



Contents lists available at ScienceDirect

International Journal of Forecasting

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

Assessing the conditional calibration of interval forecasts using decompositions of the interval score[☆]

Sam Allen^{a,*}, Julia Burnello^b, Johanna Ziegel^c

^a Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany

^b Independent researcher, Switzerland

^c Seminar for Statistics, ETH Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:

Dataset link: https://github.com/sallen12/ConditionalCal_rep

Keywords:

Prediction intervals

Forecast calibration

Conditional coverage

Isotonic distributional regression

Proper scoring rules

ABSTRACT

Forecasts for uncertain future events should be probabilistic. Probabilistic forecasts are commonly issued as prediction intervals, which provide a measure of uncertainty in the unknown outcome whilst being easier to understand and communicate than full predictive distributions. The calibration of a $(1 - \alpha)$ -level prediction interval can be assessed by checking whether the probability that the outcome falls within the interval is equal to $1 - \alpha$. However, such coverage checks are typically unconditional and therefore relatively weak. Although this is well known, there remains a lack of methods to assess the conditional calibration of interval forecasts. In this work, we demonstrate how this can be achieved via decompositions of the well-known interval (or Winkler) score. We study notions of calibration for interval forecasts and then introduce a decomposition of the interval score based on isotonic distributional regression. This decomposition provides meaningful non-negative measures of the discriminative ability and conditional miscalibration of interval forecasts, at both the population and the sample level, and additionally provides an interpretable diagnostic tool to assess conditional calibration in practice. This is illustrated on simulated data and in three applications to benchmark regression datasets.

© 2026 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Reliable forecasts are essential for effective decision making in the face of uncertainty. Forecasts are commonly issued as prediction intervals, which specify a range within which an outcome is expected to lie with a certain probability. Prediction intervals account for the uncertainty in the unknown outcome, in contrast to traditional point forecasts, whilst providing a convenient means to communicate the output from full predictive distributions to forecast users. They are therefore issued in a variety of application domains: national weather centres commonly report prediction intervals for temperature and precipitation to the general public; epidemiological forecasts often take the form of a collection of prediction intervals at several probability levels, such as those available in the COVID-19 Forecast Hub (Ray et al., 2020);

[☆] The numerical results presented in this manuscript were reproduced by the reproducibility team on 14 September 2026.

* Corresponding author.

E-mail address: sam.allen@kit.edu (S. Allen).

<https://doi.org/10.1016/j.ijforecast.2026.09.005>

0169-2070/© 2026 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Open access funding enabled and organized by Projekt DEAL.

and interval forecasts have also been requested in popular forecasting competitions (Hong et al., 2016; Makridakis et al., 2020, 2022). Throughout this paper, we use the terms ‘interval forecast’ and ‘prediction interval’ interchangeably.

Interval forecasts have gained considerable attention recently, due to the development of conformal prediction algorithms, which can generate prediction intervals that are calibrated (or valid) by construction (Vovk et al., 2022, Section 2). Generally, the calibration of a $(1 - \alpha)$ -level prediction interval is assessed by checking whether the empirical coverage of the prediction intervals is close to $1 - \alpha$. It has repeatedly been acknowledged that such unconditional checks for calibration are insufficient, since biases can cancel out in such a way that the forecasts have the correct marginal coverage despite being clearly miscalibrated in particular circumstances (Christoffersen, 1998). For example, many conformal prediction algorithms achieve unconditional calibration despite issuing prediction intervals of a fixed length, which thus do not adapt to covariates or to the complexity of the forecasting task (see e.g. the discussion in Romano et al., 2019).

It is therefore desirable to additionally assess the *conditional* calibration of interval forecasts, which essentially corresponds to the prediction intervals having the desired coverage conditional on some auxiliary information. This auxiliary information typically corresponds to the information used to generate the forecast, represented by covariates, filtration sets, or the forecast itself. In practice, conditional calibration is often evaluated by grouping together forecasts according to this auxiliary information, and then checking whether the prediction intervals exhibit the correct coverage given that each group occurred. Statistical tests for conditional calibration can similarly be derived by checking whether it is possible to predict when the outcome will fall within the prediction interval. If this is predictable given the auxiliary information, then there is evidence to suggest that forecasts are not conditionally calibrated (Gibbs et al., 2025). This is analogous to the standard framework of conditional moment testing (Andrews & Shi, 2013; Newey, 1985), which is closely related to the notions of multicalibration and multiaccuracy in fair machine learning (Hébert-Johnson et al., 2018; Kim et al., 2019), and is also often used to test for conditional calibration of quantile forecasts (or, equivalently, conditional backtesting of value-at-risk forecasts) (Engle & Manganelli, 2004; Kuester et al., 2006; Nolde & Ziegel, 2017).

In this paper, we present an alternative framework to quantify the conditional (mis)calibration of interval forecasts. This framework provides a diagnostic tool that practitioners can use to assess whether or not their interval forecasts are conditionally calibrated. While this measure of miscalibration could also be used to construct tests for conditional calibration, we do not explore this here. The framework that we introduce is based on score decompositions, which are commonly used to assess the conditional calibration of both point forecasts and probabilistic forecasts. Scoring functions and scoring rules summarise forecast quality in a single number, through which competing forecasters can be ranked and compared (Gneiting, 2011; Gneiting & Raftery, 2007). Average scores over multiple forecast cases can be decomposed into terms that measure different aspects of forecast performance. Most commonly, average scores are decomposed into measures of forecast uncertainty, information content, and (conditional) miscalibration. Such decompositions were first proposed by Sanders (1963) and Murphy (1973) for the squared error. DeGroot and Fienberg (1982) and Bröcker (2009) then generalised this decomposition to any scoring rule, thus facilitating an assessment of the conditional calibration of probabilistic forecasts. Bentzen and Friederichs (2014) applied this decomposition to the quantile score (or pinball loss), allowing the conditional calibration of quantile forecasts to be quantified. Gneiting and Resin (2023) introduced a general framework for decomposing scoring functions when the forecast is an arbitrary univariate functional of a predictive distribution. A detailed history of score decompositions is provided by Mitchell (2020) and Pohle (2020).

However, despite the importance of prediction intervals in many application domains, no score decompositions have been proposed to assess the conditional calibration of interval forecasts.¹ In this work, we demonstrate how a conditional notion of interval forecast calibration can be assessed via decompositions of the interval score (also referred to as the Winkler score), which is the canonical scoring function with which to assess central prediction intervals (Brehmer & Gneiting, 2021; Gneiting & Raftery, 2007; Winkler, 1972). The decomposition of the interval score provides a measure of conditional calibration that is non-negative and equal to zero if and only if the interval forecasts satisfy a relatively strong notion of conditional calibration, namely auto-calibration (Gneiting & Ranjan, 2013; Tsyplakov, 2011). The decomposition additionally provides a measure of discriminative ability that quantifies how informative the interval forecasts are, which is not available from existing tests for conditional calibration.

Having introduced the interval score decomposition at the population level, we then demonstrate how the decomposition terms can be estimated in practice using isotonic distributional regression (IDR; Henzi et al., 2021). This follows a recent strand of literature on the estimation of score decomposition terms using isotonic regression: Dimitriadis et al. (2021) originally proposed this when evaluating probability forecasts for binary outcomes; Gneiting and Resin (2023) extended it to forecasts for particular functionals, which was applied to quantile forecasts by Gneiting et al. (2023); Arnold et al. (2024) employed IDR to estimate the decomposition terms for the continuous ranked probability score (CRPS) when evaluating the conditional calibration of full predictive distributions. Like their population counterparts, the estimated interval score decomposition terms are guaranteed to be non-negative. In a series of applications to simulated data and

¹ Braun et al. (2025) recently independently proposed a similar decomposition-based approach to estimate the conditional calibration of interval forecasts. While we focus here on decompositions of the interval score, Braun et al. (2025) considered decompositions of scoring rules for binary outcomes, such as the Brier score or log score, which are used to assess predictions for whether the outcome will fall within the interval forecast. One benefit of this latter approach is that the decomposition terms are arguably easier to estimate in practice. However, by decomposing the interval score, the decomposition terms are more directly connected to the overall performance of the interval forecasts, since this is typically quantified using the interval score. In particular, we can quantify how much of a forecast’s interval score is attributable to conditional miscalibration.

benchmark regression datasets, we demonstrate that the resulting decomposition terms provide information that is not available from simply comparing the average length of unconditionally calibrated prediction intervals, which is commonly performed for example in the conformal prediction literature (Fontana et al., 2023; Shafer & Vovk, 2008).

The remainder of the paper is organised as follows. Notions of coverage and calibration of interval forecasts are introduced and compared in the following section. Section 3 demonstrates how these can be assessed using decompositions of the interval score. These decompositions are applied to simulated data in Section 4, and to benchmark regression datasets in Section 5. All proofs are deferred to Appendix A, while R code to reproduce the results herein is available at https://github.com/sallen12/CondIntCal_rep.

2. Calibration of interval forecasts

2.1. Notation

Suppose that we wish to forecast a real-valued variable $Y \in \mathbb{R}$, and that our forecast takes the form of a central (or equal-tailed) $1 - \alpha$ prediction interval for some arbitrary fixed $\alpha \in (0, 1)$. We treat the interval forecast here as a random interval, with lower and upper bounds denoted by $L \in \mathbb{R}$ and $U \in \mathbb{R}$, respectively, with $L \leq U$ almost surely. In practice, forecast evaluation is performed over a set of multiple forecasts and observations, which can be interpreted as realisations of $[L, U]$ and Y .

Following Brehmer and Gneiting (2021), we denote the set of possible intervals as $\mathcal{I} = \{[\ell, u] : \ell, u \in \mathbb{R}, \ell \leq u\}$. We treat $[L, U]$ as a central prediction interval, in the sense that the forecaster assumes that the probability that Y is lower than L is equal to the probability that Y is larger than U . Central prediction intervals are by far the most common form of interval forecast in practical applications, though extensions to non-central prediction intervals and modal interval forecasts are discussed in Appendix B and Appendix D, respectively.

Central prediction intervals are often generated from the quantiles of a predictive distribution F . We denote by $Q_\alpha(F)$ the set of α -quantiles of F , and by $q_\alpha(F)$ an arbitrary element of this set. The set of quantiles $Q_\alpha(F)$ is either a singleton, making $q_\alpha(F)$ the unique α -quantile of F , or a closed interval between the lower and upper α -quantiles of F . We simplify the notation to $Q_\alpha(Y)$ and $q_\alpha(Y)$ to denote the α -quantiles of the marginal distribution of Y , and $Q_\alpha(Y | Z)$ and $q_\alpha(Y | Z)$ for the α -quantiles of the conditional distribution of Y given an arbitrary random variable Z .

Regardless of how the forecasts are constructed, we assume that the interval forecasts implicitly or explicitly depend on some covariate information $X \in \mathcal{X}$. That is, L and U can be interpreted as functions of X . More generally, X can be interpreted as the information available to the forecaster at the time the forecast $[L, U]$ is issued. We assume that X characterises the information used in the prediction, in the sense that $\sigma(L, U) \subseteq \sigma(X)$. We denote by $\mathbf{P}$ the joint distribution of the tuple (L, U, Y, X) , and, where relevant, assume that equalities involving these random variables hold almost surely.

2.2. Unconditional calibration

The calibration of a central prediction interval is typically assessed by checking whether the outcome falls within the interval with the desired coverage level,

$$\mathbf{P}(Y \in [L, U]) = 1 - \alpha. \quad (1)$$

While this definition ignores that $[L, U]$ is a central prediction interval, in that it neglects the probabilities with which the outcome falls above and below the interval, it is common to additionally assess whether $\mathbf{P}(Y < L) = \mathbf{P}(Y > U) = \alpha/2$, which would imply that (1) is satisfied. Requiring equality in (1) may also be prohibitive, especially when Y is discrete. To account for this, we therefore define an interval forecast as unconditionally calibrated if it satisfies the following requirement.

Definition 1. An interval forecast $[L, U]$ is *unconditionally calibrated* if

$$\mathbf{P}(Y < L) \leq \frac{\alpha}{2} \leq \mathbf{P}(Y \leq L), \quad \text{and} \quad \mathbf{P}(Y > U) \leq \frac{\alpha}{2} \leq \mathbf{P}(Y \geq U). \quad (2)$$

If a forecast is unconditionally calibrated according to (2), then

$$\mathbf{P}(Y \in (L, U)) \leq 1 - \alpha \leq \mathbf{P}(Y \in [L, U]), \quad (3)$$

which simplifies to (1) when $\mathbf{P}(Y \in \{L, U\}) = 0$. This generalised notion of unconditional calibration accounts for discontinuities in the joint distribution of (L, U, Y) . In the literature, accommodating discontinuities in the joint distribution is often achieved by weakening (1) to $\mathbf{P}(Y \in [L, U]) \geq 1 - \alpha$, allowing conservative prediction intervals to be calibrated. The additional lower bound in (3) strengthens the requirement for calibration by limiting how conservative a calibrated prediction interval can be.

