Statistical methods for probabilistic forecasts of real-valued outcomes

Karlsruhe Institute of Technology

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Eva-Maria Walz

Tag der mündlichen Prüfung: 07 Februar 2024

Referent: Prof. Dr. Tilmann Gneiting
1. Korreferent: Prof. Dr. Peter Knippertz
2. Korreferent: Prof. Dr. Mathias Trabs

Acknowledgement

I would like to express my sincere gratitude and appreciation to my advisor Tilmann Gneiting who supported me throughout my master thesis and the entire period of my PhD program. I enjoyed the collaboration, his guidance, the many fruitful discussions and I am very thankful that I had the chance to work and learn in such an inspiring environment. I greatly appreciate being given a lot of freedom when implementing ideas and projects and I value the trust placed in me.

I further thank Peter Knippertz and Mathias Trabs for agreeing to serve as a reviewer on my Ph.D. committee.

The research presented in this thesis has been funded by the Deutsche Forschungsgemeinschaft (DFG) through the subproject C2 of the Transregional Collaborative Research Center SFB / TRR 165 “Waves to Weather” and is gratefully acknowledged.

Moreover, I want to thank Johanna Ziegel for collaboration, support, her honest feedback and her encouragement. A special thanks to Sebastian Arnold for a very pleasant and enjoyable collaboration. Moreover, I want to thank Benedikt Schulz, Sebastian Lerch, Johannes Resin, Alexander Jordan, Timo Dimitriadis and all other members of the Computational Statistics group at HITS.

For my interdisciplinary work, I was greatly supported by Andreas Fink and Peter Knippertz. Without their expertise, their significant contributions and their enthusiam the projects could not have been implemented. Moreover, I want to thank Marlon Maranan for consistently offering assistance and all other members of the meteorological working group and all early career scientists and members of Waves to Weather.

Thanks to Gregor Köhler and Luisa Walz for guidance, ideas and discussion.

Finally, I am grateful for the great support of my family and friends.

Contents

1 | Introduction

Making accurate forecasts is of wide practical relevance in a large number of fields. With the growing share of renewable energy sources in Germany, for instance, accurate solar and wind forecasts are of increasing importance to balance the power grid. The power output of photovoltaic cells and wind farms exhibits large fluctuations due to current weather conditions, and accurate forecasting reduces the risk of negative impacts on the grid and improves unit commitment optimization (Ren et al., 2015). Another illustrative example is the insurance industry, where accurate forecasting is crucial to ensure rigorous risk assessment, pricing precision, and strategic financial planning. In this context, shortcomings in adequate forecasting carry a high financial risk which can ultimately lead to an insurance bankruptcy or low financial rating (Van Gestel et al., 2007). In a similar line of reasoning, correct forecasting is essential for various business divisions such as finance and personnel but also in broader contexts such as our economic and political systems.

As argued in Petropoulos et al. (2022), this wide range of forecast applications has led to the development of an enormous variety of methods, principles, and tools. With the advances in computing power, more sophisticated models detecting complex relationships within the data could be implemented. This progress led to a strong improvement of machine learning (ML) models, and consequently, ML approaches have gained strong popularity for all kinds of forecasting problems. Combining the available tools from classical statistics and machine learning yields a large toolbox, summarized in an encyclopedic (non-exhaustive) overview in Petropoulos et al. (2022). A specific method is usually selected based on the type of data involved and the problem that needs to be solved.

In this thesis, the focus is on continuous forecasting problems, where the target or response is real-valued and the corresponding forecasts are probabilistic taking the form of predictive densities or predictive cumulative distribution functions. While there exist a wide range of methods for binary problems and probability forecasts the number of suitable analogs for real-valued problems is considerable smaller. The contributions in this paper, reduce this gap by presenting three newly developed approaches for probabilistic forecasts of real-valued outcomes which have been motivated by binary or deterministic counterparts.

Against this background, this thesis is structured in the following way: After providing an introduction to the fundamentals and key concepts of statistical forecasting in Chapters 2 and 3, the main part of this work focuses on the presentation of the newly developed approaches. To provide a clear structure and research focus for these methods, the forecasting cycle framework introduced by Hyndman and Athanasopoulos (2021) was utilized. This framework divides the general forecasting task into five essential steps: (I) Problem definition, (II) Information gathering, (III) Preliminary (exploratory) data analysis, (IV) Choosing and fitting models, and finally (V) Evaluation. Since the preparation steps (I) and (II) do not require the usage of statistical or ML approaches, this thesis has developed new approaches for real-valued outcomes within the scope and context of the process steps (III) to (V):

- Preliminary (exploratory) data analysis
- Choosing and fitting models
- Evaluation

For each of these three steps, a new method is developed to apply the forecasting cycle for real-valued forecasting problems.
Chapter 4 presents the coefficient of predictive ability (CPA) measure, a new tool to assess asymmetric relationships between variables. It can be used to perform feature screening or variable selection and, thus, can be applied in the data analysis step (III). In Chapter 5, Easy Uncertainty Quantification (EasyUQ) is introduced, a method that transforms deterministic forecasts for real-valued response variables into probabilistic forecasts and conducts a detailed comparison between EasyUQ and state-of-the art alternative approaches. This method supports the choosing and fitting models step (IV). To complete the process steps of the forecasting cycle, Chapter 6 develops a decomposition of the continuous ranked probability score (CRPS) into three informative components which allows for a more detailed comparison between different forecasts and, hence, can be applied in the evaluation step (V). Finally, this thesis illustrates and motivates the usage of the newly developed tools by applying the relevant

steps of the forecasting cycle to a challenging weather forecasting task in Chapter 7. Specifically, this chapter investigates the issue of producing probabilistic forecasts for accumulated precipitation over northern tropical Africa. In chapter 8, this thesis concludes by consolidating the main contributions, highlighting key results and presenting potential avenues for future research.

1.1 Relation to published work

This thesis comprises work that has been developed in collaboration with several other authors. In the following paragraph, the amount and type of individual contributions are defined.

Gneiting and Walz (2022) Gneiting, T. and Walz, E.-M. (2022). Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Machine Learning*, **111**, 2769-2797.

This research article was originally developed in my master thesis and has been extended by providing a more detailed investigation of the CPA measure (Section 4.4.2), a comparison to other relevant measures (Section 4.4.4), and by including real world data examples (Section 4.5). Both authors contributed to equal parts to this publication. Chapter 4 is almost identical to this research article.

Walz et al. (2024) Walz, E.-M., Henzi, A., Gneiting, T. and Ziegel, J. (2024). Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output. *SIAM Review*. In press, `arXiv:2212.08376`

The new approach in this paper was developed jointly by all authors. Chapter 5 is almost identical to this research article. In addition small parts of the paper contributed to Chapter 2 and Chapter 3.

Arnold et al. (2023b) Arnold, S., Walz, E.-M., Gneiting T. and Ziegel, J. (2023). Decompositions of the mean continuous ranked probability score. Preprint, `arXiv:2311.14122`

The new approach in this paper was developed jointly by all authors. The first two authors contributed to equal parts. Chapter 6 is almost identical to this research article.

In addition small parts of the paper contributed to Chapter 2.

The work presented in Chapter 7 is based on joint, ongoing research with Gregor Köhler, Andreas H. Fink, Peter Knippertz, and Tilmann Gneiting.

During the work on this thesis, I developed an R package for the implementation of ROC movies, UROC curves, and CPA:

`https://github.com/evwalz/urocc`,

and an R package for isotonic distributional regression (IDR) with probabilistic input and the CRPS decomposition:

`https://github.com/evwalz/isodisregSD`.

Moreover, I developed a python package for IDR:

`https://github.com/evwalz/isodisreg`.

2 Basics of forecast verification

For the study of statistical forecasts with a primary emphasis on probabilistic forecasts which are the main focus in this thesis, we review theoretical fundamentals and describe the concepts of a prediction space (Gneiting and Ranjan, 2013), calibration and sharpness (Gneiting et al., 2007), and scoring rules and scoring functions (Gneiting and Raftery, 2007; Gneiting, 2011).

2.1 Prediction space

A general framework considering the joint distribution of point forecast and observation was introduced in Murphy and Winkler (1987) and extended by Gneiting and Ranjan (2013) to potentially multiple probabilistic forecasts and observations taking values in just any space. Consider a prediction space, i.e., a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where the elements of the sample space Ω correspond to realizations of random tuples

$$(F_1, \ldots, F_k, Y, U)$$

where Y is a real-valued outcome, $F_1, \ldots, F_k$ are probability measures on the real line, which are identified with their associated cumulative distribution function (CDF) F or corresponding probability density function (PDF) f, and U is uniformly distributed on the unit interval. Let $\mathcal{A}_0 \subset \mathcal{A}$ denote a forecaster's information basis, i.e., a sub σ-algebra $\mathcal{A}_0$ such that F is measurable with respect to $\mathcal{A}_0$. Then F is ideal relative to $\mathcal{A}_0$ if

$$F(y) = \mathbb{P}(Y \leq y|\mathcal{A}_0) \text{ almost surely, for all } y \in \mathbb{R}.$$

Note that $\mathbb{P}(Y \leq y|X)$ generate the conditional law of Y given $\mathcal{A}_0 = \sigma(X)$, which in short we denote by $P_{Y|X}$. In a prediction space setting with deterministic forecasts, the probabilistic forecasts $F_1, \ldots, F_k$ are replaced by point forecasts $X_1, \ldots, X_k$.

2.2 Calibration and sharpness

As argued in Gneiting et al. (2007) probabilistic forecasts should be as sharp as possible subject to calibration. Informally, predictive distributions are calibrated if they provide a statistically coherent explanation of the outcomes. Sharpness, on the other hand, quantifies how well one can discriminate different scenarios for future events according to the forecast and is a property of the forecast only. The next section defines several notions of calibration. A comprehensive (non-exhaustive) overview of the main types of calibration and their properties can be found in Gneiting and Resin (2023).

2.2.1 Notions of calibration

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

$$F(x) = \mathbb{P}(Y \leq x \mid F) \quad \text{almost surely for all } x \in \mathbb{R}. \tag{2.1}$$

For any threshold value $x \in \mathbb{R}$, we may condition on the random variable $F(x)$ instead of the random distribution F in (2.1), to obtain the weaker notion of threshold calibration. Specifically, the forecast F is called *threshold calibrated* (Henzi et al., 2021) if

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

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

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

where for any $\alpha \in (0,1)$, $q_\alpha(Y \mid F^{-1}(\alpha))$ denotes the lower-α-quantile of the conditional law of Y given $F^{-1}(\alpha)$.

In a recent publication, Arnold and Ziegel (2023) introduce *isotonic calibration*. The definition of this new notion of calibration and an overview of relevant concepts is deferred to Section 6.4.2. By Proposition 5.3 of Arnold and Ziegel (2023), auto-calibration

$$\begin{tikzcd}
 & \text{AC} \arrow[d, Rightarrow] & \\
 & \text{IC} \arrow[dl, Rightarrow] \arrow[dr, Rightarrow] & \\
\text{TC} \arrow[dr, hook] & & \text{QC} \arrow[dl, hook] \\
 & \text{PC} &
\end{tikzcd}$$

Figure 2.1: Implications between auto-calibration (AC), isotonic calibration (IC), threshold calibration (TC), and quantile calibration (QC). Implications with respect to probabilistic calibration (PC) are indicated by hooked arrows and hold under Assumption 2.15 of Gneiting and Resin (2023).

implies isotonic calibration, and isotonic calibration implies threshold calibration and quantile calibration. The probability integral transform (PIT) of the CDF-valued random quantity F is the random variable $Z_F = F(Y-) + U(F(Y) - F(Y-))$, where $F(y-) = \lim_{x \uparrow y} F(x)$ denotes the left-hand limit of F at $y \in \mathbb{R}$, with a random variable U that is standard uniform and independent of F and Y. The PIT of a continuous CDF F simplifies to $Z_F = F(Y)$. The forecast F is *probabilistically calibrated* if Z_F is uniformly distributed on the unit interval (Gneiting and Ranjan, 2013). Originally suggested by Dawid (1984), checks for probabilistic calibration, and for the uniformity of the closely related rank histogram, constitute a cornerstone of forecast evaluation (Diebold et al., 1998; Hamill, 2001; Gneiting et al., 2007). Under regularity conditions, a threshold calibrated or quantile calibrated forecast is probabilistically calibrated; details and a direct implication from isotonic calibration to a weak form of probabilistic calibration are available in Gneiting and Resin (2023, Section 3.3) and Arnold and Ziegel (2023, Appendix D), respectively. Figure 2.1 summarizes relationships between the notions of calibration discussed in this section and in Chapter 6.

2.3 Scoring functions

A performance criterion to evaluate point forecasts $x_1, \ldots, x_n \in \mathbb{R}$ for corresponding observations $y_1, \ldots, y_n \in \mathbb{R}$ typically takes the form

$$s_n = \frac{1}{n} \sum_{i=1}^{n} s(x_i, y_i),$$

where $s : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is denoted as scoring function. To avoid misguided inferences, the forecasted functional and the scoring function must be matched carefully (Gneiting, 2011). A scoring function is consistent for a target functional T, which for simplicity we assume to be single-valued, relative to the class $\mathcal{P}$ of predictive distributions at hand if

$$\mathbb{E}_{Y\sim P} s(T(P), Y) \leq \mathbb{E}_{Y\sim P} s(x, Y) \tag{2.2}$$

for all $x \in \mathbb{R}$ and $P \in \mathcal{P}$. It is strictly consistent if equality in (2.2) implies $x = T(P)$. Popular scoring functions are the squared error

$$\mathrm{se}(x, y) = (x - y)^2, \tag{2.3}$$

which is consistent for the mean functional relative to the class of probability measures with finite first moment and the absolute error

$$\mathrm{ae}(x, y) = |x - y| \tag{2.4}$$

which is consistent for the median functional. A scoring function strictly consistent for an α-quantile, $\alpha \in (0, 1)$, relative to any class of probability measures with finite first moment is the piecewise linear quantile score

$$\mathrm{qs}_\alpha(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(x - y). \tag{2.5}$$

2.4 Scoring rules

A widely accepted principle in the generation of predictive distributions is that sharpness ought to be maximized subject to calibration (Gneiting et al., 2007). Maximizing sharpness requires forecasters to provide informative, concentrated predictive distributions, and calibration posits that probabilities derived from these distributions conform with actual observed frequencies. This is in line with and generalizes the classical goal of prediction intervals being as narrow as possible while attaining nominal coverage.

A key tool for evaluating and comparing predictive distributions under this principle is that of proper scoring rules (Gneiting and Raftery, 2007; Matheson and Winkler, 1976) which are functions $S : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ mapping a predictive distribution $P \in \mathcal{P}$ and the outcome y to a numerical score such that

$$\mathbb{E}_{Y\sim P}[\mathrm{S}(P, Y)] \leq \mathbb{E}_{Y\sim P}[\mathrm{S}(Q, Y)] \tag{2.6}$$

for all distributions P, Q in a given class $\mathcal{P}$ of probability measures on $\mathbb{R}$. A scoring rule is strictly proper if (2.6) holds with equality only if $P = Q$. Here $\mathbb{E}_{Y\sim P}[\cdot]$ denotes the expected value of the quantity in parentheses when Y follows the distribution P. From a decision-theoretic point of view, proper scoring rules encourage truthful forecasting, since forecasters minimize their expected score if they issue predictive distributions that correspond to their true beliefs.

Arguably the most widely used strictly proper scoring rules for real-valued observations are the logarithmic score for a predictive CDF F with density f,

$$\mathrm{LogS}(F, y) = -\log(f(y)). \tag{2.7}$$

and the continuous ranked probability score (CRPS)

$$\begin{aligned} \mathrm{CRPS}(\mathrm{F}, \mathrm{y}) &= \mathbb{E}|Y - y| - \frac{1}{2}\mathbb{E}|Y - Y'| \\ &= \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \le z\})^2 \,\mathrm{d}z \\ &= 2\int_0^1 (\mathbb{1}\{y < F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y) d\alpha, \end{aligned} \tag{2.8}$$

where Y and Y' are independent random variables with distribution F and finite first moment and F^{-1} denotes the quantile function, defined as $F^{-1}(\alpha) = \inf\{z \in \mathbb{R} \mid F(z) \ge \alpha\}$ for $\alpha \in (0, 1)$ (Matheson and Winkler, 1976; Gneiting and Raftery, 2007; Laio and Tamea, 2007). The popularity of the CRPS is due to the facts that it allows arbitrary types of predictive distributions (e.g., discrete, continuous, mixed discrete-continuous), is reported in the same unit as the outcome, and reduces to the absolute error if F assigns probability one to a point $x \in \mathbb{R}$. The LogS is (save for a change of sign) the ubiquitous loss function in maximum likelihood estimation. Closed form expressions for the CRPS and LogS are available for the most commonly used parametric distributions and have been implemented in software packages (Jordan et al., 2019).

In practice, forecast methods are compared in terms of their average score over a collection (F_j, y_j) for $j = 1, \ldots, n$,

$$\bar{\mathrm{S}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{S}(F_j, y_j), \tag{2.9}$$

and the method achieving the lowest average score is considered superior. For example, a popular strictly proper scoring rule for binary observation $y \in \{0, 1\}$ and corresponding predictive probability $p \in [0,1]$ for the outcome $y = 1$, is the Brier score (BS)

$$\mathrm{BS}(p, y) = (p - y)^2. \tag{2.10}$$

The empirical average BS for a sequence of forecast–observation pairs $(p_1, y_1), \ldots, (p_n, y_n)$, where $p_i \in [0,1]$ and $y_i \in \{0,1\}$, equals

$$\overline{\mathrm{BS}} = \frac{1}{n} \sum_{1=1}^{n} (p_i - y_i)^2 \, . \tag{2.11}$$

3 | Regression models

Regression models are used to model the relation between a response variable Y and one or more explanatory variables $X_1, \ldots, X_m$ by approximating the conditional distribution of Y, or certain characteristics of it, given $X_1, \ldots, X_m$. The relation between the response and the covariates is described by a suitable class of regression functions which follow specific shape or order restrictions. A regression function is fitted by minimizing a suitable loss function based on training data

$$\{(x_{i1}, \ldots, x_{im}, y_i) : i = 1, \ldots, n\},$$

with covariate values (explanatory variables) $x_{i1}, \ldots, x_{im}$ and corresponding observation y_i for $i = 1, \ldots, n$.

3.1 Linear regression

In linear regression, the regression function $f : \mathbb{R}^m \to \mathbb{R}$ is assumed to be a linear function of the covariates,

$$f(x_1, \ldots, x_m) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

Based on the functional that one wants to estimate, a suitable scoring function needs to be selected to estimate the regression coefficients $\beta_0, \beta_1, \ldots, \beta_m$. The most common regression model is ordinary least squares regression where one seeks the conditional expectation of a general real-valued response variable given the predictor variables, namely

$$\mathbb{E}(Y | X_1 = x_1, \ldots, X_m = x_m) = f(x_1, \ldots, x_m),$$

Since the squared error scoring function at (2.3) is consistent for the mean functional, the parameters are estimated by

$$\operatorname*{arg\,min}_{\beta_0,\beta_1,\ldots,\beta_m} \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2.$$

Instead of modeling the conditional expectation, quantile regression seeks to model the α-quantile q_α for some fixed level $\alpha \in (0,1)$ of a real-valued response variable conditional on the explanatory variables with

$$q_\alpha(Y|X_1 = x_1, \ldots, X_m = x_m) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

To estimate the regression coefficients any consistent scoring function for the α-quantile can be used. In practice, one uses the asymmetric piecewise linear scoring function at (2.5) leading to the estimate

$$\operatorname*{arg\,min}_{\beta_0,\beta_1,\ldots,\beta_m} \sum_{i=1}^{n}(\mathbb{1}\{y_i \le \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}\} - \alpha)(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} - y_i),$$

which can be rewritten to a linear programming problem and thus be solved efficiently.

3.2 Logistic regression

If the response variable Y is binary, then the conditional expectation equals the conditional event probability p which is constrained to the unit interval:

$$p = \mathbb{E}(Y|X_1 = x_1, \ldots, X_m = x_m) = \mathbb{P}(Y = 1|X_1 = x_1, \ldots, X_m = x_m)$$

Since the linear combination of the explanatory variables can attain any value on the real line, one employs a suitable transformation function such as in logistic regression which models the relation

$$\operatorname{logit}(p) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots \beta_m x_m$$

or equivalently

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_m x_m)}.$$

To estimate the regression coefficients, any scoring function consistent for the expectation functional can be used or any proper scoring rule for probability forecasts for binary events. In practice, one typically uses the logarithmic scoring rule, which yields optimal score estimates

$$\underset{\beta_0,\beta_1,\dots,\beta_m}{\arg\max} \sum_{i=1}^{n} \Bigg[y_i \log\left(\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots\beta_m x_{im})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots\beta_m x_{im})}\right) + (1-y_i)\log\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots\beta_m x_{im})}\right)\Bigg].$$

Typically, the optimization needs to be performed numerically.

3.3 Isotonic regression

For isotonic regression, one assumes that the regression functions satisfies certain order constraints; i.e., inequality restrictions that constrain the values of these functions. To this end, the covariate space $\mathcal{X} = \mathbb{R}^m$ is equipped with a partial order $\preceq$, e.g.,

$$x_i \preceq x_j \text{ if } x_{ik} \leq x_{jk}, \text{ for all } k = 1,\dots,m,$$

where $i, j = 1, \dots, n$. The isotonic regression problem can be stated as

$$\min_{\beta_1,\dots,\beta_n} \sum_{i=1}^{n} h(y_i, \beta_i)$$

with the constraint

$$\beta_i \leq \beta_j \text{ if } x_i \preceq x_j,$$

where h is a convex loss function. For

$$h(y,\beta) = |y-\beta|^p$$

and $p = 2$ or $p = 1$, we obtain the least squares problem and the least absolute values problem of isotonic regression. A consistent estimator for isotonic quantile regression for a general quantile q_α at level $\alpha \in (0,1)$ is obtained using the loss function

$$h(y,\beta) = \alpha \max(y-\beta, 0) + (1-\alpha)\max(\beta - y, 0).$$

The different optimization problems can be solved via the generalized pool adjacent violator (PAV) algorithm (de Leeuw et al., 2009).

Figure 3.1 illustrates the difference between linear and isotonic regression and the usage of different objective functionals such as the mean versus different quantiles at level α based on simulated data (X, Y), where X is uniform on $(0, 10)$ and

$$Y|X \sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{scale} = \min\{\max\{X, 1\}, 6\}) \tag{3.1}$$

as defined in equation (1) of Henzi et al. (2021) with a sample of size $n = 400$.

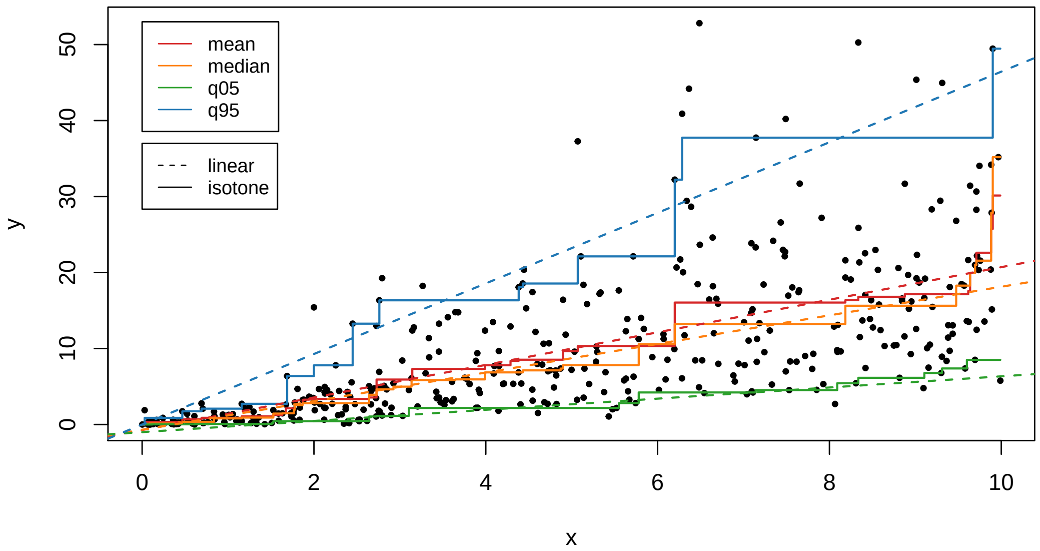

Figure 3.1: Illustration of linear regression (solid line) and isotonic regression (dashed line) for data in 3.1. The fitted functionals are the mean, the median, the 5%-quantile, and the 95%-quantile.

3.4 Distributional regression

Thus far, the focus was on regression models that estimate certain functionals of the conditional distribution of Y given $X_1, \dots, X_m$, such as the conditional expectation, median or other specific quantiles. As argued in Henzi et al. (2021): "The reduction of a conditional distribution to a single-valued functional results in tremendous loss of information. Therefore, from a perspective of both estimation and prediction, regression analysis ought to be distributional". To this end, Henzi et al. (2021) introduce isotonic distributional regression (IDR), a non-parametric technique to estimate the conditional distribution $P_{Y|X}$ of Y given a covariate vector X under isotonic constraints. Just like for isotonic regression, the covariate space $\mathcal{X}$ is equipped with a partial order $\preceq$. The notion of isotonicity for distributional regression is understood in the following way. Formally, IDR assumes that the conditional distributions of outcome Y given the covariate vector X, identified with the CDFs $F_x(y) = \mathbb{P}(Y \leq y \mid X = x)$, are increasing in stochastic order, namely, $F_x \leq_{\text{st}} F_{x'}$, i.e., $F_x(y) \geq F_{x'}(y)$ for all $y \in \mathbb{R}$, if $x \preceq x'$. As introduced in Section 2.4, proper scoring rules S are summary measures to evaluate probabilistic forecasts and thus serve as suitable loss functions in distributional regression. The IDR solution is a minimizer of empirical loss

$$\frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i)$$

with the constraint

$$F_i \leq_{st} F_j \text{ if } x_i \preceq x_j, \quad i, j = 1, \ldots, n.$$

Henzi et al. (2021) show that there exists a unique CRPS-based solution, denoted as IDR solution, which is simultaneously optimal with respect to broad classes of proper scoring rules including relevant choices in the extant literature. Theorem 2.2 in Henzi et al. (2021) implies that the IDR solution $\hat{F} = (\hat{F}_{x_1}, \ldots, \hat{F}_{x_n})$ satisfies

$$(\hat{F}_{x_1}(y), \ldots \hat{F}_{x_n}(y)) = \underset{\theta \in \mathbb{R}^n : \theta_i \geq \theta_j \text{ if } x_i \preceq x_j}{\arg\min} \sum_{i=1}^{n} (\theta_i - \mathbb{1}\{y_i \leq y\})^2$$

at every threshold $y \in \mathbb{R}$. Thus, finding the IDR solution reduces to solving a quadratic programming problem by computing IDR CDFs at any fixed threshold value. Moreover, it suffices to compute the IDR CDFs at the unique values of the response only, as the optimal solution remains constant in between. Alternatively, the IDR solution, also being optimal with respect to quantile loss, can be recovered by performing isotonic quantile regression at all α-levels (see Section 3.3) and piecing the conditional quantile functions together. While solving a minimization problem at all quantile levels l/k where $k = 1, \ldots, n$ and $l = 1, \ldots, k-1$ is computationally challenging, the quantile representation offers a valuable interpretation of IDR, illustrated in Figure 3.2, which is a replication of Figure 1 in Henzi et al. (2021) based on data defined in (3.1).

In the case of a total order the IDR solution is computed using a recursive adaption of the PAV algorithm for a considerable reduction of computing time (Henzi et al., 2022). For general partial orders see de Leeuw et al. (2009) and their active set solutions.

Thus far, the IDR solution $\hat{F} = (\hat{F}_{x_1}, \ldots, \hat{F}_{x_n})$ is defined at the covariate values $x_1, \ldots, x_n \in \mathcal{X}$ only. To make a prediction at a new covariate value $x \notin \{x_1, \ldots, x_n\}$, Henzi et al. (2021) introduce an approach for general covariate spaces which simplifies in the special case $\mathcal{X} = \mathbb{R}$ of a single real-valued covariate to the following procedure: Suppose that $x_1 \leq \cdots \leq x_n$. If $x < x_1$, we may let $F = F_{x_1}$ and if $x > x_n$, then let $F = F_{x_n}$. If $x \in (x_i, x_{i+1})$ for some $i \in \{1, \ldots, n-1\}$ linear interpolation is performed, namely,

$$F(z) = \frac{x - x_i}{x_{i+1} - x_i} F_{x_i}(z) + \frac{x_{i+1} - x}{x_{i+1} - x_i} F_{x_{i+1}}(z)$$

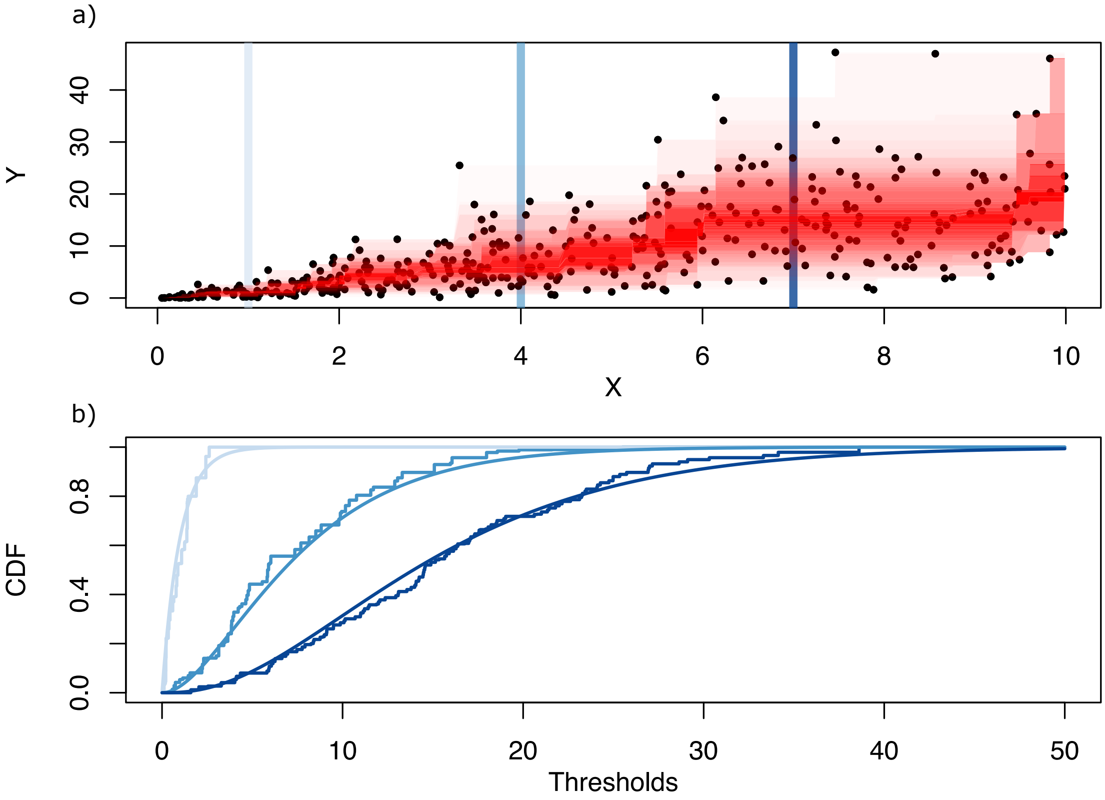

Figure 3.2: Simulated data from 3.1 of size $n = 400$ with shaded bands corresponding to central intervals (a) and true conditional CDFs and IDR estimates for selected values of X (b) which corresponding to vertical stripes displayed in (a).

for $z \in \mathbb{R}$. For details of the procedure in general covariate spaces see Henzi et al. (2021). For the choice of partial order, Henzi et al. (2021) argue that for $\mathcal{X} \subseteq \mathbb{R}^d$ the componentwise order may be suitable for many applications whereas for ordinal covariates a lexicographic order may be more appropriate. In case of covariates that are exchangeable the empirical stochastic order and the empirical increasing convex order are suggested.

4 | ROC movies, UROC curve and CPA

Throughout science and technology, receiver operating characteristic (ROC) curves and associated area under the curve (AUC) measures constitute powerful tools for assessing the predictive abilities of features, markers and tests in binary classification problems. Despite its immense popularity, ROC analysis has been subject to a fundamental restriction, in that it applies to dichotomous (yes or no) outcomes only. In this chapter, ROC movies and universal ROC (UROC) curves that apply to just any linearly ordered outcome, along with an associated coefficient of predictive ability (CPA) measure are introduced. Their usage is illustrated in data examples from biomedicine and meteorology. In addition, CPA measure is used for feature analysis in the precipitation forecasting problem discussed in Chapter 7 of this work. This highlights the practical usefulness of CPA measure in properly performing the *preliminary (exploratory) data analysis* step of the forecasting cycle introduced in Section 1.

