrm{ISO}} = \mathbb{E}\,\mathrm{crps}(F, Y) - \mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y).$$

The decomposition at (36) is analogous to the theoretically preferred Candille–Talagrand decomposition at (35), except that the performance of the forecast F is compared with the isotonic conditional law $P_{Y|\mathscr{L}(F)}$ rather than the conditional law $P_{Y|F}$. The general decompositions at (35) and (36) reduce to (12) and (25), respectively, when $\mathbb{P}$ is the empirical distribution of the data in (8).

The population version of the Brier score based decomposition at (16) is

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_{\mathrm{BS}} - \mathrm{DSC}_{\mathrm{BS}} + \mathrm{UNC}_0, \tag{37}$$

where

$$\mathrm{MCB}_{\mathrm{BS}} = \mathbb{E}\,\mathrm{crps}(F, Y) - \mathbb{E}\int \big(\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) - \mathbb{1}\{Y \leq z\}\big)^2\,\mathrm{d}z.$$

Similarly, the population version of the quantile based based decomposition at (17) is

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_{\mathrm{QS}} - \mathrm{DSC}_{\mathrm{QS}} + \mathrm{UNC}_0, \tag{38}$$

where

$$\mathrm{MCB}_{\mathrm{QS}} = \mathbb{E}\,\mathrm{crps}(F, Y) - \mathbb{E}\int_0^1 \mathrm{qs}_\alpha\big(q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))), Y\big)\,\mathrm{d}\alpha.$$

The properties of isotonic conditional expectations and isotonic conditional quantiles imply that $\mathbb{E}\int(\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) - \mathbb{1}\{Y \leq z\})^2\,\mathrm{d}z \leq \mathbb{E}\,\mathrm{crps}(F, Y) <$

∞ and $\mathbb{E}\int_0^1 \mathrm{qs}_\alpha(q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))), Y)\,\mathrm{d}\alpha \le \mathbb{E}\,\mathrm{crps}(F, Y) < \infty$. The decompositions at (37) and (38) reduce to (16) and (17), respectively, when $\mathbb{P}$ is the empirical distribution of the data in (8).

Finally, we consider the Hersbach decomposition. To this end, let ν_F be the image of the Lebesgue measure λ under F, i.e., $\nu_F(A) = \lambda(F^{-1}(A))$, and define the measures given by

$$\mu(A) = \mathbb{E}\,(\nu_F(A)) \tag{39}$$

and

$$\tau(A) = \mathbb{E}\left(\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p)\right), \tag{40}$$

respectively, where $A \in \mathcal{B}(0,1)$ is any Borel set. We are now ready to state a population version of the Hersbach decomposition from Section 2.5.

Proposition 4.1. *Let Assumption 4.1 hold, and let μ and τ be the measures defined at (39) and (40), respectively. Then τ is absolutely continuous with respect to μ; let f denote the respective Radon–Nikodym derivative. It holds that*

$$\mathbb{E}\,\mathrm{crps}(F, Y) = \mathrm{MCB}_{\mathrm{HB}} - \mathrm{DSC}_{\mathrm{HB}} + \mathrm{UNC}_0, \tag{41}$$

where UNC_0 *is given at (32),*

$$\mathrm{MCB}_{\mathrm{HB}} = \int_0^1 (p - f(p))^2\,\mathrm{d}\mu(p),$$

$$\mathrm{DSC}_{\mathrm{HB}} = \mathrm{UNC}_0 - \int_0^1 f(p)(1 - f(p))\,\mathrm{d}\mu(u) - \mathrm{MS},$$

and

$$\begin{aligned}\mathrm{MS} = {} & \mathbb{E}\left[\mathbb{1}\{F(Y) = 0\}\,(F^{-1}(0+) - Y)\right] \\ & + \mathbb{E}\left[\mathbb{1}\{F(Y) > 0\}\,(2F(Y) - 1)(Y - F^{-1}(F(Y)))\right].\end{aligned} \tag{42}$$

The MS component can only be nonzero when Y lies outside the support of F with positive probability; hence, we write MS for misspecified support. Note that MS can be negative, e.g., if $F = (\delta_0 + 3\,\delta_2)/4$ and $Y = 1$ almost surely then $\mathrm{MS} = -1/2$.

The following result is a special case of the more general statement in Corollary D.1 in Appendix D. It shows that the population decomposition nests the modified empirical Hersbach decomposition.

Corollary 4.2. *If $\mathbb{P}$ is the empirical measure of a finite collection of forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where each F_i is the empirical cdf of a sample of size m, then the population decomposition at (41) reduces to the modified empirical Hersbach decomposition at (22).*

The next result demonstrates that Proposition 4.1 subsumes the Hersbach–Lalaurette decomposition for strictly increasing forecast cdfs as given in Appendix A of Candille and Talagrand (2005).

Corollary 4.3. *Let Assumption 4.1 hold, and suppose that F^{-1} is almost surely absolutely continuous. Then* $\mathrm{MS} = 0$ *and the measure μ at (39) has density*

$$\gamma(p) = \mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right) \tag{43}$$

with respect to the Lebesgue measure on the unit interval. Furthermore, the measure τ at (40) has Radon–Nikodym derivative defined by

$$f(p) = \frac{1}{\gamma(p)}\,\mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\,\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right) \tag{44}$$

if $\gamma(p) > 0$, and $f(p) = 0$ otherwise, with respect to μ.

Considering a practically relevant case, we derive in Example D.1 in Appendix D the empirical Hersbach decomposition for probabilistic forecasts of a nonnegative quantity, assuming that the forecast distributions are mixtures of a point mass at zero and a strictly positive density on the positive halfline.

4.5. Properties of the decompositions

The following theorem summarizes properties of the Candille–Talagrand, the isotonicity-based, the Brier score based, and the quantile score based decompositions at the population level. Proofs of the theorem and all other results in this section are deferred to Appendix E.

Theorem 4.4. *Under Assumption 4.1 the following statements hold.*

(a) The Candille–Talagrand decomposition at (35) is exact and satisfies
- $\mathrm{MCB}_{\mathrm{CT}} \ge 0$ *with equality if, and only if, F is auto-calibrated;*
- $\mathrm{DSC}_{\mathrm{CT}} \ge 0$ *with equality if, and only if, $P_{Y|F} = F_{\mathrm{mg}}$ almost surely.*

(b) The isotonicity-based decomposition at (36) is exact and satisfies
- $\mathrm{MCB}_{\mathrm{ISO}} \ge 0$ *with equality if, and only if, F is isotonically calibrated;*
- $\mathrm{DSC}_{\mathrm{ISO}} \ge 0$ *with equality if, and only if, $P_{Y|\mathscr{L}(F)} = F_{\mathrm{mg}}$ almost surely.*

(c) The Brier score based decomposition at (37) is exact and satisfies
- $\mathrm{MCB}_{\mathrm{BS}} \ge 0$ *with equality if, and only if, F is threshold calibrated;*
- $\mathrm{DSC}_{\mathrm{BS}} \ge 0$ *with equality if, and only if, for all $z \in \mathbb{R}$, $\mathbb{P}(Y \le z \mid \mathscr{L}(F(z))) = \mathbb{P}(Y \le z)$ almost surely.*

(d) The quantile score based decomposition at (38) is exact and satisfies
- $\mathrm{MCB}_{\mathrm{QS}} \ge 0$ *with equality if F is quantile calibrated; conversely, if the random element $(Y, F^{-1}(\alpha))$ satisfies Assumption 6.1 in Arnold and Ziegel (2024) for all $\alpha \in (0,1)$ then* $\mathrm{MCB}_{\mathrm{QS}} = 0$ *implies quantile calibration of F;*

– $\mathrm{DSC}_{\mathrm{QS}} \geq 0$ *with equality if, and only if, for all* $\alpha \in (0,1)$, $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))) = q_\alpha(Y)$ *almost surely.*

To summarize, the population versions of the Candille–Talagrand, isotonicity-based, Brier score based, and quantile score based decompositions satisfy properties P_0, P_1, P_2, P_4, and P_5, and property P_3 with auto-calibration, isotonic calibration, threshold calibration, and quantile calibration, respectively. These findings lend theoretical support to the use of the isotonicity-based decomposition.

While in principle one would like to quantify miscalibration in terms of deviations from auto-calibration, as done by the Candille–Talagrand decomposition, the empirical version thereof is degenerate. The isotonicity-based decomposition quantifies miscalibration as deviation from isotonic calibration, which avoids degeneracy and is closer to auto-calibration than threshold or quantile calibration, as illustrated in Figure 1.

In view of known relationships between notions of calibration (Gneiting and Resin, 2023, Sections 2.2 and 2.3) the following implications hold.

Corollary 4.5. *Under Assumption 4.1, an auto-calibrated forecast yields*

$$\mathrm{MCB}_{\mathrm{CT}} = \mathrm{MCB}_{\mathrm{ISO}} = \mathrm{MCB}_{\mathrm{BS}} = \mathrm{MCB}_{\mathrm{QS}} = 0.$$

Corollary 4.6. *Under Assumption 4.1, it holds that*

$$\mathbb{E}\,\mathrm{crps}(F,Y) \geq \mathrm{MCB}_{\mathrm{CT}} \geq \mathrm{MCB}_{\mathrm{ISO}} \geq \max\{\mathrm{MCB}_{\mathrm{BS}}, \mathrm{MCB}_{\mathrm{QS}}\}. \tag{45}$$

Importantly, while formulated at the population level, the above results apply to the empirical versions of the decompositions in view of the reduction property P_0, which is obvious for the decompositions considered here. In particular, the relations in (45) nest the respective inequalities (26) for the empirical decompositions. For the isotonicity-based decomposition, if modified cdfs $F^{(a,b)}$ are used the results apply to the latter, and we refer to (30) for relationships to the respective components computed on the original cdfs.

Finally, we consider the Hersbach decomposition from Proposition 4.1, which struggles to satisfy the desirable properties from Section 4.1. By Corollary 4.2, the reduction property P_0 holds under conditions. By definition, properties P_1 and P_5 hold. The miscalibration component is clearly nonnegative. However, $\mathrm{DSC}_{\mathrm{HB}}$ may be negative as in Example F.3, i.e., property P_2 is violated. Moreover, the example in the proof of Proposition 2.3 shows that the Hersbach decomposition fails to satisfy P_4. Concerning P_3, Hersbach (2000), Candille and Talagrand (2005) and Yang and Kleissl (2024, p. 421) argue that the Hersbach reliability component is closely related to the rank histogram and hence one might expect that $\mathrm{MCB}_{\mathrm{HB}} = 0$ if, and only if, F is probabilistically calibrated. However, the examples in Appendices F.4 and F.5 show that, in general, probabilistic calibration is neither sufficient nor necessary for $\mathrm{MCB}_{\mathrm{HB}} = 0$ to hold. The following proposition collects calibration properties in relation to the Hersbach decomposition.

TABLE 3

Properties of the population versions of the Candille–Talagrand (CT), isotonicity-based (ISO), Brier score based (BS), quantile score based (QS), and Hersbach (HB) decompositions of the mean crps. For properties P_0, P_1, ..., P_5 see Section 4.1. The acronyms AC, IC, TC, QC, and PC concern modes of calibration and are expanded in the legend of Figure 1. The Hersbach decomposition satisfies P_0 under the conditions of Corollary 4.2 (UC). Regarding P_3 we indicate if a mode of calibration is necessary and sufficient (NaS); necessary and sufficient under Assumption 2.15 in Gneiting and Resin (2023) (UC); necessary but not sufficient (N); necessary under Assumption 2.15 (NuC); sufficient but not necessary (S); or sufficient under Assumption 2.15 and the conditions of Proposition 4.7(c) but not necessary (SuC); for the MCB *component to vanish.*

	P_0	P_1	P_2	P_3					P_4	P_5
				AC	IC	TC	QC	PC		
CT	✓	✓	✓	NaS	N	N	N	N	✓	✓
ISO	✓	✓	✓	S	NaS	N	N	NuC	✓	✓
BS	✓	✓	✓	S	S	NaS	UC	NuC	✓	✓
QS	✓	✓	✓	S	S	UC	NaS	NuC	✓	✓
HB	UC	✓	✗	S	SuC	SuC	SuC	SuC	✗	✓

Proposition 4.7. *Let Assumption 4.1 hold and consider the population version of the Hersbach decomposition at* (41).

(a) *If* $Y \in \text{supp}(F)$ *almost surely, then* $\text{MS} = 0$, *where* MS *is defined at* (42).
(b) *For an auto-calibrated forecast, it holds that* $\text{MS} = \text{MCB}_{\text{HB}} = 0$.
(c) *Suppose that* F *belongs to a location family, i.e., for all* $x \in \mathbb{R}$, $F(x) = F_0(x-\mu)$ *for some* $F_0 \in \mathcal{P}(\mathbb{R})$ *and random location* μ. *Suppose furthermore that* F_0 *has no jumps and* F_0^{-1} *is absolutely continuous. Then* $\text{MCB}_{\text{HB}} = 0$ *if* F *is probabilistically calibrated.*

For a succinct overview of properties of the population versions of the Candille–Talagrand, isotonicity-based, Brier score based, quantile score based, and Hersbach decompositions see Table 3, which mirrors the findings at the empirical level from Table 2.

In Appendix F we illustrate the different types of decompositions in a number of analytic examples at the population level. Figure 2 summarizes how the respective miscalibration terms relate to the theoretically preferred MCB_{CT} component.