2.3. Conditional calibration

The definition of unconditional interval calibration averages across all possible forecasts and observations, neglecting possible conditional biases and rendering the requirement in (2) fairly weak. One could argue that it therefore provides a minimum requirement for forecasts to be considered reliable, justifying checks for unconditional calibration in practice. However, it has repeatedly been remarked that forecasts should be conditionally calibrated, since this ensures that the prediction intervals exhibit the correct coverage in particular forecasting scenarios, not just on average (see e.g. [Christoffersen, 1998](#); [Fontana et al., 2023](#); [Shafer & Vovk, 2008](#)). Given the covariates X used to obtain the forecast, conditional calibration is typically defined as follows.

Definition 2. An interval forecast $[L, U]$ is *conditionally calibrated* given X if

$$P(Y < L | X) \leq \frac{\alpha}{2} \leq P(Y \leq L | X), \quad \text{and} \quad P(Y > U | X) \leq \frac{\alpha}{2} \leq P(Y \geq U | X). \quad (4)$$

As defined in (4), conditional calibration given X implies that

$$P(Y \in (L, U) | X) \leq 1 - \alpha \leq P(Y \in [L, U] | X), \quad (5)$$

which simplifies to the more familiar requirement that $P(Y \in [L, U] | X) = 1 - \alpha$ when $P(Y \in \{L, U\} | X) = 0$. As in (1), the inequalities in (4) and (5) constitute sensible bounds on how conservative a conditionally calibrated interval forecast can be.

The requirement in (4) is often referred to simply as conditional calibration. We make explicit in the definition that this corresponds to conditional calibration given X , since calibration could more generally be defined conditional on other information sets. For example, in the probabilistic forecasting community, it is common to condition on the information provided by the forecast itself. This is typically referred to as auto-calibration ([Gneiting & Ranjan, 2013](#); [Tsyplakov, 2011](#)).

Definition 3. An interval forecast $[L, U]$ is *auto-calibrated* if

$$P(Y < L | L, U) \leq \frac{\alpha}{2} \leq P(Y \leq L | L, U), \quad \text{and} \quad P(Y > U | L, U) \leq \frac{\alpha}{2} \leq P(Y \geq U | L, U). \quad (6)$$

Remark 4. Both conditional calibration and auto-calibration can alternatively be defined in terms of quantiles of the conditional distribution of Y . In particular, (4) is satisfied if and only if

$$L \in Q_{\frac{\alpha}{2}}(Y | X), \quad \text{and} \quad U \in Q_{1-\frac{\alpha}{2}}(Y | X).$$

That is, the interval forecast is conditionally calibrated if and only if L and U provide calibrated forecasts for the $\alpha/2$ and $1 - \alpha/2$ quantiles of Y . Similarly, the forecast is auto-calibrated if and only if

$$L \in Q_{\frac{\alpha}{2}}(Y | L, U), \quad \text{and} \quad U \in Q_{1-\frac{\alpha}{2}}(Y | L, U).$$

Auto-calibration again implies a conditional coverage requirement, which simplifies to $P(Y \in [L, U] | L, U) = 1 - \alpha$ when the joint distribution of (Y, L, U) is continuous. While it is more common to assess the conditional calibration of interval forecasts given covariates, as defined in (4), auto-calibration constitutes a very intuitive notion of calibration, since it implies that the forecast can be taken at face value, which is the minimum requirement for forecasts to be useful for decision making. Auto-calibration is analogous to the notion of conditional calibration typically used to assess probabilistic forecasts for binary events, where it is required that the conditional probability that the event occurs given the forecast is equal to the forecast itself ([Dawid, 1982](#); [Murphy & Winkler, 1977](#)). This is often referred to as ‘perfect calibration’ in the machine learning literature (e.g. [Guo et al., 2017](#)). Similarly, the calibration of mean forecasts is often assessed conditionally on the forecast value ([Gneiting & Resin, 2023](#); [Mincer & Zarnowitz, 1969](#)).

Moreover, when X is high-dimensional, auto-calibration also has the advantage that it is generally easier to assess than conditional calibration. Conditional calibration assessment requires estimating the conditional quantiles of Y given X , which is essentially equivalent to the original forecasting problem. In contrast, assessing auto-calibration requires estimating the conditional quantiles of Y given the prediction interval, which is typically easier in practice. Essentially, the forecast $[L, U]$ represents a lower-dimensional proxy for the information provided by X . A similar argument is sometimes made to justify assessing the calibration of probabilistic classifiers conditional on forecast probabilities rather than covariates (e.g. [Huang et al., 2023](#)). If the forecast leverages all information provided by the covariates, in the sense that $\sigma(L, U) = \sigma(X)$, then auto-calibration and conditional calibration are equivalent. More generally, this is the case if $[L, U]$ is a sufficient statistic for Y relative to X . On the other hand, if the forecast does not efficiently utilise the information provided by the covariates X , then auto-calibration will generally be a weaker notion of calibration than conditional calibration. Recall that we want our forecasts to satisfy the strongest notion of calibration possible, since this implies that they exhibit the correct coverage conditional on a larger information set.

If the forecast fails to use any of the information in the covariates, then $\sigma(L, U)$ is trivial, in which case auto-calibration is equivalent to unconditional calibration. Conditional calibration also simplifies to unconditional calibration if Y and

X are independent. In practice, Y is generally independent of neither X nor $[L, U]$, so that conditional calibration and auto-calibration are much stronger than unconditional calibration. Similarly, when the forecasts are obtained using state-of-the-art statistical and machine learning models, $[L, U]$ contains much of the information provided by X , so that auto-calibration is not much weaker than conditional calibration. These relationships between the different notions of calibration are summarised in the following proposition.

Proposition 5. *The following relationships hold between the notions of interval calibration:*

- (i) *Conditional calibration given X implies unconditional calibration. Unconditional calibration implies conditional calibration given X if Y and X are independent.*
- (ii) *Auto-calibration implies unconditional calibration. Unconditional calibration implies auto-calibration if Y and $[L, U]$ are independent.*
- (iii) *Conditional calibration given X implies auto-calibration. Auto-calibration implies conditional calibration given X if $[L, U]$ is a sufficient statistic for Y relative to X .*

Remark 6. We assume throughout that the information provided by the forecasts is no larger than the information provided by the covariates, i.e. $\sigma(L, U) \subseteq \sigma(X)$. In this case, Proposition 5(iii) states that if an interval forecast is conditionally calibrated, then it is also auto-calibrated; put differently, conditional calibration is the stronger notion of calibration. If we were to consider conditional calibration with respect to a smaller subset of covariates, for which $\sigma(L, U) \not\subseteq \sigma(X)$, then conditional calibration will generally not imply auto-calibration. Conversely, if $\sigma(X) \subseteq \sigma(L, U)$, then auto-calibration would imply conditional calibration, making it the stronger notion of calibration.

Example 7. Let $X = (X_1, X_2)$, where $X_1, X_2 \sim N(0, 1)$ independently. Suppose $Y \sim N(X_1 + X_2, 1)$, and let τ be equal to 1 and -1 each with probability $1/2$. Consider the predictive distributions $F_0 = N(0, 3)$, $F_1 = N(X_1, 2)$, $F_2 = (1/2)N(X_1 + X_2, 1) + (1/2)N(X_1 + X_2 + \tau, 1)$, and $F_3 = N(X_1 + X_2, 1)$, with corresponding prediction intervals $I_i = [q_{\alpha/2}(F_i), q_{1-\alpha/2}(F_i)]$, for $i \in \{0, 1, 2, 3\}$. The forecast I_0 does not use the information provided by X_1 and X_2 , issuing the central prediction interval of the unconditional distribution of Y . This forecast is unconditionally calibrated and auto-calibrated but not conditionally calibrated given X . The forecast I_1 uses the information provided by X_1 but not X_2 , issuing the central prediction interval of the conditional distribution of Y given X_1 . This forecast is unconditionally calibrated and auto-calibrated, and is conditionally calibrated given X_1 but not given $X = (X_1, X_2)$. The forecast I_2 uses the information provided by X but not perfectly, exhibiting a positive or negative bias with probability $1/2$. These biases average out so that the forecast is unconditionally calibrated, though the forecast is neither auto-calibrated nor conditionally calibrated given X . Finally, the forecast I_3 issues the central prediction interval of the conditional distribution of Y given X_1 and X_2 . This forecast is conditionally calibrated given X , and is therefore auto-calibrated and unconditionally calibrated. These results are demonstrated in a simulation experiment in Section 4. In practice, a typical model for Y in this example would be to fit two linear quantile regressions for the lower and upper interval bounds—that is, to assume that L and U are linear functions of both X_1 and X_2 . Since $q_\alpha(Y | X)$ is a linear function of X_1 and X_2 for any $\alpha \in (0, 1)$, it follows that auto-calibration and conditional calibration are equivalent for such a forecast.

3. Interval score decompositions

In this section, we demonstrate how the auto-calibration of interval forecasts can be assessed using decompositions of the interval score. After introducing the interval score, we present a population-level decomposition of the expected interval score into terms of forecast uncertainty, discriminative ability, and conditional (auto-) miscalibration. We then illustrate how these decomposition terms can be estimated in practice using isotonic distributional regression (IDR; Henzi et al., 2021).

3.1. The interval score

Calibration is a necessary requirement for interval forecasts to be considered trustworthy. However, checks for calibration do not directly provide a means to compare competing forecasts. Instead, interval forecasts should be compared using consistent scoring functions (Brehmer & Gneiting, 2021; Gneiting, 2011). A scoring function takes a forecast and an observation as inputs, and outputs a numerical value that quantifies the forecast accuracy.

A scoring function is *consistent* for a functional of a distribution if it is minimised in expectation when the functional of the true outcome distribution is issued as the forecast. If forecasters know the true data generating process, then they are therefore incentivised to issue this as their prediction (Gneiting, 2011). Consistent scoring functions are closely related to proper scoring rules for full predictive distributions, with a lower score being indicative of a more accurate forecast (Gneiting & Raftery, 2007; Waghamare & Ziegel, 2025).

For example, a scoring function $S : \mathcal{X} \times \mathbb{R} \rightarrow [0, \infty)$ is consistent for the central $1 - \alpha$ prediction interval if

$$E[S([\ell^*, u^*], Y)] \leq E[S([\ell, u], Y)] \quad (7)$$

for all $\ell^* \in Q_{\alpha/2}(Y)$, $u^* \in Q_{1-\alpha/2}(Y)$, and $[\ell, u] \in \mathcal{I}$. That is, the expected score is minimised when the interval forecast is derived from the relevant quantiles of the outcome variable Y . The scoring function is strictly consistent if equality holds in (7) if and only if $\ell \in Q_{\alpha/2}(Y)$ and $u \in Q_{1-\alpha/2}(Y)$. Note that the interval forecast $[\ell, u]$ in (7) is a fixed (not random) element of $\mathcal{I}$.

Interval forecasts are typically assessed using the interval score, or Winkler score (Gneiting & Raftery, 2007; Winkler, 1972),

$$IS_{\alpha}([\ell, u], y) = |u - \ell| + \frac{2}{\alpha} \mathbb{1}\{y < \ell\}(\ell - y) + \frac{2}{\alpha} \mathbb{1}\{y > u\}(y - u) \quad (8)$$

$$= \frac{2}{\alpha} \left[QS_{\frac{\alpha}{2}}(\ell, y) + QS_{1-\frac{\alpha}{2}}(u, y) \right], \quad (9)$$

where

$$QS_{\alpha}(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(x - y) \quad (10)$$

is the quantile score, or pinball loss, at level α . The quantile score at level α is consistent for the α -quantile, and, as a sum of quantile scores, the interval score is therefore consistent for the central $1 - \alpha$ prediction interval.

The interval score is equal to the length of the interval forecast plus a penalty if the outcome falls outside the interval. The penalty increases as α decreases, recognising that the $1 - \alpha$ prediction interval should contain the outcome more often when α is close to zero. The interval score therefore implicitly rewards interval forecasts that are as short as possible, subject to containing the outcome with the desired coverage probability. This is made more explicit in the following section using decompositions of the interval score.

The interval score can be extended by applying non-decreasing transformations of ℓ , u , and y , thereby changing how outcomes outside of the interval are penalised (Brehmer & Gneiting, 2021), and can also be generalised to non-central prediction intervals. These two extensions are considered in Appendix B.

3.2. Population-level decomposition of the interval score

While scoring functions condense forecast performance into a single numeric value, expected scores can be decomposed into terms that quantify different aspects of forecast performance (Bröcker, 2009; Murphy, 1973). Most commonly, forecasts are decomposed into three terms, quantifying the inherent uncertainty in the forecast scenario, the discriminative ability of the forecast, and the forecast's conditional (auto-) miscalibration. While decompositions of quantile scores have previously been studied (Bentzen & Friederichs, 2014; Gneiting et al., 2023), decompositions of the interval score have not yet been considered.