4.1 Introduction

Originating from signal processing and psychology, popularized in the 1980s (Hanley and McNeil, 1982; Swets, 1998), and witnessing a surge of usage in machine learning (Bradley, 1997; Huang and Ling, 2005; Fawcett, 2006; Flach, 2016), receiver operating characteristic or relative operating characteristic (ROC) curves and area under the ROC curve (AUC) measures belong to the most widely used quantitative tools in science and technology. Strikingly, a Web of Science topic search for the terms “receiver operating characteristic” or “ROC” yields well over 15,000 scientific papers published in calendar year 2019 alone. In a nutshell, the ROC curve quantifies the potential value of a real-valued classifier score, feature, marker, or test as a predictor of a binary outcome. To give a classical example, Figure 4.1 illustrates the initial levels of two biomedical markers, serum albumin and serum bilirubin, in a Mayo Clinic trial on primary biliary

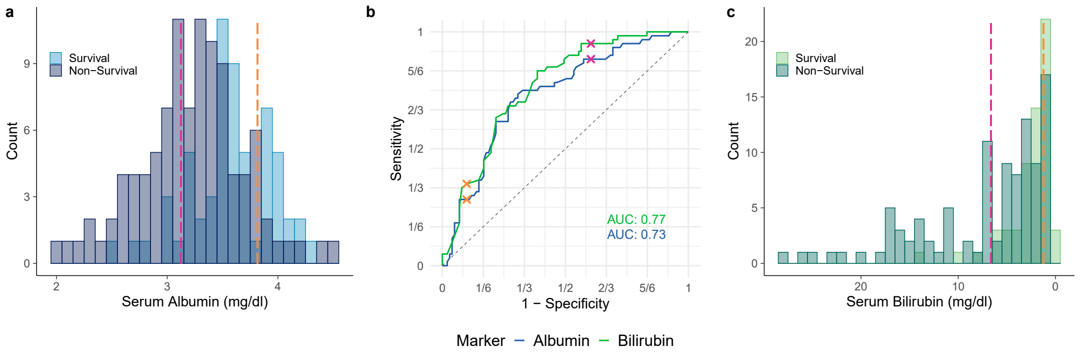

Figure 4.1: Traditional ROC curves for two biomedical markers, serum albumin and serum bilirubin, as predictors of patient survival beyond a threshold value of 1462 days (four years) in a Mayo Clinic trial. (a, c) Bar plots of marker levels conditional on survival or non-survival. The stronger shading results from overlap. For bilirubin, we reverse orientation, as is customary in the biomedical literature. (b) ROC curves and AUC values. The crosses correspond to binary classifiers at the feature thresholds indicated in the bar plots.

cirrhosis (PBC), a chronic fatal disease of the liver (Dickson et al., 1989). While patient records specify the duration of survival in days, traditional ROC analysis mandates the reduction of the outcome to a binary event, which here we take as survival beyond four years. Assuming that higher marker values are more indicative of survival, we can take any threshold value to predict survival if the marker exceeds the threshold, and non-survival otherwise. This type of binary classifier yields true positives, false positives (erroneous predictions of survival), true negatives, and false negatives (erroneous predictions of non-survival). The ROC curve is the piecewise linear curve that plots the true positive rate, or sensitivity, versus the false positive rate, or one minus the specificity, as the threshold for the classifier moves through all possible values.

Despite its popularity, ROC analysis has been subject to a fundamental shortcoming, namely, the restriction to binary outcomes. Real-valued outcomes are ubiquitous in scientific practice, and investigators have been forced to artificially make them binary if the tools of ROC analysis are to be applied. In this light, researchers have been seeking generalizations of ROC analysis that apply to just any type of ordinal or real-valued outcomes in natural ways (Etzioni et al., 1999; Heagerty et al., 2000; Bi and Bennett, 2003; Pencina and D'Agostino, 2004; Heagerty and Zheng, 2005; Rosset et al., 2005; Mason and Weigel, 2009; Hernández-Orallo, 2013). Still, notwithstanding decades of scientific endeavor, a fully satisfactory generalization has been elusive.

In this chapter, we propose a powerful generalization of ROC analysis, which overcomes extant shortcomings, and introduce data science tools in the form of the ROC movie, the universal ROC (UROC) curve, and an associated, rank based coefficient of (potential) predictive ability (CPA) measure — tools that apply to just any linearly ordered outcome, including both binary, ordinal, mixed discrete-continuous, and continuous variables. The ROC movie comprises the sequence of the traditional, static ROC curves as the linearly ordered outcome is converted to a binary variable at successively higher thresholds. The UROC curve is a weighted average of the individual ROC curves that constitute the ROC movie, with weights that depend on the class configuration, as induced by the unique values of the outcome, in judiciously predicated, well-defined ways. CPA is a weighted average of the individual AUC values in the very same way that the UROC curve is a weighted average of the individual ROC curves that constitute the ROC movie. Hence, CPA equals the area under the UROC curve. This set of generalized tools reduces to the standard ROC curve and AUC when applied to binary outcomes. Moreover, key properties and relations from conventional ROC theory extend to ROC movies, UROC curves, and CPA in meaningful ways, to result in a coherent toolbox that properly extends the standard ROC concept. For a graphical preview, we return to the survival data example from Section 4.1, where the outcome was artificially made binary. Equipped with the new set of tools we no longer need to transform survival time into a specific dichotomous outcome. Figure 4.2 shows the ROC movie, the UROC curve, and CPA for the survival dataset.

The remainder of the chapter is organized as follows. Section 4.2 provides a brief review of conventional ROC analysis for dichotomous outcomes. The key technical development is in Sections 4.3 and 4.4, where we introduce and study ROC movies, UROC curves, and the rank based CPA measure. To illustrate practical usage and relevance, real data examples from survival analysis and weather prediction are presented in Section 4.5. We monitor recent progress in numerical weather prediction (NWP) and shed new light on a recent comparison of the predictive abilities of convolutional neural networks (CNNs) vs. traditional NWP models. A final discussion is found in Section 4.6.

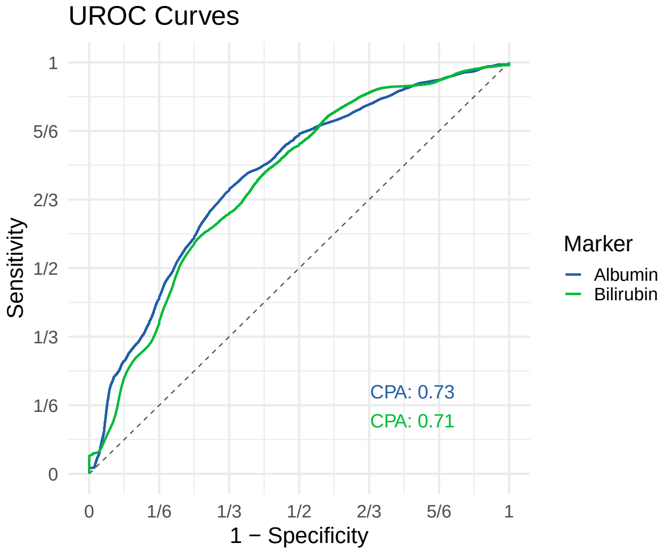

Figure 4.2: ROC movies, UROC curves, and CPA for two biomedical markers, serum albumin and serum bilirubin, as predictors of patient survival (in days) in a Mayo Clinic trial. The ROC movies show the traditional ROC curves for binary events that correspond to patient survival beyond successively higher thresholds. The numbers at upper left show the current value of the threshold in days, at upper middle the respective relative weight, and at bottom right the AUC values. The threshold value of 1462 days recovers the traditional ROC curves in Figure 4.1. The video ends in a static screen with the UROC curves and CPA values for the two markers.

4.2 Receiver operating characteristic (ROC) curves and area under the curve (AUC) for binary outcomes

Before we introduce ROC movies, UROC curves, and CPA, it is essential that we establish notation and review the classical case of ROC analysis for binary outcomes, as described in review articles and monographs by Hanley and McNeil (1982), Swets (1998), Bradley (1997), Pepe (2003), Fawcett (2006), and Flach (2016), among others.

4.2.1 Binary setting

Throughout this section we consider bivariate data of the form

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \{0, 1\}, \tag{4.1}$$

where $x_i \in \mathbb{R}$ is a real-valued classifier score, feature, marker, or covariate value, and $y_i \in \{0, 1\}$ is a binary outcome, for $i = 1, \ldots, n$. Following the extant literature, we refer to $y = 1$ as the positive outcome and to $y = 0$ as the negative outcome, and we assume that higher values of the feature are indicative of stronger support for the positive outcome. Throughout we assume that there is at least one index $i \in \{1, \ldots, n\}$ with $y_i = 0$, and a further index $j \in \{1, \ldots, n\}$ with $y_j = 1$.

4.2.2 Receiver operating characteristic (ROC) curves

We can use any threshold value $x \in \mathbb{R}$ to obtain a hard classifier, by predicting a positive outcome for a feature value $> x$, and predicting a negative outcome for a feature value $\leq x$. If we compare to the actual outcome, four possibilities arise. True positive and true negative cases correspond to correctly classified instances from class 1 and class 0, respectively. Similarly, false positive and false negative cases are misclassified instances from class 1 and class 0, respectively.

Considering the data (4.1), we obtain the respective *true positive rate*, *hit rate* or *sensitivity* (se),

$$\mathrm{se}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i > x, y_i = 1\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = 1\}},$$

and the *false negative rate*, *false alarm rate* or one minus the *specificity* (sp),

$$1 - \mathrm{sp}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i > x, y_i = 0\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = 0\}},$$

at the threshold value $x \in \mathbb{R}$, where the indicator $\mathbb{1}\{A\}$ equals one if the event A is true and zero otherwise.

Evidently, it suffices to consider threshold values x equal to any of the unique values of $x_1, \ldots, x_n$ or some $x_0 < x_1$. For every x of this form, we obtain a point

$$(1 - \mathrm{sp}(x), \mathrm{se}(x))$$

in the unit square. Linear interpolation of the respective discrete point set results in a piecewise linear curve from $(0, 0)$ to $(1, 1)$ that is called the *receiver operating characteristic* (ROC) *curve*. For a mathematically oriented, detailed discussion of the construction see Section 2 of Gneiting and Vogel (2022).

4.2.3 Area under the curve AUC

The area under the ROC curve is a widely used measure of the predictive potential of a feature and generally referred to as the *area under the curve* (AUC).

In what follows, a well-known interpretation of AUC in terms of probabilities will be useful. To this end, we define the function

$$s(x, x') = \mathbb{1}\{x < x'\} + \frac{1}{2}\mathbb{1}\{x = x'\}, \tag{4.2}$$

where $x, x' \in \mathbb{R}$. For subsequent use, note that if x and x' are ranked within a list, and ties are resolved by assigning equal ranks within tied groups, then $s(x, x') = s(\mathrm{rk}(x), \mathrm{rk}(x'))$, where $\mathrm{rk}(x)$ and $\mathrm{rk}(x')$ are the ranks of x and x'.

We now change notation and refer to the feature values in class $i \in \{0, 1\}$ as x_{ik} for $k = 1, \ldots, n_i$, where $n_0 = \sum_{i=1}^{n} \mathbb{1}\{y_i = 0\}$ and $n_1 = \sum_{i=1}^{n} \mathbb{1}\{y_i = 1\}$, respectively. Thus, we have rewritten (4.1) as

$$(x_{01}, 0), \ldots, (x_{0n_0}, 0), (x_{11}, 1), \ldots, (x_{1n_1}, 1) \in \mathbb{R} \times \{0, 1\}. \tag{4.3}$$

Using the new notation, Result 4.10 of Pepe (2003) states that

$$\mathrm{AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} s(x_{0i}, x_{1j}). \tag{4.4}$$

In words, AUC equals the probability that under random sampling a feature value from a positive instance is greater than a feature value from a negative instance, with any ties resolved at random. Expressed differently, AUC equals the tie-adjusted probability of concordance in feature–outcome pairs, where we define instances $(x, y) \in \mathbb{R}^2$ and $(x', y') \in \mathbb{R}^2$ with $y \neq y'$ to be *concordant* if either $x > x'$ and $y > y'$, or $x < x'$ and $y < y'$. Similarly, instances (x, y) and (x', y') with $y \neq y'$ are *discordant* if either $x > x'$ and $y < y'$, or $x < x'$ and $y > y'$.

Further investigation reveals a close connection to Somers' D, a classical measure of ordinal association (Somers, 1962). This measure is defined as

$$D = \frac{n_c - n_d}{n_0 n_1},$$

where $n_0 n_1$ is the total number of pairs with distinct outcomes that arise from the data in (4.3), n_c is the number of concordant pairs, and n_d is the number of discordant pairs. Finally, let n_e be the number of pairs for which the feature values are equal. The relationship (4.4) yields

$$\mathrm{AUC} = \frac{n_c}{n_0 n_1} + \frac{1}{2} \frac{n_e}{n_0 n_1},$$

and as $n_0 n_1 = n_c + n_d + n_e$, it follows that

$$\mathrm{AUC} = \frac{1}{2}(D + 1) \tag{4.5}$$

relates linearly to Somers' D.

To give an example, suppose that the real-valued outcome Y and the features X, X' and X'' are jointly Gaussian. Specifically, we assume that the joint distribution of (Y, X, X', X'') is multivariate normal with covariance matrix

$$\begin{pmatrix} 1 & 0.8 & 0.5 & 0.2 \\ 0.8 & 1 & 0.8 & 0.5 \\ 0.5 & 0.8 & 1 & 0.8 \\ 0.2 & 0.5 & 0.8 & 1 \end{pmatrix}. \tag{4.6}$$

In order to apply classical ROC analysis, the real-valued outcome Y needs to be converted to a binary variable, namely, an event of the type $Y_\theta = \mathbb{1}\{Y \geq \theta\}$ of Y being greater than or equal to a threshold value θ. Figure 4.3 shows ROC curves for the features X, X' and X'' as a predictor of the binary variable Y_1, based on a sample of size $n = 400$. The AUC values for X, X' and X'' as a predictor of Y_1 are .91, .72 and .61, respectively.

4.2.4 Key properties

A key requirement for a persuasive generalization of classical ROC analysis is the reduction to ROC curves and AUC if the outcomes are binary. Furthermore, well established desirable properties from ROC analysis ought to be retained. To facilitate judging whether the generalization in Sections 4.3 and 4.4 satisfies these desiderata, we summarize key properties of ROC curves and AUC in the following (slightly informal) listing.

(1) The ROC curve and AUC are straightforward to compute and interpret, in the (rough) sense of *the larger the better*.

(2) AUC attains values between 0 and 1 and relates linearly to Somers' D. For a perfect feature, $\mathrm{AUC} = 1$ and $D = 1$; for a feature that is independent of the binary outcome, $\mathrm{AUC} = \frac{1}{2}$ und $D = 0$.

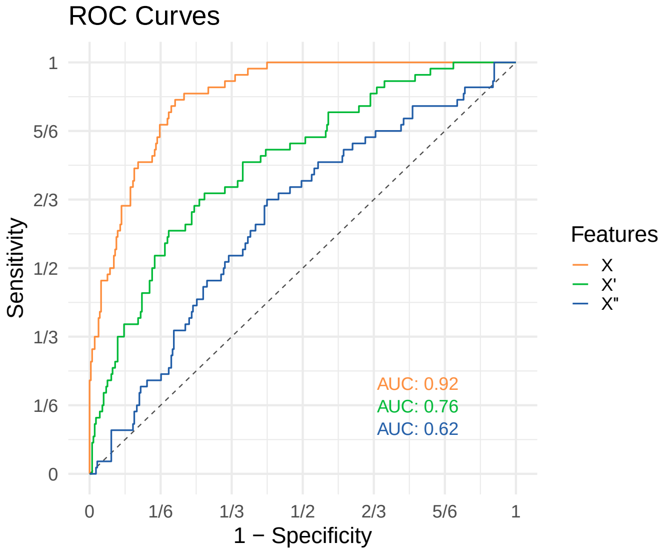

Figure 4.3: Traditional ROC curves and AUC values for the features X, X' and X'' as predictors of the binary outcome $Y_1 = \mathbb{1}\{Y \geq 1\}$ in the simulation example of Section 4.2.3, based on a sample of size $n = 400$.

(3) The numerical value of AUC admits an interpretation as the probability of concordance for feature–outcome pairs.

(4) The ROC curve and AUC are purely rank based and, therefore, invariant under strictly increasing transformations. Specifically, if $\varphi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function, then the ROC curve and AUC computed from

$$(\varphi(x_1), y_1), \ldots, (\varphi(x_n), y_n) \in \mathbb{R} \times \{0, 1\} \tag{4.7}$$

are the same as the ROC curve and AUC computed from (4.1).

As an immediate consequence of the latter property, ROC curves and AUC assess the discrimination ability or *potential* predictive ability of a classifier, feature, marker, or test (Wilks, 2019). Distinctly different methods are called for if one seeks to evaluate a classifier's *actual* value in any given applied setting (Adams and Hand, 1999; Hernández-Orallo et al., 2012; Ehm et al., 2016).

4.3 ROC movies and universal ROC (UROC) curves for real-valued outcomes

As noted, traditional ROC analysis applies to binary outcomes only. Thus, researchers working with real-valued outcomes, and desiring to apply ROC analysis, need to convert and reduce to binary outcomes, by thresholding artificially at a cut-off value. Here we propose a powerful generalization of ROC analysis, which overcomes extant shortcomings, and introduce data analytic tools in the form of the ROC movie, the universal ROC (UROC) curve, and an associated rank based coefficient of (potential) predictive ability (CPA) measure — tools that apply to just any linearly ordered outcome, including both binary, ordinal, mixed discrete-continuous, and continuous variables.

4.3.1 General real-valued setting

Generalizing the binary setting in (4.1), we now consider bivariate data of the form

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}, \tag{4.8}$$

where x_i is a real-valued point forecast, regression output, feature, marker, or covariate value, and y_i is a real-valued outcome, for $i = 1, \ldots, n$. Throughout we assume that there are at least two unique values among the outcomes $y_1, \ldots, y_n$.

The crux of the subsequent development lies in a conversion to a sequence of binary problems. To this end, we let

$$z_1 < \cdots < z_m$$

denote the $m \leq n$ unique values of $y_1, \ldots, y_n$, and we define

$$n_c = \sum_{i=1}^{n} \mathbb{1}\{y_i = z_c\}$$

as the number of instances among the outcomes $y_1, \ldots, y_n$ that equal z_c, for $c = 1, \ldots, m$, so that $n_1 + \cdots + n_m = n$. We refer to the respective groups of instances as *classes*.

Next we transform the real-valued outcomes $y_1, \ldots, y_n$ into binary outcomes $\mathbb{1}\{y_1 \geq \theta\}, \ldots, \mathbb{1}\{y_n \geq \theta\}$ relative to a threshold value $\theta \in \mathbb{R}$. Thus, instead of analysing the original problem in (4.8), we consider a series of binary problems. By construction, only values of θ equal to $z_2, \ldots, z_m$ result in nontrivial, unique sets of binary outcomes.

Therefore, we consider $m-1$ derived classification problems with binary data of the form

$$(x_1, \mathbb{1}\{y_1 \geq z_{c+1}\}), \ldots, (x_n, \mathbb{1}\{y_n \geq z_{c+1}\}) \in \mathbb{R} \times \{0,1\}, \tag{4.9}$$

where $c = 1, \ldots, m-1$. As the derived problems are binary, all the tools of traditional ROC analysis apply.

In the remainder of the section we describe our generalization of ROC curves for binary data to ROC movies and universal ROC (UROC) curves for real-valued data. First, we argue that the $m-1$ classical ROC curves for the derived data in (4.9) can be merged into a single dynamical display, to which we refer as a ROC movie (Definition 1). Then we define the UROC curve as a judiciously weighted average of the classical ROC curves of which the ROC movie is composed (Definition 2).

Finally, we introduce a general measure of potential predictive ability for features, termed the coefficient of predictive ability (CPA). CPA is a weighted average of the AUC values for the derived binary problems in the very same way that the UROC curve is a weighted average of the (classical) ROC curves that constitute the ROC movie. Hence, CPA equals the area under the UROC curve (Definition 3). Alternatively, CPA can be interpreted as a weighted probability of concordance (Theorem 1) or in terms of rank based covariances (Theorem 2). CPA reduces to AUC if the outcomes are binary, and relates linearly to Spearman's rank correlation coefficient if the outcomes are continuous (Theorems 3 and 4).

4.3.2 ROC movies

We consider the sequence of $m-1$ classification problems for the derived binary data in (4.9). For $c = 1, \ldots, m-1$, we let ROC_c denote the associated ROC curve, and we let AUC_c be the respective AUC value.

Definition 1. For data of the form (4.8), the ROC *movie* is the sequence $(\mathrm{ROC}_c)_{c=1,\ldots,m-1}$ of the ROC curves for the induced binary data in (4.9).

If the original problem is binary there are $m = 2$ classes only, and the ROC movie reduces to the classical ROC curve. In case the outcome attains $m \geq 3$ distinct values the ROC movie can be visualized by displaying the associated sequence of $m-1$ ROC curves. In medical survival analysis, the outcomes $y_1, \ldots, y_n$ in data of the form (4.8) are survival times, and the analysis is frequently hampered by censoring, as patients drop out of studies. In this setting, Etzioni et al. (1999) and Heagerty et al. (2000)

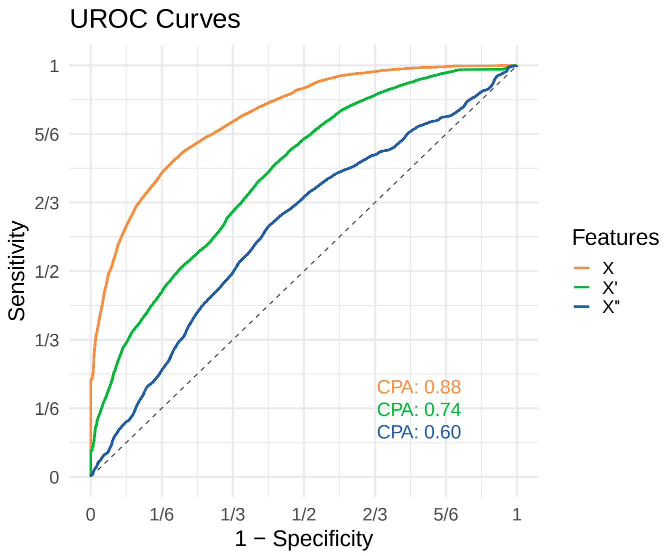

Figure 4.4: ROC movies and UROC curves for the features X, X' and X'' as predictors of the real-valued outcome Y in the simulation example of Section 4.2.3, based on the same sample as in Figure 4.3. In the ROC movies, the number at upper left shows the threshold under consideration, the number at upper center the relative weight $w_c / \max_{l=1,\ldots,m-1} w_l$ from (4.11), and the numbers at bottom right the respective AUC values.

introduced the notion of time-dependent ROC curves, which are classical ROC curves for the binary indicator $\mathbb{1}\{y_i \geq t\}$ of survival through (follow-up) time t, with censoring being handled efficiently. For an example see Figure 2 of Heagerty et al. (2000), where the ROC curves concern survival beyond follow-up times of 40, 60, and 100 months, respectively. If the thresholds considered correspond to the unique values of the outcomes, the sequence of time-dependent ROC curves becomes a ROC movie in the sense of Definition 1, save for the handling of censored data. When the number $m \leq n$ of classes is small or modest, the generation of the ROC movie is straightforward. Adaptations might be required as m grows, and we tend to this question in Section 4.5.2.

We have implemented ROC movies, UROC curves, and CPA within the `uroc` package for the statistical programming language R (R Core Team, 2021) where the `animation` package of Xie (2013) provides functionality for converting R images into a GIF animation, based on the external software `ImageMagick`. The `uroc` package can be downloaded from `https://github.com/evwalz/uroc`. In addition, a Python (Python Software Foundation, 2021) implementation is available at `https://github.com/e` `vwalz/urocc`. Returning to the example of Section 4.2.3, Figure 4.4 compares the features X, X' and X'' as predictors of the real-valued outcome Y in a joint display of the three ROC movies and UROC curves, based on the same sample of size $n = 400$ as in Figure 4.3. In the ROC movies, the threshold $z = 1.00$ recovers the traditional ROC curves in Figure 4.3.

4.3.3 Universal ROC (UROC) curves

Next we propose a simple and efficient way of subsuming a ROC movie for data of the form (4.8) into a single, static graphical display. As before, let $z_1 < \cdots < z_m$ denote the unique values of $y_1, \ldots, y_n$, let $n_c = \sum_{i=1}^{n} \mathbb{1}\{y_i = z_c\}$, and let ROC_c denote the (classical) ROC curve associated with the binary problem in (4.9), for $c = 1, \ldots, m-1$.

By Theorem 5 of Gneiting and Vogel (2022), there is a natural bijection between the class of the ROC curves and the class of the cumulative distribution functions (CDFs) of Borel probability measures on the unit interval. In particular, any ROC curve can be associated with a non-decreasing, right-continuous function $R : [0,1] \to [0,1]$ such that $R(0) = 0$ and $R(1) = 1$. Hence, any convex combination of the ROC curves $\mathrm{ROC}_1, \ldots, \mathrm{ROC}_{m-1}$ can also be associated with a non-decreasing, right-continuous function on the unit interval. It is in this sense that we define the following; in a nutshell, the UROC curve averages the traditional ROC curves of which the ROC movie is composed.

Definition 2. For data of the form (4.8), the *universal receiver operating characteristic* (UROC) *curve* is the curve associated with the function

$$\sum_{c=1}^{m-1} w_c \, \mathrm{ROC}_c \tag{4.10}$$

on the unit interval, with weights

$$w_c = \left(\sum_{i=1}^{c} n_i \sum_{i=c+1}^{m} n_i \right) \Big/ \left(\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i) n_i n_j \right) \tag{4.11}$$

for $c = 1, \ldots, m-1$.

Importantly, the weights in (4.11) depend on the data in (4.8) via the outcomes $y_1, \ldots, y_n$ only. Thus, they are independent of the feature values and can be used meaningfully in order to compare and rank features. Their specific choice is justified in Theorems

1 and 2 below. Clearly, the weights are nonnegative and sum to one. If $m = n$ then $n_1 = \cdots = n_m = 1$, and (4.11) reduces to

$$w_c = 6\,\frac{c(n-c)}{n(n^2-1)} \quad \text{for} \quad c = 1, \ldots, n-1; \tag{4.12}$$

so the weights are quadratic in the rank c and symmetric about the inner most rank(s), at which they attain a maximum. As we will see, our choice of weights has the effect that in this setting the area under the UROC curve, to which we refer as a general coefficient of predictive ability (CPA), relates linearly to Spearman's rank correlation coefficient, in the same way that AUC relates linearly to Somers' D.

In Figure 4.4 the UROC curves appear in the final static screen, subsequent to the ROC movies. Within each ROC movie, the individual frames show the ROC curve ROC_c for the feature considered. Furthermore, we display the threshold z_c, the *relative* weight from (4.11) (the actual weight normalized to the unit interval, i.e., we show $w_c / \max_{l=1,\ldots,m-1} w_l$), and AUC_c, respectively, for $c = 1, \ldots, m-1$. Once more we emphasize that the use of ROC movies, UROC curves, and CPA frees researchers from the need to select — typically, arbitrary — threshold values and binarize, as mandated by classical ROC analysis.

Of course, if specific threshold values are of particular substantive interest, the respective ROC curves can be extracted from the ROC movie, and it can be useful to plot AUC_c versus the associated threshold value z_c. Displays of this type have been introduced and studied by Rosset et al. (2005).

4.4 Coefficient of predictive ability (CPA)

We proceed to define the coefficient of predictive ability (CPA) as a general measure of potential predictive ability, based on notation introduced in Sections 4.3.2 and 4.3.3.

Definition 3. For data of the form (4.8) and weights $w_1, \ldots, w_{m-1}$ as in (4.11), the *coefficient of predictive ability* (CPA) is defined as

$$\mathrm{CPA} = \sum_{c=1}^{m-1} w_c \, \mathrm{AUC}_c. \tag{4.13}$$

In words, CPA equals the area under the UROC curve.

Importantly, ROC movies, UROC curves, and CPA satisfy a fundamental requirement on any generalization of ROC curves and AUC, in that they reduce to the classical notions when applied to a binary problem, whence $m = 2$ in (4.10) and (4.13), respectively. Another requirement that we consider essential is that, when both the feature values $x_1, \ldots, x_n$ and the outcomes $y_1, \ldots, y_n$ are pairwise distinct, the value of a performance measure remains unchanged if we transpose the roles of the feature and the outcome. As we will see, this is true under our specific choice (4.11) of the weights w_c in the defining formula (4.13) for CPA, but is not true under other choices, such as in the case of equal weights.

4.4.1 Interpretation as a weighted probability

We now express CPA in terms of pairwise comparisons via the function s in (4.2). To this end, we usefully change notation for the data in (4.8) and refer to the feature values in class $c \in \{1, \ldots, m\}$ as x_{ck}, for $k = 1, \ldots, n_c$. Thus, we rewrite (4.8) as

$$(x_{11}, z_1), \ldots, (x_{1n_1}, z_1), \ \ldots, \ (x_{m1}, z_m), \ldots, (x_{mn_m}, z_m) \ \in \ \mathbb{R} \times \mathbb{R}, \tag{4.14}$$

where $z_1 < \cdots < z_m$ are the unique values of $y_1, \ldots, y_n$ and $n_c = \sum_{i=1}^{n} \mathbb{1}\{y_i = z_c\}$, for $c = 1, \ldots, m$.

Theorem 1. *For data of the form* (4.14),

$$\mathrm{CPA} = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (j-i)\, s(x_{ik}, x_{jl})}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i)\, n_i n_j}. \tag{4.15}$$

Proof. By (4.4), the individual AUC values satisfy

$$\mathrm{AUC}_c = \frac{1}{\sum_{i=1}^{c} n_i \sum_{i=c+1}^{m} n_i} \sum_{i=1}^{c} \sum_{j=c+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} s(x_{ik}, x_{jl})$$

for $c = 1, \ldots, m-1$. In view of (4.11) and (4.13), summation yields

$$\begin{aligned} \mathrm{CPA} &= \sum_{c=1}^{m-1} w_c \, \mathrm{AUC}_c \\ &= \frac{\sum_{c=1}^{m-1} \sum_{i=1}^{c} \sum_{j=c+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} s(x_{ik}, x_{jl})}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i)\, n_i n_j} \\ &= \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (j-i)\, s(x_{ik}, x_{jl})}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i)\, n_i n_j}, \end{aligned}$$

as claimed. □

Thus, CPA is based on pairwise comparisons of feature values, counting the number of concordant pairs in (4.14), adjusting to a count of $\frac{1}{2}$ if feature values are tied, and weighting a pair's contribution by a class based distance, $j - i$, between the respective outcomes, $z_j > z_i$. In other words, CPA equals a weighted probability of concordance, with weights that grow linearly in the class based distance between outcomes.

The specific form of CPA in (4.15) invites comparison to a widely used measure of discrimination in biomedical applications, namely, the C *index* (Harrell et al., 1996; Pencina and D'Agostino, 2004)

$$\mathrm{C} = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} s(x_{ik}, x_{jl})}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} n_i n_j}. \tag{4.16}$$

If the outcomes are binary, both the C index and CPA reduce to AUC. While CPA can be interpreted as a weighted probability of concordance, C admits an interpretation as an unweighted probability, whence Mason and Weigel (2009) recommend its use for administrative purposes. However, the weighting in (4.15) may be more meaningful, as concordance between feature–outcome pairs with outcomes that differ substantially in rank tends to be of greater practical relevance than concordance between pairs with alike outcomes. While CPA admits the appealing, equivalent interpretation (4.13) in terms of binary AUC values and the area under the UROC curve, relationships of this type are unavailable for the C index.

Subject to conditions, the C index relates linearly to Kendall's rank correlation coefficient (Somers, 1962; Pencina and D'Agostino, 2004; Mason and Weigel, 2009). In Section 4.4.3 we demonstrate the same type of relationship for CPA and Spearman's rank correlation coefficient, thereby resolving a problem raised by Heagerty and Zheng (2005, p. 95). Just as the C index bridges and generalizes AUC and Kendall's coefficient, CPA bridges and nests AUC and Spearman's coefficient, with the added benefit of appealing interpretations in terms of the area under the UROC curve and rank based covariances.

4.4.2 Representation in terms of covariances

The key result in this section represents CPA in terms of the covariance between the class of the outcome and the mid rank of the feature, relative to the covariance between the class of the outcome and the mid rank of the outcome itself.

The mid rank method handles ties by assigning the arithmetic average of the ranks involved (Woodbury, 1940; Kruskal, 1958). For instance, if the third to seventh positions in a list are tied, their shared *mid rank* is $\frac{1}{5}(3+4+5+6+7) = 5$. This approach treats equal values alike and guarantees that the sum of the ranks in any tied group is unchanged from the case of no ties. As before, if $y_i = z_j$, where $z_1 < \cdots < z_m$ are the unique values of $y_1, \ldots, y_n$ in (4.8), we say that the *class* of y_i is j. In brief, we express this as $\mathrm{cl}(y_i) = j$. Similarly, we refer to the mid rank of x_i within $x_1, \ldots, x_n$ as $\overline{\mathrm{rk}}(x_i)$.