5. Case studies

We now illustrate the use of the isotonicity-based decomposition from Section 3 in case studies on weather forecasts and benchmark regression tasks from machine learning, respectively. For simplicity, we use an abbreviated notation for the components of the mean score $\overline{\text{CRPS}}$ throughout this section, namely, $\overline{\text{MCB}} = \overline{\text{MCB}}_{\text{ISO}}$, $\overline{\text{DSC}} = \overline{\text{DSC}}_{\text{ISO}}$, and $\overline{\text{UNC}} = \overline{\text{UNC}}_0$, respectively. Note the

FIG 2. *The graphic indicates for the population level examples F.1, . . . , F.5 in Appendix F whether the* $\mathrm{MCB}_\bullet$ *term, where* $\bullet$ *stands for* CT, ISO, BS, QS, *or* HB, *respectively, agrees with the theoretically preferred quantity* $\mathrm{MCB}_{\mathrm{CT}}$ *(green), is smaller than* $\mathrm{MCB}_{\mathrm{CT}}$ *but remains positive (orange), or deceptively equals zero (red). Connected segments indicate equality of corresponding terms. For analytic results, see Table 4.*

opposite orientation of $\overline{\mathrm{MCB}}$ and $\overline{\mathrm{DSC}}$, in that higher $\overline{\mathrm{DSC}}$ corresponds to better discrimination ability, whereas lower $\overline{\mathrm{MCB}}$ indicates better calibration.

When one seeks to simultaneously compare $\overline{\mathrm{CRPS}}$, $\overline{\mathrm{MCB}}$, and $\overline{\mathrm{DSC}}$ between larger numbers of forecast methods, tables get cumbersome. Therefore, we suggest a graphical display, namely, the miscalibration–discrimination ($\overline{\mathrm{MCB}}$–$\overline{\mathrm{DSC}}$) plot, which is motivated by similar displays in Gneiting et al. (2023) and Dimitriadis et al. (2024). In this type of graphic, $\overline{\mathrm{DSC}}$ is plotted against $\overline{\mathrm{MCB}}$, and isolines correspond to specific values of the mean score $\overline{\mathrm{CRPS}}$, which is constant along parallel lines. The uncertainty component $\overline{\mathrm{UNC}}$ is independent of the forecast method, and we display it in the upper left or upper right corner of the plot.

5.1. Probabilistic quantitative precipitation forecasts

Ensemble prediction systems have tremendously improved weather forecasts over the past decades (Bauer, Thorpe and Brunet, 2015). However, ensemble forecasts remain subject to biases and dispersion errors, and hence require some form of statistical postprocessing (Gneiting and Raftery, 2005; Vannitsem, Wilks and Messner, 2018). Here we consider the case study in Henzi, Ziegel and Gneiting (2021), which compares the performance of raw and postprocessed ensemble forecasts for 24-hour accumulated precipitation in terms of the mean score $\overline{\mathrm{CRPS}}$, which we decompose into $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$, respectively.

Following Henzi, Ziegel and Gneiting (2021), we consider forecasts and observations for 24-hour accumulated precipitation from 6 January 2007 to 1 January 2017 at Brussels, Frankfurt, London, and Zurich in millimeters. The 52 member raw ensemble (ENS) forecast operated by the European Centre for Medium-Range Weather Forecasts comprises a high resolution member, a control member at lower resolution, and 50 perturbed members at the same lower resolution but with perturbed initial conditions (Molteni et al., 1996). We use data from

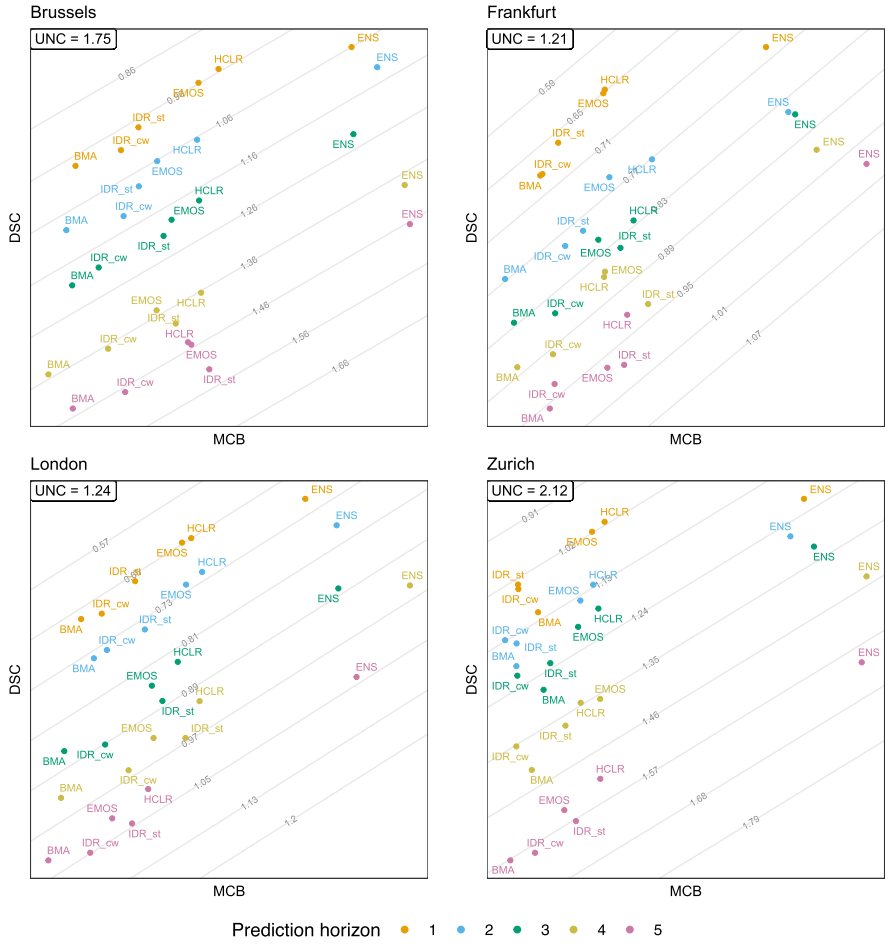

FIG 3. $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ *plots for forecasts of 24-hour accumulated precipitation at Brussels, Frankfurt, London, and Zurich, at prediction horizons of one to five days ahead. The mean score* $\overline{\text{CRPS}}$ *is constant along the parallel lines and shown in the unit of millimeters. Acronyms are defined in the text, and details of the forecast methods are documented in Henzi, Ziegel and Gneiting (2021, Section 5).*

2007 to 2014 to train the postprocessing techniques Bayesian model averaging (BMA; Sloughter et al., 2007), ensemble model output statistics (EMOS; Scheuerer, 2014), heteroscedastic censored logistic regression (HCLR; Messner et al., 2014) and two versions, IDR_{cw} and IDR_{st}, of isotonic distributional regression (IDR; Henzi, Ziegel and Gneiting, 2021), where IDR_{cw} is documented in Henzi, Ziegel and Gneiting (2021) and IDR_{st} uses the stochastic order on the ensemble cdfs. For further implementation details we refer the reader to Henzi, Ziegel and Gneiting (2021). The years 2015 and 2016 form the evaluation period.

The ENS and IDR forecast distributions have finite support and we apply

the isotonicity-based decomposition of $\overline{\text{CRPS}}$ in its pure form from Section 3.1. For the other forecasts, which employ mixtures of a point mass at zero (for no precipitation) and a density at positive accumulations as predictive distributions, we fix $a = 0$ and use Algorithm 1 to determine the upper bound b, which generally is identical to, or very slightly higher than, the highest accumulation observed in the test data; then we compute stochastic order relations on an equidistant grid of size 5000 over $[a, b]$ and apply the isotonicity-based decomposition in its approximate form from Section 3.2.

The respective $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ plots for Brussels, Frankfurt, London, and Zurich are shown in Figure 3. We note an increase of the mean score $\overline{\text{CRPS}}$ values with the prediction horizon, which is due to a decrease in discrimination ability. The raw ensemble (ENS) forecasts discriminate very well, but are poorly calibrated. The postprocessing methods yield considerable improvement in $\overline{\text{CRPS}}$, subject to a trade-off between $\overline{\text{MCB}}$ and $\overline{\text{DSC}}$. The EMOS and HCLR techniques, which employ inflexible parametric densities with fixed shape, excel in terms of discrimination, but lack in calibration. In contrast, the BMA and IDR techniques, which are much more flexible, are better calibrated, but inferior in terms of discrimination ability.

5.2. Benchmark regression problems from machine learning

A sizable strand of recent literature in machine learning is concerned with methods for uncertainty quantification for neural networks, where the task is the transformation of single-valued neural network output into predictive distributions (Gawlikowski et al., 2023). In this literature, performance is typically evaluated in terms of the mean logarithmic score (Gneiting and Raftery, 2007, Section 4.1) which, in sharp contrast to the crps, can only be applied to methods that generate predictive densities. Furthermore, extant measures for the assessment of calibration and discrimination ability tend to be ad hoc. In this section, we demonstrate the use of the mean score $\overline{\text{CRPS}}$ and its isotonicity-based decomposition into $\overline{\text{MCB}}$, $\overline{\text{DSC}}$, and $\overline{\text{UNC}}$ in this context.

We adopt the benchmark regression tasks setting originally proposed by Hernandéz-Lobato and Adams (2015) and consider the datasets and methods from the middle block of Table 6 in Walz et al. (2024), except that we skip results for the Naval and Year datasets, for which there are missing entries. The experimental setting is based on single-valued output from a neural network, which learns a regression function based on a collection of covariates or features. In this setting, Walz et al. (2024) compare competing methods for uncertainty quantification, including the popular Monte Carlo Dropout approach (MC Dropout; Gal and Ghahramani, 2016) and a scalable Laplace approximation based technique (Laplace; Immer et al., 2021; Ritter, Botev and Barber, 2018) that operate within the neural network learning pipeline. Their competitors include output-based methods that learn on training data of previous single-valued model output and outcomes only, without accessing feature values, namely, the Single Gaussian technique, conformal prediction (CP; Vovk et al.,

2020), and the EasyUQ technique (Walz et al., 2024), which is based on IDR (Henzi, Ziegel and Gneiting, 2021). Furthermore, we consider smoothed versions of the discrete CP and EasyUQ distributions, termed Smooth CP and Smooth EasyUQ, respectively. For implementation details, we refer the reader to Walz (2022).

The CP and EasyUQ distributions have finite support, and the Single Gaussian incurs normal distribution with a fixed variance, but varying mean. For these three methods, we use the isotonicity-based decomposition of $\overline{\text{CRPS}}$ in the pure form from Section 3.1. The Laplace method also employs normal distributions, but with varying mean and variances. The MC Dropout technique yields mixtures of normal distributions, and the Smooth CP and Smooth EasyUQ distributions are mixtures of Student-t distributions (or normal distributions as a limit case). For these methods, we use the approximations described in Section 3.2.

The $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ plots in Figure 4 illustrate the mean score $\overline{\text{CRPS}}$ and the $\overline{\text{MCB}}$, $\overline{\text{DSC}}$, and $\overline{\text{UNC}}$ components for the eight datasets and seven methods, respectively. The MC Dropout technique yields predictive distributions that are poorly calibrated, a finding that is well documented in the machine learning literature (Gawlikowski et al., 2023), though with high discrimination ability. The predictive distributions generated by the Laplace method trade better calibration for diminished discrimination ability. The simplistic Single Gaussian technique performs surprisingly well, typically with both the $\overline{\text{MCB}}$ and the $\overline{\text{DSC}}$ component being small relative to the competitors. The EasyUQ and CP distributions generally are well calibrated, with low $\overline{\text{MCB}}$ components throughout, and often superior overall performance. Smoothing of the discrete EasyUQ and CP distributions has only small effects. The only exception is for the EasyUQ forecast for the Wine dataset, which has only ten unique outcomes that correspond to quality levels, thus favoring the discrete basic EasyUQ distributions, which place all probability mass on this small set of outcomes.

6. Discussion

In line with the general idea of the CORP approach of Dimitriadis et al. (2024) and Gneiting and Resin (2023), we have developed an isotonicity-based decomposition of the mean score $\overline{\text{CRPS}}$ into miscalibration ($\overline{\text{MCB}}$), discrimination ($\overline{\text{DSC}}$), and uncertainty ($\overline{\text{UNC}}$) components. Both theoretically and computationally, the isotonicity-based decomposition serves as an attractive alternative to the Candille–Talagrand decomposition, which is of theoretical appeal, but yields degenerate decompositions in practice. Remarkably, Proposition 3.2 ensures that theoretical guarantees for the pure form from Section 3.1 very nearly carry over to the approximate implementation described in Section 3.2. In typical practice, interest focuses on the relative contributions of miscalibration and discrimination to the mean score. Competing forecast methods can be compared in $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ plots, which visualize these contributions and showcase the methods' strengths and deficiencies in succinct ways.