For the interval score, we can write

$$E[IS_{\alpha}([L, U], Y)] = \text{UNC} - \text{DSC}_A + \text{MCB}_A, \quad (11)$$

where

$$\text{UNC} = E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y), q_{1-\frac{\alpha}{2}}(Y)], Y)],$$

$$\text{DSC}_A = E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y), q_{1-\frac{\alpha}{2}}(Y)], Y)] - E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y | L, U), q_{1-\frac{\alpha}{2}}(Y | L, U)], Y)],$$

$$\text{MCB}_A = E[IS_{\alpha}([L, U], Y)] - E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y | L, U), q_{1-\frac{\alpha}{2}}(Y | L, U)], Y)].$$

The expectation in (11) is taken over both the interval forecast $[L, U]$ and the outcome Y . These decomposition terms can be defined using arbitrary $\alpha/2$ and $1 - \alpha/2$ quantiles of Y (i.e. any $q_{\alpha/2} \in Q_{\alpha/2}$ and $q_{1-\alpha/2} \in Q_{1-\alpha/2}$), since all result in the same expected interval score.

The first term, UNC, is the expected interval score with the unconditional $\alpha/2$ and $1 - \alpha/2$ quantiles of Y used to construct the prediction interval. This term is independent of the forecast and depends only on the outcome, providing a measure of uncertainty in the forecasting scenario. The second term, DSC_A , is the expected improvement in score obtained by using the conditional quantiles of Y given the prediction interval $[L, U]$ to define the interval forecast rather than the unconditional quantiles of Y , providing a measure of the discriminative ability inherent in the forecast $[L, U]$. This term acts negatively on the score, so that more informative forecasts yield lower interval scores. The final term, MCB_A , is the expected improvement in score obtained by using the conditional quantiles of Y given the prediction interval $[L, U]$ to define the interval forecast rather than $[L, U]$ itself, providing a measure of the miscalibration in the forecast $[L, U]$. This term acts positively on the score, penalising forecasts that are miscalibrated.

Since the interval score is non-negative and strictly consistent for the central prediction interval, the decomposition in (11) satisfies several desirable properties.

Proposition 8. *The decomposition of the expected interval score in (11) satisfies the following properties:*

- (i) $\text{DSC}_A \geq 0$ with equality if and only if $q_{\alpha/2}(Y | L, U) \in Q_{\alpha/2}(Y)$ and $q_{1-\alpha/2}(Y | L, U) \in Q_{1-\alpha/2}(Y)$.
- (ii) $\text{MCB}_A \geq 0$ with equality if and only if the interval forecast $[L, U]$ is auto-calibrated.

Of particular interest here is the miscalibration term, MCB_A . This term provides a measure of how far the interval score for $[L, U]$ differs from that of an auto-calibrated forecast interval based on the same information, thereby quantifying the extent to which the forecast $[L, U]$ is not auto-calibrated. In particular, the measure of miscalibration obtained from the score decomposition can be interpreted as the amount by which the average interval score would be improved if the forecast were auto-calibrated.

Remark 9. An analogous decomposition could also be used to measure conditional calibration instead of auto-calibration by replacing the quantiles $q_{\alpha/2}(Y \mid L, U)$ and $q_{1-\alpha/2}(Y \mid L, U)$ with $q_{\alpha/2}(Y \mid X)$ and $q_{1-\alpha/2}(Y \mid X)$, respectively. Analogous theoretical properties to [Proposition 8](#) would hold, with the miscalibration being equal to zero if and only if the interval forecast is conditionally calibrated. This decomposition is discussed in detail in [Appendix C](#). However, such a decomposition requires knowledge of the conditional quantiles of Y given X , which is essentially the original forecasting problem, limiting the applicability of such a decomposition in practice. For this reason, we restrict our attention in the following to auto-calibration and the decomposition in [\(11\)](#).

3.3. Sample-level decomposition of the interval score

In practice, an evaluation is performed over a set of n interval forecasts $[\ell_1, u_1], \dots, [\ell_n, u_n]$ and corresponding outcomes $y_1, \dots, y_n$, which can be interpreted as realisations of the random interval $[L, U]$ and the outcome variable Y . To estimate the terms of the auto-calibration decomposition in [\(11\)](#), we require estimates of the conditional quantiles of Y given $[L, U]$. However, assuming that the forecasts $[\ell_1, u_1], \dots, [\ell_n, u_n]$ are pairwise distinct, we only have one realisation of Y given each observed interval forecast. Estimation of the conditional quantiles therefore requires making some assumptions about the relationship between Y and $[L, U]$, which should ideally be as weak as possible.

Several recent studies have argued that it is often reasonable to assume that there is an isotonic relationship between the forecasts and the observations. That is, a larger forecast generally corresponds to a larger observation. In this case, the decomposition terms can be estimated using variants of isotonic regression ([Arnold et al., 2024](#); [Dimitriadis et al., 2021](#); [Gneiting & Resin, 2023](#); [Gneiting et al., 2023](#)). Following this reasoning, we estimate the conditional quantiles $q_{\alpha/2}(Y \mid L, U)$ and $q_{1-\alpha/2}(Y \mid L, U)$ using isotonic distributional regression (IDR; [Henzi et al., 2021](#)), a flexible non-parametric distributional regression procedure that requires very few implementation choices. IDR essentially corresponds to fitting isotonic quantile regressions at every quantile level simultaneously, under the constraint that the quantiles do not cross. [Gneiting et al. \(2023\)](#) employ isotonic quantile regression to obtain decompositions of the quantile score when evaluating quantile forecasts, and the framework adopted herein can therefore be interpreted as an extension of this approach to interval forecasts, which additionally accounts for possible quantile crossing.

We apply IDR to the observations y_i using the corresponding interval forecast $[\ell_i, u_i]$ as covariate, for $i = 1, \dots, n$. IDR assumes that a larger covariate corresponds to a larger observation, and hence, to fit IDR, we require a formal definition of when one interval forecast is larger than another. In other words, we require a partial order on $\mathcal{I}$. We say that $[\ell_i, u_i] \leq [\ell_j, u_j]$ if $\ell_i \leq \ell_j$ and $u_i \leq u_j$. That is, an interval forecast is larger (smaller) than another if the bounds of the interval are larger (smaller) than those of the other forecast. In this case, we say that the two forecasts are comparable; if they are not comparable, the intervals are nested. In theory, other orderings on the space of interval forecasts could also be considered, though not all orderings will result in meaningful decomposition terms; this is discussed further in [Appendix E](#).

Applying IDR to the interval forecasts and observations yields discrete predictive distributions $\tilde{F}_1, \dots, \tilde{F}_n$, and we denote by $\tilde{Q}_{\alpha,i}$ the set of α -quantiles of $\tilde{F}_i$, with $\tilde{q}_{\alpha,i} = \min \tilde{Q}_{\alpha,i}$ being the lower α -quantile of $\tilde{F}_i$, for $i = 1, \dots, n$. We similarly denote by $\tilde{Q}_\alpha$ the set of α -quantiles of $y_1, \dots, y_n$, and $\tilde{q}_\alpha = \min \tilde{Q}_\alpha$. In assuming an isotonic relationship between the covariates and the outcome, the resulting IDR prediction intervals satisfy $[\tilde{q}_{\alpha/2,i}, \tilde{q}_{1-\alpha/2,i}] \leq [\tilde{q}_{\alpha/2,j}, \tilde{q}_{1-\alpha/2,j}]$ when $y_i \leq y_j$, for all $i, j = 1, \dots, n$, where the ordering of the intervals is understood as defined above.

The general idea is that IDR provides recalibrated versions of the original interval forecasts, with $\tilde{q}_{\beta,i}$ an estimate of $q_\beta(Y \mid L = \ell_i, U = u_i)$, for $\beta \in \{\alpha/2, 1 - \alpha/2\}$. The empirical quantile $\tilde{q}_\beta$ similarly provides an estimate of $q_\beta(Y)$. Using this, we can estimate the terms of [\(11\)](#) using the sample decomposition

$$\hat{\text{IS}} = \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\ell_i, u_i], y_i) = \widehat{\text{UNC}} - \widehat{\text{DSC}}_A + \widehat{\text{MCB}}_A, \quad (12)$$

where

$$\begin{aligned} \widehat{\text{UNC}} &= \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\tilde{q}_{\frac{\alpha}{2}}, \tilde{q}_{1-\frac{\alpha}{2}}], y_i), \\ \widehat{\text{DSC}}_A &= \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\tilde{q}_{\frac{\alpha}{2}}, \tilde{q}_{1-\frac{\alpha}{2}}], y_i) - \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\tilde{q}_{\frac{\alpha}{2},i}, \tilde{q}_{1-\frac{\alpha}{2},i}], y_i), \\ \widehat{\text{MCB}}_A &= \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\ell_i, u_i], y_i) - \frac{1}{n} \sum_{i=1}^n \text{IS}_\alpha([\tilde{q}_{\frac{\alpha}{2},i}, \tilde{q}_{1-\frac{\alpha}{2},i}], y_i). \end{aligned}$$

Each of these terms corresponds to a sample estimate of the respective terms in (11). Arnold and Ziegel (2025) demonstrated that, for infinite sample sizes, the IDR distribution can be interpreted as the isotonic conditional law of Y given the covariates. Hence, it is additionally possible to study this isotonic decomposition at the population level, facilitating a greater understanding of its theoretical properties. This population-level decomposition and its relation to interval forecast calibration are discussed in detail in Appendix F.

The decomposition in (12) is exact, in that the terms of the decomposition cancel out to recover the average interval score of the forecasts $[\ell_1, u_1], \dots, [\ell_n, u_n]$. Since the interval score is non-negative, it follows that $\widehat{\text{UNC}} \geq 0$. As in the population decomposition, this term depends only on the observations $y_1, \dots, y_n$. Moreover, Henzi et al. (2021, Theorem 2) demonstrated that IDR implicitly learns the interval forecasts that optimise the interval score over its training data, subject to the constraint that a larger interval corresponds to a larger observation. Since the original interval forecasts and the unconditional quantiles are also admissible solutions to this isotonic regression problem, it follows that the recalibrated prediction intervals receive a lower interval score, so that the estimated discrimination and miscalibration terms of the decomposition are non-negative. This ensures that the estimated decomposition terms are interpretable. This is formalised in the following proposition.

Proposition 10. *The sample decomposition of the interval score in (12) has the following properties:*

- (i) $\widehat{\text{DSC}}_A \geq 0$ with equality if and only if $\tilde{q}_{\alpha/2} = \tilde{q}_{\alpha/2,i}$ and $\tilde{q}_{1-\alpha/2} = \tilde{q}_{1-\alpha/2,i}$ for all $i = 1, \dots, n$.
- (ii) $\widehat{\text{MCB}}_A \geq 0$ with equality if and only if $\ell_i \in \tilde{Q}_{\alpha/2,i}$ and $u_i \in \tilde{Q}_{1-\alpha/2,i}$ for all $i = 1, \dots, n$ and $(\ell_1, \dots, \ell_n), (u_1, \dots, u_n)$ are valid solutions to the respective isotonic quantile regression problems.

Remark 11. The last condition—that $(\ell_1, \dots, \ell_n)$ and $(u_1, \dots, u_n)$ are valid solutions to the respective isotonic quantile regression problems—means that $(\ell_1, \dots, \ell_n)$ needs to be a minimiser of the criterion

$$\sum_{i=1}^n \text{QS}_{\alpha/2}(x_i, y_i) \quad (13)$$

over all sequences $(x_1, \dots, x_n)$ such that $x_i \leq x_j$ if $[l_i, u_i] \leq [l_j, u_j]$, where the quantile score is defined in (10), and analogously for $(u_1, \dots, u_n)$. Any such solution satisfies $\ell_i \in \tilde{Q}_{\alpha/2,i}$ for all $i = 1, \dots, n$. Conversely, while it is tempting to think that all sequences that satisfy $\ell_i \in \tilde{Q}_{\alpha/2,i}$, $i = 1, \dots, n$ and the ordering constraints are valid solutions, this is unfortunately not true; see Jordan et al. (2022, Section 4.1). However, the sequence of lower quantiles, upper quantiles, or any convex combination thereof is always a valid solution.

Proposition 10 states that the terms of the sample decomposition in (12) are guaranteed to be non-negative, with the discrimination component equal to zero if and only if the estimated conditional quantiles are equal to the unconditional quantiles of $y_1, \dots, y_n$, providing a sample version of the requirement in Proposition 8(i). Similarly, the miscalibration term is equal to zero if and only if the forecasts are equal to the estimated conditional quantiles (up to an unimportant technical requirement on the exact choice of the quantiles), which corresponds to a sample analogue of Proposition 8(ii).

Remark 12. The properties listed in Proposition 10 hold without any requirements on the data. However, for the terms of the sample decomposition to be practically meaningful, three things are required:

1. Sufficient data are available. IDR typically requires a moderate (effective) sample size to provide reasonable estimates of the conditional quantiles and thus the decomposition terms. From our experience, we find that a sample size of $n = 500$ is usually sufficient, with a larger n resulting in more accurate quantile decomposition estimates.
2. The interval forecasts are largely comparable. That is, $[\ell_i, u_i] \leq [\ell_j, u_j]$ or $[\ell_i, u_i] \geq [\ell_j, u_j]$ holds for a large proportion of the data. Incomparable pairs of forecasts reduce the effective sample size used to train IDR, which can violate the previous requirement.
3. There is an isotonic relationship between the interval forecasts $[\ell_1, u_1], \dots, [\ell_n, u_n]$ and the observations $y_1, \dots, y_n$. That is, $[\ell_i, u_i] \leq [\ell_j, u_j]$ generally implies that $y_i \leq y_j$. This is not necessary for the theoretical properties of the decomposition terms derived in this section to hold, but the decomposition will essentially measure a stronger notion of conditional calibration when there is a stronger isotonic relationship. This is discussed further in Appendix C.