Theorem 2. *Let the random vector (X, Y) be drawn from the empirical distribution of the data in* (4.8) *or* (4.14). *Then*

$$\mathrm{CPA} = \frac{1}{2}\left(\frac{\mathrm{cov}(\mathrm{cl}(Y), \overline{\mathrm{rk}}(X))}{\mathrm{cov}(\mathrm{cl}(Y), \overline{\mathrm{rk}}(Y))} + 1\right). \tag{4.17}$$

Proof. Suppose that the law of the random vector (X, Y) is the empirical distribution of the data in (4.8). Based on the equivalent representation in (4.14), we find that

$$\frac{\mathrm{cov}(\mathrm{cl}(Y), \overline{\mathrm{rk}}(X))}{\mathrm{cov}(\mathrm{cl}(Y), \overline{\mathrm{rk}}(Y))} = \frac{\sum_{i=1}^m \sum_{k=1}^{n_i} i\,\overline{\mathrm{rk}}(x_{ik}) - \frac{1}{2}(n+1)\sum_{i=1}^m i n_i}{\sum_{i=1}^m i n_i \left(\sum_{j=0}^{i-1} n_j + \frac{1}{2}(n_i+1)\right) - \frac{1}{2}(n+1)\sum_{i=1}^m i n_i},$$

where $n_0 = 0$. Consequently, we can rewrite (4.17) as

$$\mathrm{CPA} = \frac{\sum_{i=1}^m \sum_{k=1}^{n_i} i\,\overline{\mathrm{rk}}(x_{ik}) + \sum_{i=1}^m i n_i \left(\sum_{j=0}^{i-1} n_j + \frac{1}{2}n_i - n - \frac{1}{2}\right)}{\sum_{i=1}^m i n_i \left(2\sum_{j=0}^{i-1} n_j + n_i - n\right)}. \tag{4.18}$$

We proceed to demonstrate that the numerator and denominator in (4.15) equal the numerator and denominator in (4.18), respectively. To this end, we first compare feature values within classes and note that

$$\sum_{i=1}^m \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} i s(x_{il}, x_{ik}) = \sum_{i=1}^m i \sum_{k=1}^{n_i} \left(n_i - k + \frac{1}{2}\right) = \frac{1}{2}\sum_{i=1}^m i n_i^2;$$

for if the feature values in class i are all distinct, the largest one exceeds $n_i - 1$ others, the second largest exceeds $n_i - 2$ others, and so on, and analogously in case of ties.

We now show the equality of the numerators in (4.15) and (4.18), in that

$$
\begin{aligned}
&\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (j-i)\, s(x_{ik}, x_{jl}) \\
&= \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} j\, s(x_{ik}, x_{jl}) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} i\, s(x_{ik}, x_{jl}) \\
&\qquad\qquad + \sum_{j=1}^{m-1} \sum_{i=j+1}^{m} \sum_{k=1}^{n_j} \sum_{l=1}^{n_i} j\, s(x_{ik}, x_{jl}) - \sum_{j=1}^{m-1} \sum_{i=j+1}^{m} \sum_{k=1}^{n_j} \sum_{l=1}^{n_i} j\, s(x_{ik}, x_{jl}) \\
&= \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} j\, s(x_{ik}, x_{jl}) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} i \left(s(x_{jl}, x_{ik}) + s(x_{ik}, x_{jl}) \right) \\
&= \sum_{j=1}^{m} \sum_{l=1}^{n_j} j \left(\overline{\mathrm{rk}}(x_{jl}) - \frac{1}{2} \right) - \sum_{i=1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} i\, s(x_{il}, x_{ik}) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} i n_i n_j \\
&= \sum_{i=1}^{m} \sum_{k=1}^{n_i} i \overline{\mathrm{rk}}(x_{ik)} - \frac{1}{2} \sum_{i=1}^{m} i n_i - \frac{1}{2} \sum_{i=1}^{m} i n_i^2 - n \sum_{i=1}^{m-1} i n_i + \sum_{i=1}^{m-1} i n_i \sum_{j=0}^{i} n_j \\
&= \sum_{i=1}^{m} \sum_{k=1}^{n_i} i \overline{\mathrm{rk}}(x_{ik)} - \frac{1}{2} \sum_{i=1}^{m} i n_i - \frac{1}{2} \sum_{i=1}^{m} i n_i^2 - n \sum_{i=1}^{m} i n_i + \sum_{i=1}^{m} i n_i \sum_{j=0}^{i} n_j \\
&= \sum_{i=1}^{m} \sum_{k=1}^{n_i} i \overline{\mathrm{rk}}(x_{ik)} + \sum_{i=1}^{m} i n_i \left(\sum_{j=0}^{i-1} n_j + \frac{1}{2} n_i - n - \frac{1}{2} \right).
\end{aligned}
$$

As for the denominators,

$$
\begin{aligned}
\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i)\, n_i n_j &= \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} j n_i n_j - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} i n_i n_j \\
&= \sum_{i=1}^{m} i n_i \sum_{k=0}^{i-1} n_k - n \sum_{i=1}^{m-1} i n_i + \sum_{i=1}^{m-1} i n_i \sum_{k=1}^{i} n_k \\
&= 2 \sum_{i=1}^{m} i n_i \sum_{k=0}^{i-1} n_k - n \sum_{i=1}^{m-1} i n_i + \sum_{i=1}^{m-1} i n_i^2 + \sum_{i=1}^{m-1} i n_i \sum_{k=0}^{i-1} n_k - \sum_{i=1}^{m} i n_i \sum_{k=0}^{i-1} n_k \\
&= 2 \sum_{i=1}^{m} i n_i \sum_{k=0}^{i-1} n_k - n \sum_{i=1}^{m-1} i n_i + \sum_{i=1}^{m-1} i n_i^2 - n m n_m + m n_m^2 \\
&= 2 \sum_{i=1}^{m} i n_i \sum_{k=0}^{i-1} n_k - n \sum_{i=1}^{m} i n_i + \sum_{i=1}^{m} i n_i^2 \\
&= \sum_{i=1}^{m} i n_i \left(2 \sum_{j=0}^{i-1} n_j + n_i - n \right),
\end{aligned}
$$

whence the proof is complete. □

Interestingly, the representation (4.17) in terms of rank and class based covariances appears to be new even in the special case when the outcomes are binary, so that CPA reduces to AUC. The representation also sheds new light on the asymmetry of CPA, in that, in general, the value of CPA changes if we transpose the roles of the feature and the outcome. In contrast to customarily used measures of bivariate association and dependence, which are necessarily symmetric (Nešlehová, 2007; Reshef et al., 2011; Weihs et al., 2018), CPA is directed when the outcome is binary or ordinal. Thus, CPA avoids a technical issue with the use of rank-based correlation coefficients in discrete settings, namely, that perfect classifiers do not reach the optimal values of the respective performance measures (Nešlehová, 2007, p. 565). However, in the case of no ties at all, to which we tend now, CPA becomes symmetric, as one would expect, given that the feature and the outcome are on equal footing then.

4.4.3 Relationship to Spearman's rank correlation coefficient

Spearman's rank correlation coefficient ρ_{S} for data of the form (4.8) is generally understood as Pearson's correlation coefficient applied to the respective ranks (Spearman, 1904). In case there are no ties in either $x_1, \ldots, x_n$ nor $y_1, \ldots, y_n$, the concept is unambiguous, and Spearman's coefficient can be computed as

$$\rho_{\mathrm{S}} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \left(\mathrm{rk}(x_i) - \mathrm{rk}(y_i)\right)^2 , \tag{4.19}$$

where $\mathrm{rk}(x_i)$ denotes the rank of x_i within $x_1, \ldots, x_n$, and $\mathrm{rk}(y_i)$ the rank of y_i within $y_1, \ldots, y_n$,

In this setting CPA relates linearly to Spearman's rank correlation coefficient ρ_{S}, in the very same way that AUC relates to Somers' D in (4.5).

Theorem 3. *In the case of no ties,*

$$\mathrm{CPA} = \frac{1}{2}\left(\rho_{\mathrm{S}} + 1\right). \tag{4.20}$$

Indeed, in case there are no ties, both mid ranks and classes reduce to ranks proper, and then (4.20) is readily identified as a special case of (4.17). For an alternative proof, in the absence of ties the weights w_c in (4.11) are of the form (4.12). The stated result then follows upon combining the defining equation (4.10), the equality stated at the bottom of the left column of page 4 in Rosset et al. (2005), and equation (5) in the same reference.

Note that CPA becomes symmetric in this case, as its value remains unchanged if we transpose the roles of the feature and the outcome. Furthermore, if the joint distribution of a bivariate random vector (X, Y) is continuous, and we think of the data in (4.8) as a sample from the respective population, then, by applying Definition 3 and Theorem 3 in the large sample limit, and taking (4.12) into account, we (informally) obtain a population version of CPA, namely,

$$\mathrm{CPA} = 6 \int_0^1 \alpha(1-\alpha)\, \mathrm{AUC}_\alpha \, \mathrm{d}\alpha = \frac{1}{2}\left(\rho_\mathrm{S} + 1\right), \tag{4.21}$$

where AUC_α is the population version of AUC for $(X, \mathbb{1}\{Y \geq q_\alpha\})$, with q_α denoting the α-quantile of the marginal law of Y. We defer a rigorous derivation of (4.21) to future work and stress that, as both X and Y are continuous here, their roles can be interchanged.

Under the assumption of multivariate normality, the population version of Spearman's ρ_S relates to Pearson's correlation coefficient r as

$$\rho_\mathrm{S} = \frac{6}{\pi} \arcsin \frac{r}{2}; \tag{4.22}$$

see, e.g., Kruskal (1958). Returning to the example in Section 4.2.3, where (Y, X, X', X'') is jointly Gaussian with covariance matrix (4.6), Table 4.1 states, for each feature, the population values of Pearson's correlation coefficient r, CPA, and the C index relative to the real-valued outcome Y, as derived from (4.21) and (4.22) and the respective relationships for the C index and Kendall's rank correlation coefficient τ_K, namely

$$\mathrm{C} = \frac{1}{2}\left(\tau_\mathrm{K} + 1\right) \tag{4.23}$$

and

$$\tau_\mathrm{K} = \frac{2}{\pi} \arcsin r. \tag{4.24}$$

These results imply that for a bivariate Gaussian population with Pearson correlation coefficient $r \in (0, 1)$ it is true that $\tau_\mathrm{K} > \rho_\mathrm{S} > 0$ and $\mathrm{CPA} > \mathrm{C} > 1/2$. In fact, under positive dependence it always holds that $\tau_\mathrm{K} \geq \rho_\mathrm{S} \geq 0$, as demonstrated by Capéraà and Genest (1993), whence $\mathrm{CPA} \geq \mathrm{C} \geq 1/2$. However, there are also settings where these inequalities get violated (Schreyer et al., 2017). In Figure 4.4 the CPA values for the features appear along with the UROC curves in the final static screen, subsequent

Table 4.1: Population values of Pearson's correlation coefficient r, CPA, and the C index for the features X, X', and X'' relative to the real-valued outcome Y, where (Y, X, X', X'') is Gaussian with covariance matrix (4.6).

Feature	r	CPA	C
X	0.800	0.893	0.795
X'	0.500	0.741	0.667
X''	0.200	0.596	0.564

to the ROC movie. The empirical values show the expected approximate agreement with the population quantities in the table.

Suppose now that the values $y_1, \dots, y_n$ of the outcomes are unique, whereas the feature values $x_1, \dots, x_n$ might involves ties. Let $p \geq 0$ denote the number of tied groups within $x_1, \dots, x_n$. If $p = 0$ let $V = 0$. If $p \geq 1$, let v_j be the number of equal values in the jth group, for $j = 1, \dots, p$, and let

$$V = \frac{1}{12} \sum_{j=1}^{p} \left(v_j^3 - v_j \right).$$

Then Spearman's *mid rank adjusted* coefficient ρ_{M} is defined as

$$\rho_{\mathrm{M}} = 1 - \frac{6}{n(n^2-1)} \left(\sum_{i=1}^{n} \left(\overline{\mathrm{rk}}(x_i) - \mathrm{rk}(y_i) \right)^2 + V \right), \tag{4.25}$$

where $\overline{\mathrm{rk}}$ is the aforementioned mid rank. As shown by Woodbury (1940), if one assigns all possible combinations of integer ranks within tied sets, computes Spearman's ρ_{S} in (4.19) on every such combination and averages over the respective values, one obtains the formula for ρ_{M} in (4.25).

The following result reduces to the statement of Theorem 3 in the case $p = 0$ when there are no ties in $x_1, \dots, x_n$ either.

Theorem 4. *In case there are no ties within $y_1, \dots, y_n$,*

$$\mathrm{CPA} = \frac{1}{2} \left(\rho_{\mathrm{M}} + 1 \right). \tag{4.26}$$

Proof. As noted, ρ_{M} arises from ρ_{S} if one assigns all possible combinations of integer ranks within tied sets, computes ρ_{S} on every such combination and averages over the respective values. In view of (4.18), if there are no ties in $y_1, \ldots, y_n$, averaging $\frac{1}{2}\left(\rho_{\mathrm{S}}+1\right)$ over the combinations yields $\frac{1}{2}\left(\rho_{\mathrm{M}}+1\right)$, which equals CPA by (4.17). □

The relationships (4.5), (4.20) and (4.26) constitute but special cases of the general, covariance based representation (4.17). In this light, CPA provides a unified way of quantifying potential predictive ability for the full gamut of dichotomous, categorical, mixed discrete-continuous and continuous types of outcomes. In particular, CPA bridges and generalizes AUC, Somers' D and Spearman's rank correlation coefficient, up to a common linear relationship.

4.4.4 Comparison of CPA to the C index and related measures

We proceed to a more detailed comparison of the CPA measure (4.13) to the C index (4.16) and measures studied by Waegeman et al. (2008).[1] As noted, both CPA and the C index are rank-based, reduce to AUC when the outcome is binary, and become symmetric when both the features and the outcomes are pairwise distinct. We relax these conditions slightly and restrict attention to measures that use ranks only, reduce to AUC when the outcome is binary *and* there are no ties in the feature values, and become symmetric when there are no ties at all. This excludes measures based on the receiver error characteristic (REC, Bi and Bennett, 2003) and the regression receiver operating characteristic (RROC, Hernández-Orallo, 2013) curve, which are neither rank based nor reduce to AUC. The U_{cons} measure of Waegeman et al. (2008) averages consecutive AUC values in the same fashion as CPA in (4.13), but uses constant weights, as opposed to the class dependent weights (4.11) for CPA, and does not become symmetric when there are no ties at all.[2] The U_{pairs} and U_{ovo} measures of Waegeman et al. (2008) satisfy our criteria, relate closely to the C index, and in the simulation setting of Figure 4.5 it holds that $U_{\text{ovo}} = U_{\text{pairs}} = \mathrm{C}$.[3]

[1] We denote the measures $\widehat{U}$, $\widehat{U}_{\text{pairs}}$, $\widehat{U}_{\text{ovo}}$, and $\widehat{U}_{\text{cons}}$ in equations (8), (16), (17), and (18) of Waegeman et al. (2008) by U, U_{pairs}, U_{ovo}, and U_{cons}, respectively.

[2] To see that U_{cons} does not become symmetric when there are no ties in $x_1, \ldots, x_n$ nor $y_1, \ldots, y_n$, consider a dataset of size $n \geq 4$, where $y_1 < \cdots < y_n$ and $x_3 < x_1 < x_2 < x_4 < \cdots < x_n$. Then $\mathrm{AUC}_1 = (n-3)/(n-1)$, $\mathrm{AUC}_2 = (2n-5)/(2n-4)$, and $\mathrm{AUC}_c = 1$ for $c = 3, \ldots, n-1$, whereas if we interchange the roles of the feature and the outcome, then $\mathrm{AUC}_1 = (n-2)/(n-1)$, $\mathrm{AUC}_2 = (2n-6)/(2n-4)$, and $\mathrm{AUC}_c = 1$ for $c = 3, \ldots, n-1$, resulting in distinct unweighted sums.

[3] The U_{pairs} measure corresponds to a performance criterion proposed by Herbrich et al. (2000, equation (7.11)) and equals the proportion of correctly ranked pairs of instances. Except for the treatment of ties in the feature, U_{pairs} equals the C index. In particular, if the feature values are pairwise

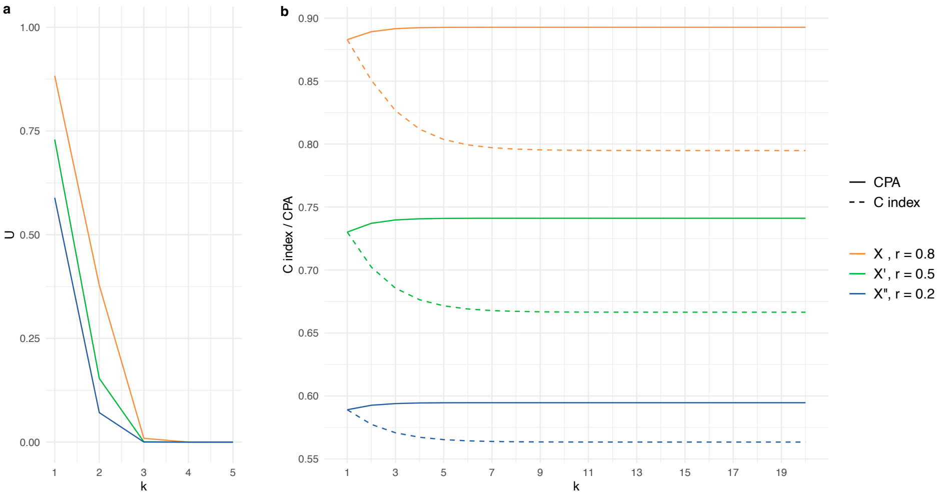

Figure 4.5: Rank based performance measures for the features X, X' and X'' as predictors of the real-valued outcome Y in the simulation example of Section 4.2.3, with Pearson correlation coefficient $r = 0.8$, 0.5 and 0.2, respectively, based on a sample of size $n = 2^{20}$. We discretize the continuous outcome into 2^k consecutive blocks of size 2^{20-k} each, and plot (a) U, and (b) CPA and the C index as functions of the discretization level $k = 1, \ldots, 20$. Note that $k = 1$ yields a binary outcome and $k = 20$ a continuous outcome.

In view of the above requirements and properties, we restrict the subsequent comparison to CPA, the C index, and the U measure introduced by Waegeman et al. (2008). For a dataset with m classes U equals the proportion of sequences of m instances, one of each class, that align correctly with the feature values. As noted, these measures are rank based and reduce to AUC when the outcome is binary and there are no ties in the feature values. In the continuous case with no ties in the feature values nor in the outcomes, they become symmetric, U attains the value 1 under a perfect ranking and the value 0 otherwise, $\mathrm{C} = \frac{1}{2}\,(1 - \tau_{\mathrm{K}})$, and $\mathrm{CPA} = \frac{1}{2}\,(1 - \rho_{\mathrm{S}})$.

In Figure 4.5 we report on a simulation experiment where we draw samples of 2^{20} instances from the joint Gaussian distribution of the random vector (Y, X, X', X'') with covariance matrix (4.6), so that the features have Pearson correlation coefficient $r = 0.8, 0.5$, and 0.2 with the continuous outcome Y. By discretizing the outcome into

distinct then $U_{\mathrm{pairs}} = \mathrm{C}$. The measure U_{ovo} represents the Hand and Till (2001) approach of averaging the $\binom{m}{2}$ one-versus-one AUC values in an m-class problem. It has been compared to U_{pairs} by Waegeman et al. (2008) and relates to the C index as well. In particular, if the feature values are pairwise distinct and the dataset furthermore is balanced with class memberships $n_1 = \cdots = n_m$, as in the simulation setting that we report on in Figure 4.5, then $U_{\mathrm{ovo}} = U_{\mathrm{pairs}} = \mathrm{C}$.

2^k consecutive blocks of size 2^{20-k} each, where $k = 1, \ldots, 20$, and computing CPA, the C index and the U measure as a function of k, all discretization levels are considered, ranging from a binary variable for $k = 1$ to continuous outcomes for $k = 20$. When $k = 1$ the three measures coincide and equal AUC, essentially at the population value of

$$\mathrm{AUC}_{1/2} = \frac{2}{\pi} \arcsin \frac{r}{\sqrt{2}} + \frac{1}{2}, \tag{4.27}$$

in the sense stated subsequent to (4.21). The U measure is tailored to ordinal outcomes with a few classes only and degenerates rapidly with k. When $k = 20$, CPA and the C index are rescaled versions of Spearman's ρ_{S} and Kendall's τ_{K}, essentially at the population values in Table 4.1.

Throughout, the measures lie in between their common value for $k = 1$, which equals AUC, and the respective values for $k = 20$. For all features and all $k > 1$, the C index is smaller than CPA, and CPA varies considerably less with the discretization level than the C index. To supplement these experiments with an analytic demonstration, suppose that X and Y are bivariate Gaussian with nonnegative Pearson correlation r. If we convert Y to a balanced binary outcome, then both CPA and the C index reduce to a common value, namely, $\mathrm{AUC}_{1/2}$ in (4.27). As a function of r, the ratio of the C index for the continuous vs. the balanced binary outcome attains values between 0.8996 and 1, whereas for CPA the respective ratio remains between 1 and 1.0156, as illustrated in Figure 4.6. These findings along with results in Capéraà and Genest (1993) and Schreyer et al. (2017) suggest that, quite generally, CPA and the C index yield qualitatively similar results in practice, with CPA being less sensitive to quantization effects, and the value of CPA typically being larger than for the C index.

4.4.5 Computational issues

We turn to a discussion of the computational costs of generalized ROC analysis for a dataset of the form (4.8) or (4.14) with n instances and $m \leq n$ classes.

It is well known that a traditional ROC curve can be generated from a dataset with n instances in $O(n \log n)$ operations (Fawcett, 2006, Algorithm 1). A ROC movie comprises $m - 1$ traditional ROC curves, so in a naïve approach, ROC movies can be computed in $O(mn \log n)$ operations. However, our implementation takes advantage of recursive relations between consecutive component curves ROC$_{i-1}$ and ROC$_i$. While a formal analysis will need to be left to future work, we believe that our algorithm has computational costs of $O(n \log n)$ operations only. If the number m of unique values of

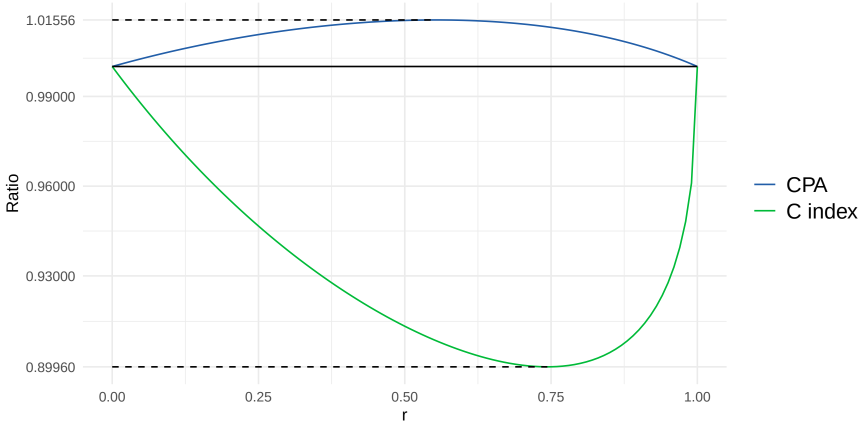

Figure 4.6: Ratio of CPA (blue curve) respectively the C index (green curve) for the feature X as a predictor of the continuous outcome Y over AUC for X and the balanced binary outcome $\mathbb{1}\{Y \geq 0\}$, where X and Y are bivariate Gaussian with Pearson correlation $r \in [0, 1]$. The solid horizontal line is at a ratio of 1, which is attained when $r = 0$ and $r = 1$.

the outcome is large, then for all practical purposes the ROC movie can be shown at a modest number m_0 of distinct values only, at a computational cost of $O(m_0 n \log n)$ operations. For example, in the setting of Figure 4.8 in the meteorological case study in Section 4.5.2 there are $m = 35,993$ unique values of the outcome, whereas the ROC movie uses $m_0 = 401$ frames only. For the vertical averaging of the component curves in the construction of UROC curves, we partition the unit interval into 1,000 equally sized subintervals.

Importantly, CPA can be computed in $O(n \log n)$ operations, without any need to invoke ROC analysis, by sorting $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, computing the respective mid ranks and classes, and plugging into the rank based representation (4.18). Similarly, there are algorithms for the computation of the C index in $O(n \log n)$ operations (Knight, 1966; Christensen, 2005).

4.4.6 Key properties: Comparison to traditional ROC analysis

We are now in a position to judge whether the proposed toolbox of ROC movies, UROC curves, and CPA constitutes a proper generalization of traditional ROC analysis. To facilitate the assessment, the subsequent statements admit immediate comparison with the key insights of classical ROC analysis, as summarized in Section 4.2.4.

We start with the trivial but important observation that the new tools nest the notions of traditional ROC analysis. This is not to be taken for granted, as extant generalizations do not necessarily share this property.

(0) In the case of a binary outcome, both the ROC movie and the UROC curve reduce to the ROC curve, and CPA reduces to AUC.

(1) ROC movies, the UROC curve and CPA are straightforward to compute and interpret, in the (rough) sense of *the larger the better*.

(2) CPA attains values between 0 and 1 and relates linearly to the covariance between the class of the outcome and the mid rank of the feature, relative to the covariance between the class and the mid rank of the outcome. In particular, if the outcomes are pairwise distinct, then $\mathrm{CPA} = \frac{1}{2}(\rho_{\mathrm{M}} + 1)$, where ρ_{M} is Spearman's mid rank adjusted coefficient (4.25). If the outcomes are binary, then $\mathrm{CPA} = \frac{1}{2}(D + 1)$ in terms of Somers' D. For a perfect feature, $\mathrm{CPA} = 1$, $\rho_{\mathrm{M}} = 1$ under pairwise distinct and $D = 1$ under binary outcomes. For a feature that is independent of the outcome, $\mathrm{CPA} = \frac{1}{2}$, $\rho_{\mathrm{M}} = 0$ under pairwise distinct and $D = 0$ under binary outcomes.

(3) The numerical value of CPA admits an interpretation as a weighted probability of concordance for feature–outcome pairs, with weights that grow linearly in the class based distance between outcomes.

(4) ROC movies, UROC curves, and CPA are purely rank based and, therefore, invariant under strictly increasing transformations. Specifically, if $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ are strictly increasing, then the ROC movie, UROC curve, and CPA computed from

$$(\varphi(x_1), \psi(y_1)), \ldots, (\varphi(x_n), \psi(y_n)) \in \mathbb{R} \times \mathbb{R} \tag{4.28}$$

are the same as the ROC movie, UROC curve, and CPA computed from the data in (4.8).

We iterate and emphasize that, as an immediate consequence of the final property, ROC movies, UROC curves, and CPA assess the discrimination ability or *potential* predictive ability of a point forecast, regression output, feature, marker, or test. Markedly different techniques are called for if one seeks to assess a forecast's *actual* value in any given applied problem (Ben Bouallègue et al., 2015; Ehm et al., 2016).

4.5 Real data examples

In the following examples from survival analysis and numerical weather prediction the usage of ROC movies, UROC curves, and CPA is demonstrated. We start by returning to the survival example from Figure 4.1, where the new set of tools frees researchers form the need to artificially binarize the outcome. Then the use of CPA is highlighted in a study of recent progress in numerical weather prediction (NWP), and in a comparison of the predictive performance of NWP models and convolutional neural networks.

4.5.1 Survival data from Mayo Clinic trial

In the introduction, Figs. 4.1 and 4.2 serve to illustrate and contrast traditional ROC curves, ROC movies and UROC curves. They are based on a classical dataset from a Mayo Clinic trial on primary biliary cirrhosis (PBC), a chronic fatal disease of the liver, that was conducted between 1974 and 1984 (Dickson et al., 1989). The data are provided by various R packages, such as `SMPracticals` and `survival`, and have been analyzed in textbooks (Fleming and Harrington, 1991; Davison, 2003). The outcome of interest is survival time past entry into the study. Patients were randomly assigned to either a placebo or treatment with the drug D-penicillamine. However, extant analyses do not show treatment effects (Dickson et al., 1989), and so we follow previous practice and study treatment and placebo groups jointly.

We consider two biochemical markers, namely, serum albumin and serum bilirubin concentration in mg/dl, for which higher and lower levels, respectively, are known to be indicative of earlier disease stages, thus supporting survival. Hence, for the purposes of ROC analysis we reverse the orientation of the serum bilirubin values. Given our goal of illustration, we avoid complications and remove patient records with censored survival times, to obtain a dataset with $n = 161$ patient records and $m = 156$ unique survival times. The proper handling of censoring is beyond the scope of our study, and we leave this task to subsequent work. For a discussion and comparison of extant approaches to handling censored data in the context of time-dependent ROC curves see Blanche et al. (2013).

The traditional ROC curves in Figure 4.1 are obtained by binarizing survival time at a threshold of 1462 days, which is the survival time in the data record that gets closest to four years. The ROC movies and UROC curves in Figure 4.2 are generated directly from the survival times, without any need to artificially pick a threshold. The CPA values for serum albumin and serum bilirubin are 0.73 and 0.77, respectively, and contrary to the ranking in Figure 4.1, where bilirubin was deemed superior, based on outcomes that were artificially made binary. Our tools free researchers from the need to binarize, and still they allow for an assessment at the binary level, if desired. For example, the ROC curves and AUC values from Figure 4.1 appear in the ROC movie at a threshold value of 1462 days. In line with current uses of AUC in a gamut of applied settings, CPA is particularly well suited to the purposes of feature screening and variable selection in statistical and machine learning models (Guyon and Elisseeff, 2003). Here, AUC and CPA demonstrate that both albumin and bilirubin contribute to prognostic models for survival (Dickson et al., 1989; Fleming and Harrington, 1991).

4.5.2 Monitoring progress in numerical weather prediction (NWP)

Here we illustrate the usage of CPA in the assessment of recent progress in numerical weather prediction (NWP), which has experienced tremendous advance over the past few decades (Bauer et al., 2015; Alley et al., 2019; Ben Bouallègue et al., 2019). Specifically, we consider forecasts of surface (2-meter) temperature, surface (10-meter) wind speed and 24-hour precipitation accumulation initialized at 00:00 UTC at lead times from a single day (24 hours) to five days (120 hours) ahead from the high-resolution model operated by the European Centre for Medium-Range Weather Forecasts (ECMWF Directorate, 2012), which is generally considered the leading global NWP model. The forecast data are available at `https://confluence.ecmwf.int/display/TIGGE`. As observational reference we take the ERA5 reanalysis product (Hersbach et al., 2018). We use forecasts and observations from $279 \times 199 = 55,521$ model grid boxes of size $0.25° \times 0.25°$ each in a geographic region that covers Europe from 25.0° W to 44.5° E in latitude and 25.0° N to 74.5° N in longitude. The time period considered ranges from January 2007 to December 2018.

In Figure 4.7 we apply CPA and the C index to compare forecasts from the ECMWF high-resolution run to a reference technique, namely, the persistence forecast. The persistence forecast is simply the most recent available observation for the weather quantity of interest; as such, the forecast value does not depend on the lead time. CPA and the C index are computed on rolling twelve-month periods that correspond to January–December, April–March, July–June or October–September, typically comprising $n = 365 \times 55,521 = 20,265,165$ individual forecast cases. The ECMWF fore-

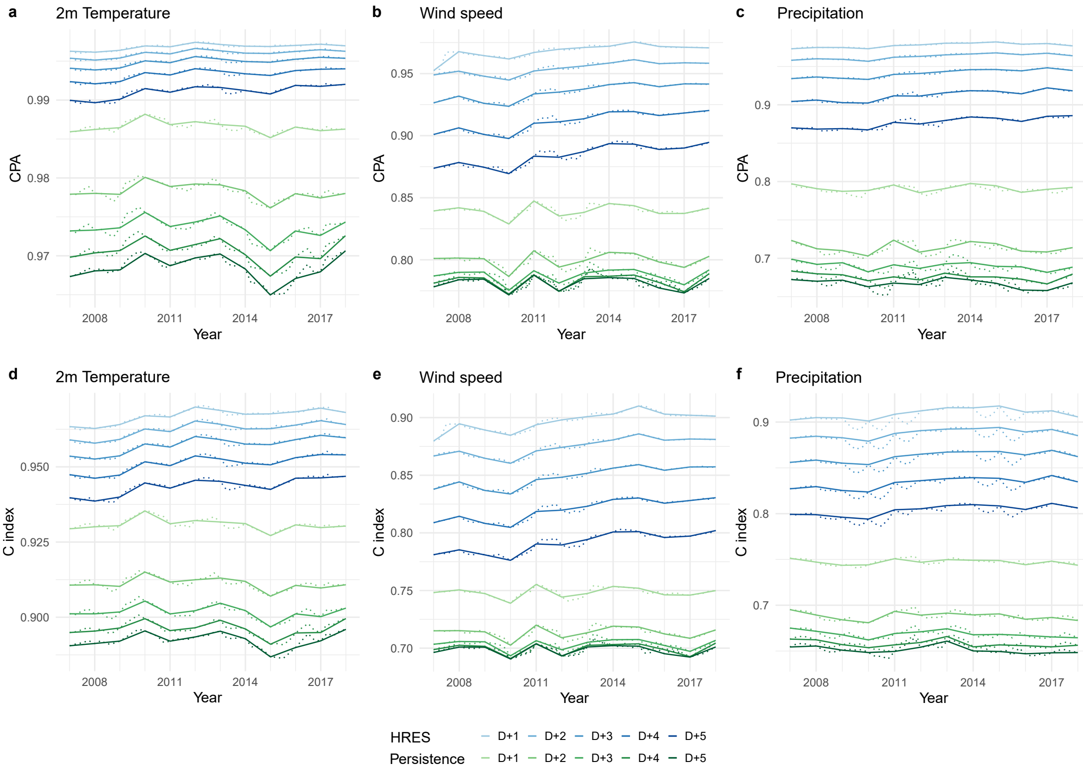

Figure 4.7: Temporal evolution of CPA and the C index for forecasts from the ECMWF high-resolution model at lead times of one to five days in comparison to the simplistic persistence forecast in terms of CPA (a, b, c) and the C index (d, e, f). The weather variables considered are (a, d) surface (2-meter) temperature, (b, e) surface wind speed and (c, f) 24-hour precipitation accumulation. The measures refer to a domain that covers Europe and twelve-month periods that correspond to January–December (solid and dotted lines), April–March, July–June and October–September (dotted lines only), based on gridded forecast and observational data from January 2007 through December 2018.

cast has considerably higher CPA and C index than the persistence forecast for all lead times and variables considered. For the persistence forecast the measures fluctuate around a constant level; for the ECMWF forecast they improve steadily, attesting to continuing progress in NWP (Bauer et al., 2015; Alley et al., 2019; Ben Bouallègue et al., 2019; Haiden et al., 2021).