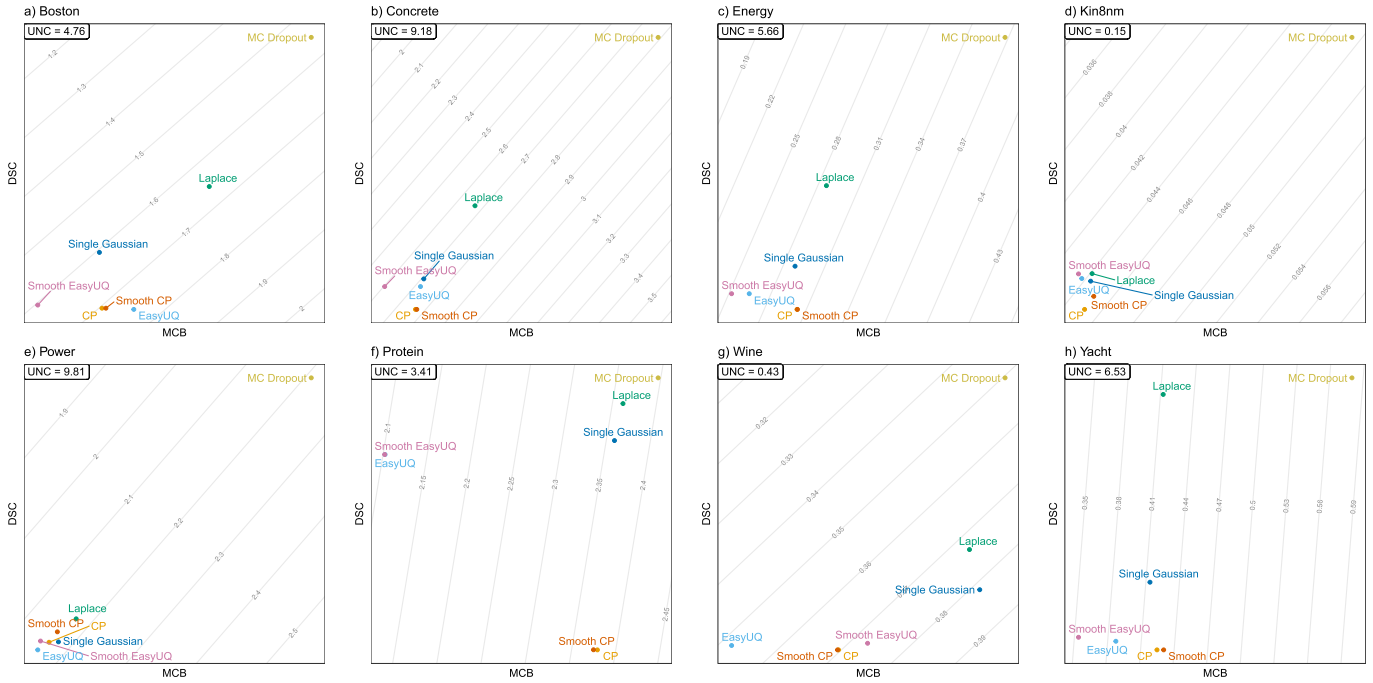

FIG 4. $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ *plots for methods of uncertainty quantification for neural network based regression from the middle block in Table 6 of Walz et al. (2024). The mean score* $\overline{\text{CRPS}}$ *is constant along the parallel lines.*

The isotonicity-based decomposition may degenerate if the data at (8) lack pairs of predictive distributions that can be compared in stochastic order. Our approximate implementation from Section 3.2 aims to increase the number of pairs of comparable cdfs, which addresses this limitation and yields appealing results in our experience. If further robustness in the decomposition is sought, convex combinations of the isotonicity-based decomposition with the Brier score based or the quantile score based decomposition can be employed.

Due to its linear computational complexity, the Hersbach decomposition is a viable option for decomposing $\overline{\text{CRPS}}$ for ensemble forecasts with a moderate number m of members, even when the size n of the evaluation set at (8) is very large and the isotonicity-based approach with its quadratic complexity is not feasible.[2] We recommend that it be used in the modified form described in Section 2.5, which allows for extensions beyond the case of ensemble forecasts, as described in Appendix D. A useful facet of the Hersbach decomposition is that it applies to general (nonnegatively) weighted sums (rather than simple averages only) of crps scores (Hersbach, 2000). The isotonicity-based decomposition generalizes to weighted sums as well, as the theoretical guarantees for IDR (Henzi, Ziegel and Gneiting, 2021) continue to apply in weighted case, and software developed by Alexander Henzi (`https://github.com/AlexanderHenzi/isodistrreg`) handles the extension. We leave details to future work.

As noted, the desirable properties $E_1, \ldots, E_5$ in the empirical case and P_0, $P_1, \ldots, P_5$ in the population case remain valid for decomposition of the mean score under proper scoring rules other than the crps. For instance, in various applications a certain region of the potential range of the outcome is of particular interest, and predictive performance might then be assessed with emphasis on these regions. In such settings, one may use versions of the crps as proposed by Gneiting and Ranjan (2013), namely,

$$\text{crps}_w(F, y) = \int_{-\infty}^{\infty} w(x)\, \text{s}_{\text{B}}(F(x), \mathbb{1}\{y \le x\})\, \mathrm{d}x$$

and

$$\text{crps}_v(F, y) = \int_0^1 v(\alpha)\, \text{qs}_\alpha(F^{-1}(\alpha), y)\, \mathrm{d}\alpha,$$

where w and v, respectively, are nonnegative weight functions. In view of the universality property of IDR (Henzi, Ziegel and Gneiting, 2021, Theorem 2), the isotonicity-based decomposition extends naturally to means of these types of scores, while preserving its desirable properties.

However, the isotonicity-based approach fails if a mean of logarithmic scores (Gneiting and Raftery, 2007, Section 4.1) is sought to be decomposed, for the logarithmic score, which allows for the comparison of density forecasts only, cannot be applied to the discrete IDR distributions. While in principle isotonic recalibration by IDR, on which isotonicity-based decompositions are based, could

[2] As a rule of thumb, the empirical isotonicity-based decomposition can be employed when the size n of the evaluation set is below 100 000. In Section 5.2 we apply it to the Protein dataset, which is of size $n = 45\,730$.

be replaced by recalibration with other methods, it is not at all evident what type of technique ought to be used, and we are unaware of any such method that would share the optimality properties of IDR that underlie the theoretical guarantees enjoyed by the isotonicity-based approach.

Various authors have pondered the use of the crps, which is favored by the meteorological and renewable energy literatures, as opposed to the logarithmic score, which is of particular popularity in econometrics and machine learning, with the choice arising both in the context of estimation via empirical score minimization and in the evaluation of predictive performance (Gneiting and Raftery, 2007). For example, D'Isanto and Polsterer (2018, Appendix B) argue that in neural network learning empirical score minimization in terms of the mean crps is preferable to optimization of the logarithmic score. In the evaluation of predictive performance, the availability of the theoretically supported and practically feasible isotonicity-based decomposition, in concert with the applicability of the score to discrete forecast distributions, strengthens arguments in favor of the crps.

Appendix A: Technical details for the Brier score and quantile score based decompositions

In this appendix we describe the Brier score (BS) and quantile score (QS) based decompositions from Sections 2.3 and 2.4 for the mean score $\overline{\mathrm{CRPS}}$ of the forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$ at (8). Both decompositions build on a general version of the pool-adjacent-violators (PAV) algorithm for nonparametric isotonic regression (Ayer et al., 1955). While historically work on the PAV algorithm has focused on the mean functional (Barlow et al., 1972; Robertson, Wright and Dykstra, 1988; de Leeuw, Hornik and Mair, 2009), the algorithm yields optimal isotonic fits under any identifiable functional; see, e.g., Jordan, Mühlemann and Ziegel (2022) and Gneiting and Resin (2023, Section 3.1).

A.1. Brier score based decomposition

For each threshold value $z \in \mathbb{R}$, we interpret $F_1(z), \ldots, F_n(z)$ as probability forecasts for the binary event $\xi_i(z) = \mathbb{1}\{y_i \le z\}$, where $i = 1, \ldots, n$. We obtain calibrated forecasts $\acute{F}_1(z), \ldots, \acute{F}_n(z)$ by applying the PAV algorithm for the mean functional on $\xi_1(z), \ldots, \xi_n(z)$ with respect to the order induced by $F_1(z), \ldots, F_n(z)$. This yields the CORP decomposition of the mean Brier score

$$\overline{\mathrm{BS}}_{F(z)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{s}_{\mathrm{B}}\big(F_i(z), \xi_i(z)\big)$$

as proposed by Dimitriadis, Gneiting and Jordan (2021), namely,

$$\overline{\mathrm{BS}}_{F(z)} = \underbrace{\Big(\overline{\mathrm{BS}}_{F(z)} - \overline{\mathrm{BS}}_{\acute{F}(z)}\Big)}_{\overline{\mathrm{MCB}}_{\mathrm{BS},z}} - \underbrace{\Big(\overline{\mathrm{BS}}_{\acute{F}(z)} - \overline{\mathrm{BS}}_{\hat{F}_{\mathrm{mg}}(z)}\Big)}_{\overline{\mathrm{DSC}}_{\mathrm{BS},z}} + \underbrace{\overline{\mathrm{BS}}_{\hat{F}_{\mathrm{mg}}(z)}}_{\overline{\mathrm{UNC}}_{\mathrm{BS},z}},$$

where $\hat{F}_{\mathrm{mg}}(z) = \frac{1}{n}\sum_{i=1}^{n} \xi_i(z)$ for $z \in \mathbb{R}$,

$$\overline{\mathrm{BS}}_{\acute{F}(z)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_{\mathrm{B}}\big(\acute{F}_i(z), \xi_i(z)\big) \quad \text{and} \quad \overline{\mathrm{BS}}_{\hat{F}_{\mathrm{mg}}(z)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_{\mathrm{B}}\big(\hat{F}_{\mathrm{mg}}(z), \xi_i(z)\big).$$

Integration of the $\overline{\mathrm{MCB}}_{\mathrm{BS},z}, \overline{\mathrm{DSC}}_{\mathrm{BS},z}$ and $\overline{\mathrm{UNC}}_{\mathrm{BS},z}$ components over $z \in \mathbb{R}$ yields the Brier score based score components and decomposition at (14), (15), and (16), respectively.

Computationally, it suffices to run the PAV algorithm at $z \in \{y_1, \dots, y_n\}$ and at the crossing points of the cdfs $F_1, \dots, F_n$.

Proof of Proposition 2.1. We note that

$$\begin{aligned}\overline{\mathrm{UNC}}_{\mathrm{BS}} &= \int \overline{\mathrm{BS}}_{\hat{F}_{\mathrm{mg}}(z)}\,\mathrm{d}z = \int \frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_{\mathrm{B}}\big(\hat{F}_{\mathrm{mg}}(z), \xi_i(z)\big)\,\mathrm{d}z \\ &= \frac{1}{n}\sum_{i=1}^{n} \int \big(\hat{F}_{\mathrm{mg}}(z) - \xi_i(z)\big)^2\,\mathrm{d}z = \frac{1}{n}\sum_{i=1}^{n} \mathrm{crps}(\hat{F}_{\mathrm{mg}}, y_i) = \overline{\mathrm{UNC}}_0,\end{aligned}$$

which implies that E_5 is satisfied. Property E_1 is immediate. Dimitriadis, Gneiting and Jordan (2021) show that $\overline{\mathrm{MCB}}_{\mathrm{BS},z}$ and $\overline{\mathrm{DSC}}_{\mathrm{BS},z}$ are nonnegative for all $z \in \mathbb{R}$ and thus E_2 is satisfied. Example F.3 implies that the decomposition is not degenerate, so E_3 is satisfied. Finally, suppose that $F_1 = \cdots = F_n$. Then for each $z \in \mathbb{R}$, the PAV algorithm for the mean functional on $\xi_1(z), \dots, \xi_n(z)$ with respect to the order induced by $F_1(z) = \cdots = F_n(z)$ yields the constant calibrated forecast $\hat{F}_{\mathrm{mg}}(z)$. Hence $\overline{\mathrm{DSC}}_{\mathrm{BS}} = 0$, so that (E_4) is satisfied. □

Remark A.1. The functions $\acute{F}_1, \dots, \acute{F}_n$ are not necessarily increasing and hence they generally fail to be cdfs. For instance, let $n = 2$ and $z < z'$. If $F_1(z) < F_2(z)$, $F_1(z') = F_2(z')$ and $y_2 \le z < z' < y_1$, then $\acute{F}_2(z) = 1 > 1/2 = \acute{F}_2(z')$, so $\acute{F}_2$ is not increasing.

A.2. Quantile score based decomposition

For each level $\alpha \in (0,1)$, we consider $F_1^{-1}(\alpha), \dots, F_n^{-1}(\alpha)$ as point forecasts in the form of the α-quantile. We apply the PAV algorithm for the α-quantile functional on $y_1, \dots, y_n$ with respect to the order induced by $F_1^{-1}(\alpha), \dots, F_n^{-1}(\alpha)$ to yield calibrated α-quantile forecasts $\grave{F}_1^{-1}(\alpha), \dots, \grave{F}_n^{-1}(\alpha)$. This induces the CORP decomposition of the mean quantile score

$$\overline{\mathrm{QS}}_{F^{-1}(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{qs}_{\alpha}\big(F_i^{-1}(\alpha), y_i\big)$$

as described by Gneiting and Resin (2023, Section 3.3) and Gneiting et al. (2023, Section 3.3), namely,

$$\overline{\mathrm{QS}}_{F^{-1}(\alpha)} = \underbrace{\big(\overline{\mathrm{QS}}_{F^{-1}(\alpha)} - \overline{\mathrm{QS}}_{\grave{F}^{-1}(\alpha)}\big)}_{\overline{\mathrm{MCB}}_{\mathrm{QS},\alpha}} - \underbrace{\big(\overline{\mathrm{QS}}_{\grave{F}^{-1}(\alpha)} - \overline{\mathrm{QS}}_{\hat{F}_{\mathrm{mg}}^{-1}(\alpha)}\big)}_{\overline{\mathrm{DSC}}_{\mathrm{QS},\alpha}} + \underbrace{\overline{\mathrm{QS}}_{\hat{F}_{\mathrm{mg}}^{-1}(\alpha)}}_{\overline{\mathrm{UNC}}_{\mathrm{QS},\alpha}},$$

where $\hat{F}_{\mathrm{mg}}^{-1}(\alpha)$ is the quantile function of the marginal empirical law of the outcomes $y_1, \dots, y_n$,

$$\overline{\mathrm{QS}}_{\grave{F}^{-1}(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{qs}_\alpha\big(\grave{F}_i^{-1}(\alpha), y_i\big), \qquad \overline{\mathrm{QS}}_{\hat{F}_{\mathrm{mg}}^{-1}(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{qs}_\alpha\big(\hat{F}_{\mathrm{mg}}^{-1}(\alpha), y_i\big).$$

Integration of the $\overline{\mathrm{MCB}}_{\mathrm{QS},\alpha}, \overline{\mathrm{DSC}}_{\mathrm{QS},\alpha}$ and $\overline{\mathrm{UNC}}_{\mathrm{QS},\alpha}$ components over $\alpha \in (0,1)$ yields the quantile score based decomposition at (17).