We argue that these three properties are satisfied in many practical applications. If a larger interval forecast does not imply a larger observation, then an assessment of conditional calibration may be superfluous, since this could suggest the forecasts are not informative (depending on the application). If the interval forecasts are often not comparable, then an alternative approach is to fit two separate isotonic quantile regression models for the lower and upper quantiles, where the model for the lower quantile depends only on the lower bounds ℓ_i , and the model for the upper quantile depends only on the upper bounds u_i . In theory, this assesses a slightly weaker notion of calibration than when both ℓ_i and u_i are covariates, thus providing a worse estimate of auto-calibration. Nonetheless, it may circumvent issues related to a reduced

effective sample size in practice, since it removes the need for interval forecasts to be comparable, thus alleviating the second requirement in [Remark 12](#).

In some applications, the required sample size to fit IDR will be prohibitive, though we argue that any method to assess the conditional calibration of interval forecasts will require a moderate amount of data. IDR can also be computationally expensive if it is fit to large datasets, in which case it may be useful to combine IDR with subsample aggregation (subbagging) strategies designed to reduce computational cost ([Henzi et al., 2021](#), Section 2.3). In the following sections, we demonstrate the practical utility of the isotonic decomposition of the interval score when sufficient data are available.

Of course, there will be some applications where the requirements in [Remark 12](#) are not satisfied, in which case the estimated decomposition terms may not be informative. For example, as pointed out by a reviewer, if we are interested in forecasting a heteroskedastic process with a constant mean, then the assumption of isotonicity will generally be unreasonable. In this case, other methods could be used to estimate the conditional quantiles in [\(11\)](#). In the economics literature, it is common to assess conditional calibration using linear regression rather than isotonic regression ([Mincer & Zarnowitz, 1969](#)), which has been used by [Dimitriadis and Puke \(2026\)](#) to decompose the quantile score. Since the interval score is a sum of quantile scores, non-negativity of the estimated decomposition terms ([Dimitriadis & Puke, 2026](#), Theorem 3.2) should extend readily to decompositions of the interval score. However, details for this would need to be worked out. In any case, we argue that the isotonicity assumption adopted here is appropriate in many applications.

4. Simulation examples

Following [Gneiting et al. \(2007\)](#), suppose $Y \mid \mu \sim N(\mu, 1)$, where $\mu \sim N(0, 1)$, and let τ be equal to 1 and -1 each with probability $1/2$, independently of (μ, Y) . Consider the following forecast distributions:

- The *unconditional* forecast: $N(0, 2)$
- The *ideal* forecast: $N(\mu, 1)$
- The *unfocused* forecast: $(1/2)N(\mu, 1) + (1/2)N(\mu + \tau, 1)$
- The *mean-biased* forecast: $N(\mu + \tau, 1)$
- The *sign-biased* forecast: $N(-\mu, 1)$
- The *mixed* forecast: $\mathbb{1}\{\tau = 1\}N(0, 2) + \mathbb{1}\{\tau = -1\}N(-\mu, 1)$

We generated central 0.9-level prediction intervals from these six distributions, and evaluated their calibration using the decomposition of the interval score proposed in the previous section. The unconditional and ideal prediction intervals are auto-calibrated, and therefore also unconditionally calibrated. However, while the unconditional forecast contains no more information than the marginal distribution of the response Y , the ideal forecast has knowledge of μ , and should therefore be more discriminative. The unfocused prediction intervals are unconditionally calibrated, but not auto-calibrated. The mean-biased, sign-biased, and mixed forecasts satisfy no notion of calibration considered here, but each contain different amounts of information. The calibration properties of these forecasts are summarised in [Table 1](#).

We simulated $n = 1000$ realisations of μ , τ , and Y , and evaluated the calibration of the corresponding prediction intervals. For each method, IDR was applied to the 1000 prediction intervals using the `isodistrreg` package in R ([Henzi, 2021](#)), from which the (lower) $\alpha/2$ and $1 - \alpha/2$ quantiles were obtained in order to calculate the interval score decomposition terms. While we chose $\hat{q}_{\alpha/2,i}$ and $\hat{q}_{1-\alpha/2,i}$ to be the lower $\alpha/2$ and $1 - \alpha/2$ quantiles of the corresponding IDR predictive distribution, respectively, this choice is somewhat arbitrary and should not significantly affect the resulting decomposition terms, especially when n is large. The decomposition terms for the six forecasts are displayed in [Fig. 1](#). Miscalibration-discrimination plots display the discrimination term of the interval score decomposition against the miscalibration term, with parallel grey isolines corresponding to different mean interval scores ([Dimitriadis et al., 2024](#)). Methods that issue more accurate forecasts, as assessed using the interval score, will be closer to the top-left corner of the plot.

As expected, the unconditional forecast is found to be auto-calibrated yet completely uninformative, with $\widehat{MCB}_A \approx \widehat{DSC}_A \approx 0$. The unconditional prediction intervals are constant, yielding constant IDR distributions $\hat{F}_1 = \dots = \hat{F}_n$ whose quantiles correspond to the empirical quantiles of the realised observations. The ideal forecast is also auto-calibrated, but is considerably more discriminative than the unconditional forecast, resulting in the lowest average interval score. The fact that $\widehat{MCB}_A$ is not exactly zero for the ideal forecast is due to the finite sample size used to calculate the decomposition terms; the miscalibration term tends towards zero as the sample size increases (not shown). The mean-biased forecast receives a higher average score, since it performs worse than the ideal forecast with respect to both miscalibration and discrimination. As a mixture between the ideal and the mean-biased forecasts, the unfocused forecast produces miscalibration and discrimination terms between those of its two component forecasts. Similarly, the mixed forecast is equal to the unconditional and sign-biased forecasts with equal probability, and therefore also has zero discriminative ability, as well as a miscalibration term halfway between those of its two component forecasts.

Although the sign-biased forecast contains the information provided by μ , it receives a discrimination term of zero. The reason for this is that there is an antitonic relationship between the sign-biased prediction intervals and the outcome, meaning the forecast contains no isotonic information, which is what is measured by the discrimination term estimated using IDR. This corresponds to a case where the requirements in [Remark 12](#) are not satisfied. As discussed in the previous

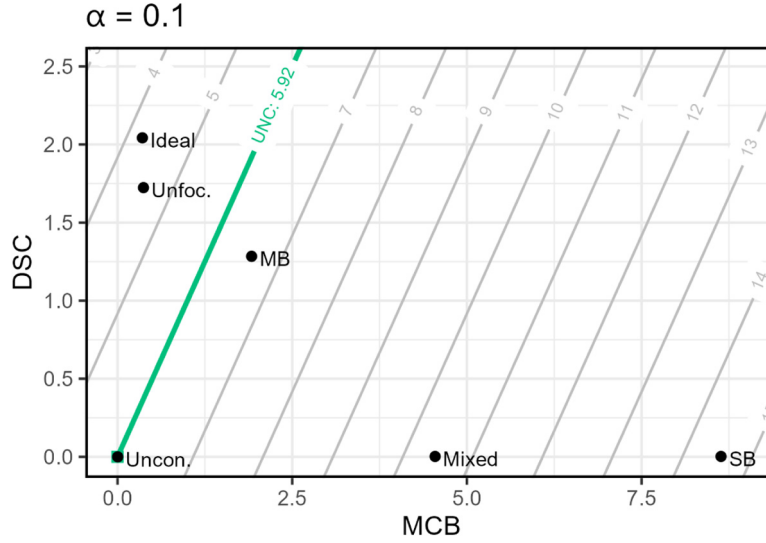


Fig. 1. Miscalibration-discrimination plot for the ideal, the unconditional, the unfocused, the mean-biased (MB), the sign-based (SB), and the mixed forecaster with $\alpha = 0.1$ and $n = 1000$. The miscalibration and discrimination terms of the isotonic interval score decomposition in (F.1) are shown on the x- and y-axis, respectively. The diagonal grey lines are isolines at different values of the interval score, with the green line indicating the estimated uncertainty term of the decomposition, when $\widehat{MCB}_A = \widehat{DSC}_A = 0$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

section, more accurate estimates of the decomposition terms would be obtained by using a model with other assumptions in this case. Such an antitonic relationship between the prediction intervals and the outcomes is unlikely to materialise in reality, but it is included here to highlight the behaviour of the estimated decomposition terms in this case. Aside from this, the estimated decomposition terms capture the theoretical properties of the different forecasts.

Table 1 displays the average interval score for the six forecasts, as well as the (unconditional) empirical frequency with which the outcome falls below and above the prediction interval bounds, before and after being recalibrated using IDR. The unconditional calibration of the unconditional, ideal, and unfocused forecasters is seen by the original forecasts having lower and upper coverages close to the nominal 0.05 level. For the recalibrated prediction intervals, we show these coverages corresponding to both the open and closed interval forecasts, i.e. $P(Y < L)$, $P(Y > U)$ and $P(Y \leq L)$, $P(Y \geq U)$. Since IDR yields prediction intervals that are unconditionally calibrated, the nominal frequency $\alpha/2 = 0.05$ is situated between these two coverages (see (2)). The open and closed coverages are the same for the prediction intervals before recalibration, since $P(Y \in \{L, U\}) = 0$ for the original interval forecasts.

The average length of the interval forecasts can either increase or decrease after the recalibration procedure, as only the order of the intervals is relevant when fitting IDR. Hence, to determine whether the isotonic decomposition could reasonably be applied to a collection of interval forecasts in practice, one could count how many pairs of prediction intervals are ordered. If only a small number of prediction intervals are ordered, then the resulting IDR distributions will be coarse step functions, with sets of quantiles that correspond to relatively large intervals. In this case, the difference between $P(Y \in (L, U))$ and $P(Y \in [L, U])$ can be large, making the notion of calibration in (3) somewhat weaker. This relates to the second requirement in Remark 12.

5. Applications

5.1. Data

In this section, we apply the interval score decomposition to prediction intervals obtained from several prediction algorithms applied to three benchmark regression datasets. The data and methods are a subset of those considered by Romano et al. (2019). While Romano et al. (2019) study the average length and coverage of prediction intervals across 11 benchmark datasets, we present results separately for three of these datasets of different sizes, allowing us to analyse how the amount of data affects the interval score decomposition terms. We consider predictions for the Student-Teacher Achievement Ratio (STAR) in Tennessee (Mosteller, 1995), for which $n = 433$ test data points are available; the usage of a bike sharing system in Washington D.C. (Fanaee-T & Gama, 2014), with $n = 2178$ test data points; and the number of comments that a Facebook post will receive (Singh, 2016), for which $n = 8190$.

For each of the three datasets, prediction intervals were obtained from eight different methods. These methods include three versions of the original split conformal prediction algorithm (Vovk et al., 2018) using random forests (RF), ridge

Table 1

The performance of the 0.9-level prediction intervals from the six forecasts in the simulation study. The second and third columns denote whether each forecast is unconditionally calibrated (UC) and auto-calibrated (AC) in theory. The fourth column presents the average interval scores of the original forecasts. The remaining columns show the unconditional calibration and length of the prediction intervals before and after recalibration using IDR. Lower and upper coverage correspond to the empirical estimates of $P(Y \leq L)$ and $P(Y \geq U)$, respectively, which should be close to 0.05 if the forecasts are unconditionally calibrated. For the IDR-recalibrated intervals, the coverage is shown for both the open (L, U) and closed $[L, U]$ intervals. The unconditional calibration of IDR ensures that 0.05 lies between these two coverages (see (2)).

Forecast	UC	AC	Interval score	Original			Recalibrated		
				Lower cov.	Upper cov.	Length	Lower cov.	Upper cov.	Length
Unconditional	✓	✓	5.92	0.046	0.049	4.65	0.049 - 0.050	0.050 - 0.051	4.56
Ideal	✓	✓	4.23	0.062	0.061	3.29	0.040 - 0.067	0.040 - 0.067	3.29
Unfocused	✓	✗	4.56	0.055	0.057	3.68	0.041 - 0.063	0.043 - 0.063	3.57
Mean-biased	✗	✗	6.55	0.135	0.135	3.29	0.042 - 0.063	0.041 - 0.063	3.86
Sign-biased	✗	✗	14.55	0.220	0.217	3.29	0.049 - 0.053	0.050 - 0.051	4.56
Mixed	✗	✗	10.46	0.133	0.142	3.95	0.049 - 0.051	0.050 - 0.053	4.56

regression (Ridge), and neural networks (Net), as well as locally adaptive variants of each. The local variants apply the conformal prediction algorithms to normalised residuals, allowing the methods to adapt to heteroskedasticity (Lei et al., 2018; Papadopoulos et al., 2008). We additionally consider the conformalised quantile regression approach of Romano et al. (2019) in combination with a neural network (CQR_Net), as well as a non-conformal quantile regression method combined with a neural network (Q_Net). Additional details regarding the methods can be found in the supplementary material of Romano et al. (2019). The methods were implemented using code from <https://github.com/yromano/cqr>. Two additional methods based on quantile random forests were also considered by Romano et al. (2019), though these are omitted from the analysis here, since no working code is available to reproduce them. IDR was implemented using the `isodistrreg` package in R, and all prediction intervals were generated at level $\alpha = 0.1$.