To place these findings further into context, recall that CPA is a weighted average of AUC values for binarized outcomes at individual threshold values, as have been used for performance monitoring by weather centers (Ben Bouallègue et al., 2019; Haiden et al., 2021). The CPA measure preserves the spirit and power of classical

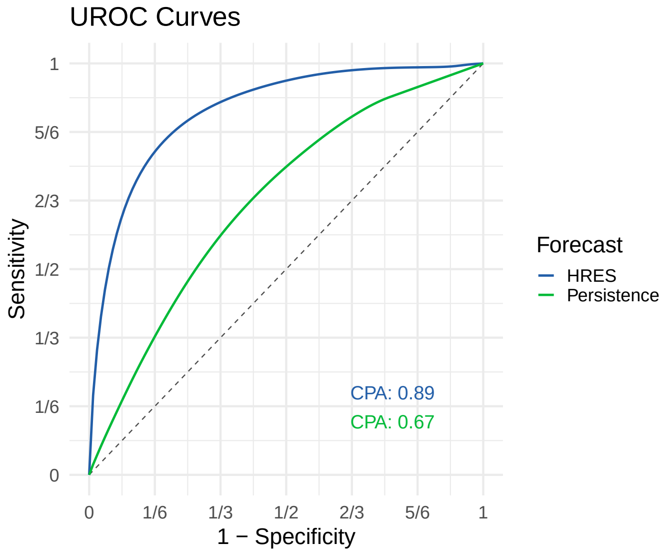

Figure 4.8: ROC movies, UROC curves, and CPA for ECMWF high-resolution (HRES) and persistence forecasts of 24-hour precipitation accumulation over Europe at a lead time of five days in calendar year 2018. In the ROC movies, the number at upper left shows the threshold at hand in the unit of millimeter, the number at upper center the relative weight $w_c / \max_{l=1,\ldots,m-1} w_l$ from (4.11), and the numbers at bottom right the respective AUC values.

ROC analysis, and frees researchers from the need to binarize real-valued outcomes. Results in terms of the C index are qualitatively similar, with the numerical value of CPA being higher than for the C index.

The ROC movies, UROC curves, and CPA values in Figure 4.8 compare the ECMWF high-resolution forecast to the persistence forecast for 24-hour precipitation accumulation at a lead time of five days in calendar year 2018. As noted, this record comprises more than 20 million individual forecast cases, and there are $m = 35{,}993$ unique values of the outcome. We certainly lack the patience to watch the full sequence of $m-1$ screens in the ROC movie. A pragmatic solution is to consider a subset $\mathcal{C} \subseteq \{1, \ldots, m-1\}$ of indices, so that ROC_c is included in the ROC movie (if and) only if $c \in \mathcal{C}$. Specifically, we set positive integer parameters $a \leq m-1$ and b such that the ROC movie comprises at least a and at most $a+b$ curves. Let the integer s be defined such that $1+(a-1)s \leq m-1 < 1+as$, and let $\mathcal{C}_a = \{1, 1+s, \ldots, 1+(a-1)s\}$, so that $|\mathcal{C}_a| = a$. Let $\mathcal{C}_b = \{c : n_c \geq n/b\}$; evidently, $|\mathcal{C}_b| \leq b$. Finally, let $\mathcal{C} = \mathcal{C}_a \cup \mathcal{C}_b$ so that $a \leq |\mathcal{C}| \leq a+b$. We have made good experiences with choices of $a = 400$ and $b = 100$, which in Figure 4.8 yield a ROC movie with 401 screens.

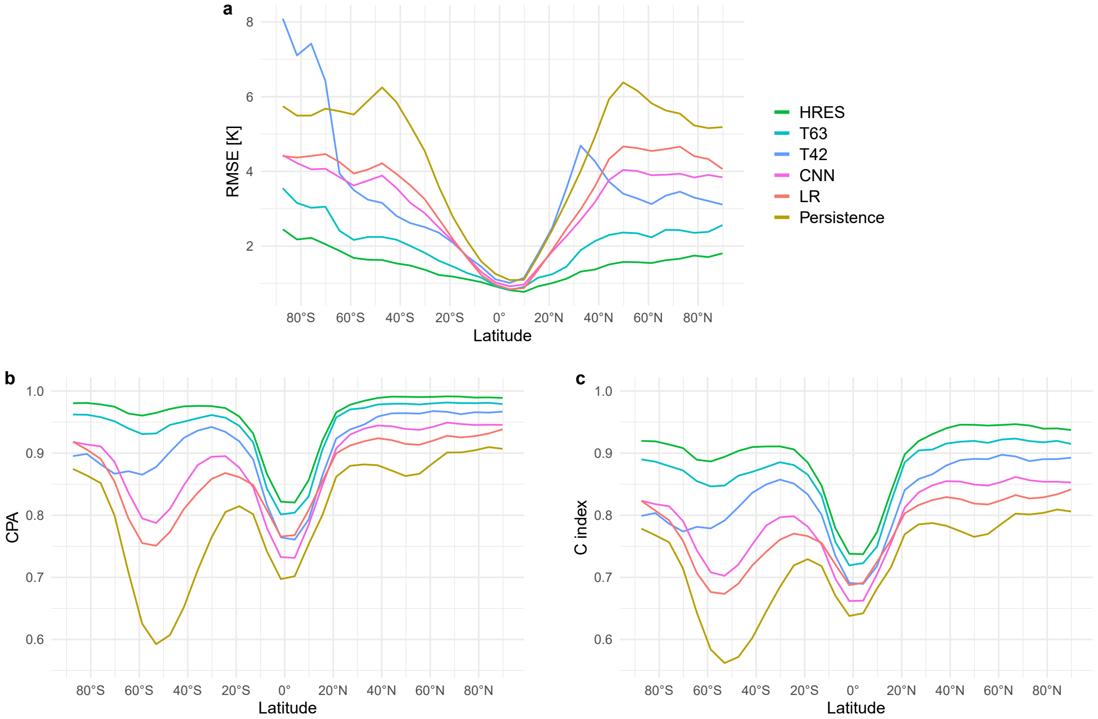

Figure 4.9: Predictive ability of WeatherBench three days ahead forecasts of 850 hPa temperature in 2017 and 2018 at different latitudes in terms of (a) RMSE, (b) CPA, and (c) the C index. HRES, T63, and T42 indicate NWP models run at decreasing grid resolution that are compared to the CNN, linear regression (LR), and persistence forecasts (Rasp et al., 2020). Note that RMSE is negatively oriented (the smaller the better), whereas CPA and the C index are positively oriented.

4.5.3 WeatherBench: Convolutional Neural Networks (CNNs) vs. NWP models

As noted, operational weather forecasts are based on the output of global NWP models that represent the physics of the atmosphere. However, the grid resolution of NWP models remains limited due to finite computing resources (Bauer et al., 2015). Spurred by the ever increasing popularity and successes of machine learning models, alternative, data-driven approaches are in vigorous development, with convolutional neural networks (CNNs; LeCun et al., 2015) being a particularly attractive starting point, due to their ease of adaptation to spatio-temporal data. Rasp et al. (2020) introduce WeatherBench, a ready-to use benchmark dataset for the comparison of data-driven approaches, such as CNNs and a classical linear regression (LR) based technique, to NWP models, such as the aforementioned HRES model and simplified versions thereof, T63 and T42, which run at successively coarser resolutions. Furthermore, WeatherBench supplies baseline methods, including both the persistence forecast and climatological forecasts.

As evaluation measure for the various types of point forecasts, WeatherBench uses the root mean squared error (RMSE). In related studies, the RMSE is accompanied by the anomaly correlation coefficient (ACC), i.e., the normalized product moment between the difference of the forecast at hand and the climatological forecast, and the difference between the outcome and the climatological forecast (Weyn et al., 2020). However, as noted by Rasp et al. (2020), results in terms of RMSE and ACC tend to be very similar. Here we argue that a rank based measure, such as CPA or the C index, would be a more suitable companion measure to RMSE than ACC.

Figure 4.9 compares WeatherBench forecasts three days ahead for temperature at 850 hPa pressure, which is at around 1.5 km height, in terms of RMSE (in Kelvin), CPA, and the C index. With reference to Table 2 of Rasp et al. (2020), we consider the persistence forecast, the (direct) linear regression (LR) forecast, the (direct) CNN forecast, the Operational IFS (HRES) forecast, and successively coarser versions thereof (T63 and T42). The panels display the performance measures as functions of latitude bands, from the South Pole at 90°S to the equator at 0° and the North Pole at 90°N, for the WeatherBench final evaluation period of the years 2017 and 2018. The measures are initially computed grid cell by grid cell, and then averaged across the grid cells in a latitude band, which is compatible with the latitude based weighting that is employed in WeatherBench. Note that RMSE is negatively oriented (the smaller, the better), whereas the rank based measures are positively oriented (the closer to the ideal value of 1 the better).

With respect to RMSE (Figure 4.9a) marked geographical differences are visible. In equatorial regions, where day-to-day temperature variations are generally low, all forecasts have a low RMSE and the range between the best-performing HRES forecast and the simplistic persistence forecast is small. The HRES forecast remains best for all latitudes, followed by the T63 forecast. The coarsest dynamical model forecast, T42, shows a further deterioration as expected, but with large outliers in the high latitudes of the southern hemisphere and in the 30s of the northern hemisphere. It is likely that the lack of model orography creates large errors in areas of high terrain such as the Antarctic plateau and the Himalayas. Among the data-driven forecasts, CNN is better than LR for all extratropical latitudes. Finally, persistence performs worst through all latitudes with prominent peaks near 50°S and 50°N. These are the midlatitude storm track regions, where day-to-day changes are large and impede good forecasts based on persistence.

The corresponding results in terms of CPA and the C index (Figure 4.9b–c) resemble each other, but show remarkable differences to the RMSE based analysis. Most notable are their low values in the tropics, which indicate poor performance of all forecasts, well in line with recent findings in meteorology (Kniffka et al., 2020). In contrast, the low RMSE suggests superior performance in this region. The rank based measures are independent of magnitude and thus provide a scale free assessment of predictability. Another striking difference to RMSE is the large drop in the Furious Fifties of the southern hemisphere, creating a large asymmetry with the northern midlatitudes. This area is almost entirely oceanic and characterized by mobile low-pressure systems, the dynamical behaviour of which appears to be difficult to learn under data-driven approaches.

In Figure 4.10 we compare CPA and the C index, both for individual grid cells and for measures that have been averaged over latitude bands. The scatterplots illustrate the findings from Sections 4.4.3 and 4.4.4, in that the value of CPA throughout is larger than for the C index, in remarkably close agreement with the respective theoretical relationship under the assumption of bivariate Gaussianity.

We conclude that RMSE and the rank based measures bring orthogonal facets of predictive performance to researchers' attention, and encourage the usage of of CPA or the C index to supplement RMSE as key performance measures in WeatherBench. While ACC is scale free as well, it is moment based rather than rank based, and thus is more closely aligned with RMSE than a rank based measure. Similar recommendations apply in many practical settings, where predictions of a real-valued outcome are evaluated, and a magnitude dependent measure, such as RMSE, is usefully accompanied by a rank based criterion of predictive performance. In the special case of probabilistic classifiers for binary outcomes, this corresponds to reporting both the Brier mean squared error measure and AUC. See Hernández-Orallo et al. (2012) for a detailed, theoretically oriented comparison of these and other performance measures under binary outcomes.

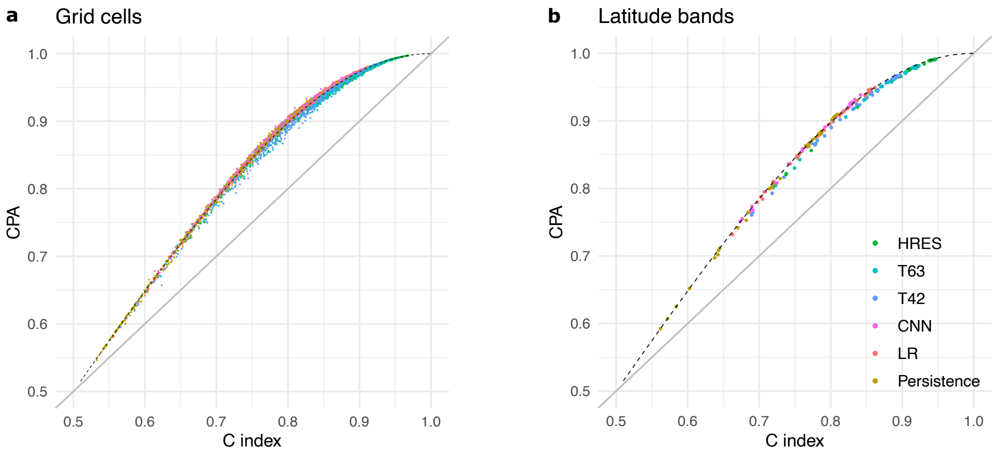

Figure 4.10: Comparison of CPA and the C index for WeatherBench three days ahead forecasts of 850 hPa temperature in 2017 and 2018. The points in the scatterplots of CPA vs. the C index correspond to (a) measures for individual grid cells and (b) averages of measures over latitude bands. The dashed curves show the theoretical relationship between CPA and the C index in bivariate Gaussian populations.

4.6 Discussion

We have addressed a long-standing challenge in data analytics, by introducing a set of tools — comprising receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and a coefficient of predictive ability (CPA) measure — for generalized ROC analysis, thereby freeing researchers from the need to artificially binarize real-valued outcomes, which often is associated with undesirable effects (Altman and Royston, 2006). Throughout this chapter, we have assumed that predictors and features are linearly ordered, thereby covering binary, ordinal, and continuous outcomes simultaneously. While our motivating example uses data from a clinical trial, our approach does not account for censored data, as typically encountered in survival analysis. We strongly encourage extensions of ROC movies, UROC curves and CPA that apply to censored data, perhaps along the lines of Blanche et al. (2013). For generalizations of ROC analysis to multi-class problems with categorical outcomes that cannot be linearly ordered see Hand and Till (2001), Ferri et al. (2003), and Section 9 of Fawcett (2006).

ROC movies, UROC curves, and CPA reduce to the classical ROC curve and AUC when applied to binary data. Moreover, attractive properties of ROC curves, such as invariance under strictly increasing transformations and straightforward interpretability are maintained by ROC movies and UROC curves. In contrast to customarily used measures of bivariate association and dependence (Reshef et al., 2011; Weihs et al., 2018), CPA is asymmetric, i.e., in general, its value changes if the roles of the feature and the outcome are transposed. However, when both the feature and the outcome are continuous, CPA becomes symmetric, and relates linearly to Spearman's rank correlation coefficient. Thus, CPA bridges and generalizes AUC, Somers' D and Spearman's rank correlation coefficient, up to a linear relationship, just like the C index connects and generalizes AUC, Somers' D and Kendall's rank correlation coefficient. While in typical practice the two measures yield qualitatively similar results, under positive dependence CPA is larger than the C index, and CPA tends to be less affected by discretization effects.

In view of the advent of dynamic graphics in mainstream scientific publishing, we contend that ROC movies, UROC curves, and CPA are bound to supersede traditional ROC curves and AUC in a wealth of applications. Open source code for their implementation in Python (Python Software Foundation, 2021) and the R language and environment for statistical computing (R Core Team, 2021) is available on GitHub at `https://github.com/evwalz/urocc` and `https://github.com/evwalz/uroc`.

5 Easy Uncertainty Quantification (EasyUQ)

How can we quantify uncertainty if our favorite computational tool — be it numerical, statistical, or machine learning approach, or just any computer model — provides single-valued output only? This chapter introduces the Easy Uncertainty Quantification (EasyUQ) technique, which transforms real-valued model output into calibrated statistical distributions, based solely on training data of model output–outcome pairs, without any need to access model input. In its basic form, EasyUQ is a special case of the recently introduced Isotonic Distributional Regression (IDR) technique which is described in its general form in Section 3.4. The smooth EasyUQ approach supplements IDR with kernel smoothing, to yield continuous predictive distributions that preserve key properties of the basic form, including stochastic monotonicity with respect to the original model output. In the final chapter of this work, the EasyUQ approach is used to perform the step *choosing and fitting models* of the forecasting cycle introduced in Section 1.

5.1 Introduction

In an editorial that remains topical and relevant (Trefethen, 2012), SIAM President Nick Trefethen noted a decade ago that

> "An answer that used to be a single number may now be a statistical distribution."

Indeed, with the increasing reliance of real-world decisions on the output of computer models – which might be numerical or statistical, parametric or nonparametric, simple or complex – and the advent of uncertainty quantification as a scientific field of its own, there is a growing consensus in the computational sciences community that decisions ought to be informed by full predictive distributions, rather than single-valued model output. For recent perspectives on these issues and uncertainty quantification in general, we refer to topical monographs (Ghanem et al., 2017; Smith, 2014; Sullivan, 2015) and review articles (Abdar et al., 2021; Berger and Smith, 2019; Gneiting and Katzfuss, 2014; Roy and Oberkampf, 2011).

How can we quantify uncertainty if the computational model at hand provides single-valued output only? With Nick Trefethen's comment in mind, we address the following problem: Given single-valued, univariate model output, how can we generate a prediction interval or, more generally, a probabilistic forecast in the form of a full statistical distribution? In this work, we introduce the Easy Uncertainty Quantification (EasyUQ) technique that serves this task, based solely on a training archive of model output–outcome pairs. The single-valued, univariate model output can be of any type — e.g., it might stem from a physics-based numerical model, might arise from a purely statistical or machine learning model, or might be based on human expertise. In a nutshell, EasyUQ applies the recently introduced Isotonic Distributional Regression (IDR, Henzi et al., 2021) approach to generate discrete, calibrated predictive distributions, conditional on the model output at hand. The name stems from the three-fold reasons that EasyUQ operates on the final model output only, without any need for access to the original model input, that the method honors a natural assumption of isotonicity, namely, that higher values of the model output entail predictive distributions that are larger in stochastic order, and that the basic version of EasyUQ does not involve any tuning parameters, and thus does not require user intervention. The more elaborate Smooth EasyUQ approach introduced in this chapter subjects the EasyUQ distribution to kernel smoothing, to yield predictive probability densities that preserve key properties of the basic approach. Prediction intervals are readily extracted; e.g., the equal-tailed 90% interval forecast is framed by the quantiles at level 0.05 and 0.95 of the predictive distribution.

As the EasyUQ approach requires training data, it addresses general "weather-like" tasks (Berger and Smith, 2019, p. 441), which are characterized by frequent repetition of the task — e.g., hourly, daily, monthly, at numerous spatial locations, or for a range of customers or patients — in concert with short to moderate lead times of the forecasts, thus enabling the development of a sizeable archive of forecast–outcome pairs. EasyUQ makes the best possible use of single-valued model output in the sense of empirical score minimization on the training data, subject to the natural constraint of isotonicity. Specifically, the larger the model output, the larger the predictive distribution, in the technical sense of the familiar stochastic order (Shaked and Shanthikumar, 2007), i.e., the respective cumulative distribution functions (CDFs) do not intersect and their graphs move to the right as the model output increases. Subject to the isotonicity constraint, the EasyUQ distributions are optimal with respect to a large class of loss functions that includes the popular continuous ranked probability score (CRPS, Gneiting and Raftery, 2007; Matheson and Winkler, 1976), all proper scoring rules for binary events, and all proper scoring rules for quantile forecasts, among others (Henzi et al., 2021, Thm. 2). For prediction, the EasyUQ and Smooth EasyUQ distributions are interpolated to the value of the model output at hand, while respecting isotonicity.

Figure 5.1 illustrates the EasyUQ approach on WeatherBench (Rasp et al., 2020), a benchmark dataset for weather prediction that serves as a running example in this chapter and is also used in Section 4.5.3 of Chapter 4. Panel a) shows single-valued forecasts of upper air temperature from the HRES numerical weather prediction model run by the European Centre for Medium-Range Weather Forecasts (ECMWF, Molteni et al. (1996)) along with the associated observed temperatures in February 2017. The training data for EasyUQ, which converts the single-valued HRES model output into conditional predictive distributions, comprise the forecast–outcome pairs from 2010 through 2016, as illustrated in the scatter plot in panel c). Panel d) shows the EasyUQ predictive distributions for February 2017, which derive from the single-valued HRES forecasts in panel a), and can be compared to the computationally much more expensive ECMWF ensemble forecasts in panel b). To facilitate the comparison, panel c) includes inset diagrams with the ECMWF ensemble and EasyUQ predictive CDFs for two particular days. Panels e) and f) show EasyUQ predictive CDFs and Smooth EasyUQ predictive densities when the HRES model output equals 263, 268, and 273 degrees Kelvin, respectively. The isotonicity property of the EasyUQ distributions is reflected by the non-intersecting CDFs. The boxes in panels b) and d) range from the 25th to the 75th percentile of the distribution and generate 50% prediction intervals, whereas the whiskers range from the 5th to the 95th percentile and form 90% intervals.

The remainder of the chapter is organized as follows. Section 5.2 provides comprehensive descriptions of IDR and the basic EasyUQ method, and gives details, background information, and a comparison to conformal prediction (Vovk et al., 2022, 2020b) for both the WeatherBench temperature forecast challenge and a precipitation forecast example. In Section 5.3, we introduce the Smooth EasyUQ technique and show that it retains the isotonocity property of the basic method.For the selection of kernel parameters, we introduce multiple one-fit grid search, a computationally much less demanding approximate version of cross-validation. In Section 5.4, we demonstrate that

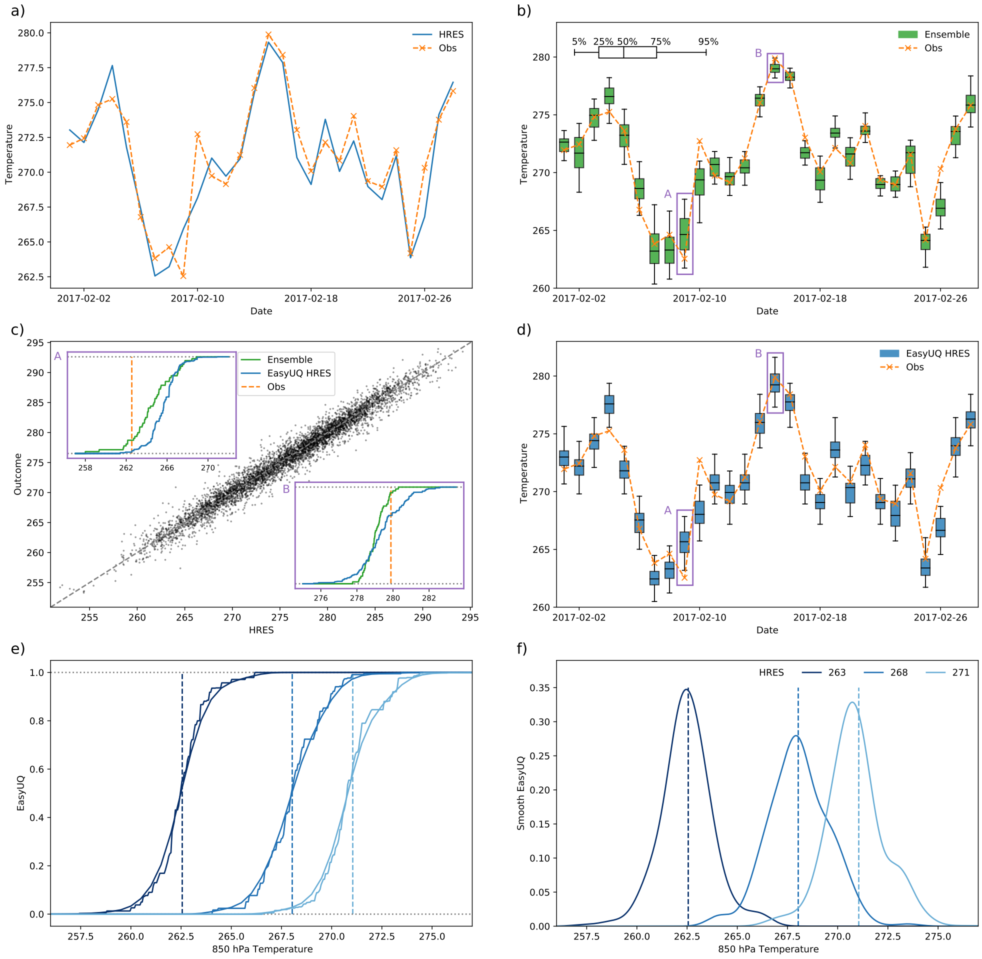

Figure 5.1: EasyUQ illustrated on WeatherBench data. Time series of three days ahead a) single-valued HRES model forecasts, b) state of the art ECMWF ensemble forecasts, and d) basic EasyUQ predictive distributions based on the single-valued HRES forecast along with associated outcomes of upper air temperature in February 2017 at a grid point over Poland, in degrees Kelvin. The boxplots show the quantiles at levels 0.05, 0.25, 0.50, 0.75, and 0.95 of the predictive distributions. c) Scatterplot of HRES model output and associated outcomes in 2010 through 2016, which serve as training data. The inset diagrams show the ECMWF and EasyUQ predictive CDFs for (A) 9 February 2017 and (B) 15 February 2017, respectively. e) Basic and Smooth EasyUQ predictive CDFs and f) Smooth EasyUQ predictive densities at selected values of the single-valued HRES forecast. For further details see Section 5.2.2.

EasyUQ can be integrated into the workflow of neural network learning and hyperparameter tuning, and we use benchmark problems to compare its predictive performance to state-of-the-art techniques from machine learning and conformal prediction. The chapter closes with remarks in Section 5.5, where we return to the discussion of input-based vs. output-based uncertainty quantification.

While the basic version of EasyUQ arises as a special case of the extant IDR technique (Henzi et al., 2021), we take the particular perspective of the conversion of single-valued model output into predictive distributions. Original contributions in this chapter include the development of the Smooth EasyUQ method (Sections 5.3.1 and 5.3.2), a detailed comparison to conformal prediction in case studies (Sections 5.2.2, 5.2.3, and 5.3.3) and from computational and methodological perspectives (Sections 5.3.4 and 5.5), and the integration and benchmarking of EasyUQ and Smooth EasyUQ for neural networks (Section 5.4).

5.2 Basic EasyUQ

We begin the section with a prelude on the evaluation of predictions in the form of full statistical distributions. Then we describe the IDR and EasyUQ techniques, and we illustrate EasyUQ on the WeatherBench data from Rasp et al. (2020) and on precipitation forecasts from Henzi et al. (2021). Generally, EasyUQ depends on the availability of training data of the form

$$(x_i, y_i), \quad i = 1, \ldots, n, \tag{5.1}$$

where $x_i \in \mathbb{R}$ is the single-valued model output and $y_i \in \mathbb{R}$ is the respective real-world outcome, for $i = 1, \ldots, n$. For subsequent discussion, we note the contrast to more elaborate, input-based ways of uncertainty quantification that require access to the features or covariates from which the model output x_i is generated. In the WeatherBench example from Figure 5.1, we have training data comprising twice daily HRES forecasts and the associated observed temperatures in 2010 through 2016 as illustrated in panel c), where $n = 5,114$, but we do not have access to the excessively high-dimensional input to the HRES model. In practice, one needs to find a predictive distribution given the value x of the model output at hand, which may or may not be among the training values $x_1 \leq \cdots \leq x_n$, and some form of interpolation is needed, while retaining isotonicity. In panel e) of Figure 5.1 we illustrate predictive CDFs when x equals 263, 268, and 273 degrees Kelvin, respectively.

Extensions of this setting to situations where single-valued output from multiple computational models is available can be handled within the IDR framework, as we discuss below. If model output and real-world outcome are vector-valued — e.g., when temperature is predicted at multiple sites simultaneously — EasyUQ can be applied to each component independently, and the EasyUQ distributions for the components can be merged by exploiting dependence structures in the training data, based on empirical copula techniques such as the Schaake shuffle (Schefzik et al., 2013).

5.2.1 Basic EasyUQ: Leveraging the Isotonic Distributional Regression (IDR) technique

In this section, it will be instructive to think of the quantities involved as random variables, which we emphasize by using the upper case in the notation. If model output X serves to predict a future quantity Y, then one typically assumes that Y tends to attain higher values as X increases; in fact, the isotonicity assumption can be regarded as a natural requirement for X to be a useful forecast for Y. Isotonic Distributional Regression (IDR), described in Section 3.4, is a recently introduced, nonparametric method for estimating the conditional distributions of a real-valued outcome Y given a covariate or feature vector X from a partially ordered space under general assumptions of isotonicity (Henzi et al., 2021). EasyUQ leverages the basic special case of IDR where X is the single-valued model output at hand. In this chapter, we review the construction and the most relevant properties of IDR for uncertainty quantification; a more general description is provided in Section 3.4 and for detailed formulations and proofs we refer the reader to Henzi et al. (2021).

Formally, EasyUQ assumes that the conditional distributions of the outcome Y given the model output X, which we identify with the CDFs $F_x(y) = \mathbb{P}(Y \leq y \mid X = x)$, are increasing in stochastic order (Shaked and Shanthikumar, 2007) in x, i.e., $F_x(y) \geq F_{x'}(y)$ for all $y \in \mathbb{R}$ if $x \leq x'$, or equivalently $q_x(\alpha) \leq q_{x'}(\alpha)$ for all $\alpha \in (0,1)$, where $q_x(\alpha) = F_x^{-1}(\alpha)$ is the conditional lower α-quantile. In plain words, the probability of the outcome Y exceeding any threshold y increases with the model output x. Isotonicity in this sense is a natural assumption that one expects to hold, to a reasonable degree of approximation, in many types of applications. An important exception arises for location-scale families. Specifically, the arguments in the proof of Proposition 1 in Gneiting and Vogel (2022) imply that isotonicity is violated when the true predictive distributions come from a location-scale family with varying scale.[1] However, the practical impact of this result is limited, due to the fact that in typical practice the scale parameter varies only mildly (Gneiting et al., 2005) and violations remain minor. Crucially, estimators that enforce isotonicity tend to be superior to estimators that do not, even when the key assumption is violated, provided the deviation from isotonicity remains modest. For an illustration in a simulation setting see the non-isotonic scenario (25) in Table 1 of Henzi et al. (2021), where IDR retains acceptable performance relative to its competitors, despite the key assumption being violated. For a rigorous result, Thm. 7 of El Barmi and Mukerjee (2005) demonstrates that, in the special case of discrete model output, EasyUQ has smaller large sample estimation error than non-isotonic alternatives even under mild violations of the isotonicity assumption.

EasyUQ assumes isotonicity with respect to the usual stochastic order. In situations where this assumption is severely violated it may be worthwhile to consider isotonicity with respect to a weaker requirement for distributions to be ordered. An analogous method to IDR under increasing concave and convex stochastic ordering constraints has been introduced by Henzi (2023). An extension of EasyUQ in this direction is left for future work.

To estimate conditional CDFs under the given stochastic order constraints from training data of the form (5.1), we define

$$(\hat{F}_{x_1}(y), \dots, \hat{F}_{x_n}(y))' = \arg \min_{\theta \in \mathbb{R}^n \,:\, \theta_i \geq \theta_j \text{ if } x_i \leq x_j} \sum_{i=1}^{n} (\theta_i - \mathbb{1}\{y_i \leq y\})^2 \tag{5.2}$$

at $y \in \mathbb{R}$. If $x_1 < \cdots < x_n$, then by classical results about isotonic regression,

$$\hat{F}_{x_j}(y) = \min_{k=1,\dots,j} \max_{l=j,\dots,n} \frac{1}{l-k+1} \sum_{i=k}^{l} \mathbb{1}\{y_i \leq y\}, \quad j = 1, \dots, n. \tag{5.3}$$

At any single threshold y, the computation can be performed efficiently in $\mathcal{O}(n \log(n))$ complexity with the well-known pool-adjacent-violators (PAV) algorithm. Since the loss function in (5.2) is constant for y in between the unique values $\tilde{y}_1 < \cdots < \tilde{y}_k$ of $y_1, \dots, y_n$, it suffices to compute (5.3) at the unique values, for which efficient recursive algorithms are available (Henzi et al., 2022). An estimate $\hat{F}_x$ for the conditional CDF at model output $x \in (x_i, x_{i+1})$ is obtained by pointwise linear interpolation in x.