For an exact computation, the PAV algorithm needs to be run at all quantile levels l/k, where $k = 1, \dots, n$ and $l = 1, \dots, k-1$, and at all crossing points of the quantile functions $F_1^{-1}, \dots, F_n^{-1}$. In practice, it suffices to apply the PAV algorithm on a fine grid of quantile levels.

Proof of Proposition 2.2. In analogy to the proof of Proposition 2.1, we find that

$$\begin{aligned}\overline{\mathrm{UNC}}_{\mathrm{QS}} &= \int_0^1 \overline{\mathrm{QS}}_{\hat{F}_{\mathrm{mg}}^{-1}(\alpha)}\,\mathrm{d}\alpha = \int_0^1 \frac{1}{n}\sum_{i=1}^{n} \mathrm{qs}_\alpha\big(\hat{F}_{\mathrm{mg}}^{-1}(\alpha), y_i\big)\,\mathrm{d}\alpha \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathrm{qs}_\alpha\big(\hat{F}_{\mathrm{mg}}^{-1}(\alpha), y_i\big)\,\mathrm{d}\alpha = \frac{1}{n}\sum_{i=1}^{n} \mathrm{crps}(\hat{F}_{\mathrm{mg}}, y_i) = \overline{\mathrm{UNC}}_0,\end{aligned}$$

and hence E_5 is satisfied. Property E_1 is clear by definition. Theorem 3.3 of Gneiting and Resin (2023) implies that $\overline{\mathrm{MCB}}_{\mathrm{QS},\alpha}$ and $\overline{\mathrm{DSC}}_{\mathrm{QS},\alpha}$ are nonnegative for all $\alpha \in (0,1)$ and thus E_2 is satisfied. Example F.3 shows that the decomposition is not degenerate, i.e., E_3 is satisfied. Finally, suppose that $F_1 = \cdots = F_n$. Then for each $\alpha \in (0,1)$, applying the PAV algorithm on $y_1, \dots, y_n$ with respect to the order induced by $F_1^{-1}(\alpha) = \cdots = F_n^{-1}(\alpha)$ yields the constant calibrated forecast $\grave{F}^{-1}(\alpha) = \hat{F}_{\mathrm{mg}}^{-1}(\alpha)$ and hence $\overline{\mathrm{DSC}}_{\mathrm{QS}} = 0$, i.e., (E_4) is satisfied. □

Remark A.2. In analogy to the statements in Remark A.1, the functions $\grave{F}_1^{-1}$, $\dots$, $\grave{F}_n^{-1}$ are not necessarily increasing and hence may not be quantile functions. For example, let $n = 2$ and $\alpha < \alpha' < 1/2$, and suppose that $y_1 < y_2$, $F_1^{-1}(\alpha) < F_2^{-1}(\alpha)$, and $F_1^{-1}(\alpha') = F_2^{-1}(\alpha')$. Then $\grave{F}_2^{-1}(\alpha) = y_2 > y_1 = \grave{F}_2^{-1}(\alpha')$ whence $\grave{F}_2^{-1}$ is not increasing.

Appendix B: Technical details for the original and modified Hersbach decompositions

As in Section 2.5, we consider a collection of the form at (8) of forecast–outcome pairs $(F_1, y_1), \dots, (F_n, y_n)$, where for $i = 1, \dots, n$, the forecast F_i is the empirical cdf of a fixed number m of numbers $x_1^i \le \cdots \le x_m^i$. Hersbach (2000) implicitly assumes that $y_i \notin \{x_1^i, \dots, x_m^i\}$ for $i = 1, \dots, n$. If this condition is not satisfied, the extension of the original Hersbach decomposition at (20), which is implemented in the R function `crpsDecomposition` from the verification package (`https://rdrr.io/cran/verification/`), is problematic. Our

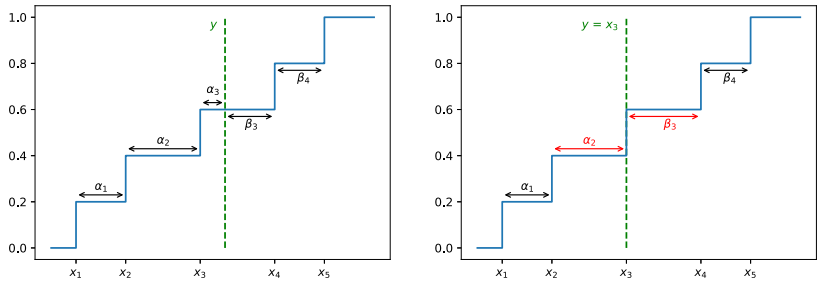

FIG 5. *Adaptation of Figure 2 from Hersbach (2000) with the empirical* cdf *of* $x_1 < \cdots < x_5$ *and outcome* y. *Hersbach (2000) assumes that* $y \notin \{x_1, \ldots, x_5\}$ *and divides the quantity* $x_{\ell+1} - x_\ell$ *for* $\ell = 1, \ldots, m-1$ *into* α_ℓ *and* β_ℓ, *as illustrated in the left panel. When* $y = x_3$ *the original decomposition sets* $\alpha_2 = \beta_3 = 0$. *However, according to display (26) in Hersbach (2000), if* $y \uparrow x_3$ *then* $\alpha_2 \to x_3 - x_2$, $\beta_2 \to 0$, *and* $\beta_3 = x_4 - x_3$, *and if* $y \downarrow x_3$ *then* $\alpha_2 = x_3 - x_2$, $\alpha_3 \to 0$, *and* $\beta_3 \to x_4 - x_3$. *This suggests that* $\alpha_2 = x_3 - x_2$, $\alpha_3 = 0$, $\beta_2 = 0$, *and* $\beta_3 = x_4 - x_3$ *when* $y = x_3$, *as indicated in the right panel and in accordance with the quantity* $\bar{f}_3$ *in the modified Hersbach decomposition.*

suggested modified Hersbach decomposition at (22) resolves this issue, as illustrated graphically in Figure 5.

We proceed to a comparison of the orginal with the modified Hersbach decomposition. For $i = 1, \ldots, n$, Hersbach (2000) defines the quantities

$$\alpha_\ell^i = (x_{\ell+1}^i - x_\ell^i)\,\mathbb{1}\{y_i > x_{\ell+1}\} + (y_i - x_\ell)\,\mathbb{1}\{x_\ell^i < y_i < x_{\ell+1}^i\},$$
$$\beta_\ell^i = (x_{\ell+1}^i - x_\ell^i)\,\mathbb{1}\{y_i < x_\ell^i\} + (x_{\ell+1}^i - y_i)\,\mathbb{1}\{x_\ell^i < y_i < x_{\ell+1}^i\},$$

for $\ell = 1, \ldots, m-1$, and

$$\alpha_m^i = (y_i - x_m^i)\,\mathbb{1}\{y_i > x_m^i\} \quad \text{and} \quad \beta_0^i = (x_1^i - y_i)\,\mathbb{1}\{y_i < x_1^i\}.$$

For $\ell = 1, \ldots, m-1$, let $\bar{\alpha}_\ell = (1/n)\sum_{i=1}^n \alpha_\ell^i$, $\bar{\beta}_\ell = (1/n)\sum_{i=1}^n \beta_\ell^i$, $\bar{g}_\ell = \bar{\alpha}_\ell + \bar{\beta}_\ell$, and $\bar{o}_\ell = \bar{\beta}_\ell / \bar{g}_\ell$. To complete the specification, let $\bar{o}_0 = (1/n)\sum_{i=1}^n \mathbb{1}\{y_i < x_1^i\}$, $\bar{g}_0 = \mathbb{1}\{\bar{o}_0 \neq 0\}\bar{\beta}_0/\bar{o}_0$, $\bar{o}_m = (1/n)\sum_{i=1}^n \mathbb{1}\{x_m^i < y_i\}$, and $\bar{g}_m = \mathbb{1}\{\bar{o}_m \neq 0\}\bar{\alpha}_m/(1-\bar{o}_m)$, where $\bar{\beta}_0 = (1/n)\sum_{i=1}^n \beta_0^i$ and $\alpha_m = (1/n)\sum_{i=1}^n \alpha_m^i$.

As before, let $p_\ell = \ell/m$ for $\ell = 0, \ldots, m$. Hersbach (2000) defines the miscalibration component as

$$\overline{\mathrm{MCB}}_{\mathrm{HBo}} = \sum_{\ell=0}^{m} \bar{g}_\ell \,(p_\ell - \bar{o}_\ell)^2 .$$

In contrast, we let

$$\overline{\mathrm{MCB}}_{\mathrm{HB}} = \sum_{\ell=1}^{m-1} \bar{g}_\ell \left(p_\ell - \bar{f}_\ell\right)^2 ,$$

where $\bar{f}_\ell = (1/n)\sum_{i=1}^n \bar{f}_\ell^i$ with $\bar{f}_\ell^i = (1/\bar{g}_\ell)\,\mathbb{1}\{y_i < x_{\ell+1}^i\}(\alpha_\ell^i + \beta_\ell^i)$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, m-1$. In other words, Hersbach (2000) includes terms

for $\ell = 0$ and $\ell = m$ in the miscalibration component and compares the nominal level p_ℓ with the quantity $\bar{o}_\ell$, which approximates the frequency of an outcome below the midpoint between x^i_ℓ and $x^i_{\ell+1}$. In contrast, we omit the outer terms and compare p_ℓ with $\bar{f}_\ell$, which approximates the frequency of an outcome y_i less than or equal to $x^i_{\ell+1}$.

Proof of Proposition 2.3. By definition, both decompositions are exact and the uncertainty component $\overline{\mathrm{UNC}}_0$ depends only on the outcomes, i.e., E_1 and E_5 are satisfied. Example F.3 shows that E_3 is satisfied, and that E_2 fails to hold for the modified Hersbach decomposition. Consider the sample $(F, y_1), (F, y_2)$ with $F = (\delta_{-1/2} + \delta_{1/2})/2$, $y_1 = -1/6$ and $y_2 = 1/6$. Then $\overline{\mathrm{CRPS}} = 1/4$ and $\overline{\mathrm{UNC}}_0 = 1/12$. Moreover, $\bar{g}_1 = 1$, $\bar{g}_0 = \bar{g}_2 = 0$, $\bar{o}_1 = 1/2$, $\bar{o}_0 = \bar{o}_2 = 0$, and $\bar{f}_1 = 1$. Thus $\overline{\mathrm{MCB}}_{\mathrm{HBo}} = 0$, $\overline{\mathrm{MCB}}_{\mathrm{HB}} = 1/4$, $\overline{\mathrm{DSC}}_{\mathrm{HBo}} = -1/6$, and $\overline{\mathrm{DSC}}_{\mathrm{HB}} = 1/12$. This demonstrates that the original Hersbach decomposition does not satisfy E_2 and E_4 and that E_4 fails to hold for the modified decomposition as well. Numerical examples in Hersbach (2000) show that property E_3 is satisfied for the original Hersbach decomposition. □

Appendix C: Relaxations of the stochastic order

Consider any partial order $\leq'$ on $\mathcal{P}(\mathbb{R})$, which is weaker than the stochastic order in the sense that $G \leq_{\mathrm{st}} H$ implies $G \leq' H$ for $G, H \in \mathcal{P}(\mathbb{R})$. Possible choices include the almost-first-stochastic-dominance order proposed by Leshno and Levy (2002) or stochastic dominance of order $(1+\gamma)$ as proposed by Müller et al. (2017). If there are only few forecasts in a sample $(F_1, y_1), \ldots, (F_n, y_n) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$ that are comparable with respect to $\leq_{\mathrm{st}}$, one could think of applying IDR with respect to $\leq'$ instead of $\leq_{\mathrm{st}}$ in order to obtain more comparable forecasts. In this appendix, we explain why such an approach is bound to fail.