5.2. Results

Table 2 presents the average interval score, unconditional lower and upper coverages, and average length of the prediction intervals obtained from the eight methods for each of the three datasets. Every method has approximately the correct unconditional coverage in all datasets, with both the lower and upper coverage close to the nominal level 0.05. The different methods yield intervals of very similar lengths for the STAR dataset, resulting in similar interval scores. The lengths and scores vary more between the methods in the other two datasets, where the CQR and quantile neural networks always yield the narrowest intervals and lowest interval scores.

The length and unconditional coverage of the IDR-recalibrated prediction intervals are also shown in Table 2. Due to the in-sample calibration guarantees of IDR, the coverage inequalities in (1) are satisfied by construction. To illustrate this, the coverages of the open and closed prediction intervals are shown, which always contain $\alpha/2 = 0.05$. A small difference between these open and closed coverages suggests a smoother IDR fit, resulting in smaller sets of IDR quantiles and more reliable estimates of the interval score decomposition terms. The Local Ridge approach tends to have a larger difference between these coverages, since it often produces nested interval forecasts, reducing the effective sample size to fit IDR (see Remark 12). Similarly, there is an even larger difference between the lower coverages for the Facebook dataset. This is because the variable of interest is censored below at zero, resulting in a high probability that $Y = L$ when the lower bound $L = 0$, in which case $Y \notin (L, U)$ but $Y \in [L, U]$.

Fig. 2 presents miscalibration-discrimination plots obtained from the sample decomposition of the interval score in (12). The decomposition terms allow us to understand why certain prediction methods issue more accurate forecasts than others. For example, for the STAR dataset, the Neural Network and Quantile Neural Network prediction intervals exhibit similar (conditional) miscalibration, but the Quantile Neural Network is more discriminative. Thus, the difference in interval score between these methods is attributable to differences in their information content. Conversely, Random Forests and the CQR Neural Net yield interval forecasts with similar discriminative ability, suggesting that the more accurate Random Forests predictions are due to improved auto-calibration.

However, for the STAR dataset, the miscalibration term of most methods is larger than the discrimination term, meaning they perform worse than a hypothetical unconditional forecast, whose score is equal to the uncertainty term of the decomposition. The reason for this is likely the smaller amount of data used to train the models. The more complex neural network approaches therefore overfit the training data, resulting in poorer conditional calibration and thus less accurate predictions. The high discrimination and miscalibration of the Local Ridge method is likely due to the comparatively low percentage of comparable interval forecasts (see Table 2), resulting in a worse estimation of the decomposition terms.

The forecasting methods are more useful when applied to the Bike and Facebook datasets, where they are ranked similarly: the more accurate forecasts are obtained from the Quantile and CQR Neural Networks, followed in general

Table 2

The performance of the 0.9-level prediction intervals from the eight prediction methods applied to the STAR (top), Bike (middle), and Facebook (bottom) datasets. The second column denotes the proportion of comparable interval forecasts for each prediction method. The third column presents the average interval scores of the original forecasts. The remaining columns show the unconditional calibration and length of the prediction intervals before and after recalibration using IDR. Lower and upper coverage correspond to the empirical estimates of $P(Y \leq L)$ and $P(Y \geq U)$, respectively, which should be close to 0.05 if the forecasts are unconditionally calibrated. For the IDR-recalibrated intervals, the coverage is shown for both the open (L, U) and closed $[L, U]$ intervals. The unconditional calibration of IDR ensures that 0.05 lies between these two coverages (see (2)).

STAR	Comp. (%)	Interval score	Original			Recalibrated		
			Lower cov.	Upper cov.	Length	Lower cov.	Upper cov.	Length
Ridge	100.0	0.22	0.06	0.07	0.17	0.04 - 0.06	0.04 - 0.07	0.17
Local Ridge	51.0	0.23	0.03	0.08	0.19	0.03 - 0.14	0.03 - 0.11	0.15
Random Forests	100.0	0.22	0.04	0.07	0.18	0.04 - 0.06	0.04 - 0.06	0.17
Local Random Forests	99.1	0.22	0.04	0.07	0.18	0.04 - 0.06	0.04 - 0.06	0.17
Neural Net	100.0	0.25	0.03	0.07	0.20	0.04 - 0.06	0.04 - 0.06	0.19
Local Neural Net	97.6	0.28	0.02	0.09	0.22	0.04 - 0.06	0.04 - 0.08	0.18
CQR Neural Net	91.0	0.26	0.03	0.08	0.19	0.04 - 0.07	0.04 - 0.06	0.18
Quantile Neural Net	88.7	0.23	0.04	0.05	0.20	0.03 - 0.09	0.04 - 0.07	0.17

Bike	Comp. (%)	Interval score	Original			Recalibrated		
			Lower cov.	Upper cov.	Length	Lower cov.	Upper cov.	Length
Ridge	100.0	3.30	0.02	0.09	2.21	0.04 - 0.07	0.04 - 0.06	2.11
Local Ridge	60.7	2.99	0.02	0.08	2.18	0.03 - 0.11	0.04 - 0.09	1.87
Random Forests	100.0	1.19	0.05	0.06	0.71	0.03 - 0.07	0.03 - 0.08	0.60
Local Random Forests	99.3	1.16	0.06	0.06	0.68	0.04 - 0.08	0.03 - 0.08	0.59
Neural Net	100.0	1.30	0.03	0.08	0.73	0.03 - 0.07	0.04 - 0.07	0.67
Local Neural Net	99.1	1.15	0.05	0.06	0.65	0.03 - 0.07	0.04 - 0.07	0.60
CQR Neural Net	96.1	0.93	0.05	0.06	0.59	0.03 - 0.09	0.03 - 0.09	0.58
Quantile Neural Net	94.5	0.79	0.05	0.04	0.59	0.03 - 0.10	0.03 - 0.10	0.49

Facebook	Comp. (%)	Interval score	Original			Recalibrated		
			Lower cov.	Upper cov.	Length	Lower cov.	Upper cov.	Length
Ridge	100.0	12.59	0.06	0.04	3.72	0.00 - 0.55	0.04 - 0.08	3.05
Local Ridge	59.5	8.41	0.05	0.05	3.46	0.00 - 0.56	0.03 - 0.09	2.86
Random Forests	100.0	9.26	0.06	0.04	1.87	0.01 - 0.56	0.03 - 0.22	2.09
Local Random Forests	96.9	7.56	0.06	0.04	1.62	0.01 - 0.56	0.03 - 0.24	2.06
Neural Net	100.0	11.52	0.05	0.05	2.34	0.00 - 0.56	0.04 - 0.08	2.72
Local Neural Net	85.3	8.39	0.04	0.05	1.78	0.00 - 0.56	0.04 - 0.09	2.58
CQR Neural Net	59.9	6.17	0.02	0.08	1.11	0.01 - 0.56	0.03 - 0.18	2.12
Quantile Neural Net	58.5	4.34	0.06	0.05	1.57	0.01 - 0.56	0.03 - 0.21	1.94

by the localised and then non-localised versions of Random Forests and the Neural Network, with the Ridge regression forecasts performing worst. For these datasets, the decomposition terms suggest that the ridge-based methods are less informative than the other methods, resulting in a lower discrimination term. In contrast, the flexibility of the Quantile and CQR Neural Networks generally allows them to produce better-calibrated forecasts, particularly when the sample size n is large. However, the miscalibration term is non-zero for all methods, suggesting that none of the generated forecasts are auto-calibrated.

The decomposition terms again allow us to better understand the behaviour of the different methods in these datasets. Most forecasting methods (except for the ridge regression approaches) are found to be similarly discriminative, suggesting that differences in the interval score between these methods are mainly attributable to differences in miscalibration. For example, in the Facebook dataset, the Quantile Neural Net predictions contain the same amount of information as the Random Forest predictions ($\widehat{DSC}_A = 11.32$ versus $\widehat{DSC}_A = 11.27$) but are considerably better auto-calibrated ($\widehat{MCB}_A = 1.23$ versus $\widehat{MCB}_A = 6.10$). This improvement in conditional calibration results in a 50% improvement in the average interval score (4.34 versus 9.26).

Finally, the decomposition terms in the miscalibration-decomposition plots also provide information that is not available by simply comparing the average lengths of unconditionally calibrated prediction intervals, which is a commonly used method for comparing competing interval forecasts. For example, in the Facebook dataset, Local Random Forests generate relatively short prediction intervals while achieving perfect unconditional coverage of 0.9. However, Fig. 2 and Table 2 demonstrate that the shorter intervals are indicative of more informative forecasts (seen by a large discrimination term), but not of better conditional calibration (seen by a large miscalibration term). Local Random Forests therefore achieve a worse overall interval score than other methods, despite their shorter intervals and perfect unconditional coverage. This demonstrates that a forecast can be unconditionally calibrated with prediction intervals that are very

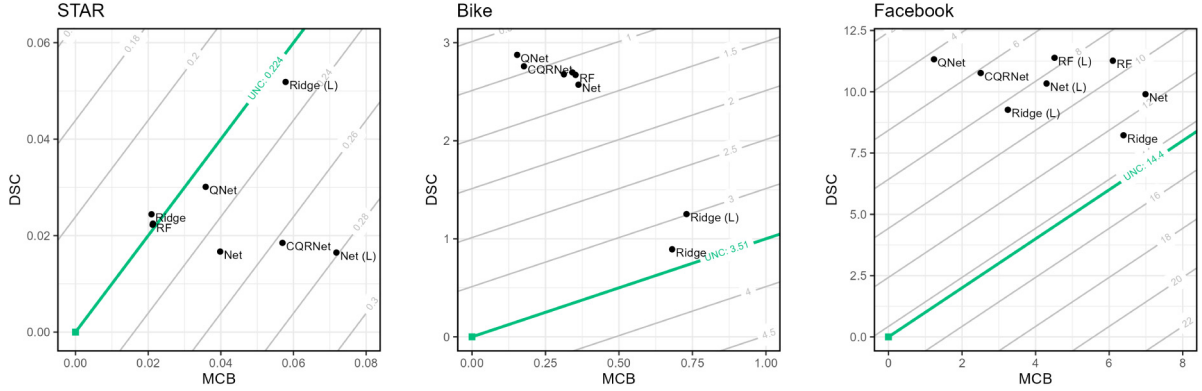


Fig. 2. Isotonic miscalibration-discrimination plots for the eight prediction methods applied to the STAR (left), Bike (middle), and Facebook (right) datasets. See Fig. 1 for details. For the STAR dataset, the decomposition terms of RF and RF (L) are almost equivalent. For the Bike dataset, the RF, Net, RF (L), and Net (L) are clustered together. Some of these labels have therefore been omitted for ease of presentation.

narrow, despite not being conditionally calibrated. This poor conditional calibration is reflected in the interval score, highlighting why such scoring functions are informative. However, from the interval score alone, it is not possible to diagnose why the forecast performs poorly. For this, we additionally require the estimated decomposition terms.

6. Conclusion

This paper provided a general framework with which to assess the conditional coverage of interval forecasts. The framework leverages a decomposition of the interval score into terms quantifying the discriminative ability and conditional miscalibration of the forecast, as well as the inherent uncertainty of the forecasting problem. While this decomposition has received much attention in the literature, it has not previously been applied to the interval score. In addition to providing a measure of conditional forecast calibration, the decomposition allows differences in forecast accuracy to be attributed to differences in information content and miscalibration, facilitating a more comprehensive understanding of forecast performance.

The score decompositions discussed herein assess auto-calibration, which is an intuitive and relatively strong notion of forecast calibration that is popular in the probabilistic forecasting community. However, the interval score decomposition can also be extended to assess stronger notions of conditional calibration (see e.g. Allen et al., 2023; Ehm & Ovcharov, 2017). For example, the conditional calibration of interval forecasts often refers to the intervals having the correct coverage conditional on the covariates used to construct the forecast. In Appendix C, we demonstrated that this notion of conditional calibration can be quantified using a decomposition of the interval score analogous to that studied here. However, this decomposition is not particularly useful in practice, since it relies on estimates of the conditional quantiles of the outcome variable given the covariates, which is essentially the original forecasting problem. Auto-calibration can instead be interpreted as a meaningful proxy for conditional calibration that is easier to estimate.

Throughout this paper, we focused on central (or equal-tailed) prediction intervals. However, the score decomposition framework can also readily be extended to generalised interval scores, and to non-central prediction intervals; these extensions were detailed in Appendix B. While different interval scores will generally lead to different measures of conditional miscalibration, all will return a value of zero when the forecasts are auto-calibrated. Brehmer and Gneiting (2021) additionally introduced a consistent scoring function for modal interval forecasts, defined as the interval that maximises the probability of containing the outcome, subject to having a fixed length. This scoring function similarly admits a decomposition into discrimination and miscalibration terms, allowing us to quantify the conditional calibration of modal interval forecasts. This analysis is presented in Appendix D. However, this notion of calibration for modal interval forecasts is not directly related to conditional coverage, which was the main focus of this paper.