[1] For example, if $F_1 = \mathcal{L}(Y|X = x_1) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $F_2 = \mathcal{L}(Y|X = x_2) = \mathcal{N}(\mu_2, \sigma_2^2)$, where $x_1 \neq x_2$ and $\sigma_1 \neq \sigma_2$, then F_1 and F_2 are incomparable in stochastic order, whence isotonicity is violated. However, if σ_1 and σ_2 are close to each other, the CDFs of F_1 and F_2 cross in the far (left or right) tail only (Gneiting and Vogel (2022), proof of Proposition 1), so violations remain minor.

For $x \leq x_1$ and $x \geq x_n$, we use $\hat{F}_{x_1}$ and $\hat{F}_{x_n}$, respectively. The EasyUQ conditional CDFs are step functions that correspond to discrete predictive distributions with mass at (a subset of) the unique values $\tilde{y}_1 < \cdots < \tilde{y}_k$ only.

The IDR approach has desirable properties that make it suitable for uncertainty quantification. By (5.2), the EasyUQ CDFs depend on the order of $x_1, \ldots, x_n$, but not on their values, and hence the solution is invariant under strictly monotone transformations of the model output, except for interpolation choices when $x \notin \{x_1, \ldots, x_n\}$. Furthermore, the EasyUQ distributions are in-sample calibrated (Henzi et al., 2021, Thm. 2). Importantly, a comparison of the loss function in (5.2) and the definition of the CRPS in (2.8) reveals that EasyUQ minimizes the CRPS over all conditional distributions satisfying the stochastic order constraints. Furthermore, the EasyUQ solution is universal, in the sense that it is simultaneously in-sample optimal with respect to comprehensive classes of proper scoring rules in terms of conditional CDFs or conditional quantiles, such as, e.g., weighted forms of the CRPS with the Lebesgue measure in (2.8) replaced by a general measure (Henzi et al., 2021, Thm. 2). Other approaches to estimating conditional CDFs, e.g., based on parametric models, nearest neighbors, or kernel regression, do not share the universality property, and estimates change depending on the loss function at hand.

In Figure 5.1 we illustrate EasyUQ predictive CDFs in the empirical WeatherBench example. Simulation examples, to which we turn now, have the advantage that the true conditional CDFs are available, so we can compare with them. Figure 5.2 illustrates the construction of the discrete EasyUQ predictive distributions step by step, based on a training archive of the form (5.1) with $n = 500$ simulated from a bivariate distribution, where the model output X is uniform on $(0, 10)$ and the outcome Y satisfies

$$Y \mid X \sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{scale} = \min\{\max\{X, 2\}, 8\}). \tag{5.4}$$

EasyUQ converts the single-valued model output X into conditional predictive CDFs close to the right-skewed true ones. Indeed, IDR, and hence, EasyUQ are asymptotically consistent: As the training archive size n grows, the estimated EasyUQ CDFs converge to the true conditional CDFs (El Barmi and Mukerjee, 2005; Henzi et al., 2021; Mösching and Dümbgen, 2020). Of particular relevance to EasyUQ is the following recent result (Henzi et al., 2023, Thm. 5.1): If $x_1, \ldots, x_n$ themselves are not fixed but are predictions from a statistical model that is estimated on the same training data then IDR is a consistent estimator of the true conditional distributions, subject to mild regularity conditions.

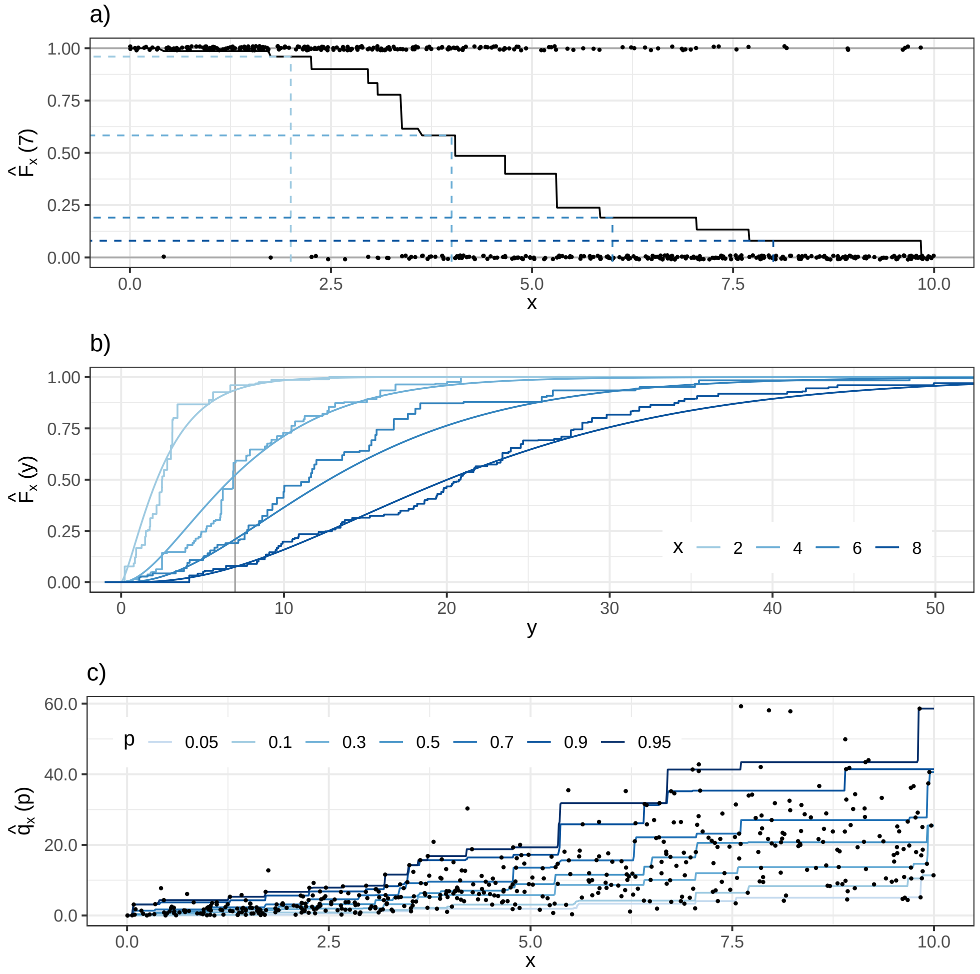

Figure 5.2: Computation of EasyUQ predictive distributions from a training archive of n = 500 model output–outcome pairs simulated according to (5.4). a) The minimizer $\hat{F}_x(y)$ of (5.3) at y = 7, interpolated linearly in x. The jiggled dots show the indicators $\mathbb{1}\{y_i \leq y\}$. b) EasyUQ conditional CDFs $\hat{F}_x$ (step functions) and the respective true conditional CDFs (smooth curves) at selected values of x. The vertical line at y = 7 highlights the values marked in the top panel. c) Training data (x_i, y_i) for i = 1, …, n, and conditional quantile curves $\hat{q}_x(p)$ resulting from inversion of the EasyUQ CDFs $\hat{F}_x$. The lowest and highest quantile curves (levels 0.05 and 0.95) together delineate equal-tailed 90% prediction intervals.

The basic EasyUQ method extends readily to vector-valued model output. If $x_1, \ldots, x_n$ are vectors in a space with a partial order $\preceq$, then the same approach (5.2) applies with the usual inequality $\leq$ replaced by the partial order $\preceq$. This allows more flexibility in the sense that distributions F_x and $F_{x'}$ are allowed to be incomparable in stochastic order if x and x' are incomparable in the partial order. A prominent example concerns ensemble weather forecasts (Gneiting and Raftery, 2005; Leutbecher and Palmer, 2008; Palmer, 2000), where a numerical model is run several times under distinct conditions, and the partial order $\preceq$ that underlies IDR can be tailored to this setting (Henzi et al., 2021).

To summarize, the basic EasyUQ method provides a data driven, theoretically principled, and fully automated approach to uncertainty quantification that is devoid of any need for implementation choices. Based on training data, EasyUQ converts single-valued model output into calibrated predictive distributions that reflect the uncertainty in the model output and training data, as opposed to tuning intense methods, where uncertainty quantification might reflect implementation decisions and user choices. The EasyUQ predictive solution is invariant under strictly monotone transformations of the model output, it is in-sample calibrated, it is in-sample optimal with respect to comprehensive classes of loss functions, and subject to mild conditions it is asymptotically consistent for both output from deterministic models and output from statistical or machine learning models, even when the model is learned on the same data.[2]

5.2.2 Illustration on WeatherBench challenge

In a notable development, WeatherBench (Rasp et al., 2020) introduces a benchmark dataset (which is also used in Section 4.5.3 of Chapter 4) for the comparison of purely data driven and numerical weather prediction (NWP) model based approaches to weather forecasting. Following up on the illustration in Figure 5.1, where we consider a grid point at (latitude, longitude) values of (53.4375, 16.875), we now provide background information and quantitative results at grid points worldwide.

Our experiments are based on the setup in WeatherBench and consider forecasts of

[2] By Thm. 2 of Henzi et al. (2021), the fitted EasyUQ distributions are threshold calibrated, i.e., the predicted non-exceedance probabilities equal their empirical counterparts in the training data. Furthermore, the fitted distributions are empirical score minimizers under a large class of proper scoring rules.

upper air temperature at a vertical level of 850 hPa pressure. The forecasts are issued twice daily at 00 and 12 Coordinated Universal Time (UTC) at lead times of three and five days ahead. The single-valued HRES forecast is from the high-resolution model operated by the European Centre for Medium-Range Weather Forecasts (ECMWF), which represents the physics and chemistry of the atmosphere and is generally considered to be the leading global NWP model. To reduce the amount of data, WeatherBench regrids the HRES model output and the respective outcomes, which originally are on a 0.25 degree latitude–longitude grid (72×144), to coarser resolution (32×64) via bilinear interpolation. The CNN forecast is also single-valued; it is purely data driven and based on a Convolutional Neural Network (CNN), with trained weights being available in WeatherBench. The single-valued Climatology forecast is the best performing baseline model from WeatherBench; it is obtained as the arithmetic mean of the observed upper air temperature in the training data, stratified by 52 calendar weeks.

Conformal Prediction (CP, Vovk et al., 2022, 2020b) is an increasingly popular, general technique for the construction of predictive distributions from single-valued model output. For a comparison with EasyUQ, we employ CP in the form of the studentized Least Squares Prediction Machine (LSPM, Vovk et al., 2022, Algorithm 7.2) with the single-valued model output as sole covariate. We consider CP to be a key competitor, as it is an output-based method that shares desirable properties of EasyUQ. Specifically, the LSPM supplements a least squares based point prediction of the outcome with a conformal predictive system for uncertainty quantification. Based on training data (x_i, y_i), where $i = 1, \ldots, n-1$, Algorithm 7.2 returns a fuzzy predictive distribution (Vovk et al., 2022, eq. (7.7)) that is defined in terms of quantities $C_1, \ldots, C_{n-1}$. Comparative evaluation requires a crisp predictive distribution, for which we use the empirical distribution of $C_1, \ldots, C_{n-1}$, which adheres to the bounds imposed by the fuzzy distribution.[3] For moderate to large training sets and x the value of the model output at hand, C_i typically is very close to $\hat{y} + y_i - \hat{y}_i$, where $\hat{y}$ and $\hat{y}_i$ are least squares point predictions based on x and x_i, respectively (Vovk et al., 2022, Section 7.3.4).

Finally, we consider the state-of-the-art approach to uncertainty quantification in

[3] Here and in Section 5.3.4, we adopt the convention in Vovk et al. (2022, Section 7.2) and assume that the size of the training set is $n-1$, rather than n, to allow for direct references to material therein. The respective crisp CDF is given by $F(y) = i/n$ for $y \in (C_{(i)}, C_{(i+1)})$ and $i = 0, 1, \ldots, n-1$, and $F(y) = i''/n$ for $y = C_{(i)}$ and $i = 1, \ldots, n-1$, where $C_{(0)} = -\infty$, $C_{(1)} \leq \cdots \leq C_{(n-1)}$ are the order statistics of $C_1, \ldots, C_{n-1}$, $C_{(n)} = \infty$ and $i'' = \max\{ j : C_{(j)} = C_{(i)}\}$. For related discussion and alternative choices of a crisp CDF that is compatible with the fuzzy CDF, see Boström et al. (2021, Section 2) and Vovk et al. (2020a, Section 5).

Table 5.1: Predictive performance in terms of mean CRPS for WeatherBench forecasts of upper air temperature at lead times of three and five days, in degrees Kelvin. The evaluation period comprises calendar years 2017 and 2018. CP and EasyUQ generate predictive CDFs that are fitted at each grid point individually, based on training data from 2010 through 2016. Forecasts are issued twice daily, and scores are averaged over 32 × 64 grid points, for a total of 2,990,080 forecast cases.

Forecast		CRPS	
Type	Method	Three Days	Five Days
Single-valued	Climatology	2.904	2.904
	CNN	2.365	2.782
	HRES	0.998	1.543
Distributional	CP on Climatology	2.055	2.055
	CP on CNN	1.673	1.955
	CP on HRES	0.731	1.123
Distributional	EasyUQ on Climatology	2.038	2.038
	EasyUQ on CNN	1.671	1.949
	EasyUQ on HRES	0.736	1.122
Distributional	ECMWF Ensemble	0.696	0.998

weather prediction, namely, ensemble forecasts (Gneiting and Raftery, 2005; Leutbecher and Palmer, 2008; Palmer, 2000), which are input-based methods. Specifically, we use the world leading ECMWF Integrated Forecast System (IFS, `https://www.ecmwf.int/en/forecasts`), which comprises 51 NWP runs, namely, a control run and 50 perturbed members (Molteni et al., 1996). The control run is based on the best estimate of the initial state of the atmosphere, and the perturbed members start from slightly different states that represent uncertainty. Even a single NWP model run, such as the HRES run, is computationally very expensive, and computing power is the limiting factor to improving model resolution. Despite having coarser resolution, an ensemble typically requires 10 to 15 times more computing power than a single run (Bauer et al., 2015). In contrast, the implementation of the output-based CP and EasyUQ methods is fast, with hardly any resources needed beyond a single NWP model run.

To compare CP and EasyUQ predictive CDFs to the respective single-valued forecasts we use the CRPS from (2.8) and recall that for single-valued forecasts the mean CRPS reduces to the mean absolute error. As evaluation period, we take calendar years 2017 and 2018; for estimating the CP and EasyUQ predictive distributions, we use training data from calendar years 2010 through 2016 and proceed grid point by grid point. The corresponding results are provided in Table 5.1. Not surprisingly, the ECMWF ensemble forecast has the lowest mean CRPS. However, CP and EasyUQ based on the HRES model output result in promising CRPS values, even though the methods require considerably less computing time and resources.

The CP and EasyUQ predictive distributions show nearly identical predictive performance. To understand this behavior, we note that in the case of temperature, Gaussian predictive distributions with fixed variance are typically very adequate (see, e.g., Gneiting et al. (2005, Table 3)). In this light, key requirements of CP in the form of the LSPM (namely, fixed spread and fixed shape of the predictive distributions) and EasyUQ (namely, isotonicity) are reasonably met. While EasyUQ generates predictive distribution that vary in spread and shape, the variations remain modest (Figure 5.1c–f), and the CP distributions, which essentially are translates of each other, are competitive.

The subsequent case study turns to a weather variable that is not covered by the WeatherBench challenge, but which serves to illuminate and highlight differences between the CP and EasyUQ techniques.

5.2.3 Illustration on precipitation forecasts

Precipitation accumulation is generally considered the “most difficult weather variable to forecast” (Ebert-Uphoff and Hilburn, 2023). Indeed, the uncertainty quantification for deterministic forecasts of precipitation is more challenging than for temperature, since precipitation accumulation follows a mixture distribution with a point mass at zero — for no precipitation — and a continuous part on the positive real numbers. Applying CP without corrections is bound to transfer mass to negative values of precipitation accumulation. Taking advantage of knowledge about the outcome distribution, a natural remedy is to censor at zero and use the CDF

$$G(y) = \begin{cases} 0, & y < 0, \\ F(y), & y \geq 0, \end{cases}$$

Table 5.2: Predictive performance in terms of mean CRPS for forecasts of daily precipitation accumulation at Frankfurt airport at lead times from one to five days, in millimeters. CP and EasyUQ generate predictive CDFs based on training data from 2007 through 2014. The evaluation period comprises calendar years 2015 and 2016.

Forecast		CRPS				
Type	Method	1 Day	2 Days	3 Days	4 Days	5 Days
Single-valued	Climatology	2.187	2.187	2.187	2.187	2.187
	HRES	1.125	1.294	1.412	1.478	1.686
Distributional	CP on Climatology	1.382	1.382	1.382	1.382	1.382
	CP on HRES	0.886	0.966	1.063	1.081	1.129
	Censored CP on Climatology	1.324	1.324	1.324	1.324	1.324
	Censored CP on HRES	0.850	0.925	1.031	1.050	1.100
Distributional	EasyUQ on Climatology	1.242	1.242	1.242	1.242	1.242
	EasyUQ on HRES	0.732	0.803	0.876	0.945	1.001
Distributional	ECMWF Ensemble	0.752	0.847	0.856	0.918	0.981

in lieu of F.[4] In contrast, the EasyUQ predictive distributions reflect the nonnegativity of the outcomes in the training data, without any need for adaptation.

We now investigate the performance of CP and EasyUQ within the experimental setup from Henzi et al. (2021), taking forecasts and observations of 24-hour accumulated precipitation from 6 January 2007 through 1 January 2017 at Frankfurt airport, Germany. Just as in the WeatherBench example, we consider a weekly climatology, the HRES forecast, and the 51 member NWP ensemble from ECMWF. The weekly climatology is computed over the period 2007 to 2014, which is the same period that is used for CP and EasyUQ training. The evaluation period comprises calendar years 2015 and 2016. Table 5.2 shows the mean CRPS over the evaluation period for the various types of forecasts at lead times from one to five days. Evidently, the climatological forecasts, along with their scores, do not depend on the lead time. In contrast to the

[4] In our experiments, we train without consideration of censoring, and we censor at zero ex post. For a nonnegative outcome, such a procedure guarantees improvement, in the technical sense that $\mathrm{CRPS(G, y)} \leq \mathrm{CRPS(F, y)}$ for all $y \geq 0$. Alternatively, one might take censoring into account during training. However, methods of this latter type are more complex to implement, and improvements in CRPS cannot be guaranteed out-of-sample.

WeatherBench temperature example, EasyUQ outperforms CP for both Climatology and the HRES model output, and at all lead times. While censoring improves the distributional forecasts from CP, the performance gap to EasyUQ remains pronounced. EasyUQ on the HRES model output even outperforms the raw ECMWF ensemble at lead times of one and two days.[5]

Figure 5.3 provides a graphical comparison of CP on HRES, Censored CP on HRES, EasyUQ on HRES, and ECMWF ensemble forecasts at small ($x = 0.38$), moderate ($x = 3.40$), and large ($x = 11.93$) values of the HRES model output x. We see that the CP predictive distributions are essentially translates of each other, with mass potentially being transferred to negative values of precipitation accumulation, and censoring shifting any such mass to zero. In contrast, the ECMWF ensemble and EasyUQ distributions do not have mass at negative values, and they vary in shape and scale. However, while the ECMWF ensemble tends to show forecast distributions that are too narrow, as is frequently observed in practice (Gneiting and Raftery, 2005) and illustrated by the right-hand example, the EasyUQ distributions, which are based on the single-valued HRES forecast only, show what appears to be adequate spread. Remarkably, and unlike any other method that we are aware of, EasyUQ achieves this desirable performance in its very basic form, without any need for implementation decisions, parameter tuning, or other forms of adaptation and intervention.

[5] This is largely due to the fact that gridded ensemble predictions are compared against station observations. To counter these effects, the ensemble forecast itself can be subjected to statistical postprocessing, i.e., the application of statistical methods to correct for biases and dispersion errors (Gneiting et al., 2005; Raftery et al., 2005). Parametric methods based on distributional regression (Messner et al., 2014; Scheuerer, 2014) model the distribution of precipitation accumulation with censored logistic or censored generalized extreme value distributions. An alternative approach is taken in Bayesian model averaging (Sloughter et al., 2007), which posits separate parametric forms for the probability of zero precipitation and the density at positive amounts. Evidently, discrete-continuous mixture distributions considerably complicate model building and estimation, and great efforts are made to find suitable parametric families for specific weather variables. For a detailed performance comparison on the data on hand see Henzi et al. (2021, Figure 5), whose study also includes versions of IDR with multivariate covariates derived from the full ECMWF ensemble and suitable partial orders on them, an option alluded to at the end of Section 5.2.1. These yield improvements compared to both the raw ensemble forecast and EasyUQ on HRES, at the price of higher conceptual complexity, higher computational costs, and the need for access to the full ensemble, rather than single-valued HRES model output.

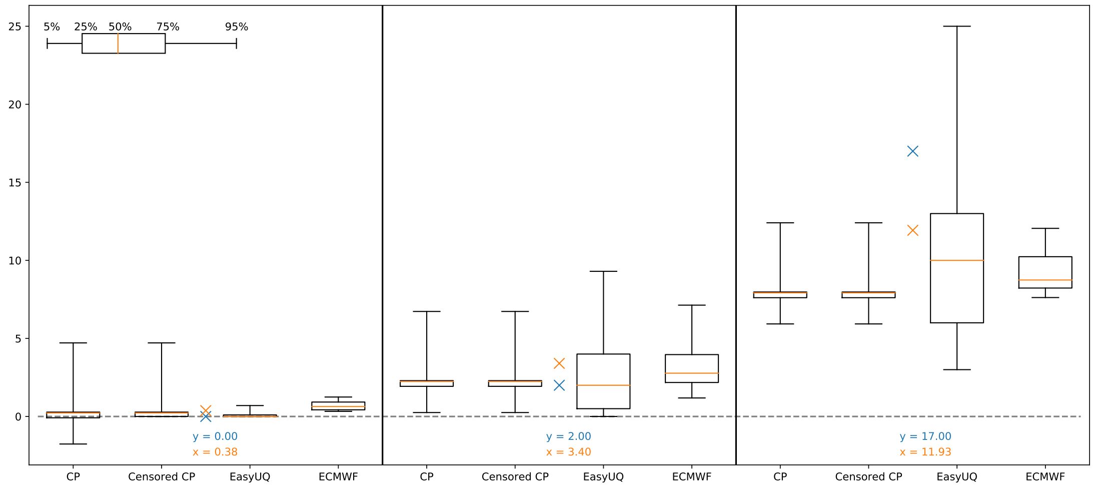

Figure 5.3: One-day ahead forecasts of daily precipitation accumulation at Frankfurt airport valid 23 January 2015 (left, HRES model output x equal to 0.38, as indicated by the blue cross), 14 January 2015 (middle, x = 3.40), and 21 February 2016 (right, x = 11.93), in millimeters. The predictive distributions for CP on HRES, Censored CP on HRES, EasyUQ on HRES, and ECMWF ensemble techniques are shown. The observed precipitation accumulation was at y = 0, y = 2, and y = 17 millimeters, respectively.

5.3 Smooth EasyUQ

EasyUQ provides discrete predictive distributions with positive probability mass at the outcomes from the training archive. For genuinely discrete outcomes, the variable of interest attains a small number of unique values only, which is a desirable property. For genuinely continuous variables, it is preferable to use continuous predictive distributions. We now describe the Smooth EasyUQ technique, which turns the discrete basic EasyUQ CDFs into continuous Smooth EasyUQ CDFs with Lebesgue densities, while preserving isotonicity. To achieve this, Smooth EasyUQ applies kernel smoothing, which requires implementation choices, unlike basic EasyUQ which does not require any tuning. However, we provide default options.

5.3.1 Smooth EasyUQ: Kernel smoothing under isotonicity preservation

Our goal is to transform the discrete basic EasyUQ CDFs $\hat{F}_x$ from (5.3) into smooth predictive CDFs $\check{F}_x$ that admit Lebesgue densities $\check{f}_x$, without abandoning the order relations honored by the basic technique. To this end, we define the Smooth EasyUQ CDF as

$$\check{F}_x(y) = \int_{-\infty}^{\infty} \hat{F}_x(t)\, K_h(y-t)\, \mathrm{d}t, \tag{5.5}$$

where $K_h(u) = (1/h)\,\kappa(u/h)$ for a smooth probability density function or kernel κ, such as a standardized Gaussian or Student-t density, with bandwidth $h > 0$. While the convolution approach in (5.5) is perfectly general for the smoothing of CDFs, we henceforth focus the presentation on EasyUQ. The choice of the kernel and the bandwidth are critical, and we tend to their selection in the next section, where we introduce multiple one-fit grid search as a computationally efficient alternative to cross-validation.

For now, recall that $\hat{F}_x(y)$ from (5.3) is a step function with possible jumps at the unique values $\tilde{y}_1 < \cdots < \tilde{y}_k$ of the outcomes $y_1, \ldots, y_n$ in the training set. Hence, we can write (5.5) as

$$\check{F}_x(y) = \sum_{j=1}^{k} \hat{F}_x(\tilde{y}_j) \int_{\tilde{y}_j}^{\tilde{y}_{j+1}} K_h(y-t)\, \mathrm{d}t,$$

where $\tilde{y}_{k+1} = \infty$. To compute the density $\check{f}_x = \check{F}_x'$, we set $\tilde{y}_0 = -\infty$, note that $\hat{F}_x$ assigns mass $w_j(x) = \hat{F}_x(\tilde{y}_j) - \hat{F}_x(\tilde{y}_{j-1})$ to $\tilde{y}_j$, and find that

$$\check{f}_x(y) = \sum_{j=1}^{k} \hat{F}_x(\tilde{y}_j)\, [K_h(y-\tilde{y}_j) - K_h(y-\tilde{y}_{j+1})] = \sum_{j=1}^{k} w_j(x)\, K_h(y-\tilde{y}_j). \tag{5.6}$$

In other words, the Smooth EasyUQ density $\check{f}_x$ from (5.6) arises as a kernel smoothing of the discrete probability measure that corresponds to $\hat{F}_x$ and assigns weight $w_j(x)$ to $\tilde{y}_j$. Consequently, $\check{f}_x$ is a probability density function, $\check{F}_x$ is a proper CDF, and, notably, Smooth EasyUQ preserves the stochastic ordering of the basic EasyUQ estimates. In Figure 5.4 we illustrate the interpretation of the Smooth EasyUQ density as a kernel smoothing of the EasyUQ point masses $w_j(x)$ on the WeatherBench example.

5.3.2 Choice of kernel and bandwidth: Multiple one-fit grid search

In order to compute the Smooth EasyUQ density $\check{f}_x$ from (5.6), one needs to choose a kernel κ and a bandwidth $h > 0$ to yield a mixture of translates of the density $K_h(u) = (1/h)\,\kappa(u/h)$. While there is a rich literature on bandwidth selection for kernel density estimation and kernel regression (see, e.g., Köhler et al. (2014) and Silverman (1986)), caution is needed when applying established approaches to Smooth EasyUQ, due to the fact that smoothing is applied to estimated conditional CDFs rather than raw data.

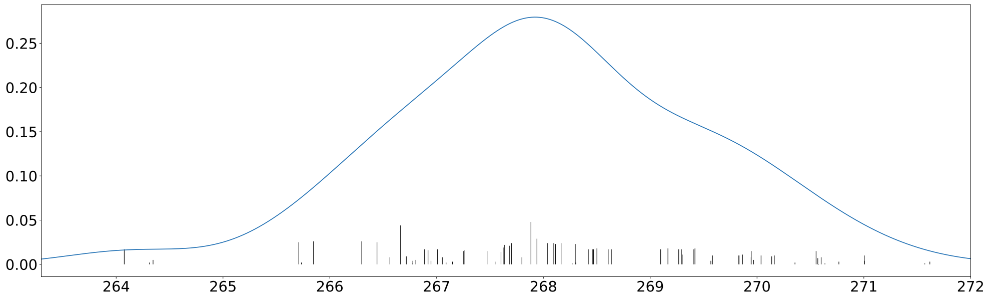

Figure 5.4: Smooth EasyUQ predictive density (5.6) in the WeatherBench example from Figure 5.1f) at HRES model output x equal to 268 degrees Kelvin. The vertical bars show the weights $w_1(x), \ldots, w_k(x)$ that the discrete EasyUQ distribution $\hat{F}_x$ assigns to the unique values $\tilde{y}_1 < \cdots < \tilde{y}_k$ of the outcomes in the training set, where k = 76.

Furthermore, while the extant literature focuses on bandwidth selection for a fixed kernel, approaches of this type are restrictive for our purposes. The Smooth EasyUQ density from (5.6) inherits the tail behavior of the kernel κ, and so the properties of the kernel are of critical importance to the quality of the uncertainty quantification in the tails of the conditional distributions. To allow for distinct tail behavior, we use the Student-t family and set $K_{\nu,h}(u) = (1/h)\,\kappa_\nu(u/h)$, where

$$\kappa_\nu(y) = \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\,\Gamma(\nu/2)} \left(1 + \frac{y^2}{\nu}\right)^{-(\nu+1)/2} \tag{5.7}$$

is a standardized Student-t probability density function with $\nu > 0$ degrees of freedom. It is well known that the Student-t distribution has a finite first moment if $\nu > 1$ and a finite variance if $\nu > 2$. In the limit as $\nu \to \infty$, we find that $\kappa_\nu(y) \to \kappa_\infty(y)$ uniformly in y, where $\kappa_\infty(y) = (2\pi)^{-1/2} \exp(-y^2/2)$ is the standard Gaussian density function, so the ubiquitous Gaussian kernel emerges as a limit case in (5.7).

Turning to the choice of the tail parameter $\nu \in (0, \infty]$ and the bandwidth $h > 0$, we begin by discussing the latter. A popular approach for bandwidth selection, in both kernel regression and kernel density estimation, is leave-one-out cross-validation. Here the target criterion in terms of the bandwidth is

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(\check{F}_{x_i,-i,h}, y_i), \tag{5.8}$$

where S is a proper scoring rule, and $\check{F}_{x_i,-i,h}$ is the Smooth EasyUQ CDF with covariate x_i and bandwidth h, estimated with all data from (5.1) except for the i-th instance. The optimization of the target criterion (5.8) uses either the CRPS as loss function S, as is (implicitly) suggested for the estimation of conditional CDFs and quantile functions (see, e.g., Bowman et al. (1998, p. 801) and Li et al. (2013, p. 58)) and yielding a target that is asymptotically equivalent to the integrated mean squared error (Henzi et al., 2021, Section S4), or the LogS, as is proposed for ensemble smoothing (Bröcker and Smith, 2008). We take the latter as the default choice, since the LogS is much more sensitive to the choice of the bandwidth h than the more robust CRPS.

However, there are a number of caveats. Empirical data are typically discrete to some extent, and might contain ties in the response variable, such as in the setting of Figure 5.4, where there are only $m = 76$ unique values among the outcomes $y_1, \ldots, y_n$, even though $\check{f}_x$ is estimated from a training archive of size $n = 5,114$. In such cases, the optimal cross-validation bandwidth under the LogS may degenerate to $h = 0$, a problem that is also known in density estimation (Silverman, 1986, pp. 51–55), in the estimation of Student-t regression models (Fernandez and Steel, 2009) and, in related form, in performance evaluation for forecast contests (Kohonen and Suomela, 2006; Quiñonero-Candela et al., 2006). Another issue is that leave-one-out cross-validation is computationally expensive, as for each value of h it requires the computation of n distinct IDR solutions. While a potential remedy is to remove a higher percentage of observations in each cross-validation step, we use a considerably faster approach, which we term one-fit grid search, that addresses both issues simultaneously.

One-fit grid search avoids repeated fits of IDR and computes EasyUQ only once, namely, on the full sample from (5.1). Specifically, given any fixed kernel κ, one-fit grid search finds the optimal bandwidth h in terms of the target criterion

$$\mathrm{OF}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(\bar{F}_{x_i,-i,h}, y_i), \tag{5.9}$$

where $\bar{F}_{x_i,-i,h}$ removes the unique value $\tilde{y}_j = y_i$ from the support of $\check{F}_{x_i}$ in (5.5), by setting $w_j(x)$ in (5.6) to zero and rescaling the remaining weights. We choose the LogS as the default option for the loss function S in the one-fit criterion (5.9), and we use Brent's algorithm (Brent, 1973) for optimization. Effectively, one-fit grid search is a fast approximation to cross-validation, and when n is small, leave-one-out cross-validation and the original criterion in (5.8) can be used instead, of course. To choose a Student-t kernel, we repeat the procedure, i.e., we consider values of $\nu \in \{2, 3, 4, 5, 10, 20, \infty\}$ in (5.7), with $\nu = \infty$ yielding the Gaussian limit, apply one-fit grid search for each of these values, to find the respective optimal bandwidth h, and select the combination of ν and h for which the target criterion (5.9) is smallest overall. While being highly effective in our experience, multiple one-fit grid search is a crude approach, and we encourage further development.