Let Y be a random variable and F be a random forecast defined on the same probability space. Recall from Section 4.2 that ICL forms the population version of IDR (Arnold and Ziegel, 2024, Proposition 4.1). In analogy to Definition 3.1 of Arnold and Ziegel (2024), one could define the σ-lattice generated by F with respect to the weaker order $\leq'$ as $\mathscr{L}'(F) = \{F^{-1}(B) \mid B \in \mathcal{B}(\mathcal{P}(\mathbb{R})) \cap \mathcal{U}'\}$, where $\mathcal{U}'$ denotes the family of all upper sets in $\mathcal{P}(\mathbb{R})$ with respect to $\leq'$. However, if the space $\mathcal{P}(\mathbb{R})$ equipped with the partial order $\leq'$ and the topology of weak convergence satisfies Assumption C.1 of Arnold and Ziegel (2024), the corresponding notion of isotonic calibration, namely, $P_{Y|\mathscr{L}'(F)} = F$, fails to be intuitive for two reasons. First, auto-calibration does not imply the respective notion of calibration. Second, $G \leq' H$ already implies $G \leq_{\mathrm{st}} H$ for all G and H in the support of F by Theorem 3.3 of Arnold and Ziegel (2024). Clearly, this implication may only hold if $\leq'$ equals $\leq_{\mathrm{st}}$ on the support of F, which is violated for any $\leq'$ that is strictly weaker than $\leq_{\mathrm{st}}$, contrary to the scope of a relaxation. Moreover, there is no theoretical guarantee that the corresponding miscalibration term $\mathrm{MCB}_{\mathrm{ISO}'} = \mathbb{E}\,\mathrm{crps}(F, Y) - \mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}'(F)}, Y)$ is nonnegative.

Appendix D: Proofs for Section 4.4 and extensions

Proof of Proposition 4.1. Following the argument in Appendix A of Candille and Talagrand (2005), we apply the change of variable $z \mapsto p = F(z)$ to demonstrate that $\mathbb{E}\,\mathrm{crps}(F, Y)$ can be represented as

$$\mathbb{E}\int_S (F(z) - \mathbb{1}\{F(Y) \le F(z)\})^2\,\mathrm{d}z + \mathbb{E}\int_S (2F(z)-1)(\mathbb{1}\{F(Y) \le F(z)\} - \mathbb{1}\{Y \le z\})\,\mathrm{d}z,$$

where $S = \{z \in \mathbb{R} \mid (F(z) - \mathbb{1}\{Y \le z\})^2 > 0\}$. The indicator is essential, since if $F(Y) = 0$ then $\mathbb{1}\{F(Y) \le F(z)\} = 1$ and the integrals may not exist. We decompose S into the disjoint sets $S_1 = S \cap \{z \in \mathbb{R} \mid F(z) > 0\}$ and $S_2 = S \cap \{z \in \mathbb{R} \mid F(z) = 0\} = \{z \in \mathbb{R} \mid Y \le z, F(z) = 0\}$, and use the equivalence $\mathbb{1}\{F(Y) \le F(z)\} - \mathbb{1}\{Y \le z\} = \mathbb{1}\{Y > z, F(Y) = F(z)\}$ to show that

$$\begin{aligned}\mathbb{E}\,\mathrm{crps}(F, Y) &= \mathbb{E}\int_{S_1} (F(z) - \mathbb{1}\{F(Y) \le F(z)\})^2\,\mathrm{d}z \\ &\quad + \mathbb{E}\int_{S_2} \mathbb{1}\{Y \le z, F(z) = 0\}\,\mathrm{d}z \\ &\quad + \mathbb{E}\int_S (2F(Y)-1)\,\mathbb{1}\{Y > z, F(Y) = F(z)\}\,\mathrm{d}z \\ &= \mathbb{E}\int_{S_1} (F(z) - \mathbb{1}\{F(Y) \le F(z)\})^2\,\mathrm{d}z + \mathrm{MS},\end{aligned}$$

where MS is given at (42).

We have $\tau(A) \le \mathbb{E}\int_A 1\,\mathrm{d}\nu_F(u) = \mathbb{E}(\nu_F(A)) = \mu(A)$ for $A \in \mathcal{B}(0,1)$, i.e., τ is absolutely continuous with respect to μ. Hence τ has a density f with respect to μ, and we find that

$$\begin{aligned}\mathbb{E}\,\mathrm{crps}(F, Y) &= \mathbb{E}\int_S (F(z) - \mathbb{1}\{F(Y) \le F(z)\})^2\,\mathrm{d}z + \mathrm{MS} \\ &= \mathbb{E}\int_0^1 (p - \mathbb{1}\{F(Y) \le p\})^2\,\mathrm{d}\nu_F(p) + \mathrm{MS} \\ &= \int_0^1 p^2\,\mathrm{d}\mu(p) - \int_0^1 (2p-1)\;\mathrm{d}\tau(p) + \mathrm{MS} \\ &= \int_0^1 p^2\,\mathrm{d}\mu(p) - \int_0^1 (2p-1)\,f(p)\,\mathrm{d}\mu(p) + \mathrm{MS} \\ &= \int_0^1 (p - f(p))^2\,\mathrm{d}\mu(p) + \int_0^1 f(p)\,(1-f(p))\,\mathrm{d}\mu(p) + \mathrm{MS},\end{aligned}$$

which yields the claimed decomposition. □

In the following corollary to Proposition 4.1, which is a more general result than Corollary 4.2, we consider forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where for each $i = 1, \ldots, n$, F_i is a distribution with a finite number m_i of support points $x^i_1 < \cdots < x^i_{m_i}$ and (cumulative) probability values $p^i_1 < \cdots < p^i_{m_i}$, so that $F_i(x^i_\ell) = p^i_\ell$ for $\ell = 1, \ldots, m_i$. Let $0 < \hat{p}_1 < \ldots < \hat{p}_M = 1$ be the unique probability values from the set $\{p^i_\ell \mid i = 1, \ldots, n;\ \ell = 1, \ldots, m_i\}$. For $i = 1, \ldots, n$ and $j = 1, \ldots, M-1$, we define

$$\sigma^i_j = \begin{cases} \ell & \text{if} \quad \hat{p}_j = p^i_\ell, \\ 0 & \text{if} \quad \hat{p}_j \notin \{p^i_1, \ldots, p^i_{m_i}\}. \end{cases}$$

Corollary D.1. *Assume that $\mathbb{P}$ is the empirical measure of forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where each F_i is a distribution with finite support as described above. Then*

$$\mathrm{MCB}_{\mathrm{HB}} = \sum_{j=1}^{M-1} \hat{g}_j (\hat{p}_j - \hat{f}_j)^2 \tag{46}$$

where, for $j = 1, \ldots, M-1$,

$$\hat{g}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\sigma^i_j \neq 0\} \left(x^i_{\sigma^i_j + 1} - x^i_{\sigma^i_j} \right), \tag{47}$$

$$\hat{f}_j = \frac{1}{n \hat{g}_j} \sum_{i=1}^{n} \mathbb{1}\{F_i(y_i) \leq \hat{p}_j\} \mathbb{1}\{\sigma^i_j \neq 0\} \left(x^i_{\sigma^i_j + 1} - x^i_{\sigma^i_j} \right). \tag{48}$$

Proof. For $i = 1, \ldots, n$, let ν_i be the image measure of F_i with respect to the Lebesgue measure, i.e.,

$$\nu_i = \sum_{j=1}^{M-1} \delta_{\hat{p}_j} \mathbb{1}\{\sigma^i_j \neq 0\} \left(x^i_{\sigma^i_j + 1} - x^i_{\sigma^i_j} \right),$$

and thus, $\mu = \sum_{j=1}^{M-1} \delta_{\hat{p}_j} \hat{g}_j$, where $\hat{g}_j$ is given at (47). Therefore, for any $A \in \mathcal{B}(0,1)$, we have

$$\begin{aligned} \tau(A) &= \mathbb{E} \int_A \mathbb{1}\{F(Y) \leq u\} \, \mathrm{d}\nu_F(u) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M-1} \delta_{\hat{p}_j}(A) \mathbb{1}\{F_i(y_i) \leq \hat{p}_j\} \mathbb{1}\{\sigma^i_j \neq 0\} \left(x^i_{\sigma^i_j + 1} - x^i_{\sigma^i_j} \right) \\ &= \sum_{j=1}^{M-1} \delta_{\hat{p}_j}(A) \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{F_i(y_i) \leq \hat{p}_j\} \mathbb{1}\{\sigma^i_j \neq 0\} \left(x^i_{\sigma^i_j + 1} - x^i_{\sigma^i_j} \right) \\ &= \sum_{j=1}^{M-1} \delta_{\hat{p}_j}(A) \hat{f}_j \hat{g}_j. \end{aligned}$$

We conclude that the Radon–Nikodym derivative of τ with respect to μ is $f(\hat{p}_j) = \hat{f}_j$ for $j = 1, \ldots, M-1$, where $\hat{f}_j$ is given at (48). □

To specialize Corollary D.1 to the ensemble setting of Corollary 4.2, let $m_i = m$ and $p_\ell^i = \ell/m$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, m-1$. Then $M = m$, $\hat{p}_\ell = \ell/m$, and the quantities in (18) and (47), and in (19) and (48), respectively, coincide.

Proof of Corollary 4.3. Since F^{-1} is almost surely absolutely continuous, for any $0 < a < b < 1$, we have almost surely

$$\nu_F([a,b)) = \lambda(F^{-1}([a,b))) = F^{-1}(b) - F^{-1}(a) = \int_a^b \frac{\mathrm{d}}{\mathrm{d}p} F^{-1}(p)\,\mathrm{d}p.$$

That is, the random measure ν_F almost surely possesses a density $(\mathrm{d}/\mathrm{d}p)\,F^{-1}(p)$ with respect to the Lebesgue measure, and it follows that the measure μ has density γ at (43) with respect to the Lebesgue measure. Since for $A \in \mathcal{B}(0,1)$,

$$\tau(A) = \mathbb{E}\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p) = \int_A \mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\,\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right)\,\mathrm{d}p,$$

the density f of the measure τ with respect to μ is given as stated at (44). □

The following example relates to the case study on probabilistic quantitative precipitation forecasts in Section 5.1, where it applies to the BMA, EMOS, and HCLR forecasts, respectively.

Example D.1. Let $(F_1, y_1), \ldots, (F_n, y_n)$ be forecast–observation pairs for a nonnegative (possibly, censored) quantity, so that $y_i \ge 0$ for $i = 1, \ldots, n$. Suppose that, for $i = 1, \ldots, n$,

$$F_i(x) = \begin{cases} 0 & \text{for} \quad x < 0, \\ p_0^i + \int_0^x f_i(t)\,\mathrm{d}t & \text{for} \quad x \ge 0, \end{cases}$$

for some $0 \le p_0^i < 1$ and a strictly positive continuous function $f_i : (0,\infty) \to \mathbb{R}_+$ with $\int_0^\infty f_i(t)\,\mathrm{d}t = 1 - p_0^i$. Then F_i^{-1} is absolutely continuous and has derivative $f_i(F_i^{-1}(p))^{-1}$ for $p \in (p_0^i, 1)$ and zero otherwise. Hence, $\overline{\mathrm{MCB}}_{\mathrm{HB}} = \int_0^1 (p - f(p))^2\,\gamma(p)\,\mathrm{d}p$ by Corollary 4.3, where

$$\gamma(p) = \frac{1}{n}\sum_{i=1}^n \frac{1}{f_i(F_i^{-1}(p))}\,\mathbb{1}_{(p_0^i,1)}(p)$$

and

$$f(p) = \frac{1}{n\gamma(p)}\sum_{i=1}^n \mathbb{1}\{F_i(y_i) \le p\}\frac{1}{f_i(F_i^{-1}(p))}\,\mathbb{1}_{(p_0^i,1)}(p)$$

for $p \in (0,1)$ with $\gamma(p) > 0$, and $f(p) = 0$ otherwise.

Appendix E: Proofs for Section 4.5

Proof of Theorem 4.4. Concerning part (a), we consider the Brier score based decomposition of $\overline{\mathrm{CRPS}}$ and apply Fubini's theorem to obtain

$$\mathrm{MCB}_{\mathrm{CT}} = \int \Big(\mathbb{E}\big(F(z) - \mathbb{1}\{Y \le z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y \le z \mid F) - \mathbb{1}\{Y \le z\}\big)^2 \Big)\, \mathrm{d}z, \tag{49}$$

$$\mathrm{DSC}_{\mathrm{CT}} = \int \Big(\mathbb{E}\big(F_{\mathrm{mg}}(z) - \mathbb{1}\{Y \le z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y \le z \mid F) - \mathbb{1}\{Y \le z\}\big)^2 \Big)\, \mathrm{d}z. \tag{50}$$

Recall that for any $z \in \mathbb{R}$, the expectation $\mathbb{E}\,(\mathbb{1}\{Y \le z\} - p)^2$ is minimized by $\mathbb{P}(Y \le z \mid F)$ over all $\sigma(F)$-measurable random variables p, and this minimizer is $\mathbb{P}$-almost surely unique. Since $F(z)$ and the constant $F_{\mathrm{mg}}(z)$ are $\sigma(F)$-measurable, it follows from (49) and (50) that $\mathrm{MCB}_{\mathrm{CT}} \ge 0$ and $\mathrm{DSC}_{\mathrm{CT}} \ge 0$, respectively. Equality in (49) holds if, and only if, F is auto-calibrated. Equality in (50) holds if, and only if, $P_{Y|F} = F_{\mathrm{mg}}$, i.e., $\mathbb{P}(Y \le z \mid F) = F_{\mathrm{mg}}(z)$ for all $z \in \mathbb{R}$.