Moreover, the approach could be used to evaluate the conditional miscalibration of central prediction intervals at several coverage levels simultaneously. For example, epidemiological forecasts typically take the form of collections of many prediction intervals, and Bracher et al. (2021) suggested evaluating these forecasts via a weighted interval score that corresponds to a weighted sum of the standard interval score at different coverage levels. This weighted interval score could similarly be decomposed using the methods proposed herein, allowing for an assessment of the conditional calibration of these collections of prediction intervals. As the number of prediction intervals increases, the decomposition terms tend towards those for the CRPS decomposition proposed by Arnold et al. (2024).

We estimated the score decomposition terms using isotonic distributional regression (IDR), which assumes that there is a (roughly) isotonic relationship between the forecasts to be assessed and the corresponding observations. That is, a higher interval forecast generally corresponds to a higher outcome. This isotonicity assumption has become standard when estimating score decomposition terms (Arnold et al., 2024; Dimitriadis et al., 2021; Gneiting & Resin, 2023; Gneiting et al.,

2023), and should be appropriate in many applications of interval forecasts. If this isotonicity assumption is not valid, then other quantile regression models could be used to obtain more accurate estimates of the decomposition terms. [Dimitriadis and Puke \(2026\)](#), for example, used linear quantile regression to decompose the average quantile score, which should extend easily to the interval score. However, details for other methods would need to be worked out. As [Dimitriadis and Puke \(2026\)](#) remark, ‘the choice of estimation method remains an active area of research,’ and further research is therefore required to understand what methods yield meaningful decomposition terms.

One advantage of using IDR is that it provides forecast distributions that simultaneously optimise a large class of scoring rules over the training data, subject to the isotonicity assumption ([Henzi et al., 2021](#), Theorem 2). This optimality ensures that the estimated decomposition terms are non-negative, which is required for them to have a meaningful interpretation. The quantile score, interval score, weighted interval score, and CRPS are all examples of scoring rules within this class. This class additionally contains weighted versions of the CRPS ([Gneiting & Ranjan, 2011](#)). By applying an analogous decomposition to the threshold-weighted CRPS, we would obtain a framework with which to assess conditional notions of tail calibration of probabilistic forecasts ([Allen et al., 2025](#)).

One disadvantage of IDR, however, is that it is generally not straightforward to perform inference on the estimated decomposition terms. Following [Dimitriadis et al. \(2024\)](#), the estimated decomposition terms were here used to produce informative diagnostic plots that simultaneously displayed the discriminative ability and the conditional miscalibration of interval forecasts. These miscalibration-discrimination plots would be more informative if they additionally provided a measure of uncertainty in the estimated terms. This could then be used to derive methods to formally test whether or not a forecast issues auto-calibrated predictions, which would complement the existing literature on conditional coverage tests. Tests could also be derived to assess whether there is a significant difference between the discrimination or miscalibration terms of competing forecast methods, along the lines of [Dimitriadis and Puke \(2026\)](#). This would allow us to determine whether one forecast issues more calibrated or more informative forecasts than another. Such a test would have to take into account that different forecast methods may yield a different number of comparable prediction intervals (see [Table 2](#)).

Finally, we remark that the score decomposition framework is more theoretically appealing than simply comparing the average length of prediction intervals that are unconditionally calibrated, which is the common procedure in conformal prediction, for example. Even if interval forecasts are auto-calibrated, there is no guarantee that a shorter interval corresponds to a more informative prediction, in which case no direct connection between the length of the prediction interval and conditional calibration or forecast accuracy can be established.

CRedit authorship contribution statement

Sam Allen: Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Data curation, Conceptualization. **Julia Burnello:** Writing – review & editing, Writing – original draft, Software, Methodology. **Johanna Ziegel:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This paper has greatly benefited from comments by Timo Dimitriadis and two anonymous referees. Sam Allen gratefully acknowledges funding from the Vector Stiftung through the Young Investigator Group “Artificial Intelligence for Probabilistic Weather Forecasting”.

Appendix A. Proofs

Proof of Proposition 5. For parts (i) and (ii), it suffices to take expectations in (4) and (6), respectively. For the reverse directions, the unfocused forecast in Section 4 is unconditionally calibrated but neither conditionally calibrated nor auto-calibrated. For part (iii), it suffices to take conditional expectations in (4). For the reverse direction, the unconditional forecast in Section 4 is auto-calibrated but not conditionally calibrated. If $[L, U]$ is a sufficient statistic for Y relative to X , then $Q_{\alpha/2}(Y | X) = Q_{\alpha/2}(Y | L, U)$ and $Q_{1-\alpha/2}(Y | X) = Q_{1-\alpha/2}(Y | L, U)$, in which case the definitions of auto-calibration and conditional calibration are equivalent (see [Remark 4](#)). $\square$

Proof of Propositions 8, 16 and 22. We only prove the properties of the decomposition assessing isotonic interval calibration ([Proposition 22](#)), as the statements for the decompositions evaluating auto-calibration ([Proposition 8](#)) and conditional calibration ([Proposition 16](#)) follow immediately, since $\sigma(L, U)$ and $\sigma(X)$ are also σ -lattices, and all $\mathcal{A}(L, U)$ -measurable random variables are also $\sigma(L, U)$ -measurable and thus $\sigma(X)$ -measurable. Here, $\mathcal{A}(L, U)$ denotes the σ -lattice generated by $[L, U]$ (see [Appendix F](#) for details).

We recall Proposition 6.3 in [Arnold et al. \(2024\)](#) on the characterisation of conditional quantiles given a general σ -lattice. It states that for all $\alpha \in (0, 1)$, a random variable belongs to the conditional quantile $Q_\alpha(Y | \mathcal{A})$ given some σ -lattice $\mathcal{A}$ if and only if it minimises the expected quantile score $\mathbb{E}[QS_\alpha(Z, Y)]$ over all $\mathcal{A}$ -measurable random variables Z .

Since $IS_\alpha([L, U], Y) = \frac{2}{\alpha} [QS_{\alpha/2}(L, Y) + QS_{1-\alpha/2}(U, Y)]$, the discrimination term can be written as

$$\begin{aligned} DSC_1 &= \frac{2}{\alpha} \left[\mathbb{E}[QS_{\alpha/2}(q_{\alpha/2}(Y), Y)] - \mathbb{E}[QS_{\alpha/2}(q_{\alpha/2}(Y | \mathcal{A}(L, U)), Y)] \right] \\ &\quad + \frac{2}{\alpha} \left[\mathbb{E}[QS_{1-\alpha/2}(q_{1-\alpha/2}(Y), Y)] - \mathbb{E}[QS_{1-\alpha/2}(q_{1-\alpha/2}(Y | \mathcal{A}(L, U)), Y)] \right]. \end{aligned} \quad (A.1)$$

This satisfies $DSC_1 \geq 0$ because $q_{\alpha/2}(Y)$ and $q_{1-\alpha/2}(Y)$ are both $\mathcal{A}(L, U)$ -measurable, since they are constant. Moreover, both terms in the large square brackets are non-negative. Therefore, $DSC_1 = 0$ if and only if $q_{\alpha/2}(Y | \mathcal{A}(L, U)) \in Q_{\alpha/2}(Y)$ and $q_{1-\alpha/2}(Y | \mathcal{A}(L, U)) \in Q_{1-\alpha/2}(Y)$ almost surely.

The miscalibration term can be rewritten in a similar way, and since L and U are both $\mathcal{A}(L, U)$ -measurable by definition, it follows that $MCB_1 \geq 0$ with equality if and only if $L \in Q_{\alpha/2}(Y | \mathcal{A}(L, U))$ and $U \in Q_{1-\alpha/2}(Y | \mathcal{A}(L, U))$ almost surely. $\square$

Proof of Proposition 10.

For the proof of part (i), suppose that $\widehat{DSC}_A = 0$. This implies that $\bar{q}_{\alpha/2}$ is a minimiser of the expression in (13). Hence, $\bar{q}_{\alpha/2} \geq \tilde{q}_{\alpha/2,i}$ for all $i = 1, \dots, n$. The lower isotonic quantiles $\tilde{q}_{\alpha/2,i}$ induce a partition of $\{1, \dots, n\}$ such that $\tilde{q}_{\alpha/2,i}$ is the lower quantile of the empirical distribution of y_j for $j \in B$, where $B \subset \{1, \dots, n\}$ is the partition element that contains i (Henzi et al., 2021, proof of Theorem 2.2). The partitions do not need to be unique. The stochastic order of the empirical distributions implies that $\bar{q}_{\alpha/2} \leq \tilde{q}_{\alpha/2,i}$ for all $i = 1, \dots, n$. This implies that $\bar{q}_{\alpha/2} = \tilde{q}_{\alpha/2,i}$ for all $i = 1, \dots, n$. Part (ii) is immediate. $\square$

Proof of Propositions 17 and 23.

Since $\sigma(L, U) \subseteq \sigma(X)$, a conditional quantile given $\sigma(L, U)$ is $\sigma(X)$ -measurable. Therefore,

$$\begin{aligned} \mathbb{E}[QS_{\alpha/2}(q_{\alpha/2}(Y | X), Y)] &\leq \mathbb{E}[QS_{\alpha/2}(q_{\alpha/2}(Y | L, U), Y)] \\ \mathbb{E}[QS_{1-\alpha/2}(q_{1-\alpha/2}(Y | X), Y)] &\leq \mathbb{E}[QS_{1-\alpha/2}(q_{1-\alpha/2}(Y | L, U), Y)] \end{aligned} \quad (A.2)$$

and therefore $MCB_C \geq MCB_A$. Since the interval score is non-negative by definition, we have $\mathbb{E}[IS_\alpha([L, U], Y)] \geq MCB_C \geq MCB_A$. This proves Proposition 17. Proposition 23 follows analogously to above, but with $\sigma(L, U)$ replaced by $\mathcal{A}(L, U)$ and $\sigma(X)$ replaced by $\sigma(L, U)$. $\square$

Proof of Proposition 18.

The expected modal interval score can be written as

$$\mathbb{E}[MIS_c(Z, Y)] = -P(Y \in [Z - c, Z + c]) = -\mathbb{E}[P(Y \in [Z - c, Z + c] | Z)].$$

By definition of the midpoint functional, $m_c(Y | Z) \in \arg \max_{a \in \mathbb{R}} P(Y \in [a - c, a + c] | Z)$. Hence,

$$\begin{aligned} MCB_{Mod} &= \mathbb{E}[MIS_c(Z, Y)] - \mathbb{E}[MIS_c(m_c(Y | Z), Y)] \\ &= \mathbb{E}[P(Y \in [m_c(Y | Z) - c, m_c(Y | Z) + c] | Z)] - \mathbb{E}[P(Y \in [Z - c, Z + c] | Z)] \\ &= \mathbb{E}[\arg \max_{a \in \mathbb{R}} P(Y \in [a - c, a + c] | Z) - P(Y \in [Z - c, Z + c] | Z)]. \end{aligned}$$

The term $\arg \max_{a \in \mathbb{R}} P(Y \in [a - c, a + c] | Z) - P(Y \in [Z - c, Z + c] | Z) \geq 0$ almost surely, with equality if and only if $Z \in M_c(Y | Z)$ almost surely, and thus this holds also for the miscalibration term itself. An analogous result holds for the discrimination term DSC_{Mod} by replacing Z above with $m_c(Y)$. $\square$

Proof of Proposition 20. Part (i) follows from Arnold and Ziegel (2025, Proposition 5.3). For the reverse direction, see Arnold and Ziegel (2025, Examples 5.1 and 5.2). Part (ii) follows from Arnold and Ziegel (2025, Proposition 6.3) and the fact that L and U are $\mathcal{A}(L, U)$ -measurable. The condition in the reverse direction of part (ii) implies that $Q_{\beta}(Y | L, U) = Q_{\beta}(Y | \mathcal{A}(L, U))$; see the last paragraph of Section 6.1 in Arnold and Ziegel (2025). In this case, the definitions of auto-calibration and isotonic calibration in (6) and (F.2) are equivalent. $\square$

Appendix B. Generalisations

B.1. Extension to non-central prediction intervals

The interval score is constructed to evaluate central prediction intervals. However, it can readily be extended to evaluate non-central prediction intervals by combining the quantile score at different quantile levels. More generally, a non-central interval score at levels $\alpha_1, \alpha_2 \in (0, 1)$, with $\alpha_1 < \alpha_2$, could be defined as

$$\begin{aligned} IS_{\alpha_1, \alpha_2}([\ell, u], y) &= \frac{1}{\alpha_1} QS_{\alpha_1}(\ell, y) + \frac{1}{1 - \alpha_2} QS_{\alpha_2}(u, y) \\ &= |u - \ell| + \frac{1}{\alpha_1} \mathbb{1}\{y < \ell\}(\ell - y) + \frac{1}{1 - \alpha_2} \mathbb{1}\{y > u\}(y - u). \end{aligned}$$

The standard interval score is a particular example of this when $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - \alpha/2$. The notions of interval forecast calibration in Section 2 can similarly be generalised to the non-central case.