5.3.3 Illustration on temperature and precipitation forecast examples

For an initial illustration, we return to the WeatherBench challenge and the EasyUQ densiies in Figs. 5.1f) and 5.4, where $n = 5,114$ and $m = 76$, and multiple one-fit grid search with respect to the LogS yields parameter values $\nu = \infty$ and $h = 0.60$ in the kernel density (5.7). Considering the $32 \times 64 = 2,048$ grid points in WeatherBench and predictions three days ahead, the value of ν selected the most frequently for Smooth EasyUQ on the HRES model output, namely, 619 times, is $\nu = 10$, with a median choice of $h = 0.49$. For Smooth EasyUQ on Climatology and CNN $\nu = \infty$ was most frequently selected, namely $1,391$ and $1,361$ times with median choices of $h = 0.85$ and $h = 1.04$, respectively.

A very simple and frequently used reference method for converting single-valued model output into a predictive density is the Single Gaussian technique (Doubleday et al., 2020). It issues a Gaussian distribution with mean equal to the single-valued model output, and a constant variance that is optimal with respect to the mean LogS on a training set, which here we take to be the same as for EasyUQ. Evidently, both Smooth EasyUQ and the Single Gaussian technique could be trained in terms of the CRPS as well. We also compare to the Smooth CP technique, which converts the discrete CP distributions to densities, as described in the next section.

In Table 5.3, we evaluate Smooth EasyUQ, Smooth CP, and Single Gaussian density temperature forecasts in the WeatherBench setting. For evaluation, we use both the CRPS and the LogS. Throughout, Smooth EasyUQ and Smooth CP outperform the Single Gaussian method, though they do not match the performance of the smoothed ECMWF ensemble forecast, which we construct as follows. Let $\tilde{z}_1 < \cdots < \tilde{z}_k$ be the unique values of the ensemble members $z_1, \ldots, z_l$ of an ensemble forecast of size l. The smoothed ensemble CDF is then of the form (5.5) with mass $w_j = \frac{1}{l} \sum_{i=1}^{l} \mathbb{1}(z_i = \tilde{z}_j)$ for $j = 1, \ldots, k$. Interestingly, this is the same as Bröcker and Smith (2008, realtions (19)–(21)) smoothing of ensemble forecasts, with parameters $a = 1$ and $r_1 = r_2 = s_2 = 0$ being fixed. However, while Bröcker and Smith (2008) use a Gaussian kernel and optimize the bandwidth parameter only, we take a more flexible approach and consider values of $\nu \in \{2, 3, 4, 5, 10, 20, \infty\}$ for a Student-t kernel, to find the optimal ν and bandwidth h in terms of the LogS. Across the $2,048$ grid points,

Table 5.3: Predictive performance in terms of mean LogS and mean CRPS for WeatherBench density forecasts of upper air temperature at lead times of three and five days, in degrees Kelvin. The evaluation period comprises calendar years 2017 and 2018. The Single Gaussian, Smooth CP, and Smooth EasyUQ methods are trained at each grid point individually, based on data from 2010 through 2016. Forecasts are issued twice daily, and scores are averaged over 32×64 grid points, for a total of 2,990,080 forecast cases.

Density Forecast	LogS		CRPS	
Days Ahead	Three	Five	Three	Five
Single Gaussian on Climatology	2.578	2.578	2.060	2.060
Single Gaussian on CNN	2.413	2.553	1.696	1.983
Single Gaussian on HRES	1.694	2.073	0.748	1.153
Smooth CP on Climatology	2.562	2.562	2.059	2.059
Smooth CP on CNN	2.384	2.519	1.672	1.952
Smooth CP on HRES	1.627	2.007	0.732	1.123
Smooth EasyUQ on Climatology	2.540	2.540	2.043	2.043
Smooth EasyUQ on CNN	2.375	2.509	1.667	1.945
Smooth EasyUQ on HRES	1.640	2.006	0.736	1.122
Smoothed ECMWF Ensemble	1.503	1.824	0.685	0.990

the most frequent choice is $\nu = 5$, namely, 743 times, with a median bandwidth value of $h = 0.50$.

While smoothing is warranted for temperature forecasts, it is problematic for forecasts of precipitation accumulation, due to the nonnegativity of the outcome and the point mass at zero. Indeed, due to the kernel smoothing, the Smooth EasyUQ and smoothed ECMWF ensemble densities have mass on the negative halfaxis, unlike the discrete (basic) EasyUQ and (raw) ECMWF distributions, which are concentrated on the nonnegative halfaxis. Nonetheless, Table 5.4 compares the predictive performance of Single Gaussian, Smooth CP, Smooth EasyUQ, and smoothed ECMWF ensemble forecasts in the setting of Section 5.2.3, in both the original and the censored variants. The results mirror the findings in Table 5.2, in that censoring yields improvement and EasyUQ outperforms CP, whereas CP outperforms the Single Gaussian technique.

Table 5.4: Predictive performance in terms of mean CRPS for density forecasts of daily precipitation accumulation at Frankfurt airport at lead times from one to five days, in millimeters. CP and EasyUQ generate predictive CDFs based on training data from 2007 through 2014. The evaluation period comprises calendar years 2015 and 2016.

Density Forecast	1 Day	2 Days	3 Days	4 Days	5 Days
Single Gaussian on HRES	1.244	1.380	1.547	1.577	1.724
Censored Single Gaussian on HRES	1.013	1.145	1.266	1.276	1.401
Smooth CP on HRES	0.886	0.971	1.064	1.087	1.132
Censored Smooth CP on HRES	0.849	0.928	1.028	1.052	1.098
Smooth EasyUQ on HRES	0.760	0.828	0.901	0.968	1.033
Censored Smooth EasyUQ on HRES	0.745	0.817	0.893	0.960	1.016
Smoothed ECMWF Ensemble	0.762	0.855	0.863	0.924	0.986
Censored Smoothed ECMWF Ensemble	0.750	0.850	0.860	0.921	0.984

5.3.4 Computational considerations

We add a brief discussion of the computational complexity of output-based methods for uncertainty quantification. For this comparison, we utilize the setting of Vovk et al. (2022, Algorithm 7.2), which requires predictive distributions for m new values of x, based on a training set of size $n-1$ with instances $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. We report upper estimates of the computational complexity for the Single Gaussian technique, CP, and EasyUQ, considering both training (i.e., initial operations on the training data only) and inference (i.e., operations to be repeated for each new value). For the simplistic Single Gaussian technique, training requires $\mathcal{O}(n)$ operations and inference is straightforward.

For EasyUQ, the main effort lies in training, where the complexity is upper bounded by $\mathcal{O}(n^2)$ operations (Henzi et al., 2022). Training the EasyUQ CDFs only on a fixed grid of ordinates guarantees a cost reduction to $\mathcal{O}(n \log n)$ operations, and Henzi et al. (2021) describe approaches based on subset aggregation that reduce the computational burden for estimation. That said, the numerical experiments in our chapter use the standard implementation throughout, without exception. For inference, each new value of x requires the determination of its position within the unique values across $x_1, \ldots, x_{n-1}$, followed by interpolation of the trained EasyUQ CDFs at the predecessor and successor values, at up to $\mathcal{O}(mn)$ operations.

For CP in the form of the studentized LSPM (Vovk et al., 2022, Algorithm 7.2) essentially no training is required, but inference incurs $\mathcal{O}(mn^2)$ operations. Residual-based approximations to CP, which are instances of split conformal predictive systems (Vovk et al., 2022, Section 7.3.4; Vovk et al., 2018), are much faster, shift the bulk of the cost to training at $\mathcal{O}(n)$ operations, and yield nearly identical predictive performance to CP in our experience, except when training sets are small.

For both, CP and EasyUQ, we have implemented smoothing in ways that avoid cross-validation and honor the aforementioned bounds. Smooth EasyUQ uses one-fit grid search as developed in this chapter. To generate the Smooth CP densities, we use kernel smoothing with a Gaussian kernel and bandwidth chosen according to Silverman's rule of thumb (Silverman, 1986), applied to the quantities $C_1, \ldots, C_{n-1}$ that arise for each new instance separately.

In the aforementioned experiments, we generally found the computational cost of EasyUQ to be nested in between the costs of CP and residual-based approximations to CP.[6] Compared to the enormous effort of running the HRES model, or even the input-based ECMWF ensemble method, which require the operational use of supercomputers, run times and computational costs for the output-based Single Gaussian, CP, and EasyUQ techniques are negligible.

[6] To provide intuition into computation times, we report mean run times for the Single Gaussian technique, CP, and EasyUQ applied to the HRES forecast in the setting of Table 5.2, where the training set is of size 2,896 and the evaluation set of size 721. The mean run time averaged over the five lead times is 0.005 seconds for the Single Gaussian technique, 0.45 seconds for CP, and 0.085 seconds for EasyUQ. We note that the computing time for CP on a CPU is 33.64 seconds, but can be reduced to 0.45 seconds on a GPU. Evidently, the comparison faces the usual challenges, given that execution times depend on factors including but not limited to hardware architecture, disk speed, memory availability, and the programming language and compiler used. Specific to the situation at hand, we use code in Python, R, and C++, run some functions on a GPU and others on a CPU, and it is unlikely that every one of our implementations, which typically are based on packages, has been coded in the most efficient way.

5.4 EasyUQ and neural networks

Neural networks and deep learning techniques have enabled unprecedented progress in predictive science. However, as they “can struggle to produce accurate uncertainties estimates [. . .] there is active research directed toward this end” (Baker et al., 2022, p. 67), which has intensified in recent years (Abdar et al., 2021; Chung et al., 2021; Duan et al., 2020; Gal and Ghahramani, 2016; Immer et al., 2021; Kuleshov et al., 2018; Lakshminarayanan et al., 2017; Marx et al., 2022; Vovk et al., 2020b). We now discuss how EasyUQ and Smooth EasyUQ can be used to yield accurate uncertainty statements from neural networks. Evidently, our methods apply in the ways described thus far, where single-valued model output is treated as given and fixed, with subsequent uncertainty quantification via EasyUQ or Smooth EasyUQ being a completely separate add-on, as illustrated using our temperature and precipitation examples. In the context of neural networks, this means that the network parameters are optimized to yield single-valued output, and only then is EasyUQ applied. We now describe a more elaborate approach where we integrate our methods within the typical workflow of neural network training and evaluation.

5.4.1 Integrating EasyUQ into the workflow of neural network learning and hyperparameter optimization

Neural networks and associated methods for uncertainty quantification are developed and evaluated in well-designed workflows that involve multiple splits of the available data into training, validation, and test sets. For each split, the training set is used to learn basic neural network parameters, the validation set is used to tune hyperparameters, and the test set is used for out-of-sample evaluation. Scores are then averaged over the tests sets across the splits, and methods with low mean score are preferred.

Algorithm 1 describes how Smooth EasyUQ can be implemented within this typical workflow of neural network learning and hyperparameter tuning. In a nutshell, we treat the kernel parameters for Smooth EasyUQ, namely, the Student-t parameter ν and the bandwidth h, as supplemental hyperparameters, and we optimize over both the neural network hyperparameters and the kernel parameters. As the evaluation occurs out-of-sample, the issues associated with the choice of the kernel parameters discussed in Section 5.3.2 are mitigated, unless a dataset is genuinely discrete, in which case even out-of-sample estimates of the bandwidth h can degenerate to zero, thereby

Algorithm 1 Integration of Smooth EasyUQ into the workflow of neural network training and hyperparameter tuning. The procedure returns the mean score of the Smooth EasyUQ predictions across data splits.

```
for split in mysplit do
    separate data into training set, validation set, and test set
    for hyperpar in myhyperpar do
        learn neural network with hyperpar on training set
        use neural network output to fit basic EasyUQ on training set
        use moderated grid search to select EasyUQ parameters $\nu$ and $h$
        save selected $(\nu, h)$ and mean score on validation set
    end for
    select best hyperpar and associated $(\nu, h)$, based on smallest mean score
    re-learn network with best hyperpar on combined training and validation sets
    use re-learned neural network output to re-fit basic EasyUQ on combined training and validation sets
    use Smooth EasyUQ based on re-fitted EasyUQ with best $(\nu, h)$ for predictions on test set
    save scores on test set
end for
return mean score across splits
```

indicating that smoothing is ill-advised. To handle even such ill-advised cases, we use a procedure that we call moderated grid search (Walz, 2023). Specifically, we first check whether using $\nu = 2$ or a Gaussian kernel results in a degeneration of the optimal bandwidth h to zero, and if so, we use the latter with bandwidth chosen according to Silverman's rule of thumb (Silverman, 1986). Otherwise, we consider values of $\nu \in \{2, 3, 4, 5, 10, 20, \infty\}$ in (5.7), with $\nu = \infty$ yielding the Gaussian limit. For each value of ν, we use Brent's method (Brent, 1973) to optimize the log score with respect to the bandwidth h on the validation set, and choose the optimal combination of ν and h. Once network hyperparameters and kernel parameters have been determined, we re-learn the neural network on the combined training and validation sets, using the optimized hyperparameters, and apply EasyUQ on the re-learned single-valued neural network output. Finally, we apply Smooth EasyUQ based on the re-learned EasyUQ solution and the selected kernel parameters, to yield density forecasts on the test set.

While optimization could be performed with respect to the CRPS, the LogS, or any other suitable proper scoring rule, we follow the machine learning literature, where benchmarking is typically done in terms of the LogS. The CRPS serves as an attractive alternative, much in line with recent developments in neural network training, where optimization is performed with respect to the CRPS (D'Isanto and Polsterer, 2018; Rasp and Lerch, 2018). Its use becomes essential in simplified versions of Algorithm 1 that work with the discrete basic EasyUQ distributions rather than Smooth EasyUQ densities.

5.4.2 Application in benchmark settings from machine learning

As noted, our intent is to compare Smooth EasyUQ in the integrated version of Algorithm 1 to extant, state of the art methods for uncertainty quantification from the statistical and machine learning literatures. The comparison is made on ten datasets for regression tasks using the experimental setup proposed and developed by Hernandéz-Lobato and Adams (2015), Gal and Ghahramani (2016), Lakshminarayanan et al. (2017), and Duan et al. (2020). Characteristics of the ten datasets are summarized in Table 5.5, including the size of the datasets, the number of unique outcomes, and the dimension of the input space for the regression problem.

Each dataset is randomly split 20 times into training (72%), validation (18%), and test (10%) sets. However, for the larger datasets, Protein and Year, the train-test split is repeated only five and a single time(s), respectively. After finding the optimal set of (hyper)parameters, methods are re-trained on the combined training and validation sets (90%) and the resulting predictions are evaluated on the held-out test set (10%). We use the same splits as in the extant literature in the implementation from `https://github.com/yaringal/DropoutUncertaintyExps`, and the final score is obtained by computing the average score over the splits.

Following the literature, we consider four techniques for the direct generation of conditional predictive distributions that do not use neural networks, namely, a semiparametric variant of the distributional forest technique (Duan et al., 2020; Schlosser et al., 2019), generalized additive models for location, scale and shape (GAMLSS, Stasinopoulos and Rigby, 2007), Gaussian process (GP) regression (Rasmussen and Williams, 2005), and natural gradient boosting (NGBoost, Duan et al., 2020). We adopt the exact implementation choices of Duan et al. (2020) for these techniques, which in some cases involve smoothing. Except for NGBoost, scores for the Year dataset are unavailable (NA), in part, because methods fail to be computationally feasible for a dataset of this size.

Table 5.5: Characteristics of datasets and predictive performance for competing methods of uncertainty quantification in regression problems, in terms of the mean logarithmic score (LogS) in a popular benchmark setting from machine learning (Duan et al., 2020; Gal and Ghahramani, 2016; Hernandéz-Lobato and Adams, 2015; Lakshminarayanan et al., 2017). For each dataset, we show size, number of unique outcomes, and dimension of the input (covariate or feature) space. Italics indicate discrete datasets where the number of unique outcomes is small. For each method, we report the mean LogS from the reference stated, with further details provided in Section 5.4.2. For each of the lower three blocks of comparable methods, the best (lowest) mean score is set in blue. Two scores are numerically infinite; missing scores are marked NA.

Method / Dataset	Boston	Concrete	Energy	Kin8nm	*Naval*	Power	Protein	*Wine*	Yacht	*Year*
Size	506	1,030	768	8,192	*11,934*	9,568	45,730	*1,599*	308	*515,345*
Unique Outcomes	229	845	586	8,191	*51*	4,836	15,903	*6*	258	*89*
Dimension Input Space	13	8	8	8	*16*	4	9	*11*	6	*90*
Distributional Forest	2.67	3.38	1.53	−0.40	−4.84	2.68	2.59	1.05	2.94	NA
GAMLSS	2.73	3.24	1.24	−0.26	−5.56	2.86	3.00	0.97	0.80	NA
GP Regression	2.37	3.03	0.66	−1.11	−4.98	2.81	2.89	0.95	0.10	NA
NGBoost	2.43	3.04	0.60	−0.49	−5.34	2.79	2.81	0.91	0.20	3.43
40 Deep Ensembles	2.41	3.06	1.38	−1.20	−5.36	2.79	2.83	0.94	1.18	3.35
40 Laplace	2.65	3.14	1.27	−1.00	NA	2.87	2.90	0.97	1.97	3.61
40 Single Gaussian	2.78	3.20	1.14	−1.03	−5.37	2.83	2.93	0.98	2.11	3.61
40 Smooth CP	2.89	3.14	1.20	−1.00	−5.52	2.85	2.88	0.97	1.88	NA
40 Smooth EasyUQ	2.83	3.04	0.79	−1.05	−6.51	2.77	2.48	0.48	1.36	3.24
400 MC Dropout	2.46	3.04	1.99	−0.95	−3.80	2.80	2.89	0.93	1.55	3.59
400 Laplace	2.61	3.07	0.80	−1.11	NA	2.83	2.87	1.04	1.18	3.61
400 Single Gaussian	3.41	3.32	0.85	−1.09	−6.32	2.81	2.87	1.38	2.04	3.61
400 Smooth CP	2.87	3.05	0.83	−1.09	−6.65	2.78	2.84	1.01	1.03	NA
400 Smooth EasyUQ	2.46	2.94	0.55	−1.13	−7.51	2.75	2.41	1.07	0.85	3.24
2L MC Dropout	2.34	2.82	1.48	−1.10	−4.32	2.67	2.70	0.90	1.37	NA
2L Laplace	2.57	2.98	0.56	−1.13	NA	2.76	2.81	1.22	1.24	3.60
2L Single Gaussian	∞	3.78	0.74	−0.96	−7.19	2.76	2.77	10.51	∞	3.61
2L Smooth CP	2.66	2.94	0.63	−1.18	−7.33	2.70	2.67	1.01	0.74	NA
2L Smooth EasyUQ	2.49	2.71	0.36	−1.21	−8.20	2.67	2.30	0.95	0.50	3.23

The remaining methods considered in Table 5.5 are based on neural networks, and we adopt the network architectures proposed by Hernandéz-Lobato and Adams (2015) and Gal and Ghahramani (2016). Specifically, we use the ReLU nonlinearity and either a single or two hidden layers, containing 50 hidden units for the smaller datasets, and 100 hidden units for the larger Protein and Year datasets. To tune the network hyperparameters, namely, the regularization parameter λ and the batch size, we use grid search. Thus, the nested hyperparameter selection in the Smooth EasyUQ Algorithm 1 finds a best combination of λ, the batch size, ν, and h by optimizing the mean LogS. Our intent is to compare EasyUQ and Smooth EasyUQ to state of the art methods for uncertainty quantification from machine learning, namely, Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017), which perform uncertainty quantification directly within the workflow of neural network fitting. Furthermore, these methods are input-based, i.e., they require access to, and operate on, the original covariate or feature vector. As seen in the table, the dimensionality of the input space in the benchmark problems varies between 4 and 90.

In contrast, EasyUQ, CP, and the Single Gaussian technique operate on the basis of the final model output only, and so can be applied without the original, potentially high-dimensional covariate or feature vector being available. For CP we adapt our previously described implementation with further refined splits into training (57.6%), calibration (14.4%), validation (18%), and test (10%) sets. Smooth CP uses the respective variant of Algorithm 1. An intermediary role between input-based and output-based methods is assumed by the recently developed Laplace approach (Immer et al., 2021; Ritter et al., 2018), which leverages scalable Laplace approximations based on weights of the trained network. For our numerical experiments we use the `laplace` software library for PyTorch (Daxberger et al., 2021).

A critical implementation decision in the intended comparisons is the number of training epochs in learning the neural network.[7] While the original setup specifies 40 training epochs (Hernandéz-Lobato and Adams, 2015), MC Dropout uses 400 or, in the 2-layer configuration, 4,000 iterations (Gal and Ghahramani, 2016). Therefore, to enable proper comparison, we apply the competing methods in three distinct neural network configurations, namely, a single-layer network with 40 training epochs (prefix 40 in Tables 5.5 and 5.6), a single-layer network with 400 training epochs (prefix 400),

[7] For the purposes of this comparison, the number of training epochs needs to be fixed. In practice, the number of epochs could be treated as a further hyperparameter and determined on the validation set.

and a 2-layer architecture with 4,000 training epochs (prefix 2L). In Tables 5.5 and 5.6, key comparisons between techniques for uncertainty quantification are then within the respective three groups of methods, for which the neural network configurations used are identical.

5.4.3 Comparison of predictive performance

We assess the predictive performance of EasyUQ, Smooth EasyUQ, and other methods for probabilistic forecasting and uncertainty quantification, by comparing the mean LogS in Table 5.5. We use the LogS from (2.7) in negative orientation, so smaller values correspond to better performance. Evidently, the use of the LogS, which is customary in machine learning, prevents comparisons to the basic versions of EasyUQ and CP, to which we turn in Table 5.6.

A first insight from Table 5.5 is that, in general, the methods in the second, third, and fourth blocks, which are based on neural networks, perform better relative to the direct, not on neural networks based methods in the first block (from top to bottom). Thus, we focus attention on the comparison of distinct methods for uncertainty quantification in neural networks, namely, Deep Ensembles (Lakshminarayanan et al., 2017) or MC dropout (Gal and Ghahramani, 2016), the Laplace approach (Ritter et al., 2018), the Single Gaussian technique, Smooth CP, and Smooth EasyUQ. The 2-layer architecture generally improves results, compared to using a single layer for the neural network. Smooth EasyUQ dominates the Single Gaussian and Smooth CP techniques and generally yields lower mean LogS than Deep Ensembles, MC Dropout, or the Laplace approach. In 24 of the $3 \times 10 = 30$ five-fold comparisons across the bottom three blocks, Smooth EasyUQ achieves or shares the top score. For eight of the ten datasets considered, the best performance across all 19 methods considered, including both neural network based approaches and not on neural networks based techniques, is achieved or shared by Smooth EasyUQ under the 2-layer network architecture. While this is not an exhaustive evaluation and no single method dominates universally, we note that Smooth EasyUQ is highly competitive with state of the art techniques for uncertainty quantification from machine learning.

To allow comparison with the basic form of EasyUQ, which generates discrete predictive distributions, we use Table 5.6 and the mean CRPS from (2.8) to assess predictive performance. Each of the three blocks in the table allows for a seven-way comparison among either Deep Ensembles or MC Dropout, the Laplace approach, the Single Gaus-

Table 5.6: Predictive performance for competing methods of uncertainty quantification in regression problems, in terms of the mean CRPS in a popular benchmark setting from machine learning (Duan et al., 2020; Gal and Ghahramani, 2016; Hernandéz-Lobato and Adams, 2015; Lakshminarayanan et al., 2017). For each dataset, we show size, number of unique outcomes, and dimension of the input (covariate or feature) space. Italics indicate discrete datasets where the number of unique outcomes is small. For Kin8mn and Naval the mean CRPS has been multiplied by factors of 10 and 1,000, respectively. For each block of comparable methods, the best (lowest) mean score is set in blue. For details, see Section 5.4.2.

Method / Dataset	Boston	Concrete	Energy	Kin8nm	*Naval*	Power	Protein	*Wine*	Yacht	*Year*
Size	506	1,030	768	8,192	*11,934*	9,568	45,730	*1,599*	308	*515,345*
Unique Outcomes	229	845	586	8,191	*51*	4,836	15,903	*6*	258	*89*
Dimension Input Space	13	8	8	8	*16*	4	9	*11*	6	*90*
40 Deep Ensembles	1.59	3.04	0.78	0.48	0.41	2.23	2.40	0.34	0.45	4.35
40 Laplace	1.71	3.02	0.45	0.49	NA	2.46	2.46	0.35	0.80	4.72
40 Single Gaussian	1.72	3.03	0.41	0.48	0.66	2.24	2.48	0.35	0.83	4.72
40 CP	1.73	3.04	0.45	0.49	0.58	2.24	2.47	0.36	0.90	NA
40 Smooth CP	1.74	3.05	0.45	0.49	0.58	2.24	2.47	0.36	0.91	NA
40 EasyUQ	1.69	2.94	0.34	0.48	0.54	2.21	2.22	0.31	0.66	4.35
40 Smooth EasyUQ	1.64	2.89	0.33	0.48	0.55	2.20	2.20	0.32	0.64	4.34
400 MC Dropout	1.56	2.79	0.37	0.48	1.22	2.21	2.40	0.35	0.57	4.73
400 Laplace	1.66	2.67	0.29	0.44	NA	2.17	2.36	0.37	0.41	4.73
400 Single Gaussian	1.61	2.72	0.29	0.44	0.27	2.17	2.36	0.38	0.41	4.73
400 CP	1.70	2.77	0.30	0.45	0.20	2.16	2.38	0.37	0.42	NA
400 Smooth CP	1.71	2.77	0.30	0.45	0.20	2.16	2.38	0.37	0.43	NA
400 EasyUQ	1.75	2.72	0.26	0.44	0.12	2.16	2.10	0.35	0.39	4.33
400 Smooth EasyUQ	1.60	2.61	0.25	0.44	0.13	2.15	2.09	0.37	0.35	4.33
2L MC Dropout	1.45	2.19	0.33	0.41	1.07	1.92	1.95	0.33	0.47	4.63
2L Laplace	1.64	2.29	0.22	0.44	NA	2.01	2.15	0.42	0.41	4.65
2L Single Gaussian	1.89	2.27	0.25	0.41	0.11	2.03	2.04	0.45	0.25	4.69
2L CP	1.70	2.47	0.24	0.42	0.11	2.02	2.02	0.38	0.36	NA
2L Smooth CP	1.71	2.48	0.24	0.42	0.11	2.03	2.02	0.38	0.36	NA
2L EasyUQ	2.07	2.40	0.24	0.42	0.03	1.98	1.83	0.42	0.30	4.30
2L Smooth EasyUQ	1.66	2.14	0.21	0.40	0.04	1.97	1.82	0.40	0.27	4.31

sian technique, Conformal Prediction in its basic (CP) and smoothed form (Smooth CP), the basic version of EasyUQ, and Smooth EasyUQ. As noted, the Naval, Wine, and Year datasets are distinctly discrete, with 51, 6, and 89 unique outcomes, respectively. For data of this type, predictive distributions ought to be discrete. Accordingly, there are no benefits of using Smooth EasyUQ for these datasets, compared to using basic EasyUQ, which adapts readily to discrete outcomes. In six of the $3 \times 3 = 9$ seven-fold comparisons for the discrete datasets, the basic version of EasyUQ achieves the lowest mean score. Across the remaining seven datasets and for all three network configurations, smoothing is beneficial, and Smooth EasyUQ outperforms the basic version of EasyUQ. In 15 of the $3 \times 7 = 21$ seven-fold comparisons on these datasets, Smooth EasyUQ achieves or shares the top score. All but one of the binary comparisons between Smooth CP and Smooth EasyUQ, all but two of the comparisons between the Single Gaussian technique and Smooth EasyUQ, all but one of the comparisons between the Laplace method and Smooth EasyUQ, and all but eight of the comparisons between Deep Ensembles or MC Dropout and Smooth EasyUQ, are in favor of the latter.

5.5 Discussion

In this chapter we have proposed EasyUQ and Smooth EasyUQ as general methods for the conversion of single-valued computational model output into calibrated predictive distributions, based on a training set of model output–outcome pairs and a natural assumption of isotonicity. Contrary to recent comments in review articles that lament an “absence of theory” (Abdar et al., 2021, p. 244), for data-driven approaches to uncertainty quantification, the basic version of EasyUQ enjoys strong theoretical support, in sharing the optimality and consistency properties of the general Isotonic Distributional Regression (IDR, Henzi et al., 2021) method. The basic EasyUQ approach is fully automated, does not require any implementation choices, and the generated predictive distributions are discrete. The more elaborate Smooth EasyUQ approach developed in this chapter generates predictive distributions with Lebesgue densities, based on a kernel smoothing of the original IDR distributions, while preserving the key properties of the basic approach. Code for the implementation of IDR in Python (Python Software Foundation, 2021) and replication material for this article are openly available (Walz, 2023).

The method is general, handling both discrete outcomes, with the basic technique being tailored to this setting, and continuous outcomes, for which Smooth EasyUQ is the method of choice. It applies whenever single-valued model output is to be converted into a predictive distribution, covering both the case of point forecasts, as in the WeatherBench example, and computational model output in all facets, such as in the machine learning example, where EasyUQ and Smooth EasyUQ convert single-valued neural network output into predictive distributions. Percentiles extracted from the predictive distributions can be used to generate prediction intervals.

The proposed term EasyUQ stems from various desirable properties. First, the basic version of EasyUQ does not involve tuning parameters nor requires user intervention. Second, EasyUQ operates on the natural, easily interpretable and communicable assumption that larger values of the computational model output yield predictive distributions that are stochastically larger. Third, EasyUQ is an output-based technique, i.e., it merely requires training data in the form of model output–outcome pairs (x_i, y_i) as in (5.1), without any need to access the potentially high-dimensional covariate or feature vector z_i, which serves as input to the computational model that generates x_i. This property is shared with the widely used Single Gaussian technique and related methods, such as the early Geostatistical Output Perturbation (GOP, Gel et al., 2004) approach and the Quantile Regression Averaging (QRA, Nowotarski and Weron, 2015) method for the generation of prediction intervals.

The term Conformal Prediction (CP, Marx et al., 2022; Vovk et al., 2020b, 2022) refers to a family of output-based methods that yield predictive distributions and prediction intervals that enjoy attractive out-of-sample coverage guarantees, but often mean that the shape and scale of the predictive distributions do not vary with the model output. In simple problems, where predictive distributions that are essentially translates of each other are appropriate, both CP and EasyUQ perform well, and typically yield very similar predictive performance, as illustrated by the temperature example in Section 5.2.2. The flexibility of EasyUQ, which allows for predictive distributions that vary in shape and/or scale, subject to the isotonicity condition, materializes in more challenging problems, where predictive distributions that are translates of each other fail. While EasyUQ adapts to such settings without any need for user intervention, CP might suffer considerable loss in predictive performance, even if adapted manually, as exemplified in the precipitation example in Section 5.2.3.

While adaptive variants of CP are available, their predictive performance in both simulated and real-data settings has been mixed, compared to standard variants (Vovk et al., 2020b). Recently, Boström et al. (2021) investigated Mondrian (i.e., covariate-conditional) CP as a flexible alternative, in which conformal predictive distributions are built on separate categories formed by binning covariates (in our case, the model output). This requires additional implementation decisions, namely, on the choice of the bins. Boström et al. (2021) take five bins with equal numbers of training instances, which improves predictive performance in their experiments. From a methodological point of view, in situations where the isotonicity assumption of IDR is met, the binning approach of Mondrian CP can be understood as an approximation to EasyUQ. EasyUQ finds optimal binnings without manual intervention (Henzi et al., 2021, Thm. 2), and training borrows strength from the entirety of the training data, whereas Mondrian CP diminishes the training sample by splitting it, which introduces a trade-off between training data size and adaptivity. A limitation of EasyUQ is that estimates under isotonicity constraints tend to be inconsistent at the boundary of the covariate domain (Guntuboyina and Sen, 2018), which raises the danger of disproportionately decreased spread of EasyUQ distributions at extreme values of the model output. In settings where this is of concern, a potential remedy is to resort to Mondrian CP at extreme values, while reaping the benefits of EasyUQ at moderate values of the model output. We leave further methodological development in these directions to future work.

In contrast to CP and EasyUQ, input-based methods such as MC Dropout (Gal and Ghahramani, 2016), Deep Ensembles (Lakshminarayanan et al., 2017), the techniques proposed by Camporeale and Caré (2021) and Chung et al. (2021), and the reference methods considered by Duan et al. (2020) require access to the covariate or feature vector z_i. Input-based methods are much more flexible than output-based methods and thus have higher potential in principle, as evidenced by the success of ensemble methods in numerical weather prediction (Bauer et al., 2015; Gneiting and Raftery, 2005). However, they tend to be more computationally intense than output-based methods, and as the machine learning example shows, they may not outperform the latter. Generally, sophisticated input-based methods for uncertainty quantification might realize their potential when applied to substantively informed, highly complex computational models, as in the case of numerical weather prediction, where predictive uncertainty varies. Output-based approaches to uncertainty quantification typically are less complex and thus easier to implement and might nonetheless yield competitive predictive performance when applied to output from data-driven models, such as the neural network models in the benchmark setting from machine learning.