For part (b), in analogy to the above, we find that

$$\mathrm{MCB}_{\mathrm{ISO}} = \int \Big(\mathbb{E}\big(\bar{F}(z) - \mathbb{1}\{Y > z\}\big)^2 - \mathbb{E}\,\big(\mathbb{P}(Y > z \mid \mathscr{L}(F)) - \mathbb{1}\{Y > z\}\big)^2 \Big) \mathrm{d}z, \tag{51}$$

$$\mathrm{DSC}_{\mathrm{ISO}} = \int \Big(\mathbb{E}\big(\bar{F}_{\mathrm{mg}}(z) - \mathbb{1}\{Y > z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y > z \mid \mathscr{L}(F)) - \mathbb{1}\{Y > z\}\big)^2 \Big) \mathrm{d}z, \tag{52}$$

where $\bar{F}(z) = 1 - F(z)$, and $\bar{F}_{\mathrm{mg}}(z) = 1 - F_{\mathrm{mg}}(z)$. Recall that for any $z \in \mathbb{R}$, the expectation $\mathbb{E}(\mathbb{1}\{Y > z\} - p)^2$ is minimized by $\mathbb{P}(Y > z \mid \mathscr{L}(F))$ over all $\mathscr{L}(F)$-measurable random variables p, and the minimizer is $\mathbb{P}$-almost surely unique. Since $\bar{F}(z)$ and the constant $\bar{F}_{\mathrm{mg}}(z)$ are $\mathscr{L}(F)$-measurable, it follows directly that $\mathrm{MCB}_{\mathrm{ISO}} \ge 0$ and $\mathrm{DSC}_{\mathrm{ISO}} \ge 0$. Equality in (51) holds if, and only if, F is isotonically calibrated, and equality in (52) holds if, and only if, $P_{Y|\mathscr{L}(F)} = F_{\mathrm{mg}}$.

To demonstrate part (c), it suffices to observe from Arnold and Ziegel (2024, Lemma 5.4) that threshold calibration is equivalent to $\mathbb{P}(Y \le z \mid \mathscr{L}(F(z))) = F(z)$ for $z \in \mathbb{R}$. The rest of the argument is analogous to the above.

Finally, for part (d), recall that for $\alpha \in (0,1)$, a random variable is a conditional quantile $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha)))$ if, and only if, it minimizes $\mathbb{E}\,\mathrm{QS}_\alpha(X,Y)$ over all $\mathscr{L}(F^{-1}(\alpha))-$ measurable random variables X, see Arnold and Ziegel (2024). It follows that $\mathrm{MCB}_{\mathrm{QS}} \ge 0$ and $\mathrm{DSC}_{\mathrm{QS}} \ge 0$. Assume that F is quantile calibrated; then $q_\alpha\big(Y \mid \mathscr{L}\big(F^{-1}(\alpha)\big)\big) = F^{-1}(\alpha)$ for $\alpha \in (0,1)$ and hence $\mathrm{MCB}_{\mathrm{QS}} = 0$. Conversely, if $\mathrm{MCB}_{\mathrm{QS}} = 0$ then Fubini's theorem implies

$$\int_0^1 \big(\mathbb{E}\,\mathrm{qs}_\alpha\big(F^{-1}(\alpha), Y\big) - \mathbb{E}\,\mathrm{qs}_\alpha\big(q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big), Y\big) \big)\, \mathrm{d}\alpha = 0.$$

Since the integrand is non-negative, it follows that $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = F^{-1}(\alpha)$ for almost all $\alpha \in (0,1)$ and, hence, there exists a Lebesgue null set $N \subseteq (0,1)$ with $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))) = F^{-1}(\alpha)$ for all $\alpha \in (0,1) \setminus N$.

Assume for a contradiction that $N \neq \emptyset$ and consider $\alpha_0 \in N$. Choose $(\alpha_n)_{n\in\mathbb{N}} \subseteq (0,1) \setminus N$ with $\alpha_n \uparrow \alpha_0$ as $n \to \infty$. Since $F^{-1}(\alpha_n) \to F^{-1}(\alpha_0)$ almost surely and $\mathrm{qs}_{\alpha_n}(\cdot, y) \to \mathrm{qs}_{\alpha_0}(\cdot, y)$ pointwise for any $y \in \mathbb{R}$, it follows that $\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y) \to \mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ almost surely, and hence, $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y) \to \mathbb{E}\,\mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ by dominated convergence. Analogously, $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(X, Y) \to \mathbb{E}\,\mathrm{QS}_{\alpha_0}(X, Y)$ for $X = q_{\alpha_0}(Y \mid \mathscr{L}(F^{-1}(\alpha_0)))$ and $\mathbb{E}\,\mathrm{qs}_{\alpha_0}(X, Y) \geq \mathbb{E}\,\mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ since $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(X, Y) \geq \mathbb{E}\,\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y)$ for all $n \in \mathbb{N}$. This shows that $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big)$ is an α-quantile of F for $\alpha \in (0,1)$. By construction in Section 6 of Arnold and Ziegel (2024), $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big)$ is the smallest possible minimizer of $\mathbb{E}\,\mathrm{qs}_\alpha(X, Y)$, so it coincides with $F^{-1}(\alpha)$ for all $\alpha \in (0,1)$ and, hence, $N = \emptyset$. Clearly, $\mathrm{DSC}_{\mathrm{QS}} = 0$ if $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = q_\alpha(Y)$ for $\alpha \in (0,1)$. Conversely, if $\mathrm{DSC}_{\mathrm{QS}} = 0$ then $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = q_\alpha(Y)$ for $\alpha \in (0,1)$. □

Proof of Corollary 4.6. For any $z \in \mathbb{R}$, $P_{Y|F}(\cdot, (z,\infty))$ minimizes $\mathbb{E}(p - \mathbb{1}\{Y > z\})^2$ over all $\sigma(F)$-measurable random variables p, and hence, also over all $\mathscr{L}(F)$-measurable random variables since any $\mathscr{L}(F)$-measurable random variable is also $\sigma(F)$-measurable, see Arnold and Ziegel (2024, Lemma 3.1). Thus, we apply Fubini to derive

$$\begin{aligned}
\mathbb{E}\,\mathrm{crps}(P_{Y|F}, Y) &= \int \mathbb{E}\,(P_{Y|F}(\cdot, (z,\infty)) - \mathbb{1}\{Y > z\})^2\,\mathrm{d}z \\
&\leq \int \mathbb{E}\,(P_{Y|\mathscr{L}(F)}(\cdot, (z,\infty)) - \mathbb{1}\{Y > z\})^2\,\mathrm{d}z \\
&= \mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y),
\end{aligned}$$

which implies $\mathrm{MCB}_{\mathrm{CT}} \geq \mathrm{MCB}_{\mathrm{ISO}}$. Moreover, for any $z \in \mathbb{R}$ we know that $\mathscr{L}(F(z)) \subseteq \overline{\mathscr{L}(F)}$, where for any σ-lattice $\mathcal{A} \subseteq \mathcal{F}$, $\bar{\mathcal{A}}$ denotes the σ-lattice which consists of all complements of elements in $\mathcal{A}$. Hence, we may argue similarly that

$$\begin{aligned}
\mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y) &= \int \mathbb{E}(1 - P_{Y|\mathscr{L}(F)}(\cdot, (z,\infty)) - \mathbb{1}\{Y \leq z\})^2\,\mathrm{d}z \\
&\leq \int \mathbb{E}(\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) - \mathbb{1}\{Y \leq z\})^2\,\mathrm{d}z,
\end{aligned}$$

which implies $\mathrm{MCB}_{\mathrm{ISO}} \geq \mathrm{MCB}_{\mathrm{BS}}$. Finally for any $\alpha \in (0,1)$, we have that $P^{-1}_{Y|\mathscr{L}(F)}(\alpha)$ minimizes $\mathbb{E}\,\mathrm{qs}_\alpha(X, Y)$ over all $\mathscr{L}(F)$-measurable random variables X. We use that $\mathscr{L}(F^{-1}(\alpha)) \subseteq \mathscr{L}(F)$, to derive that

$$\begin{aligned}
\mathbb{E}\,\mathrm{crps}(P_{Y|\mathscr{L}(F)}, Y) &= \int_0^1 \mathbb{E}\,\mathrm{qs}_\alpha(P^{-1}_{Y|\mathscr{L}(F)}(\alpha), Y)\,\mathrm{d}\alpha \\
&\leq \int_0^1 \mathbb{E}\,\mathrm{qs}_\alpha(q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha)), Y)\,\mathrm{d}\alpha
\end{aligned}$$

and hence $\mathrm{MCB}_{\mathrm{ISO}} \geq \mathrm{MCB}_{\mathrm{QS}}$. □

Proof of Proposition 4.7. The claim in part (a) follows from the definition of MS at (42). For part (b), suppose that F is auto-calibrated. Then $Y \in \operatorname{supp}(F)$ almost surely and hence MS $= 0$ by part (a). The tower property implies for any $A \in \mathcal{B}(0,1)$ that

$$\begin{aligned}\tau(A) &= \mathbb{E}\left(\mathbb{E}\left(\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p) \;\middle|\; F\right)\right)\\ &= \mathbb{E}\left(\int_A \mathbb{E}\left(\mathbb{1}\{F(Y) \le p\} \mid F\right)\,\mathrm{d}\nu_F(p)\right)\\ &= \mathbb{E}\left(\int_A F(F^{-1}(p))\,\mathrm{d}\nu_F(p)\right),\end{aligned}$$

where the last equality follows since if $Y \in \operatorname{supp}(F)$, then $F(Y) \le p$ if and only if $Y \le F^{-1}(p)$ and $\mathbb{P}(Y \le F^{-1}(p) \mid F) = F(F^{-1}(p))$ by auto-calibration. By the properties of generalized inverses (Embrechts and Hofert, 2013), we have $F(F^{-1}(p)) \ge p$ for all $p \in (0,1)$. However, if $F(F^{-1}(p)) > p$ for all $p \in B$ in some $B \in \mathcal{B}(0,1)$, then $F^{-1}(B) = \{x \in \mathbb{R} \mid F(x) \in B\} = \emptyset$ and hence $\nu_F(B) = 0$ almost surely. That is, $\nu_F(\{p \in (0,1) : F(F^{-1}(p)) > p\} = 0$ almost surely and thus

$$\tau(A) = \mathbb{E}\left(\int_A F(F^{-1}(p))\,\mathrm{d}\nu_F(p)\right) = \mathbb{E}\left(\int_A p\,\mathrm{d}\nu_F(p)\right) = \int_A p\,\mathrm{d}\mu(p).$$

We conclude that $f(p) = p$ μ-almost surely and hence $\mathrm{MCB}_{\mathrm{HB}} = 0$.

The condition in part (c) is equivalent to assuming that $\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}$ is almost surely constant for all $p \in (0,1)$. Since F is probabilistically calibrated, we have for any $p \in (0,1)$,

$$\begin{aligned}f(p) &= \frac{1}{\gamma(p)}\mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right)\\ &= \frac{\gamma(p)}{\gamma(p)}\mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\right) = \mathbb{P}(F(Y) \le p) = p\end{aligned}$$

and hence $\mathrm{MCB}_{\mathrm{HB}} = 0$. □

Appendix F: Analytic examples at the population level

In this section we compare the population level decompositions from Section 4 in a number of examples in the prediction space setting. Table 4 collects and summarizes the analytic forms of the decomposition components in these examples. Assumption 4.1 is satisfied throughout.

F.1. Auto-calibrated Gaussian

In this example, the predictive distribution F is Gaussian with mean μ_i and standard deviation $\sigma_i > 0$ with probability w_i for $i = 1, \dots, n$, where $w_1 +$

TABLE 4
Analytic form of the various different types of decomposition in population level examples F.1, . . . ,F.5. For details and supporting calculations see the text.

Example	F.1	F.2	F.3	F.4	F.5
$\mathbb{E}\,\mathrm{crps}(F,Y)$	$\sum_{i=1}^n w_i \frac{\sigma_i}{\sqrt{\pi}}$	$\frac{1}{6}$	1	$\frac{39}{80}$	$\frac{5}{24}t$
$\mathrm{MCB}_{\mathrm{CT}}$	0	$\frac{1}{30}$	1	$\frac{7}{400}$	$\frac{3}{200}t$
$\mathrm{MCB}_{\mathrm{ISO}}$	0	$\frac{1}{30}$	1	$\frac{9}{2800}$	$\frac{3}{200}t_2$
$\mathrm{MCB}_{\mathrm{QS}}$	0	$\frac{1}{30}$	$\frac{13}{16}$	$\frac{9}{2800}$	0
$\mathrm{MCB}_{\mathrm{BS}}$	0	$\frac{1}{30}$	$\frac{1}{2}$	$\frac{9}{2800}$	0
$\mathrm{MCB}_{\mathrm{HB}}$	0	0	$\frac{1}{8}$	$\frac{1}{1600}$	0
UNC_0	$\frac{1}{2}\sum_{i,j=1}^n w_i w_j\, A(\mu_i-\mu_j,\sigma_i^2+\sigma_j^2)$	$\frac{2}{5}$	$\frac{3}{4}$	$\frac{3}{2}$	$\frac{2}{9}t$

$\cdots + w_n = 1$. Conditionally on F, the outcome Y has distribution F, so F is auto-calibrated. We conclude that

$$\mathrm{MCB}_{\mathrm{CT}} = \mathrm{MCB}_{\mathrm{ISO}} = \mathrm{MCB}_{\mathrm{BS}} = \mathrm{MCB}_{\mathrm{QS}} = 0.$$

Since auto-calibration implies probabilistic calibration, Proposition 4.7 yields $\mathrm{MCB}_{\mathrm{HB}} = \mathrm{MS}_{\mathrm{HB}} = 0$. Finally, we apply formulas in Grimit et al. (2006) to obtain

$$\mathbb{E}\,\mathrm{crps}(F,Y) = \sum_{i=1}^n w_i \frac{\sigma_i}{\sqrt{\pi}} \quad \text{and} \quad \mathrm{UNC}_0 = \frac{1}{2}\sum_{i,j=1}^n w_i w_j A(\mu_i-\mu_j,\sigma_i^2+\sigma_j^2),$$

where $A(\mu,\sigma^2) = 2\sigma\varphi(\frac{\mu}{\sigma}) + \mu(2\Phi(\frac{\mu}{\sigma})-1)$, with φ and Φ denoting the density and the cdf of the standard normal distribution, respectively.