Definition 13. A non-central interval forecast $[L, U]$ at levels α_1, α_2 is *unconditionally calibrated* if

$$P(Y < L) \leq \alpha_1 \leq P(Y \leq L) \quad \text{and} \quad P(Y < U) \leq \alpha_2 \leq P(Y \leq U).$$

This simplifies to $P(Y \leq L) = \alpha_1$ and $P(Y \leq U) = \alpha_2$ when $P(Y \in \{L, U\}) = 0$.

Definition 14. A non-central interval forecast $[L, U]$ at levels α_1, α_2 is *conditionally calibrated* given X if

$$P(Y < L | X) \leq \alpha_1 \leq P(Y \leq L | X) \quad \text{and} \quad P(Y < U | X) \leq \alpha_2 \leq P(Y \leq U | X).$$

Definition 15. A non-central interval forecast $[L, U]$ at levels α_1, α_2 is *auto-calibrated* if

$$P(Y < L | L, U) \leq \alpha_1 \leq P(Y \leq L | L, U) \quad \text{and} \quad P(Y < U | L, U) \leq \alpha_2 \leq P(Y \leq U | L, U).$$

Analogously to Remark 4, conditional calibration can alternatively be written as

$$L \in Q_{\alpha_1}(Y | X) \quad \text{and} \quad U \in Q_{\alpha_2}(Y | X),$$

while auto-calibration can be expressed as

$$L \in Q_{\alpha_1}(Y | L, U) \quad \text{and} \quad U \in Q_{\alpha_2}(Y | L, U).$$

The relationships between the different notions of calibration listed in Proposition 5 extend readily to these more general definitions of calibration for non-central prediction intervals. The auto-calibration of the non-central interval forecasts can then be assessed analogously to that in Section 3, by decomposing the non-central interval score by comparing the expected score to the expected score of the non-central intervals obtained from the unconditional and conditional distributions of Y at the relevant levels. All of the desirable properties of the decomposition terms extend naturally to the non-central case.

B.2. Extension to generalised interval scores

As mentioned in Section 3, the interval score is strictly consistent for central prediction intervals, which follows from the representation of the interval score as a weighted sum of two quantile scores. However, the quantile score is not the only scoring function that is strictly consistent for quantiles. In particular, the quantile score in (10) can be extended to the generalised quantile score (or generalised piecewise linear loss), which takes the form

$$\text{GQS}_{\alpha, g}(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(g(x) - g(y)),$$

for some non-decreasing function $g : \mathbb{R} \rightarrow \mathbb{R}$ (Saerens, 2000; Thomson, 1979).

Using this, Brehmer and Gneiting (2021) define a generalised interval score of the form

$$\text{GIS}_{\alpha, g_1, g_2}([\ell, u], y) = w_1 \text{GQS}_{\frac{\alpha}{2}, g_1}(\ell, y) + w_2 \text{GQS}_{1-\frac{\alpha}{2}, g_2}(u, y),$$

for non-decreasing real-valued functions g_1, g_2 , and non-negative weights $w_1, w_2 \in (0, \infty)$. Brehmer and Gneiting (2021) remark that ‘essentially all strictly consistent scoring functions for [the central $(1 - \alpha)$ -level prediction interval] are equivalent to this form.’ The standard interval score is recovered when g_1 and g_2 are the identity function and $w_1 = w_2 = 2/\alpha$. This is the most commonly used interval score, which is also the only interval score that is symmetric and translation invariant (Brehmer & Gneiting, 2021, Theorem 3.1). The generalised interval score essentially transforms the observation and the prediction interval before calculating the standard interval score, with the transformation allowing the events $\{y > u\}$ and $\{y < \ell\}$ to be penalised (and weighted) differently.

The generalised interval score can be decomposed analogously to the standard interval score. In particular, we can replace the interval score in the decomposition terms with the generalised interval score, and apply this to the quantiles obtained from isotonic distributional regression. If the functions g_1 and g_2 are strictly increasing, then the generalised interval score is strictly consistent for the central prediction interval, and we recover exactly the same theoretical properties as outlined in Section 3.

Different interval scores will generally result in different measures of the discrimination and miscalibration terms, which may result in different orderings when these terms are compared between different forecast models. While this may seem undesirable, it aligns with the idea that the generalised interval score allows the events $\{y > u\}$ and $\{y < \ell\}$ to be penalised differently. In any case, all interval scores will result in an auto-calibrated forecast obtaining a miscalibration term of zero. Moreover, the decomposition terms allow the measure of forecast miscalibration to be directly interpreted in terms of the forecast’s overall performance, regardless of what scoring function is chosen to assess this. That is, the miscalibration term can be interpreted as the improvement in score that would be obtained if the forecast were auto-calibrated. This is not the case for existing methods to assess conditional calibration.

Appendix C. Conditional decompositions of the interval score

The decomposition of the interval score in (11) provides a miscalibration term that quantifies auto-calibration. However, the interval score can also be decomposed into a term that quantifies conditional calibration given covariates, as defined in (4). In particular, we have that

$$E[IS_{\alpha}([L, U], Y)] = \text{UNC} - \text{DSC}_C + \text{MCB}_C, \quad (\text{C.1})$$

where

$$\begin{aligned} \text{UNC} &= E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y), q_{1-\frac{\alpha}{2}}(Y)], Y)], \\ \text{DSC}_C &= E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y), q_{1-\frac{\alpha}{2}}(Y)], Y)] - E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y | X), q_{1-\frac{\alpha}{2}}(Y | X)], Y)], \\ \text{MCB}_C &= E[IS_{\alpha}([L, U], Y)] - E[IS_{\alpha}([q_{\frac{\alpha}{2}}(Y | X), q_{1-\frac{\alpha}{2}}(Y | X)], Y)]. \end{aligned}$$

This differs from (11) only in that the conditional quantiles $q_{\alpha/2}(Y | L, U)$ and $q_{1-\alpha/2}(Y | L, U)$ have been replaced with $q_{\alpha/2}(Y | X)$ and $q_{1-\alpha/2}(Y | X)$, the conditional quantiles of Y given X .

This has the same desirable properties as the auto-calibration decomposition, as summarised by the following proposition, which is analogous to Proposition 8 for auto-calibration.

Proposition 16. *The decomposition of the expected interval score in (C.1) satisfies the following properties:*

- (i) $\text{DSC}_C \geq 0$ with equality if and only if $q_{\alpha/2}(Y | X) \in Q_{\alpha/2}(Y)$ and $q_{1-\alpha/2}(Y | X) \in Q_{1-\alpha/2}(Y)$.
- (ii) $\text{MCB}_C \geq 0$ with equality if and only if the interval forecast $[L, U]$ is conditionally calibrated given X , as defined in (4).

As suggested by Proposition 5, if the conditional quantiles of Y given X are equivalent to those given $[L, U]$, then auto-calibration and conditional calibration are equivalent, and (C.1) reduces to (11).

These conditional decompositions were studied by Ehm and Ovcharov (2017) and Allen et al. (2023). Under the assumption that $\sigma(L, U) \subseteq \sigma(X)$, the miscalibration term MCB_C in (C.1) will be at least as large as the miscalibration term MCB in (11), with equality when $\sigma(L, U) = \sigma(X)$.

Proposition 17. *It holds that*

$$E[IS_{\alpha}([L, U], Y)] \geq \text{MCB}_C \geq \text{MCB}_A.$$

However, in practice, this decomposition is arguably less useful than that in (11), since it requires knowledge of the conditional quantiles of Y given X , which is exactly the original forecasting problem. In practice, estimating the conditional quantiles of Y given $[L, U]$ is generally more feasible. For example, the decomposition terms could not generally be implemented with isotonic regression, since there is no reason to assume that a higher covariate value results in a higher outcome; instead, other more restrictive assumptions would have to be made. Moreover, X could be high-dimensional, while the interval forecasts condense the covariate information into a lower-dimensional representation where the isotonicity assumption is more sensible.

Appendix D. Modal intervals

For ease of presentation, we restrict our attention here to distributions with Lebesgue density. Following Brehmer and Gneiting (2021, Theorem 3.10), the unique strictly consistent scoring function to assess modal intervals in this case is

$$\text{MIS}_c(z, y) = -\mathbb{1}\{y \in [z - c, z + c]\}, \quad (\text{D.1})$$

for prespecified $c > 0$. Since the length c is fixed, the modal interval can be parametrised via its midpoint z , rather than its lower and upper bounds. Following Brehmer and Gneiting (2021), for a probability distribution F , we define the (set-valued) midpoint functional given length c as $M_c(F) = \arg \max_{a \in \mathbb{R}} (F(a + c) - F(a - c))$. We use $m_c(F)$ to denote an arbitrary element of $M_c(F)$.

Akin to the general framework outlined by Gneiting and Resin (2023), the expected modal interval score can be decomposed into

$$E[\text{MIS}_c(Z, Y)] = \text{UNC}_{\text{Mod}} - \text{DSC}_{\text{Mod}} + \text{MCB}_{\text{Mod}}, \quad (\text{D.2})$$

where

$$\begin{aligned} \text{UNC}_{\text{Mod}} &= E[\text{MIS}_c(m_c(Y), Y)], \\ \text{DSC}_{\text{Mod}} &= E[\text{MIS}_c(m_c(Y), Y)] - E[\text{MIS}_c(m_c(Y | Z), Y)], \\ \text{MCB}_{\text{Mod}} &= E[\text{MIS}_c(Z, Y)] - E[\text{MIS}_c(m_c(Y | Z), Y)], \end{aligned}$$

with $m_c(Y)$ a midpoint functional of the marginal distribution of Y , and $m_c(Y | Z)$ a midpoint functional of the conditional distribution of Y given Z . As above, the exact choice of midpoint functional is irrelevant, since they all receive the same expected score. The interpretation of the decomposition terms in (D.2) is similar to the interpretation of those in (11): the first term measures forecast uncertainty, the second term quantifies discriminative ability, and the third term constitutes a measure of conditional miscalibration for modal intervals.

The following proposition is an analogue of Proposition 8 for modal intervals (see also Gneiting & Resin, 2023, Theorem 2.23).

Proposition 18. *The decomposition of the expected modal interval score in (D.2) satisfies the following properties:*

- (i) $\text{DSC}_{\text{Mod}} \geq 0$ with equality if and only if $m_c(Y | Z) \in M_c(Y)$.
- (ii) $\text{MCB}_{\text{Mod}} \geq 0$ with equality if and only if $Z \in M_c(Y | Z)$.

This latter requirement, that $Z \in M_c(Y | Z)$, can be interpreted as a notion of auto-calibration for modal interval forecasts. This is essentially a particular example of (conditional) T-calibration, as defined by Gneiting and Resin (2023, Definition 2.7). Hence, these score decompositions can be used to assess the auto-calibration of modal interval forecasts. Note, however, that this notion of calibration has no direct correspondence to the coverage of the interval forecast, which is more commonly of interest when assessing central prediction intervals.

Appendix E. Orderings

To estimate the decomposition of the interval score in Section 3.3, we define a partial order on the space of interval forecasts. This is necessary for the implementation of IDR when estimating the conditional quantiles, since IDR assumes that a larger observation generally occurs when a larger interval forecast is issued. Thus, we require a definition of when one interval forecast is ‘larger’ than another. We choose an ordering that defines an interval forecast as larger than another if both the lower and upper bounds are no smaller than those of the other forecast, with at least one bound being strictly larger. This is similar in essence to stochastic ordering of predictive distributions. While other partial orders could also be employed, not all orderings will yield interval score decomposition terms with the desirable properties listed in Section 3.3.

IDR is designed so that a larger covariate value, according to the chosen partial order, results in a larger predictive distribution, according to the stochastic order on distributions (denoted $\leq_{\text{st}}$ in the following). In our context, the IDR predictive distributions uniquely minimise the average interval score over all n -tuples $(F_1, \dots, F_n)$ of distribution functions that satisfy $F_i \leq_{\text{st}} F_j$ if $[\ell_i, u_i] \leq [\ell_j, u_j]$ for $i, j \in \{1, \dots, n\}$, where $[\ell_1, u_1], \dots, [\ell_n, u_n]$ represent the original interval forecasts on which IDR is applied. We additionally have that the IDR prediction intervals uniquely minimise the average interval score over all n -tuples $([\ell_1^*, u_1^*], \dots, [\ell_n^*, u_n^*])$ of interval forecasts that satisfy $\ell_i^* \leq \ell_j^*$ and $u_i^* \leq u_j^*$ if $[\ell_i, u_i] \leq [\ell_j, u_j]$ for $i, j \in \{1, \dots, n\}$. These statements hold regardless of the ordering chosen on the space of interval forecasts.

The theoretical guarantees in Proposition 10 rely on the fact that the original interval forecasts satisfy the same order relations as the prediction intervals obtained from the IDR-recalibrated predictive distributions. That is, for the partial order employed in Section 3.3, the n -tuple of original forecasts $([\ell_1, u_1], \dots, [\ell_n, u_n])$ satisfies $\ell_i \leq \ell_j$ and $u_i \leq u_j$ if $[\ell_i, u_i] \leq [\ell_j, u_j]$, since this is simply the definition of the partial order. They therefore belong to the class of n -tuples over which IDR is the optimal solution. The average interval score of the IDR-recalibrated prediction intervals is thus no worse than that of the original forecasts, which ensures that the miscalibration term of the isotonic decomposition is non-negative.