We end the chapter with speculations about the usage of EasyUQ and Smooth EasyUQ in weather prediction. The current approach to forecasts at lead times of hours to weeks rests on ensembles of physics-based numerical models (Bauer et al., 2015; Gneiting and Raftery, 2005) but it is being challenged by the advent of purely-data driven models based on ever more sophisticated neural networks (Ebert-Uphoff and Hilburn, 2023; Schultz et al., 2021). Published only recently, the WeatherBench comparison (Rasp et al., 2020) showed a huge performance gap between forecasts from physics-based numerical models and neural network based, purely data-driven forecasts, with the latter being clearly inferior, as exemplified in our Tables 5.1 and 5.3. Fast breaking developments suggest that the situation may have reversed since then, with purely data-driven approaches now outperforming physics-based forecasts of univariate weather quantities (Ben Bouallègue et al., 2023; Bi et al., 2023; Chen et al., 2023; Lam et al., 2023). There is a caveat, though, as under the new, data-driven paradigm, spatio-temporal and inter-variable dependence structures might be misrepresented, due to the lack of physical constraints in the model and a need for hierarchical temporal aggregation in the generation of weather scenarios (Bi et al., 2023; Ebert-Uphoff and Hilburn, 2023). However, the resulting neural network based forecasts can be subjected to EasyUQ and Smooth EasyUQ, and samples from the resulting predictive distributions can be merged by empirical copula techniques such as ensemble copula coupling (ECC, Schefzik et al., 2013), to adopt and transfer spatio-temporal and inter-variable dependence structures in physics-based ensemble forecasts. Hybrid approaches of this type might combine and extract the best from both traditional physics-based and emerging data-driven approaches to weather prediction, and may turn out to be superior to both.

6 | Decompositions of the mean continuous ranked probability score

The continuous ranked probability score (CRPS) is the most commonly used scoring rule in the evaluation of probabilistic forecasts for real-valued outcomes. To assess and rank forecasting methods, researchers compute the mean CRPS over given sets of forecast situations, based on the respective predictive distributions and outcomes. We propose a new, isotonicity-based decomposition of the mean CRPS into interpretable components that quantify miscalibration (MCB), discrimination ability (DSC), and uncertainty (UNC), respectively. In the final chapter of this work, the isotonicity-based decomposition is used to properly perform the step *Evaluation* of the forecasting cycle which introduced in Section 1.

6.1 Introduction

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

Historically, the first score decomposition was introduced by Murphy (1973), who proposed a decomposition of the mean Brier score (BS). For a sequence of forecast–observation pairs $(p_1, y_1), \ldots, (p_n, y_n)$, consisting of predictive probabilities $p_i \in [0, 1]$ and corresponding binary outcomes $y_i \in \{0, 1\}$, the empirical average Brier score from (2.11)

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

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

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

Siegert (2017) shows that the Murphy decomposition reads as

$$\overline{\mathrm{BS}} = \underbrace{(\overline{\mathrm{BS}} - \overline{\mathrm{BS}}_c)}_{\overline{\mathrm{MCB}}} - \underbrace{(\overline{\mathrm{BS}}_r - \overline{\mathrm{BS}}_c)}_{\overline{\mathrm{DSC}}} + \underbrace{\overline{\mathrm{BS}}_r}_{\overline{\mathrm{UNC}}} . \tag{6.1}$$

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

The rationale behind the decomposition in (6.1) can be summarized as the following recipe: Having available a calibration method that transforms the original forecasts $p_1, \ldots, p_n$ into calibrated forecasts $q_1, \ldots, q_n$, one can measure miscalibration as the difference in the mean score of the original forecasts to the calibrated ones, resulting in the $\overline{\mathrm{MCB}}$ term. The CORP (Consistent, Optimally binned, Reproducible, and PAV algorithm based) score decomposition suggested by Dimitriadis et al. (2021) uses this general recipe, where the calibrated forecasts $q_1, \ldots, q_n$ are computed by applying nonparametric isotonic regression on the vector $(y_1, \ldots, y_n)$ with respect to the order induced by $(p_1, \ldots, p_n)$. The authors argue persuasively that "the assumption of nondecreasing CEPs is natural, as decreasing estimates are counterintuitive, routinely being dismissed as artifacts by practitioners" (Dimitriadis et al., 2021, p. 4). If we consider, e.g., the conditional event probability over all events where we predicted a positive outcome with probability 0.5, then we should expect this value to be smaller than the conditional event probability over all events where we predicted a positive outcome with probability 0.6. As noted by Bentzien and Friederichs (2014), Siegert (2017), Leutbecher and Haiden (2021), and Gneiting et al. (2023a), and discussed in detail by Gneiting and Resin (2023), the recipe extends to scores other than the Brier score and general types of statistical functionals. In this chapter, we focus on the continuous ranked probability score (CRPS; Matheson and Winkler, 1976). The CRPS is one of the most prominent scoring rules for the evaluation of probabilistic forecasts for real-valued outcomes and is popular across application areas and methodological communities; see, e.g., Gneiting et al. (2005), Hothorn et al. (2014), Pappenberger et al. (2015), Rasp and Lerch (2018), and Gasthaus et al. (2019). The CRPS is defined in terms of any cumulative distribution function (CDF) F on $\mathbb{R}$ and $y \in \mathbb{R}$ (see (2.8)), and given by

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

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

$$\overline{\mathrm{CRPS}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i) \tag{6.2}$$

serves to quantify the overall performance of the forecasts. Possible decompositions of the mean score at (6.2) have been discussed in the literature, with the most prominent approaches being introduced by Hersbach (2000) and Candille and Talagrand (2005). These methods offer promising solutions but come with severe limitations. In a nutshell, the Hersbach decomposition lacks a theoretical background and the desirable property that the components of the decomposition are nonnegative, whereas the decomposition of Candille and Talagrand (2005) is not practically feasible, as acknowledged by the authors. Another approach for decomposing the mean CRPS is by exploiting its representation as an integral over Brier scores, compare (6.5), and then integrating existing decompositions of $\overline{\mathrm{BS}}$. Similarly, the CRPS can be expressed as an integral over quantile scores, see (6.6), and existing decompositions for quantile scores can be leveraged to decompose the mean score at (6.2). However, these approaches have the drawback that miscalibration and discrimination ability are not measured with respect to the full probabilistic forecasts but only with respect to individual threshold or quantile levels.

In this article, we propose a new decomposition of the mean CRPS based on Isotonic Distributional Regression (IDR; Henzi et al., 2021) described in 3.4. In the case of binary outcomes, Dimitriadis et al. (2021) argue that isotonicity between the predictive probabilities and the calibrated forecasts is a natural constraint, as violations of isotonicity indicate poor predictive performance. This argument generalizes to the real-valued setting, since it is natural to assume that the conditional law of the outcome, given the forecast, should tend to be small (large) if the predictive distribution is small (large), where notions of small and large are understood with respect to the usual stochastic order. IDR is a nonparametric distributional regression technique that honors the shape constraint of isotonicity between covariates and responses. Applying IDR to the data $(F_1, y_1), \ldots, (F_n, y_n)$ yields calibrated forecasts, whereas the marginal distribution of the outcomes $y_1, \ldots, y_n$ serves as static reference forecast r. The general recipe from (6.1) then yields mean scores $\overline{\mathrm{CRPS}}_c$ and $\overline{\mathrm{CRPS}}_r$ for the calibrated forecast and the reference forecast, respectively, and a corresponding exact decomposition,

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

of the mean CRPS at (6.2), to which we refer as the isotonicity-based decomposition. The isotonicity-based approach guarantees the nonnegativity of the three components, and the miscalibration term admits a persuasive interpretation in terms of calibration.

While auto-calibration serves as the universal notion of calibration for binary events (Gneiting and Ranjan, 2013, Theorem 2.11), for real-valued random outcomes, numerous different notions of calibration are found in the literature (Dawid, 1984; Diebold et al., 1998; Strähl and Ziegel, 2017; Arnold et al., 2023a), as reviewed by Gneiting and Resin (2023). All relevant notions of calibration for this chapter are summarized in

Section 2.2.1. The strongest notion is auto-calibration and, ideally, one would like to measure miscalibration as deviation from auto-calibration, as targeted by the decomposition of Candille and Talagrand (2005). However, the Candille–Talagrand approach yields degenerate empirical decompositions. Therefore, we quantify miscalibration as the deviation from isotonic calibration, as introduced by Arnold and Ziegel (2023) in a study of the population version of IDR. Isotonic calibration is closer to auto-calibration than the notions of calibration targeted by the Hersbach decomposition, or by the aforementioned decompositions based on Brier or quantile scores.

The remainder of the chapter is organized as follows. Section 6.2 reviews the previously proposed decompositions and their properties. In Section 6.3, we develop the empirical version of the new isotonicity-based decomposition, followed by a thorough study of the population versions of the various types of decomposition and their properties in Section 6.4, with particular emphasis on calibration. In Section 6.5, we apply the proposed isotonicity-based decomposition in case studies from meteorology and machine learning. The main part of the chapter closes with a discussion in Section 6.6. Technical comments, and a series of detailed analytic examples in population settings are available in Appendices 6.A through 6.C.

6.2 Previously proposed empirical decompositions

6.2.1 Preliminaries

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

Single-valued forecasts for functionals of an unknown quantity should be compared using consistent scoring functions (Gneiting, 2011). For example, the *quadratic score* $(x-y)^2$, and the piecewise linear *quantile score* (see (2.5))

$$\mathrm{qs}_\alpha(x,y) = (\mathbb{1}\{y \le x\} - \alpha)\,(x-y), \tag{6.3}$$

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

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

$$\mathrm{s_B}(p, y) = (p - y)^2, \tag{6.4}$$

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

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int \mathrm{s_B}(F(z), \mathbb{1}\{y \le z\}) \, \mathrm{d}z \qquad (6.5) \\ &= \int_0^1 \mathrm{qs}_\alpha(F^{-1}(\alpha), y) \, \mathrm{d}\alpha, \qquad (6.6) \end{aligned}$$

where $\mathrm{s_B}$ and qs_α are defined at (6.4) and (6.3), respectively, and where F^{-1} denotes the quantile function defined as $F^{-1}(\alpha) = \inf\{z \in \mathbb{R} \mid F(z) \ge \alpha\}$ for $\alpha \in (0, 1)$. The representation at (6.6) is due to Laio and Tamea (2007).

We consider a collection

$$(F_1, y_1), \ldots, (F_n, y_n) \tag{6.7}$$

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

$$\overline{\mathrm{CRPS}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i), \tag{6.8}$$

of the forecast–observation pairs at (6.7) into three distinct components, namely, miscalibration ($\overline{\mathrm{MCB}}$), discrimination ($\overline{\mathrm{DSC}}$), and uncertainty ($\overline{\mathrm{UNC}}$). The following desirable properties are relevant.

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

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

(E_2) The components $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$ are nonnegative.

(E_3) The decomposition is not degenerate. Here, a decomposition is *degenerate* if $\overline{\mathrm{MCB}} = \overline{\mathrm{CRPS}}$ whenever $F_1, \ldots, F_n$ are pairwise distinct.

(E_4) The $\overline{\text{DSC}}$ component vanishes if $F_1 = \cdots = F_n$.

(E_5) The $\overline{\text{UNC}}$ component can be expressed in terms of the outcomes $y_1, \ldots, y_n$ only.

These conditions do not depend on the use of any specific scoring rule; they are desirable for decompositions of mean scores in general.

An exact decomposition (E_1) is desirable, since it allows us to fully decompose the mean score. A degenerate decomposition is undesirable, as in typical practice, such as in the case studies in Section 6.5, the issued forecast distributions are pairwise distinct, and then the method is useless. A static forecast, i.e., $F_1 = \cdots = F_n$, has no discrimination ability, hence (E_4) is desirable. Requirement (E_5) is natural since intrinsic uncertainty does not depend on the activities of forecasters.

Finally, we argue that there ought to be a population version of the decomposition that applies to any admissible joint distribution $\mathbb{P}$ of tuples (F, Y). Furthermore, the population version ought to reduce to the empirical version if $\mathbb{P}$ is the empirical measure for the data at (6.7). We study decompositions at the population level in Section 6.4.

6.2.2 Candille–Talagrand decomposition

Candille and Talagrand (2005) naturally extend the idea of the Murphy decomposition at (6.1). To describe their approach, let δ_y denote the Dirac or point measure in $y \in \mathbb{R}$, and let the marginal law $\hat{F}_{\text{mg}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$ denote the empirical distribution of the outcomes $y_1, \ldots, y_n$ in (6.7). Let $\hat{F}_i$ be the auto-calibrated version of the forecast F_i in (6.7), i.e., let $\hat{F}_i$ be the normalized version of $\sum_{j=1}^{n} \mathbb{1}\{F_j = F_i\}\, \delta_{y_j}$ for $i = 1, \ldots, n$. Then

$$\overline{\text{CRPS}}_{\text{mg}} = \frac{1}{n} \sum_{i=1}^{n} \text{CRPS}(\hat{F}_{\text{mg}}, y_i) \quad \text{and} \quad \overline{\text{CRPS}}_{\text{ac}} = \frac{1}{n} \sum_{i=1}^{n} \text{CRPS}(\hat{F}_i, y_i) \tag{6.9}$$

are the mean score of the marginal forecast and the auto-calibrated forecast, respectively. Candille and Talagrand (2005) define uncertainty, miscalibration, and discrimination components as

$$\overline{\text{UNC}}_0 = \overline{\text{CRPS}}_{\text{mg}}, \tag{6.10}$$

$$\overline{\text{MCB}}_{\text{CT}} = \overline{\text{CRPS}} - \overline{\text{CRPS}}_{\text{ac}}, \qquad \overline{\text{DSC}}_{\text{CT}} = \overline{\text{CRPS}}_{\text{mg}} - \overline{\text{CRPS}}_{\text{ac}}, \tag{6.11}$$

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

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{CT}} - \overline{\mathrm{DSC}}_{\mathrm{CT}} + \overline{\mathrm{UNC}}_{0}. \tag{6.12}$$

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

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

6.2.3 Brier score based decomposition

The Brier score based representation of individual CRPS values at (6.5) implies that

$$\overline{\mathrm{CRPS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} \overline{\mathrm{BS}}_z \,\mathrm{d}z, \tag{6.13}$$

where

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

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

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

$$\overline{\mathrm{MCB}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{MCB}}_{\mathrm{BS},z} \,\mathrm{d}z, \; \overline{\mathrm{DSC}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{DSC}}_{\mathrm{BS},z} \,\mathrm{d}z, \; \overline{\mathrm{UNC}}_{\mathrm{BS}} = \int_{-\infty}^{\infty} \overline{\mathrm{UNC}}_{\mathrm{BS},z} \,\mathrm{d}z. \tag{6.14}$$

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

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{DSC}}_{\mathrm{BS}} + \overline{\mathrm{UNC}}_{\mathrm{BS}}, \tag{6.15}$$

as the *Brier score based* (BS) decomposition of $\overline{\mathrm{CRPS}}$. Details of this approach are reviewed in Appendix 6.A.1, where we prove the following result.

Proposition 6.2.1. *For the Brier score based decomposition at* (6.15) *it holds that* $\overline{\mathrm{UNC}}_{\mathrm{BS}} = \overline{\mathrm{UNC}}_0$, *and the decomposition satisfies properties* (E_1), (E_2), (E_3), (E_4), *and* (E_5).

Despite these favorable properties, the Brier score based decomposition is subject to shortcomings and inconsistencies, due to the isolated treatment of probability forecasts at fixed thresholds. For discussion, we refer the reader to Section 6.2.6 and Appendix 6.A.

6.2.4 Quantile score based decomposition

In view of the quantile score representation of the CRPS at (6.6), a natural approach to decomposing the mean score $\overline{\mathrm{CRPS}}$ leverages decompositions of the mean quantile score at (6.3). Specifically, the quantile score representation implies that

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

where

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

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

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{QS}} - \overline{\mathrm{DSC}}_{\mathrm{QS}} + \overline{\mathrm{UNC}}_{\mathrm{QS}}, \tag{6.16}$$

as the *quantile score based* (QS) decomposition of $\overline{\mathrm{CRPS}}$. For details, we refer the reader to Appendix 6.A.2 where we prove the following result.

Proposition 6.2.2. *For the quantile score based decomposition at* (6.16) *it holds that* $\overline{\mathrm{UNC}}_{\mathrm{QS}} = \overline{\mathrm{UNC}}_0$, *and the decomposition satisfies properties* (E_1), (E_2), (E_3), (E_4), *and* (E_5).

The quantile score based decomposition is subject to shortcomings in analogy to the issues with the Brier score based approach, due to the reliance on quantile forecasts at fixed levels; for further discussion see Section 6.2.6 and Appendix 6.A.

6.2.5 Hersbach decomposition

The decomposition of Hersbach (2000) applies specifically to ensemble forecasts and operates under the implicit assumption of a continuous outcome. For the data at (6.7), Hersbach's assumptions imply, without loss of generality, that for $i = 1, \ldots, n$ the forecast F_i is the empirical CDF of a fixed number m of values $x_1^i \leq \cdots \leq x_m^i$, with the outcome $y_i \notin \{x_1^i, \ldots, x_m^i\}$ being distinct from these values. However, with a view towards a generalization of the Hersbach decomposition, we (naively) allow for any real-valued outcome y_i.

In line with the other types of decomposition, Hersbach (2000) defines the uncertainty component as $\overline{\mathrm{UNC}}_0$ at (6.10). The miscalibration component, which Hersbach (2000) refers to as reliability, is

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

where $p_\ell = \ell/m$ for $\ell = 0, \ldots, m$, and $\bar{g}_\ell$ is the average width of bin i, i.e.,

$$\bar{g}_\ell = \frac{1}{n} \sum_{i=1}^{n} (x_{\ell+1}^i - x_\ell^i) \tag{6.17}$$

for $\ell = 1, \ldots, m - 1$. The term $\bar{o}_\ell$ approximates the average frequency of an outcome below the midpoint of bin ℓ; specifically,

$$\bar{o}_\ell = \bar{f}_\ell - \bar{m}_\ell,$$

where

$$\bar{f}_\ell = \frac{1}{n\bar{g}_\ell} \sum_{i=1}^{n} \mathbb{1}\{F_i(y_i) \leq p_\ell\} \, (x_{\ell+1}^i - x_\ell^i) \quad \text{and} \quad \bar{m}_\ell = \frac{1}{n\bar{g}_\ell} \sum_{i=1}^{n} \mathbb{1}\{x_\ell^i < y_i < x_{\ell+1}^i\} \, (y_i - x_\ell^i) \tag{6.18}$$

for $\ell = 1, \ldots, m-1$. For any ℓ with $x_\ell^i < x_{\ell+1}^i$ it holds that $F_i(y_i) \le p_l$ if, and only if, $y_i < x_{\ell+1}^i$. To complete the specification, we let $\bar{o}_0 = (1/n) \sum_{i=1}^n \mathbb{1}\{y_i < x_1^i\}$ and $\bar{o}_m = (1/n) \sum_{i=1}^n \mathbb{1}\{x_m^i < y_i\}$, and if these quantities are nonzero then we let $\bar{g}_0 = (1/(n\bar{o}_0)) \sum_{i=1}^n \mathbb{1}\{y_i < x_1^i\} \, (x_1^i - y_i)$ and $\bar{g}_m = (1/(n\bar{o}_m)) \sum_{i=1}^n \mathbb{1}\{x_m^i < y_i\} \, (y_i - x_m^i)$. The miscalibration component thus measures deviations from uniformity for the rank histogram (Hamill, 2001; Gneiting et al., 2007).

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

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{HBo}} - \overline{\mathrm{DSC}}_{\mathrm{HBo}} + \overline{\mathrm{UNC}}_0. \tag{6.19}$$

Towards a generalization, we introduce a slightly modified miscalibration component,

$$\overline{\mathrm{MCB}}_{\mathrm{HB}} = \sum_{\ell=1}^{m-1} \bar{g}_\ell \left(p_\ell - \bar{f}_\ell\right)^2, \tag{6.20}$$

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

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{HB}} - \overline{\mathrm{DSC}}_{\mathrm{HB}} + \overline{\mathrm{UNC}}_0. \tag{6.21}$$

The interpretation of the miscalibration component remains unchanged, as $\overline{\mathrm{MCB}}_{\mathrm{HB}}$ and $\overline{\mathrm{MCB}}_{\mathrm{HBo}}$ differ only slightly, with $\bar{f}_\ell$ in (6.20) being the approximate frequency of an outcome below the right endpoint of bin ℓ. For a more detailed comparison and the proof of the following result, we refer the reader to Appendix 6.B.

Proposition 6.2.3. *The original and modified Hersbach decompositions at* (6.19) *and* (6.21), *respectively, satisfy properties* (E_1), (E_3), *and* (E_5), *while properties* (E_2) *and* (E_4) *fail to hold.*

As discussed thus far, the Hersbach decomposition requires that the forecasts assume the form of an ensemble. Further shortcomings have been discussed in the literature (Siegert, 2017); in particular, it has been noted that the discrimination component $\overline{\mathrm{DSC}}_{\mathrm{HBo}}$ is defined “somewhat artificially” (Hersbach, 2000, p. 565) and that it can be negative, thus violating (E_2). The original Hersbach decomposition has been extended by Lalaurette so that it applies to forecasts with strictly increasing CDFs (Candille and Talagrand, 2005, Appendix A). We discuss and generalize Lalaurette’s extension in Section 6.4.3, and our analysis demonstrates that the extensions can more naturally be

Table 6.1: Candille–Talagrand (CT), quantile score based (QS), Brier score based (BS), and Hersbach (HB) decomposition of the mean score $\overline{\mathrm{CRPS}}$, as applied to the one-day ahead raw ensemble (ENS) forecast of precipitation accumulation at Frankfurt Airport (Section 6.5.1), and the EasyUQ forecast for the Boston and Wine data, respectively (Section 6.5.2).

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

interpreted as extensions of the modified Hersbach decomposition. In Section 6.4.3 we describe empirical versions that apply in the general case of forecast distributions with finite support, and to mixed discrete-continuous distributions for nonnegative quantities, respectively.

6.2.6 Numerical example and discussion

For illustration, we consider forecasts from the case studies in Section 6.5. The decompositions from Sections 6.2.2 through 6.2.5 all use the uncertainty component $\overline{\mathrm{UNC}}_0$ at (6.10), and they specify the discrimination component as

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

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

Table 6.1 displays the mean score $\overline{\mathrm{CRPS}}$, the uncertainty component UNC_0, and the various $\overline{\mathrm{MCB}}_{\bullet}$ terms for the ENS forecast of precipitation accumulation at Frankfurt Airport, as studied in our Section 6.5.1 and Henzi et al. (2021), and the EasyUQ forecasts for the Boston Housing and Wine data, as considered in our Section 6.5.2 and in Chapter 5. The ENS forecast is an ensemble forecast with $m = 52$ members and so the Hersbach decomposition at (6.19) applies; for the EasyUQ forecasts, we apply formula (6.40). For the first two examples in the table, it holds that $\overline{\mathrm{CRPS}} =$

$\overline{\text{MCB}}_{\text{CT}} > \overline{\text{MCB}}_{\text{QS}} > \overline{\text{MCB}}_{\text{BS}} > \overline{\text{MCB}}_{\text{HB}}$, where the initial equality reflects the degeneracy of the Candille–Talagrand decomposition. In our experience, the subsequent inequalities hold in many, though not all, empirical examples. However, as we state in further generality at (6.24) and in Corollary 6.4.6, it always holds that $\overline{\text{CRPS}} \geq \overline{\text{MCB}}_{\text{CT}} \geq \max\{\overline{\text{MCB}}_{\text{BS}}, \overline{\text{MCB}}_{\text{QS}}\}$.

While the Candille–Talagrand decomposition seems attractive and preferable from theoretical perspectives, the degeneracy prohibits its practical use. The Hersbach decomposition has been popular in the specific setting of ensemble forecasts, but has serious shortcomings including but not limited to the possibility of a negative discrimination component. The Brier score and quantile score based decompositions have desirable properties, but they define the components of the decomposition in terms of isolated functionals (probabilities and quantiles, respectively) rather than the entire predictive distributions, which is "unsatisfactory" (Ferro and Fricker, 2012, p. 1958) and entails the artifacts described in Remarks 6.A.1 and 6.A.2, respectively. Furthermore, it is not obvious whether the Brier score based or the quantile score based decomposition ought to be preferred. In this light, there remains the need for a decomposition that is both practically feasible and theoretically justifiable and appealing.

6.3 Empirical isotonicity-based decomposition

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

6.3.1 Empirical isotonicity-based decomposition

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

In the spirit of the Candille–Talagrand decomposition, a calibration tool ought to be applied to the assessed forecasts $F_1, \ldots, F_n$ from (6.7), and we propose that this tool be isotonic distributional regression (IDR; Henzi et al., 2021). IDR is a nonparametric distributional regression method under the shape constraint of isotonicity between covariates and responses: For training data consisting of covariates $x_1, \ldots, x_n$ in a partially ordered set $(\mathcal{X}, \preceq)$ and real-valued responses $y_1, \ldots, y_n$, Henzi et al. (2021) prove that there exists a unique minimizer of the criterion

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(P_i, y_i) \tag{6.22}$$

over all vectors of CDFs $(P_1, \ldots, P_n)$ with $P_i \leq_{\mathrm{st}} P_j$ if $x_i \preceq x_j$ for $i, j = 1, \ldots, n$, and they refer to this minimizer as the IDR solution.

The constraint of isotonicity between the assessed and the calibrated forecasts is natural, and hence, we apply IDR to the data $(F_1, y_1), \ldots, (F_n, y_n)$ at (6.7) with the stochastic order serving as the partial order on the covariate space $\mathcal{P}(\mathbb{R})$. In a number of practically relevant situations the stochastic order is too strong, since it does not allow for crossings between CDFs, and we discuss modifications that resolve this problem in the latter part of this section. For now, we assume that there are sufficiently many pairs of CDFs across $F_1, \ldots, F_n$ that can be ranked in stochastic order.

Let $\check{F}_1, \ldots, \check{F}_n$ denote the calibrated forecasts that are obtained by using IDR, let

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

denote the mean score of the calibrated forecasts, let the marginal forecast $\hat{F}_{\mathrm{mg}}$ and its mean score $\overline{\mathrm{CRPS}}_{\mathrm{mg}}$ be defined as at (6.9), and let

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

Then the *isotonicity-based* (ISO) decomposition

$$\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{ISO}} - \overline{\mathrm{DSC}}_{\mathrm{ISO}} + \overline{\mathrm{UNC}}_0 \tag{6.23}$$

differs from the Candille–Talagrand decomposition at (6.11) by the choice of the calibration method only, as it draws on the slightly weaker notion of isotonic calibration in lieu of auto-calibration. The isotonicity-based decomposition has desirable and appealing properties, as follows.

Proposition 6.3.1. *The isotonicity-based decomposition at* (6.23) *satisfies* (E_1), (E_2), (E_3), (E_4), *and* (E_5). *Furthermore,* $\overline{\mathrm{MCB}}_{\mathrm{ISO}} = 0$ *if, and only if,* $F_i = \check{F}_i$ *for* $i = 1, \ldots, n$, *and* $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$ *if, and only if,* $\check{F}_i = \hat{F}_{\mathrm{mg}}$ *for* $i = 1, \ldots, n$.

Proof of Proposition 6.3.1. By definition, the isotonicity-based decomposition satisfies properties (E_1) and (E_5). The IDR solution is the unique minimizer of the criterion (6.22) over all vectors of distributions $(P_1, \ldots, P_n)$ that are stochastically ordered with the same order relations as the covariates. Here, the covariates are $F_1, \ldots, F_n$ and the partial order on the covariate space is the stochastic order. Therefore, $(F_1, \ldots, F_n)$ is an admissible vector of distributions in the minimization problem, whence $\overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq 0$. A further admissible vector in the minimization problem is the constant vector with entries $\hat{F}_{\mathrm{mg}}$, whence $\overline{\mathrm{DSC}}_{\mathrm{ISO}} \geq 0$, so ($E_2$) is satisfied. The examples in the case study in Section 6.5 imply that the isotonicity-based decomposition satisfies (E_3). Assume now that $F_1 = \cdots = F_n$. Then we obtain $\hat{F}_{\mathrm{mg}}$ as the IDR solution, whence $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$, so ($E_4$) is satisfied. Finally, if $\overline{\mathrm{MCB}}_{\mathrm{ISO}} = 0$ then $F_i = \check{F}_i$, since IDR is the unique minimizer of the criterion at (6.22), and analogously, if $\overline{\mathrm{DSC}}_{\mathrm{ISO}} = 0$ then $\check{F}_i = \hat{F}_{\mathrm{mg}}$ for $i = 1, \ldots, n$. □

Generally, the determination of the pairwise stochastic order relations between the distributions $F_1, \ldots, F_n$ requires $\mathcal{O}(n^2)$ operations. As IDR can be implemented in at most $\mathcal{O}(n^2)$ operations (Henzi et al., 2021, 2022), the computation of the isotonicity-based decomposition is of complexity $\mathcal{O}(n^2)$. In contrast, the Brier score based and quantile score based decompositions require $\mathcal{O}(n)$ or more distinct determinations of pairwise stochastic order relations (cf. Appendices 6.A.1 and 6.A.2) and, hence, the implementation is of complexity at least $\mathcal{O}(n^2 \log n)$. The computation of the Hersbach decomposition for an ensemble forecast of size m requires $\mathcal{O}(mn)$ operations.

In its present form, the isotonicity-based decomposition is fully automated in the sense that it does not involve any tuning parameter. For the examples in Table 6.1, $\overline{\mathrm{MCB}}_{\mathrm{ISO}}$ equals 0.34, 0.80, and 0.072, respectively, and so $\overline{\mathrm{MCB}}_{\mathrm{ISO}}$ is larger than $\overline{\mathrm{MCB}}_{\mathrm{BS}}$ (which equals 0.068 in the third example) and $\overline{\mathrm{MCB}}_{\mathrm{QS}}$ and smaller than the essentially useless $\overline{\mathrm{MCB}}_{\mathrm{CT}} = \overline{\mathrm{CRPS}}$ term. As we demonstrate in Section 6.4.4, it is always true that

$$\overline{\mathrm{CRPS}} \geq \overline{\mathrm{MCB}}_{\mathrm{CT}} \geq \overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq \max\{\overline{\mathrm{MCB}}_{\mathrm{BS}}, \overline{\mathrm{MCB}}_{\mathrm{QS}}\}. \tag{6.24}$$

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

6.3.2 Computational implementation

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

In relevant applications, the stochastic order may be to strong, since it allows for no crossings of the forecasts. For example, for Gaussian forecasts $F = \mathcal{N}(\mu, \sigma^2)$ and $G = \mathcal{N}(\nu, \tau^2)$, F and G only order with respect to the stochastic order in case of $\sigma = \tau$, a condition which is rarely satisfied if parameters are estimated from data. Generally, if F and G are members of a location-scale family, they are stochastically ordered if, and only if, they have equal scale parameter, subject to minimal conditions. If only very few forecasts in the dataset are comparable with respect to the stochastic order, applying IDR results in calibrated forecast that are close to Dirac measures of the corresponding observations. Hence, in principle, the ISO-based decomposition faces the same problem as the Candille–Talagrand decomposition in this setting. However, we argue that there is a convincing remedy to the issue.

Consider settings where only few of the predictive distributions F_i in the collection at (6.7) are comparable with respect to the stochastic order. Frequently, predictive distributions fail to order due to crossings of the CDFs in a far tail. Recent work by Brehmer and Strokorb (2019) and Taillardat et al. (2023) casts doubt on the ability of the average CRPS to distinguish tail behavior of the forecast distribution, which provides support for the evaluation of the forecasts on a bounded interval only. Motivated by these findings, instead of decomposing the original mean score $\overline{\mathrm{CRPS}}$ as given in (6.8), we decompose

$$\overline{\mathrm{CRPS}}^{(a,b)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(\tilde{F}_i^{(a,b)}, y_i), \tag{6.25}$$

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

$$F_i^{(a,b)}(x) = \begin{cases} 0, & x < a, \\ F_i(x), & x \in [a, b), \\ 1, & x \ge b, \end{cases} \tag{6.26}$$

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

such that the condition

$$\left|\overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}^{(a,b)}\right| = \overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}^{(a,b)} < \epsilon \tag{6.27}$$

is satisfied, where the equality holds since $\overline{\mathrm{CRPS}} \geq \overline{\mathrm{CRPS}}^{(a,b)}$. Condition (6.27) is equivalent to

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

A simple method for determining the thresholds a and b to be used in (6.26) is described in Algorithm 2. If the support of the predictive distributions is bounded from above or below (e.g., in the case of precipitation accumulations, which are necessarily nonnegative), it is natural to set a or b equal to the respective bound (e.g., $a = 0$ for precipitation accumulations).

Algorithm 2 Thresholds a, b

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

The computation of this modified isotonicity-based decomposition remains of complexity $\mathcal{O}(n^2)$. Furthermore, the following result shows that, even with the approximation, theoretical guarantees from (6.24) continue to hold.