F.2. Example in Candille and Talagrand (2005)

In this example of Candille and Talagrand (2005, p. 2145), the forecast F is F_1, which is uniform on $(-1,0)$, or F_2, which is uniform on $(0,1)$, with equal probability. Given $F = F_1$, the conditional cdf of Y is $Q_1(z) = 1 - z^2$ for $z \in (-1,0)$, and given $F = F_2$, the conditional cdf of Y is $Q_2(z) = z^2$ for $z \in (0,1)$.

For $i = 1,2$, we denote by G_i the isotonic conditional law of Y given $F = F_i$. Since $F_1 \leq_{\mathrm{st}} F_2$ and $Q_1 \leq_{\mathrm{st}} Q_2$ it follows that $Q_i = G_i$ for $i = 1,2$ and the isotonicity-based decomposition coincides with the Candille–Talagrand decomposition. For any $z \in (-1,1)$, $F_1(z)$ and $F_2(z)$ strictly order and hence the random variable $F(z)$ already reveals the value of F. That is, $\sigma(F(z)) = \sigma(F)$ and hence $\mathbb{P}(Y \leq z \mid F(z)) = \mathbb{P}(Y \leq z \mid F) = P_{Y|F}(z)$. Since this conditional probability is already an increasing function of $F(z)$, we may conclude by Proposition 3.2. in Arnold and Ziegel (2024) that $\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) = P_{Y|F}(z)$ for all $z \in \mathbb{R}$ and hence the Brier score based decomposition correspond with the

Candille–Talagrand decomposition. Analogously the claim can be shown for the quantile score based decomposition. Thus the isotonicity-based, Brier score based, and quantile score based decompositions coincide with the Candille–Talagrand decomposition, where $\mathbb{E}\,\mathrm{crps}(F,Y) = 1/6$, $\mathrm{MCB}_{\mathrm{CT}} = 1/30$, and $\mathrm{UNC}_0 = 2/5$.

The forecasts satisfy the conditions in part (c) of Proposition 4.7, therefore $\mathrm{MCB}_{\mathrm{HB}} = 0$. Since $Y \in \mathrm{supp}(F)$ almost surely, we have $\mathrm{MS} = 0$.

F.3. Example with two atoms

This simple example illustrates that the Brier score and quantile score based decompositions do not coincide in general, that the corresponding calibration methods do not necessarily produce valid cdfs or quantile functions, respectively, and that $\mathrm{DSC}_{\mathrm{HB}}$ can be negative.

Consider the distributions $F_1 = (\delta_1 + \delta_2)/2$ and $F_2 = (\delta_0 + \delta_3)/2$, where δ_z denotes the Dirac measure at $z \in \mathbb{R}$. Assume that F is F_1 and F_2 with equal probability and that $Y = y_1$ if $F = F_1$ and $Y = y_2$ if $F = F_2$. Let $y_1 = 3$ and $y_2 = 0$, so the marginal law F_{mg} of Y is F_2. We readily compute $\mathbb{E}\,\mathrm{crps}(F,Y) = 1$ and $\mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}},Y) = \mathrm{UNC}_0 = 3/4$.

An application of the PAV algorithm for the mean functional on $(\mathbb{1}\{y_1 \le z\}, \mathbb{1}\{y_2 \le z\})$ with respect to the order induced by $(F_1(z), F_2(z))$ at threshold $z \in \mathbb{R}$ results in

$$\acute{F}_1(z) = \tfrac{1}{2}\mathbb{1}_{[1,3)}(z) + \mathbb{1}_{[3,\infty)}(z) \quad \text{and} \quad \acute{F}_2(z) = \mathbb{1}_{[0,1)}(z) + \tfrac{1}{2}\mathbb{1}_{[1,3)}(z) + \mathbb{1}_{[3,\infty)}(z),$$

and we see that $\acute{F}_2$ fails to be increasing. Similarly, an application of the PAV algorithm for the α-quantile on (y_1, y_2) with respect to the order induced by $(F_1^{-1}(\alpha), F_2^{-1}(\alpha))$ at level $\alpha \in (0,1)$ results in

$$\grave{F}_1^{-1}(\alpha) = 3 \quad \text{and} \quad \grave{F}_2^{-1}(\alpha) = 3\mathbb{1}_{(\frac{1}{2},1]}(\alpha).$$

It follows easily that $\mathrm{MCB}_{\mathrm{BS}} = 1/2 \neq 13/16 = \mathrm{MCB}_{\mathrm{QS}}$. As the conditional law of Y given F is a Dirac measure, $\mathbb{E}\,\mathrm{crps}(P_{Y|F},Y) = 0$ and $\mathrm{MCB}_{\mathrm{CT}} = 1$. Similarly, $\mathrm{MCB}_{\mathrm{ISO}} = 1$ since F_1 and F_2 do not order.

According to the formulas in Section 2.5, $\bar{g}_1 = 2$ and $\bar{f}_1 = (\mathbb{1}\{F_1(y_1) \le \frac{1}{2}\} + 3\mathbb{1}\{F_2(y_2) \le 1/2\})/(2\bar{g}_1) = 3/4$ and thus $\mathrm{MCB}_{\mathrm{HB}} = (p_1 - \bar{f}_1)^2\,\bar{g}_1 = 1/8$, whence we conclude that $\mathrm{DSC}_{\mathrm{HB}} = \mathrm{MCB}_{\mathrm{HB}} + \mathrm{UNC}_0 - \mathbb{E}\,\mathrm{crps}(F,Y) = -1/8$.

F.4. Example 2.4 a) in Gneiting and Resin (2023)

Let F be a mixture of uniform distributions on $[0,1]$, $[1,2]$, and $[2,3]$ with weights p_1, p_2, and p_3, respectively, and let Y be drawn from a mixture of these distributions with weights q_1, q_2, and q_3, respectively, where the tuple $(p_1, p_2, p_3; q_1, q_2, q_3)$ attains each of the values

$$\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}; \tfrac{5}{10}, \tfrac{1}{10}, \tfrac{4}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}; \tfrac{1}{10}, \tfrac{8}{10}, \tfrac{1}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}; \tfrac{4}{10}, \tfrac{1}{10}, \tfrac{5}{10}\right)$$

with equal probability. We note that F is probabilistically calibrated, and still we find that $\mathrm{MCB}_{\mathrm{HB}} \neq 0$.

Let F_1, F_2, and F_3 denote the distributions that F attains. For $i = 1, 2, 3$, let Q_i be the conditional law of Y given $F = F_i$, and let G_i be the isotonic conditional law of Y given $F = F_i$. The marginal law F_{mg} of Y is uniform on $[0, 3]$ and, hence,

$$\begin{aligned}\mathrm{UNC}_0 = \mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}}, Y) &= \int\int (F_{\mathrm{mg}}(x) - \mathbb{1}\{y \le x\})^2 \,\mathrm{d}x\,\mathrm{d}F_{\mathrm{mg}}(y) \\ &= \frac{1}{3}\int_0^3 \int_0^3 \Big(\frac{x}{3} - \mathbb{1}\{y \le x\}\Big)^2 \,\mathrm{d}x\,\mathrm{d}y = \frac{3}{2}.\end{aligned}$$

It holds that $F_1 \le_{\mathrm{st}} F_2 \le_{\mathrm{st}} F_3$ but only $Q_1 \le_{\mathrm{st}} Q_3$, hence $P_{Y|F} \neq P_{Y|\mathscr{L}(F)}$. Let $r = 10/7$, $s = 11/7$. On $(-\infty, r]$, we have the pointwise inequalities $Q_2 \le Q_3 \le Q_1$; on $[r, s]$, we have $Q_3 \le Q_2 \le Q_1$; and on $[s, \infty)$, we have $Q_3 \le Q_1 \le Q_2$. Consider the pooled cdfs $Q_{12} = (Q_1 + Q_2)/2$ and $Q_{23} = (Q_2 + Q_3)/2$. The G_i's may be derived by pooling the Q_i's according to the given order constraint $G_1 \le_{\mathrm{st}} G_2 \le_{\mathrm{st}} G_3$, namely,

$$\begin{aligned}G_1(z) &= Q_1(z)\mathbb{1}_{(-\infty,s]}(z) + Q_{12}(z)\mathbb{1}_{[s,\infty)}(z), \\ G_2(z) &= Q_{23}(z)\mathbb{1}_{(-\infty,r]}(z) + Q_2(z)\mathbb{1}_{[r,s]}(z) + Q_{12}(z)\mathbb{1}_{[s,\infty)}(z), \\ G_3(x) &= Q_{23}(z)\mathbb{1}_{(-\infty,r]}(x) + Q_3(z)\mathbb{1}_{[r,\infty)}(z).\end{aligned}$$

By the law of total expectation and Fubini's theorem,

$$\begin{aligned}\mathbb{E}\,\mathrm{crps}(F, Y) &= \frac{1}{3}\sum_{i=1}^3 \mathbb{E}\left(\mathrm{crps}(F, Y) \mid F = F_i\right) \\ &= \frac{1}{3}\sum_{i=1}^3 \int\int \big(F_i(x) - \mathbb{1}\{y \le x\}\big)^2 \,\mathrm{d}x\,\mathrm{d}Q_i(y) \\ &= \frac{1}{3}\sum_{i=1}^3 \int\int \big(F_i(x) - \mathbb{1}\{y \le x\}\big)^2 \,\mathrm{d}Q_i(y)\,\mathrm{d}x \\ &= \frac{1}{3}\sum_{i=1}^3 \int \big(F_i^2(x) - 2F_i(x)Q_i(x) + Q_i(x)\big)\,\mathrm{d}x.\end{aligned}$$

Similarly, we find that $\mathbb{E}\,\mathrm{crps}(G, Y) = (1/3)\sum_{i=1}^3 \int (G_i^2(x) - 2G_i(x)Q_i(x) + Q_i(x))\,\mathrm{d}x$ and $\mathbb{E}\,\mathrm{crps}(Q, Y) = (1/3)\sum_{i=1}^3 \int (Q_i(x) - Q_i^2(x))\,\mathrm{d}x$; and therefore $\mathbb{E}\,\mathrm{crps}(F, Y) = 39/80$, $\mathbb{E}\,\mathrm{crps}(G, Y) = 339/700$, and $\mathbb{E}\,\mathrm{crps}(Q, Y) = 47/100$. We conclude that

$$\mathrm{MCB}_{\mathrm{CT}} = \tfrac{39}{80} - \tfrac{47}{100} = \tfrac{7}{400} \quad \text{and} \quad \mathrm{MCB}_{\mathrm{ISO}} = \tfrac{39}{80} - \tfrac{339}{700} = \tfrac{9}{2800}.$$

Since the predictive distributions are ordered with respect to $\le_{\mathrm{st}}$, it follows that for every threshold z, the ordering of $F_i(z)$ is the same. For $z \in (-\infty, 1]$,

$F_2(z)$ and $F_3(z)$ coincide but this also holds for $G_2(z)$ and $G_3(z)$. Similarly, for $z \in [2, \infty)$, $F_1(z)$ and $F_2(z)$ coincide but this also holds for $G_1(z)$ and $G_2(z)$. This implies that the Brier score based and the isotonocity-based decompositions coincide. Since the stochastic order is equivalently characterized by pointwise orderings of lower quantile functions, the quantile score based and the isotonicity-based decompositions also coincide.