As an example of where this does not hold, consider the order on the space of interval forecasts defined by $[\ell_i, u_i] \leq [\ell_j, u_j]$ if $(\ell_i + u_i)/2 \leq (\ell_j + u_j)/2$. That is, one interval forecast is larger than another if its midpoint is larger. This corresponds to a total order rather than a partial order, which could circumvent the practical issues related to non-comparable prediction intervals discussed in Remark 12. However, with this ordering, $[\ell_i, u_i] \leq [\ell_j, u_j]$ does not necessarily imply that $\ell_i \leq \ell_j$ and $u_i \leq u_j$. Hence, while the IDR predictive intervals are still optimal with respect to the average interval score over all intervals that satisfy $\ell_i^* \leq \ell_j^*$ and $u_i^* \leq u_j^*$ if $[\ell_i, u_i] \leq [\ell_j, u_j]$ for $i, j \in \{1, \dots, n\}$, this criterion is not satisfied by the original forecasts. There is therefore no guarantee that the average interval score of the recalibrated prediction intervals is lower than the average interval score of the original interval forecasts, meaning that the miscalibration term of the decomposition could be negative. It is therefore necessary to choose an order that aligns with the stochastic order of predictive distributions that is used in IDR.

Appendix F. Isotonic calibration

To estimate the terms of the decomposition of the interval score in (11), we employ isotonic distributional regression (IDR). Arnold and Ziegel (2025) show that the population version of the IDR predictive distributions is the isotonic conditional law of the outcome variable Y given the covariates. The isotonic conditional law of Y given a random variable Z is the conditional distribution of Y given the σ -lattice generated by Z , denoted by $\mathcal{A}(Z)$. A σ -lattice shares the same definition as a σ -algebra, except that it does not need to be closed under complements. Hence, every σ -algebra is also a σ -lattice, and the σ -algebra generated by a random variable is always larger than the corresponding σ -lattice,

i.e. $\mathcal{A}(Z) \subseteq \sigma(Z)$. The isotonic conditional law of Y given Z coincides with the conditional distribution of Y given Z if and only if the latter is already isotonic in Z (Arnold & Ziegel, 2025, Theorem 3.3). We refer readers to Arnold and Ziegel (2025) and the references therein for further details on σ -lattices and isotonic conditional laws; the basic properties listed above are all that is necessary to understand the results in this section.

Using the relationship between IDR and isotonic conditional laws, we can define a population counterpart to the sample decomposition in (12) as

$$E[IS_\alpha([L, U], Y)] = \text{UNC} - \text{DSC}_I + \text{MCB}_I, \quad (\text{F.1})$$

where UNC is as defined in (11), and

$$\begin{aligned} \text{DSC}_I &= E[IS_\alpha([q_{\frac{\alpha}{2}}(Y), q_{1-\frac{\alpha}{2}}(Y)], Y)] - E[IS_\alpha([q_{\frac{\alpha}{2}}(Y | \mathcal{A}(L, U)), q_{1-\frac{\alpha}{2}}(Y | \mathcal{A}(L, U))], Y)], \\ \text{MCB}_I &= E[IS_\alpha([L, U], Y)] - E[IS_\alpha([q_{\frac{\alpha}{2}}(Y | \mathcal{A}(L, U)), q_{1-\frac{\alpha}{2}}(Y | \mathcal{A}(L, U))], Y)], \end{aligned}$$

with $\mathcal{A}(L, U)$ the σ -lattice generated by L and U , and $q_\alpha(Y | \mathcal{A}(L, U))$ an arbitrary α -quantile of the isotonic conditional law of Y given L and U .

This facilitates a definition of isotonic calibration for interval forecasts, which is analogous to the definition of isotonic calibration for probabilistic forecasts introduced by Arnold and Ziegel (2025, Definition 5.1).

Definition 19. An interval forecast $[L, U]$ is *isotonically calibrated* if

$$L \in Q_{\frac{\alpha}{2}}(Y | \mathcal{A}(L, U)) \quad \text{and} \quad U \in Q_{1-\frac{\alpha}{2}}(Y | \mathcal{A}(L, U)). \quad (\text{F.2})$$

That is, L and U are quantiles of the isotonic conditional law of Y given the interval forecast. From Remark 4, this definition is analogous to the definition of auto-calibration, except that the conditional distribution of Y given L and U has been replaced by the corresponding isotonic conditional law. However, in contrast to auto-calibration, isotonic calibration does not correspond to a conditional coverage guarantee as in (6).

Under mild conditions, isotonic calibration implies that

$$L \in Q_{\frac{\alpha}{2}}(Y | L) \quad \text{and} \quad U \in Q_{1-\frac{\alpha}{2}}(Y | U). \quad (\text{F.3})$$

See Arnold and Ziegel (2025, Lemma 6.8). That is, the bounds L and U are quantile calibrated predictions for the $\alpha/2$ and $1 - \alpha/2$ quantiles of Y , respectively, in the sense of Gneiting and Resin (2023). The requirement in (F.3) could itself be used to define a notion of quantile calibration of interval forecasts, though since this is generally strictly weaker than isotonic calibration, we do not consider it in the following.

Isotonic interval calibration additionally implies unconditional interval calibration, as defined in (1). However, isotonic calibration is implied by auto-calibration. In other words, isotonic calibration is stronger than unconditional calibration, but weaker than auto-calibration. The relationships between these different notions of interval calibration are summarised below.

Proposition 20. The following relationships hold between the notions of interval calibration:

- (i) Isotonic calibration implies unconditional calibration, but the converse is not true.
- (ii) Auto-calibration implies isotonic calibration. Isotonic calibration implies auto-calibration if, for any $g(L, U) \in Q_\beta(Y | L, U)$, $\beta \in \{\alpha/2, 1 - \alpha/2\}$, it holds that g is componentwise increasing.

Remark 21. The reverse direction in Proposition 20(ii) implies that the conditional $\alpha/2$ and $1 - \alpha/2$ quantiles of Y given L and U are monotonic; note that any conditional quantile $Z \in Q_\beta(Y | L, U)$ for some $\beta \in (0, 1)$ can be written as $Z = g(L, U)$ for some (measurable) function g of L and U .

If Y is continuous and the interval forecast is constant, then auto-calibration, isotonic calibration, and unconditional calibration are equivalent, since the interval $[L, U]$ is deterministic. More generally, auto-calibration is a stronger requirement than isotonic calibration, and both are much stronger than unconditional calibration. Isotonic and auto-calibration are equivalent if the conditional upper and lower quantiles of Y are ordered (formalised in part (ii), above), which is a natural requirement for informative predictions and should be satisfied at least approximately in many relevant forecasting situations. This is the motivation for employing IDR to estimate the decomposition terms.

Using the properties of the isotonic conditional law (see Arnold & Ziegel, 2025), we can deduce that this isotonic decomposition has desirable properties analogous to those of the decomposition in (11).

Proposition 22. The decomposition of the expected interval score in (F.1) satisfies the following properties:

- (i) $\text{DSC}_I \geq 0$ with equality if and only if $q_{\alpha/2}(Y | \mathcal{A}(L, U)) \in Q_{\alpha/2}(Y)$ and $q_{1-\alpha/2}(Y | \mathcal{A}(L, U)) \in Q_{1-\alpha/2}(Y)$.
- (ii) $\text{MCB}_I \geq 0$ with equality if and only if the interval forecast $[L, U]$ is isotonically calibrated.

In this case, the miscalibration term MCB_I measures the extent to which the interval forecast $[L, U]$ deviates from an isotonicly calibrated prediction interval. Since auto-calibration implies isotonic calibration (Proposition 20(ii)), it follows that MCB_I is also equal to zero when the forecast is auto-calibrated. More generally, the deviation from auto-calibration is always at least as large as the deviation from isotonic calibration, but these terms will often be the same since these two definitions are equivalent in many relevant forecasting situations (reverse direction of Proposition 20(ii)).

Proposition 23. *It holds that*

$$E[IS_\alpha([L, U], Y)] \geq MCB_A \geq MCB_I.$$

Data and code availability

The code and data used in this study are publicly available. Instructions to reproduce the results herein can be found at https://github.com/sallen12/CondIntCal_rep.

References

- Allen, S., Ferro, C. A. T., & Kwasniok, F. (2023). A conditional decomposition of proper scores: Quantifying the sources of information in a forecast. *Quarterly Journal of the Royal Meteorological Society*, 149, 1704–1725.
- Allen, S., Koh, J., Segers, J., & Ziegel, J. (2025). Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 120, 2796–2808.
- Andrews, D. W., & Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81, 609–666.
- Arnold, S., Walz, E.-M., Ziegel, J., & Gneiting, T. (2024). Decompositions of the mean continuous ranked probability score. *Electronic Journal of Statistics*, 18, 4992–5044.
- Arnold, S., & Ziegel, J. (2025). Isotonic conditional laws. *Bernoulli*, 31, 1140–1159.
- Bentzen, S., & Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140, 1924–1934.
- Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 17, Article e1008618.
- Braun, S., Holzmüller, D., Jordan, M. I., & Bach, F. (2025). Conditional coverage diagnostics for conformal prediction. arXiv preprint arXiv:2512.11779.
- Brehmer, J. R., & Gneiting, T. (2021). Scoring interval forecasts: Equal-tailed, shortest, and modal interval. *Bernoulli*, 27, 1993–2010.
- Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135, 1512–1519.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 841–862.
- Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77, 605–610.
- DeGroot, M., & Fienberg, S. (1982). Assessing probability assessors: Calibration and refinement. *Statistical Decision Theory and Related Topics III*, 291–314.
- Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118, Article e2016191118.
- Dimitriadis, T., Gneiting, T., Jordan, A. I., & Vogel, P. (2024). Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40, 1101–1122.
- Dimitriadis, T., & Puke, M. (2026). Statistical inference for score decompositions. arXiv preprint arXiv:2603.04275.
- Ehm, W., & Ovcharov, E. (2017). Bias-corrected score decomposition for generalized quantiles. *Biometrika*, 104, 473–480.
- Engle, R. F., & Manganello, S. (2004). CAViAR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22, 367–381.
- Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2, 113–127.
- Fontana, M., Zeni, G., & Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29, 1–23.
- Gibbs, I., Cherian, J. J., & Candès, E. J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 87, 1100–1126.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.
- Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69, 243–268.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.
- Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29, 411–422.
- Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.
- Gneiting, T., & Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17, 3226–3286.
- Gneiting, T., Wolfram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., & Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and its Application*, 10, 597–621.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330). PMLR.
- Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International conference on machine learning* (pp. 1939–1948). PMLR.
- Henzi, A. (2021). Isodistreg: Isotonic distributional regression (IDR): Distributional regression under stochastic order restrictions for numeric and binary response variables and partially ordered covariates. *The Comprehensive R Archive Network*, URL <https://cran.r-project.org/web/packages/isodistreg/index.html>.
- Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 83, 963–993.
- Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32, 896–913.

- Huang, Z., Lam, H., & Zhang, H. (2023). Conditional coverage estimation for high-quality prediction intervals. *Journal of Systems Science and Systems Engineering*, 32, 289–319.
- Jordan, A. I., Mühlemann, A., & Ziegel, J. F. (2022). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Annals of the Institute of Statistical Mathematics*, 74, 489–514.
- Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 247–254).
- Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4, 53–89.
- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113, 1094–1111.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36, 54–74.
- Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2022). The M5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38, 1365–1385.
- Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: analysis of forecasting behavior and performance* (pp. 3–46). NBER.
- Mitchell, K. (2020). *Score decompositions in forecast verification* (Ph.D. thesis), University of Exeter.
- Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 5, 113–127.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12, 595–600.
- Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 26, 41–47.
- Newey, W. K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica*, 1047–1070.
- Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11, 1833–1874.
- Papadopoulos, H., Gammernan, A., & Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED international conference on artificial intelligence and applications (AIA 2008)* (pp. 64–69).
- Pohle, M.-O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. arXiv preprint arXiv:2005.01835.
- Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., Woody, S., Wang, Y., Wang, L., Walraven, R. L., Tomar, V., Sherratt, K., Sheldon, D., Reiner, R. C., Prakash, B. A., ... Consortium, C.-. F. H. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. *MedRxiv*, 2020-2008.
- Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 3543–3553.
- Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks*, 11, 1263–1271.
- Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology and Climatology*, 2, 191–201.
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
- Singh (2016). Facebook comment volume dataset. <http://dx.doi.org/10.24432/C5Q886>, <https://archive.ics.uci.edu/dataset/363/facebook+comment+volume+dataset>.
- Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, 20, 360–380.
- Tsyplakov, A. (2011). Evaluating density forecasts: A comment. Available At SSRN 1907799.
- Vovk, V., Gammernan, A., & Shafer, G. (2022). *Algorithmic learning in a Random World* (2nd). Springer.
- Vovk, V., Nouretdinov, I., Manokhin, V., & Gammernan, A. (2018). Cross-conformal predictive distributions. In *Conformal and probabilistic prediction and applications* (pp. 37–51). *Proceedings of Machine Learning Research*.
- Waghmare, K., & Ziegel, J. (2025). Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and its Application*, 13, 271–296.
- Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67, 187–191.