As all F_i^{-1}'s are absolutely continuous, we may apply Corollary 4.3 to compute $\mathrm{MCB}_{\mathrm{HB}}$. For $p \in (0,1) \setminus \{1/4, 1/2, 3/4\}$ we find that

$$\frac{\mathrm{d}}{\mathrm{d}p} F_1^{-1}(p) = 2\mathbb{1}_{(0,\frac{1}{2})}(p) + 4\mathbb{1}_{(\frac{1}{2},1)}(p), \quad \frac{\mathrm{d}}{\mathrm{d}p} F_3^{-1}(p) = 4\mathbb{1}_{(0,\frac{1}{2})}(p) + 2\mathbb{1}_{(\frac{1}{2},1)}(p),$$
$$\frac{\mathrm{d}}{\mathrm{d}p} F_2^{-1}(p) = 4\mathbb{1}_{(0,\frac{1}{4})}(p) + 2\mathbb{1}_{(\frac{1}{4},\frac{3}{4})}(p) + 4\mathbb{1}_{(\frac{3}{4},1)}(p),$$

hence

$$\gamma(p) = \tfrac{1}{3}\textstyle\sum_{i=1}^{3} \mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}p} F^{-1}(p) \middle| F = F_i\right) = \frac{10}{3}\mathbb{1}_{(0,\frac{1}{4})}(p) + \frac{8}{3}\mathbb{1}_{(\frac{1}{4},\frac{3}{4})}(p) + \frac{10}{3}\mathbb{1}_{(\frac{3}{4},1)}(p).$$

The law of total expectation implies

$$\mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\tfrac{\mathrm{d}}{\mathrm{d}p} F^{-1}(p)\right) = \frac{10}{3}p\mathbb{1}_{(0,\frac{1}{4})}(p) + \left(\frac{3}{15} + \frac{34}{15}p\right)\mathbb{1}_{(\frac{1}{4},\frac{3}{4})}(p) + \frac{10}{3}p\mathbb{1}_{(\frac{3}{4},1)}(p),$$

and hence,

$$f(p) = p\mathbb{1}_{(0,\frac{1}{4})}(p) + \left(\tfrac{3}{40} + \tfrac{17}{20}p\right)\mathbb{1}_{(\frac{1}{4},\frac{3}{4})}(p) + p\mathbb{1}_{(\frac{3}{4},1)}(p).$$

Finally, we obtain

$$\mathrm{MCB}_{\mathrm{HB}} = \int (p - f(p))^2 \gamma(p)\,\mathrm{d}p = \textstyle\int_{\frac{1}{4}}^{\frac{3}{4}} \left(\frac{3}{20}p - \frac{3}{40}\right)^2 \frac{8}{3}\,\mathrm{d}p = \frac{1}{1600}.$$

F.5. Example 2.14 b) in Gneiting and Resin (2023)

For $y_1 < y_2 < y_3$, let F be a mixture of the Dirac measures on y_1, y_2, and y_3 with weights p_1, p_2, and p_3, and let Y be drawn from a mixture of the same Dirac measures with weights q_1, q_2, and q_3, respectively. Suppose that the tuple $(p_1, p_2, p_3; q_1, q_2, q_3)$ attains each of the values

$$\left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}; \tfrac{5}{10}, \tfrac{4}{10}, \tfrac{1}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}; \tfrac{1}{10}, \tfrac{5}{10}, \tfrac{4}{10}\right), \quad \left(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}; \tfrac{4}{10}, \tfrac{1}{10}, \tfrac{5}{10}\right)$$

with equal probability. Let $t_1 = y_2 - y_1 > 0$, $t_2 = y_3 - y_2 > 0$, and $t = t_1 + t_2$. It is immediate that $\mathbb{E}\,\mathrm{crps}(F, Y) = 5t/24$ and $\mathrm{UNC}_0 = \mathbb{E}\,\mathrm{crps}(F_{\mathrm{mg}}, Y) = 2t/9$. As Gneiting and Resin (2023) show, F is threshold and quantile calibrated, hence $\mathrm{MCB}_{\mathrm{BS}} = \mathrm{MCB}_{\mathrm{QS}} = 0$.

Let F_1, F_2, and F_3 denote the three discrete distributions that F may attain. For $i = 1, 2, 3$, denote by Q_i the conditional law of Y given $F = F_i$ and by G_i the isotonic conditional law of Y given $F = F_i$, namely,

$$G_1 = \tfrac{1}{2}\delta_{y_1} + \tfrac{4}{10}\delta_{y_2} + \tfrac{1}{10}\delta_{y_3}, \quad G_2 = \tfrac{1}{4}\delta_{y_1} + \tfrac{7}{20}\delta_{y_2} + \tfrac{4}{10}\delta_{y_3},$$

and

$$G_3 = \frac{1}{4}\delta_{y_1} + \frac{1}{4}\delta_{y_2} + \frac{1}{2}\delta_{y_3}.$$

Since the image of the random vector (F, Y) is finite and ICL is the population version of IDR (Arnold and Ziegel, 2024, Proposition 4.1), one obtains the G_i's alternatively by applying IDR on the finite sample of size $n = 30$ with five occurrences of (F_1, y_1), four of (F_1, y_2), one each of $(F_1, y_3$ and (F_2, y_1), five of (F_2, y_2), four each of (F_2, y_3) and (F_3, y_1), one of (F_3, y_2), and five of (F_3, y_3). The $\mathrm{MCB}_{\mathrm{CT}}$ and $\mathrm{MCB}_{\mathrm{ISO}}$ components may be calculated in analogy to previous examples. We obtain $\mathrm{MCB}_{\mathrm{CT}} = 3t/200$ and $\mathrm{MCB}_{\mathrm{ISO}} = 3t_2/200$.

To compute the Hersbach decomposition, let ν_i be the image of the Lebesgue measure on $(0, 1)$ under F_i where $i = 1, 2, 3$. We have $\nu_1 = t_1\delta_{1/2} + t_2\delta_{3/4}$, $\nu_2 = t_1\delta_{1/4} + t_2\delta_{3/4}$, and $\nu_3 = t_1\delta_{1/4} + t_2\delta_{1/2}$, and hence, $\mu = (1/3)(2t_1\,\delta_{1/4} + t\,\delta_{1/2} + 2t_2\,\delta_{3/4})$. For $\ell = 1, 2, 3$ and $p_\ell = \ell/4$, and for any $A \in \mathcal{B}(0, 1)$, the quantities $f_\ell = f(p_\ell)$ satisfy

$$\begin{aligned}\tau(A) &= \mathbb{E}\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p) \qquad (53)\\ &= \int_A f(p)\,\mathrm{d}\mu(p) = f_1\,\frac{2t_1}{3}\,\delta_{1/4}(A) + f_2\,\frac{t}{3}\,\delta_{1/2}(A) + f_3\,\frac{2t_2}{3}\,\delta_{3/4}(A),\end{aligned}$$

where the expectation in (53) may be calculated by the law of total expectation:

$$\begin{aligned}\mathbb{E}\textstyle\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu(p) &= \frac{1}{3}\sum_{i=1}^{3}\mathbb{E}\left(\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p) \;\middle|\; F = F_i\right)\\ &= \frac{1}{3}\sum_{i=1}^{3}\int\int_A \mathbb{1}\{F_i(y) \le p\}\,\mathrm{d}\nu_i(p)\,\mathrm{d}Q_i(y)\\ &= \frac{t_1}{6}\,\delta_{1/4}(A) + \frac{t}{6}\,\delta_{1/2}(A) + \frac{t_2}{2}\,\delta_{3/4}(A).\end{aligned}$$

We conclude that $f_\ell = p_\ell$ for $\ell = 1, 2, 3$, and hence $\mathrm{MCB}_{\mathrm{HB}} = 0$.

Acknowledgments

We thank two anonymous referees, Tim Hewson, Kai Polsterer, and Johannes Resin for comments and discussion. Computations for the weather case study have been performed on UBELIX (https://ubelix.unibe.ch/), the HPC cluster of the University of Bern.

Funding

Tilmann Gneiting is grateful for support by the Klaus Tschira Foundation. The work of Eva-Maria Walz was funded by the German Research Foundation (DFG) through grant number 257899354. Sebastian Arnold and Johanna Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation.

References

Armerin, F. (2014). The conditional quantile as a minimizer. Working paper, https://doi.org/10.13140/RG.2.2.27136.99847.

Arnold, S., Henzi, A. and Ziegel, J. F. (2023). Sequentially valid tests for forecast calibration. *Ann. Appl. Stat.* **17** 1909–1935. MR4637650

Arnold, S. and Ziegel, J. (2024). Isotonic conditional laws. *Bernoulli*. In press, pre-published at https://www.bernoullisociety.org/index.php/publications/bernoulli-journal/bernoulli-journal-papers.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silvermann, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895

Barlow, R. E., Brunk, H. D., Bartholomew, D. J. and Bremner, J. M. (1972). *Statistical Inference under Order Restrictions*. Wiley.

Bauer, P., Thorpe, A. and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** 47–55.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Q. J. R. Meteorol. Soc.* **140** 1924–1934.

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electron. J. Stat.* **13** 4015–4034. MR4015787

Brunk, H. D. (1965). Conditional expectation given a σ-lattice and applications. *Ann. Math. Stat.* **36** 1339–1350. MR0185629

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131** 2131–2150.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *J. R. Stat. Soc. Ser. A: Stat. Soc.* **147** 278–290. MR0763811

de Leeuw, J., Hornik, K. and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *J. Stat. Softw.* **32** 1–24.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* **39** 863–883.

Dimitriadis, T., Gneiting, T. and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proc. Natl Acad. Sci.* **118** e2016191118. MR4275118

Dimitriadis, T., Gneiting, T., Jordan, A. I. and Vogel, P. (2024). Evaluating probabilistic classifiers: The triptych. *Int. J. Forecast.* **40** 1101–1122.

D'Isanto, A. and Polsterer, K. (2018). Photometric redshift estimation via deep learning. *Astron. Astrophys.* **A111** 1–16.

Embrechts, P. and Hofert, M. (2013). A note on generalized inverses. *Math. Methods Oper. Res.* **77** 423–432. MR3072795

Ferro, C. A. T. and Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Q. J. R. Meteorol. Soc.* **138** 1954–1960.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning*.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V. and Januschowski, T. (2019). Probabilistic forecasting with spline quantile function RNNs. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R. and Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56** S1513–S1589.

Gneiting, T. (2011). Making and evaluating point forecasts. *J. Am. Stat. Assoc.* **106** 746–762. MR2847988

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **69** 243–268. MR2325275

Gneiting, T., Lerch, S. and Schulz, B. (2023). Probabilistic solar forecasting: Benchmarks, post-processing, verification. *Sol. Energy* **252** 72–80.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science* **310** 248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102** 359–378. MR2345548

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. MR3080409

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electron. J. Stat.* **17** 3226–3286. MR4669757

Gneiting, T. and Vogel, P. (2022). Receiver operating characteristic (ROC) curves: Equivalences, beta model, and minimum distance estimation. *Mach. Learn.* **111** 2147–2159. MR4432494

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

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K. and Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annu. Rev. Stat. Appl.* **10** 597–621. MR4567807

Grimit, E. P., Gneiting, T., Berrocal, V. J. and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.* **132** 2925–2942.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble

forecasts. *Mon. Weather Rev.* **129** 550–560.
Henzi, A., Mösching, A. and Dümbgen, L. (2022). Accelerating the pool-adjavent-violators algorithm for isotonic distributional regression. *Methodol. Comput. Appl. Probab.* **24** 2633–2645. MR4528395
Henzi, A., Ziegel, J. F. and Gneiting, T. (2021). Isotonic distributional regression. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **83** 963–969. MR4349124
Hernandéz-Lobato, J. M. and Adams, R. P. (2015). Probabilistic back-propagation for scalable learning of Bayesian neural networks. In *32nd International Conference on Machine Learning*. MR4577124
Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.
Hothorn, T., Kneib, T. and Bühlmann, P. (2014). Conditional transformation models. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **76** 3–27. MR3153931
Immer, A., Bauer, M., Fortuin, V., Rätsch, G. and Khan, M. E. (2021). Scalable marginal likelihood estimation for model selection in deep learning. In *38th International Conference on Machine Learning*.
Jordan, A. I., Mühlemann, A. and Ziegel, J. F. (2022). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Ann. Inst. Stat. Math.* **74** 489–514. MR4417369
Laio, F. and Tamea, P. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.* **11** 1267–1277.
Lauret, P., David, M. and Pinson, P. (2019). Verification of solar irradiance probabilistic forecasts. *Sol. Energy* **194** 254–271.
Leshno, M. and Levy, H. (2002). Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Manag. Sci.* **48** 1074–1085.
Leutbecher, M. and Haiden, T. (2021). Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Q. J. R. Meteorol. Soc.* **147** 425–442.
Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Manag. Sci.* **22** 1087–1096.
Messner, J. W., Mayr, G. J., Wilks, D. S. and Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Weather Rev.* **142** 3003–3014.
Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122** 73–119.
Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley. MR1889865
Müller, A., Scarsini, M., Tsetlin, I. and Winkler, R. L. (2017). Between first- and second-order stochastic dominance. *Manag. Sci.* **63** 2933–2947.
Murphy, A. H. (1973). A new vector partition of the probability score. *J. Appl. Meteorol. Climatol.* **12** 595–600.
Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. and Salomon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrologic ensemble prediction. *J. Hydrol.* **522** 697–713.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **146** 3885–3900.

Ritter, H., Botev, A. and Barber, D. (2018). A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley. MR0961262

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **140** 1086–1096.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer. MR2265633

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Q. J. R. Meteorol. Soc.* **143** 1178–1183.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.* **135** 3209–3220. MR2713575

Strähl, C. and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Stat.* **11** 608–639. MR3619318

Taillardat, M., Fougères, A.-L., Naveau, P. and de Fondeville, R. (2023). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *Int. J. Forecast.* **39** 1448–1459.

R Core Team (2023). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria.

Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Mon. Weather Rev.* **140** 2005–2017.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. Preprint, http://dx.doi.org/10.2139/ssrn.2236605.

Vannitsem, S., Wilks, D. S. and Messner, J., eds. (2018). *Statistical Postprocessing of Ensemble Forecasts*. Elsevier.

Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E. and Carlsson, L. (2020). Conformal calibration. In *Conformal and Probabilistic Prediction and Applications*.

Walz, E. M. (2022). Replication material for "Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output". Available at https://github.com/evwalz/easyuq.

Walz, E. M. (2023). Replication material for "Decompositions of the mean continuous ranked probability score". Available at https://github.com/evwalz/paper_isocrpsdeco. https://doi.org/10.5281/zenodo.13145180.

Walz, E. M., Henzi, A., Ziegel, J. and Gneiting, T. (2024). Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output. *SIAM Rev.* **66** 91–122. MR4704684

Yang, D. and Kleissl, J. (2024). *Solar Irradiance and Photovoltaic Power Forecasting*. CRC Press.