Proposition 6.3.2. *Let* $\overline{\mathrm{CRPS}} = \overline{\mathrm{MCB}}_{\mathrm{ISO}} - \overline{\mathrm{DSC}}_{\mathrm{ISO}} + \overline{\mathrm{UNC}}_0 = \overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{DSC}}_{\mathrm{BS}} + \overline{\mathrm{UNC}}_0$ *denote decompositions for data* $(F_1, y_1), \dots, (F_n, y_n)$*, and let*

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

denote the respective decompositions for modified data $(F_1^{(a,b)}, y_1), \dots, (F_n^{(a,b)}, y_n)$*, where* $F_1^{(a,b)}, \dots, F_n^{(a,b)}$ *derive from* $F_1, \dots, F_n$ *as in (6.26). Then* $I(a,b) = \overline{\mathrm{CRPS}} - \overline{\mathrm{CRPS}}^{(a,b)} < \epsilon$ *implies that*

$$\overline{\mathrm{MCB}}_{\mathrm{ISO}} \geq \overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)} \geq \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)} > \overline{\mathrm{MCB}}_{\mathrm{BS}} - \epsilon. \tag{6.28}$$

Proof of Proposition 6.3.2. The properties of the IDR solution imply $\overline{\mathrm{CRPS}}_{\mathrm{ISO}} \leq \overline{\mathrm{CRPS}}_{\mathrm{ISO}}^{(a,b)} \leq \overline{\mathrm{CRPS}}^{(a,b)} \leq \overline{\mathrm{CRPS}}$, and we conclude that

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

To complete the proof, we apply the inequality (6.24) to the modified data to yield $\overline{\mathrm{MCB}}_{\mathrm{ISO}}^{(a,b)} \geq \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)}$, and we note that $a \leq \min\{y_1, \ldots, y_n\}$ and $b \geq \max\{y_1, \ldots, y_n\}$, whence $\overline{\mathrm{MCB}}_{\mathrm{BS}} - \overline{\mathrm{MCB}}_{\mathrm{BS}}^{(a,b)} = I(a,b) < \epsilon$. □

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

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

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

In order to increase the number of comparable pairs amongst $F_1, \ldots, F_n$, it may appear natural to exchange the stochastic order with a weaker partial $\leq'$ order on $\mathcal{P}(\mathbb{R})$ in the sense that $G \leq_{\mathrm{st}} H$ implies $G \leq' H$ for $G, H \in \mathcal{P}(\mathbb{R})$, rather than restricting the support of the predictive distributions to a bounded interval $[a,b] \subseteq \mathbb{R}$. Possible choices include the almost-first-stochastic-dominance order proposed by Leshno and Levy (2002) or stochastic dominance of order $(1+\gamma)$ as proposed by Müller et al. (2017). If there are only few forecasts in a sample $(F_1, y_1), \ldots, (F_n, y_n) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$ that are comparable with respect to $\leq_{\mathrm{st}}$, one could think of applying IDR with respect to $\leq'$ instead of $\leq_{\mathrm{st}}$ in order to obtain more comparable forecasts. However, such an approach is bound to fail since isotonic calibration is generally only compatible with the stochastic order. More specifically, let Y be a random variable and F be a random forecast defined on the same probability space. Recall from Section

6.4.2 that ICL forms the population version of IDR (Arnold and Ziegel, 2023, Proposition 4.1). In analogy to Definition 3.1 of Arnold and Ziegel (2023), one could define the σ-lattice generated by F with respect to the weaker order $\leq'$ as $\mathscr{L}'(F) = \{F^{-1}(B) \mid B \in \mathcal{B}(\mathcal{P}(\mathbb{R})) \cap \mathcal{U}'\}$, where $\mathcal{U}'$ denotes the family of all upper sets in $\mathcal{P}(\mathbb{R})$ with respect to $\leq'$. However, if the space $\mathcal{P}(\mathbb{R})$ equipped with the partial order $\leq'$ and the topology of weak convergence satisfies Assumption C.1 of Arnold and Ziegel (2023), the corresponding notion of isotonic calibration, namely, $P_{Y|\mathscr{L}'(F)} = F$, fails to be intuitive for two reasons. First, auto-calibration does not imply the respective notion of calibration. Second, $G \leq' H$ already implies $G \leq_{\mathrm{st}} H$ for all G and H in the support of F by Theorem 3.3 of Arnold and Ziegel (2023). Clearly, this implication may only hold if $\leq'$ equals $\leq_{\mathrm{st}}$ on the support of F, which is violated for any $\leq'$ that is strictly weaker than $\leq_{\mathrm{st}}$, contrary to the scope of a relaxation. Moreover, there is no theoretical guarantee that the corresponding miscalibration term $\mathrm{MCB}_{\mathrm{ISO}'} = \mathbb{E}\,\mathrm{CRPS}(F, Y) - \mathbb{E}\,\mathrm{CRPS}(P_{Y|\mathscr{L}'(F)}, Y)$ is nonnegative.

Consequently, the stochastic order is the only valid choice of a partial order if IDR is applied to generate a calibrated forecast for an isotonicity-based approach in the spirit of the Candille–Talagrand decomposition.

6.4 Population level analysis

In this section, we present population level versions of all decompositions which we have discussed so far, and we analyse their relations to notions of calibration. The population quantity to be decomposed is the expected score

$$\mathbb{E}\,\mathrm{CRPS}(F, Y), \tag{6.29}$$

where the expectation is with respect to the joint law $\mathbb{P}$ of the random tuple (F, Y) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where F is a CDF-valued random quantity, which we interpret as the forecast, and the random variable Y is the real-valued outcome. For subsequent use, we assume the existence of a standard uniform variable U on $(\Omega, \mathcal{F}, \mathbb{P})$, which is independent of (F, Y). Evidently, if $\mathbb{P}$ is the empirical distribution for the data at (6.7) the expectation at (6.29) reduces to the mean score $\overline{\mathrm{CRPS}}$ from (6.8).

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

$$\mathrm{UNC}_0 = \mathbb{E}\,\mathrm{CRPS}(F_{\mathrm{mg}}, Y) \tag{6.30}$$

of the marginal law F_{mg} of Y. Again, the expectation is with respect to $\mathbb{P}$, and if $\mathbb{P}$ is the empirical distribution of the data at (6.7) then (6.30) reduces to (6.10). In this light, the decompositions at the population level read

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

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

6.4.1 Desiderata for decompositions at the population level

We adapt the desirable properties (E_1) through (E_5) for decompositions of a mean score from Section 6.2 to the population setting, as follows.

(P_1) The decomposition is exact.

(P_2) The components MCB, DSC, and UNC are nonnegative.

(P_3) The MCB component vanishes if, and only if, the forecast is calibrated in a well-defined sense.

(P_4) The DSC component vanishes if the forecast is static, i.e., there is an $F_0 \in \mathcal{P}(\mathbb{R})$ such that $F = F_0$ almost surely.

(P_5) The UNC component only depends on the unconditional distribution F_{mg} of the outcome.

Concerning (P_3), a notion of forecast calibration has to be specified. In the special case of a binary outcome, there is a unique, clear-cut notion of calibration (Gneiting and Ranjan, 2013, Theorem 2.11). Here, we consider the case of a real-valued outcome, for which numerous notions of calibration exist (Gneiting and Resin, 2023). Auto-calibration is the strongest such notion, but typically cannot be used in practice. Indeed, it turns out that (E_3) and (P_3) are competing requirements in the sense that if a decomposition satisfies (P_3) with respect to auto-calibration, then (E_3) is violated and the decomposition becomes degenerate. If a weaker notion of calibration is requested for (P_3), then (E_3) can be satisfied for the empirical counterpart of the decomposition. Requirement (P_4) is natural, since a static forecast has no discrimination ability at all. Finally, property (P_5) is motivated by the observation that intrinsic uncertainty does not depend on the forecast; evidently, the criterion is satisfied by UNC_0 at (6.30).

6.4.2 Isotonic conditional expectations and laws

The population versions of the isotonicity-based, Brier score based, and quantile score based decompositions rely on conditional expectations given σ-lattices and isotonic conditional laws. We give a short overview of the necessary concepts and refer to Arnold and Ziegel (2023) for further details. Readers not familiar with measure theory might skip the current subsection and intuitively think of the conditional expectation and the conditional law of a random variable Y given a σ-lattice $\mathcal{A}$, which we denote $\mathbb{E}(Y \mid \mathcal{A})$ and $P_{Y|\mathcal{A}}$, respectively, as classical conditional expectations and laws under the constraint of isotonicity.

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A subset $\mathcal{A} \subseteq \mathcal{F}$ is a *σ-lattice* if it is closed under countable unions and intersections and $\Omega, \emptyset \in \mathcal{A}$. Let $\mathcal{A} \subseteq \mathcal{F}$ be a σ-lattice and let X and Z be integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We call X *$\mathcal{A}$-measurable* if $\{X > x\} \in \mathcal{A}$ for all $x \in \mathbb{R}$ and define the *σ-lattice generated by X*, denoted by $\mathscr{L}(X)$, as the smallest σ-lattice which contains $\{X > x\}$ for all $x \in \mathbb{R}$. We call an $\mathcal{A}$-measurable random variable $\tilde{X}$ a *conditional expectation of X given $\mathcal{A}$*, for short $\mathbb{E}(X \mid \mathcal{A})$, if $\mathbb{E}(X \mathbb{1}_A) \le \mathbb{E}(\tilde{X} \mathbb{1}_A)$ for all $A \in \mathcal{A}$ and $\mathbb{E}(X \mathbb{1}_B) = \mathbb{E}(\tilde{X} \mathbb{1}_B)$ for all $B \in \sigma(\tilde{X})$, where $\sigma(\tilde{X})$ denotes the σ-algebra generated by $\tilde{X}$. Brunk (1965) showed that $\mathbb{E}(X \mid \mathcal{A})$ is almost surely unique and coincides with the classical conditional expectations if $\mathcal{A}$ is a σ-algebra. Conditional expectations given σ-lattices are closely connected to isotonicity as illustrated in Arnold and Ziegel (2023). In particular, for any integrable random variable X and random variable Z, there exists an increasing Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(X \mid \mathscr{L}(Z)) = f(Z)$. This result is analogous to the well-known factorization result for classical conditional expectations given σ-algebras, with the difference that, additionally, f has to be increasing.

Isotonic conditional laws can be defined in analogy to classical conditional laws. Specifically, the isotonic conditional law (ICL) of the random variable Y given $\mathcal{A}$, denoted $P_{Y|\mathcal{A}}$, is a Markov kernel from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\omega \mapsto P_{Y|\mathcal{A}}(\omega, (y, \infty))$ is a version of $\mathbb{P}(Y > y \mid \mathcal{A}) = \mathbb{E}(\mathbb{1}\{Y > y\} \mid \mathcal{A})$ for any $y \in \mathbb{R}$. Arnold and Ziegel (2023) show the existence and uniqueness of ICL. Equivalently, ICL emerges as the minimizer of an expected score, where the scoring rule may be taken from a large class of proper scoring rules that includes the CRPS.

We are particularly interested in ICL with respect to the σ-lattice generated by the forecast F. We call $B \subseteq \mathcal{P}(\mathbb{R})$ an upper set if $P \in B$ and $P \leq_{\mathrm{st}} Q$ implies $Q \in B$ for $Q \in \mathcal{P}(\mathbb{R})$, and we denote by $\mathcal{U}$ the family of all upper sets in $\mathcal{P}(\mathbb{R})$. For the forecast F, we define the σ-lattice generated by F as the family of all preimages of measurable upper sets under F, i.e., $\mathscr{L}(F) = \{F^{-1}(B) \mid B \in \mathcal{B}(\mathcal{P}(\mathbb{R})) \cap \mathcal{U}\} \subseteq \mathcal{F}$, where $\mathcal{B}(\mathcal{P}(\mathbb{R}))$ denotes the σ-algebra on $\mathcal{P}(\mathbb{R})$ with respect to the weak topology. For details, we refer the reader to Definition 3.1 of Arnold and Ziegel (2023).

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

The forecast F is called *isotonically calibrated* if F is almost surely equal to the isotonic conditional law of Y given $\mathscr{L}(F)$, i.e., $F = P_{Y|\mathscr{L}(F)}$ almost surely. Other relevant notions of calibration are described in Section 2.2.1 and Figure 2.1 summarizes relationships between them.

6.4.3 Population level decompositions

We now give generalizations of the empirical decompositions discussed in Sections 6.2 and 6.3 that apply at the population level. Recall that we consider the joint law $\mathbb{P}$ of the random tuple (F, Y) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As before, we let $\mathcal{P}(\mathbb{R})$ denote the class of the Borel probability measures on $\mathbb{R}$ that have a finite first moment. In the current and the subsequent subsections, we generally operate under the following regularity conditions.

Assumption 6.4.1. Let the marginal law F_{mg} of Y be such that $F_{\mathrm{mg}} \in \mathcal{P}(\mathbb{R})$, and suppose that

$$\mathbb{E} \int |x| \,\mathrm{d}F(x) < \infty. \tag{6.31}$$

In view of the kernel score representation of the CRPS (Gneiting and Raftery, 2007, eq. (21)), Assumption 6.4.1 implies that

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

where X and X' are independent random variables with law F. Similarly, it follows that $\mathbb{E}\,\mathrm{CRPS}(F_{\mathrm{mg}}, Y) < \infty$. Furthermore, the properties of isotonic and standard conditionals laws imply that $\mathbb{E}\,\mathrm{CRPS}(P_{Y|\mathscr{L}(F)}, Y) \le \mathbb{E}\,\mathrm{CRPS}(F, Y)$ and $\mathbb{E}\,\mathrm{CRPS}(P_{Y|F}, Y) \le \mathbb{E}\,\mathrm{CRPS}(F, Y)$, respectively. In this light, Assumption 6.4.1 ensures that $\mathbb{E}\,\mathrm{CRPS}(F, Y)$, $\mathbb{E}\,\mathrm{CRPS}(F_{\mathrm{mg}}, Y)$, $\mathbb{E}\,\mathrm{CRPS}(P_{Y|\mathscr{L}(F)}, Y)$, and $\mathbb{E}\,\mathrm{CRPS}(P_{Y|F}, Y)$ are finite.

The population version of the Candille–Talagrand decomposition at (6.12) is

$$\mathbb{E}\,\mathrm{CRPS}(F, Y) = \mathrm{MCB}_{\mathrm{CT}} - \mathrm{DSC}_{\mathrm{CT}} + \mathrm{UNC}_0, \tag{6.32}$$

where UNC_0 is defined at (6.30), and

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

Similarly, the population version of the isotonicity-based decomposition at (6.23) is

$$\mathbb{E}\,\mathrm{CRPS}(F, Y) = \mathrm{MCB}_{\mathrm{ISO}} - \mathrm{DSC}_{\mathrm{ISO}} + \mathrm{UNC}_0, \tag{6.33}$$

where

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

The decomposition at (6.33) is analogous to the theoretically preferred Candille–Talagrand decomposition at (6.32), except that the performance of the forecast F is compared with the isotonic conditional law $P_{Y|\mathscr{L}(F)}$ rather than the conditional law $P_{Y|F}$. The general decompositions at (6.32) and (6.33) reduce to (6.12) and (6.23), respectively, when $\mathbb{P}$ is the empirical distribution of the data in (6.7).

The population version of the Brier score based decomposition at (6.15) is

$$\mathbb{E}\,\mathrm{CRPS}(F, Y) = \mathrm{MCB}_{\mathrm{BS}} - \mathrm{DSC}_{\mathrm{BS}} + \mathrm{UNC}_0, \tag{6.34}$$

where

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

Similary, the population version of the quantile based based decomposition at (6.16) is

$$\mathbb{E}\,\mathrm{CRPS}(F, Y) = \mathrm{MCB}_{\mathrm{QS}} - \mathrm{DSC}_{\mathrm{QS}} + \mathrm{UNC}_0, \tag{6.35}$$

where

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

The properties of isotonic conditional expectations and isotonic conditional quantiles imply that $\mathbb{E}\int(\mathbb{P}(Y \le z \mid \mathscr{L}(F(z))) - \mathbb{1}\{Y \le z\})^2 \,\mathrm{d}z \le \mathbb{E}\,\mathrm{CRPS}(F, Y) < \infty$ and $\mathbb{E}\int_0^1 \mathrm{qs}_\alpha(q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))), Y)\,\mathrm{d}\alpha \le \mathbb{E}\,\mathrm{CRPS}(F, Y) < \infty$. The decompositions at (6.34) and (6.35) reduce to (6.15) and (6.16), respectively, when $\mathbb{P}$ is the empirical distribution of the data in (6.7).

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

$$\mu(A) = \mathbb{E}\left(\nu_F(A)\right) \tag{6.36}$$

and

$$\tau(A) = \mathbb{E}\left(\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p)\right), \tag{6.37}$$

respectively, where $A \in \mathcal{B}(0, 1)$ is any Borel set. We are now ready to state a population version of the Hersbach decomposition from Section 6.2.5.

Proposition 6.4.1. *Let Assumption 6.4.1 hold, and let μ and τ be the measures defined at* (6.36) *and* (6.37)*, respectively. Then τ is absolutely continuous with respect to μ; let f denote the respective Radon–Nikodym derivative. It holds that*

$$\mathbb{E}\,\mathrm{CRPS}(F, Y) = \mathrm{MCB_{HB}} - \mathrm{DSC_{HB}} + \mathrm{UNC}_0, \tag{6.38}$$

where UNC_0 is given at (6.30),

$$\mathrm{MCB_{HB}} = \int_0^1 (p - f(p))^2\,\mathrm{d}\mu(p), \quad \mathrm{DSC_{HB}} = \mathrm{UNC}_0 - \int_0^1 f(p)(1 - f(p))\,\mathrm{d}\mu(u) - \mathrm{MS},$$

and

$$\mathrm{MS} = \mathbb{E}\big[\mathbb{1}\{F(Y) = 0\}\,(F^{-1}(0+) - Y) + \mathbb{1}\{F(Y) > 0\}\,(2F(Y) - 1)(Y - F^{-1}(F(Y)))\big]. \tag{6.39}$$

Proof of Proposition 6.4.1. Following Appendix A in Candille and Talagrand (2005), we apply the change of variable $z \mapsto p = F(z)$ to demonstrate that $\mathbb{E}\,\mathrm{CRPS}(F, Y)$ can be represented as

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

where $S = \{z \in \mathbb{R} \mid (F(z) - \mathbb{1}\{Y \le z\})^2 > 0\}$. The indicator is essential, since if $F(Y) = 0$ then $\mathbb{1}\{F(Y) \le F(z)\} = 1$ and the integrals may not exist. We decompose

S into the disjoint sets $S_1 = S \cap \{z \in \mathbb{R} \mid F(z) > 0\}$ and $S_2 = S \cap \{z \in \mathbb{R} \mid F(z) = 0\} = \{z \in \mathbb{R} \mid Y \leq z, F(z) = 0\}$, and use the equivalence $\mathbb{1}\{F(Y) \leq F(z)\} - \mathbb{1}\{Y \leq z\} = \mathbb{1}\{Y > z, F(Y) = F(z)\}$ to show that

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

where MS is given at (6.39).

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

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

which yields the claimed decomposition. □

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

The following result shows that the population decomposition nests the modified empirical Hersbach decomposition. Therefore, we consider forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where for each $i = 1, \ldots, n$, F_i is a distribution with a finite number m_i of support points $x_1^i < \cdots < x_{m_i}^i$ and (cumulative) probability values $p_1^i < \cdots < p_{m_i}^i$, so that $F_i(x_\ell^i) = p_\ell^i$ for $\ell = 1, \ldots, m_i$. Let $0 < \hat{p}_1 < \ldots < \hat{p}_M = 1$ be the unique probability values from the set $\{p_\ell^i \mid i = 1, \ldots, n;\ \ell = 1, \ldots, m_i\}$. For

$i = 1, \ldots, n$ and $j = 1, \ldots, M-1$, we define

$$\sigma_j^i = \begin{cases} \ell & \text{if} \quad \hat{p}_j = p_\ell^i, \\ 0 & \text{if} \quad \hat{p}_j \notin \{p_1^i, \ldots, p_{m_i}^i\}. \end{cases}$$

Corollary 6.4.2. *Assume that $\mathbb{P}$ is the empirical measure of forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where each F_i is a distribution with finite support as described above. Then*

$$\mathrm{MCB}_{\mathrm{HB}} = \sum_{j=1}^{M-1} \hat{g}_j (\hat{p}_j - \hat{f}_j)^2 \tag{6.40}$$

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

$$\hat{g}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\sigma_j^i \neq 0\} \left(x^i_{\sigma_j^i + 1} - x^i_{\sigma_j^i} \right), \tag{6.41}$$

$$\hat{f}_j = \frac{1}{n \hat{g}_j} \sum_{i=1}^{n} \mathbb{1}\{F_i(y_i) \leq \hat{p}_j\} \mathbb{1}\{\sigma_j^i \neq 0\} \left(x^i_{\sigma_j^i + 1} - x^i_{\sigma_j^i} \right). \tag{6.42}$$

Proof of Corollary 6.4.2. For $i = 1, \ldots, n$, let ν_i be the image measure of F_i with respect to the Lebesgue measure, i.e.,

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

and thus, $\mu = \sum_{j=1}^{M-1} \delta_{\hat{p}_j} \hat{g}_j$, where $\hat{g}_j$ is given at (6.41). Therefore, for any $A \in \mathcal{B}(0,1)$, we have

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

We conclude that the Radon–Nikodym derivative of τ with respect to μ is $f(\hat{p}_j) = \hat{f}_j$ for $j = 1, \ldots, M-1$, where $\hat{f}_j$ is given at (6.42). □

To specialize Corollary 6.4.2 to the ensemble setting, let $m_i = m$ and $p_\ell^i = \ell/m$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, m-1$. Then $M = m$, $\hat{p}_j = j/m$, and the quantities in (6.17) and (6.41) coincide, as do the first quantity in (6.18) and that in (6.42).

The next result demonstrates that Proposition 6.4.1 subsumes the Hersbach–Lalaurette decomposition for strictly increasing forecast CDFs as given in Appendix A of Candille and Talagrand (2005).

Corollary 6.4.3. *Let Assumption 6.4.1 hold, and suppose that F^{-1} is almost surely absolutely continuous. Then $\mathrm{MS} = 0$ and the measure μ at* (6.36) *has density*

$$\gamma(p) = \mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right) \tag{6.43}$$

with respect to the Lebesgue measure on the unit interval. Furthermore, the measure τ at (6.37) *has Radon–Nikodym derivative defined by*

$$f(p) = \frac{1}{\gamma(p)}\,\mathbb{E}\left(\mathbb{1}\{F(Y) \le p\}\,\frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p)\right) \tag{6.44}$$

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

Proof of Corollary 6.4.3. Since F^{-1} is almost surely absolutely continuous, for any $0 < a < b < 1$, we have almost surely

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

That is, the random measure ν_F almost surely possesses a density $(\mathrm{d}/\,\mathrm{d}p)\,F^{-1}(p)$ with respect to the Lebesgue measure, and it follows that the measure μ has density γ at (6.43) with respect to the Lebesgue measure. Since for $A \in \mathcal{B}(0,1)$,

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

the density f of the measure τ with respect to μ is given as stated at (6.44). □

Relating to the case study on probabilistic quantitative precipitation forecasts in Section 6.5.1, the following Example derives the empirical Hersbach decomposition for probabilistic forecasts of a nonnegative quantity, assuming that the forecast distributions are mixtures of a point mass at zero and a strictly positive density on the positive halfline.

Example 6.4.1. Let $(F_1, y_1), \ldots, (F_n, y_n)$ be forecast–observation pairs for a nonnegative (possibly, censored) quantity, so that $y_i \ge 0$ for $i = 1, \ldots, n$. Suppose that, for $i = 1, \ldots, n$,

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

for some $0 \leq p_0^i < 1$ and a strictly positive continuous function $f_i : (0, \infty) \to \mathbb{R}_+$ with $\int_0^\infty f_i(t)\,\mathrm{d}t = 1 - p_0^i$. Then F_i^{-1} is absolutely continuous and has derivative $f_i(F_i^{-1}(p))^{-1}$ for $p \in (p_0^i, 1)$ and zero otherwise. Hence, $\overline{\mathrm{MCB}}_{\mathrm{HB}} = \int_0^1 (p - f(p))^2\,\gamma(p)\,\mathrm{d}p$ by Corollary 6.4.3, where

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

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

6.4.4 Properties of the decompositions

The population versions of the Candille–Talagrand, isotonicity-based, Brier score based, and quantile score based decompositions satisfy properties (P_1), (P_2), (P_4), and (P_5), and property (P_3) with auto-calibration, isotonic calibration, threshold calibration, and quantile calibration, respectively. The following theorem and its proof summarize and elaborate on property (P_3) and lend theoretical support to the use of the isotonicity-based decomposition. While in principle one would like to quantify miscalibration in terms of deviations from auto-calibration, as done by the Candille–Talagrand decomposition, the empirical version thereof is degenerate. By imposing the natural shape constraint of isotonicity between the assessed and the calibrated forecasts, a practically useful decomposition is obtained that does not rely on implementation choices, save for a possible choice of threshold values a and b in the modified CDFs $F^{(a,b)}$ at (6.26). The isotonicity-based decomposition quantifies miscalibration as deviation from isotonic calibration, which is closer to auto-calibration than threshold or quantile calibration as illustrated in Figure 2.1.

Theorem 6.4.4. *Under Assumption 6.4.1 the following statements hold.*

(a) The Candille–Talagrand decomposition at (6.32) *is exact and satisfies*

- $\mathrm{MCB}_{\mathrm{CT}} \geq 0$ *with equality if, and only if,* F *is auto-calibrated;*
- $\mathrm{DSC}_{\mathrm{CT}} \geq 0$ *with equality if, and only if,* $P_{Y|F} = F_{\mathrm{mg}}$ *almost surely.*

(b) The isotonicity-based decomposition at (6.33) *is exact and satisfies*

- $\mathrm{MCB}_{\mathrm{ISO}} \geq 0$ *with equality if, and only if,* F *is isotonically calibrated;*
- $\mathrm{DSC}_{\mathrm{ISO}} \geq 0$ *with equality if, and only if,* $P_{Y|\mathscr{L}(F)} = F_{\mathrm{mg}}$ *almost surely.*

(c) The Brier score based decomposition at (6.34) *is exact and satisfies*

- $\mathrm{MCB}_{\mathrm{BS}} \geq 0$ *with equality if, and only if, F is threshold calibrated;*
- $\mathrm{DSC}_{\mathrm{BS}} \geq 0$ *with equality if, and only if, for all $z \in \mathbb{R}$, $\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) = \mathbb{P}(Y \leq z)$ almost surely.*

(d) The quantile score based decomposition at (6.35) *is exact and satisfies*

- $\mathrm{MCB}_{\mathrm{QS}} \geq 0$ *with equality if F is quantile calibrated; conversely, if the random element $(Y, F^{-1}(\alpha))$ satisfies Assumption 6.1 in Arnold and Ziegel (2023) for all $\alpha \in (0,1)$ then $\mathrm{MCB}_{\mathrm{QS}} = 0$ implies quantile calibration of F;*
- $\mathrm{DSC}_{\mathrm{QS}} \geq 0$ *with equality if, and only if, for all $\alpha \in (0,1)$, $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))) = q_\alpha(Y)$ almost surely.*

Proof of Theorem 6.4.4. Concerning part (a), we consider the Brier score based decomposition of $\overline{\mathrm{CRPS}}$ and apply Fubini's theorem to obtain

$$\mathrm{MCB}_{\mathrm{CT}} = \int \Big(\mathbb{E}\big(F(z) - \mathbb{1}\{Y \leq z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y \leq z \mid F) - \mathbb{1}\{Y \leq z\}\big)^2 \Big) \, \mathrm{d}z, \tag{6.45}$$

$$\mathrm{DSC}_{\mathrm{CT}} = \int \Big(\mathbb{E}\big(F_{\mathrm{mg}}(z) - \mathbb{1}\{Y \leq z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y \leq z \mid F) - \mathbb{1}\{Y \leq z\}\big)^2 \Big) \, \mathrm{d}z. \tag{6.46}$$

Recall that for any $z \in \mathbb{R}$, the expectation $\mathbb{E}\left(\mathbb{1}\{Y \leq z\} - p\right)^2$ is minimized by $\mathbb{P}(Y \leq z \mid F)$ over all $\sigma(F)$-measurable random variables p, and this minimizer is $\mathbb{P}$-almost surely unique. Since $F(z)$ and the constant $F_{\mathrm{mg}}(z)$ are $\sigma(F)$-measurable, it follows from (6.45) and (6.46) that $\mathrm{MCB}_{\mathrm{CT}} \geq 0$ and $\mathrm{DSC}_{\mathrm{CT}} \geq 0$, respectively. Equality in (6.45) holds if, and only if, F is auto-calibrated. Equality in (6.46) holds if, and only if, $P_{Y|F} = F_{\mathrm{mg}}$, i.e., $\mathbb{P}(Y \leq z \mid F) = F_{\mathrm{mg}}(z)$ for all $z \in \mathbb{R}$.

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

$$\mathrm{MCB}_{\mathrm{ISO}} = \int \Big(\mathbb{E}\big(\bar{F}(z) - \mathbb{1}\{Y > z\}\big)^2 - \mathbb{E}\left(\mathbb{P}(Y > z \mid \mathscr{L}(F)) - \mathbb{1}\{Y > z\}\right)^2 \Big) \, \mathrm{d}z, \tag{6.47}$$

$$\mathrm{DSC}_{\mathrm{ISO}} = \int \Big(\mathbb{E}\big(\bar{F}_{\mathrm{mg}}(z) - \mathbb{1}\{Y > z\}\big)^2 - \mathbb{E}\big(\mathbb{P}(Y > z \mid \mathscr{L}(F)) - \mathbb{1}\{Y > z\}\big)^2 \Big) \, \mathrm{d}z, \tag{6.48}$$

where $\bar{F}(z) = 1 - F(z)$, and $\bar{F}_{\mathrm{mg}}(z) = 1 - F_{\mathrm{mg}}(z)$. Recall that for any $z \in \mathbb{R}$, the expectation $\mathbb{E}(\mathbb{1}\{Y > z\} - p)^2$ is minimized by $\mathbb{P}(Y > z \mid \mathscr{L}(F))$ over all $\mathscr{L}(F)$-measurable random variables p, and the minimizer is $\mathbb{P}$-almost surely unique. Since $\bar{F}(z)$ and the constant $\bar{F}_{\mathrm{mg}}(z)$ are $\mathscr{L}(F)$-measurable, it follows directly that $\mathrm{MCB}_{\mathrm{ISO}} \geq 0$ and $\mathrm{DSC}_{\mathrm{ISO}} \geq 0$. Equality in (6.47) holds if, and only if, F is isotonically calibrated, and equality in (6.48) holds if, and only if, $P_{Y|\mathscr{L}(F)} = F_{\mathrm{mg}}$.

To demonstrate part (c), it suffices to observe from Arnold and Ziegel (2023, Lemma 5.4) that threshold calibration is equivalent to $\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) = F(z)$ for $z \in \mathbb{R}$. The rest of the argument is analogous to the above.

Finally, for part (d), recall that for $\alpha \in (0,1)$, a random variable is a conditional quantile $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha)))$ if, and only if, it minimizes $\mathbb{E}\,\mathrm{QS}_\alpha(X,Y)$ over all $\mathscr{L}(F^{-1}(\alpha))-$ measurable random variables X, see Arnold and Ziegel (2023). It follows that $\mathrm{MCB}_{\mathrm{QS}} \geq 0$ and $\mathrm{DSC}_{\mathrm{QS}} \geq 0$. Assume that F is quantile calibrated; then $q_\alpha\big(Y \mid \mathscr{L}\big(F^{-1}(\alpha)\big)\big) = F^{-1}(\alpha)$ for $\alpha \in (0,1)$ and hence $\mathrm{MCB}_{\mathrm{QS}} = 0$. Conversely, if $\mathrm{MCB}_{\mathrm{QS}} = 0$ then Fubini's theorem implies

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

Since the integrand is non-negative, it follows that $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = F^{-1}(\alpha)$ for almost all $\alpha \in (0,1)$ and, hence, there exists a Lebesgue null set $N \subseteq (0,1)$ with $q_\alpha(Y \mid \mathscr{L}(F^{-1}(\alpha))) = F^{-1}(\alpha)$ for all $\alpha \in (0,1) \setminus N$. Assume for a contradiction that $N \neq \emptyset$ and consider $\alpha_0 \in N$. Choose $(\alpha_n)_{n\in\mathbb{N}} \subseteq (0,1) \setminus N$ with $\alpha_n \uparrow \alpha_0$ as $n \to \infty$. Since $F^{-1}(\alpha_n) \to F^{-1}(\alpha_0)$ almost surely and $\mathrm{qs}_{\alpha_n}(\cdot, y) \to \mathrm{qs}_{\alpha_0}(\cdot, y)$ pointwise for any $y \in \mathbb{R}$, it follows that $\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y) \to \mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ almost surely, and hence, $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y) \to \mathbb{E}\,\mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ by dominated convergence. Analogously, $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(X, Y) \to \mathbb{E}\,\mathrm{QS}_{\alpha_0}(X, Y)$ for $X = q_{\alpha_0}(Y \mid \mathscr{L}(F^{-1}(\alpha_0)))$ and, hence, $\mathbb{E}\,\mathrm{qs}_{\alpha_0}(X, Y) \geq \mathbb{E}\,\mathrm{qs}_{\alpha_0}(F^{-1}(\alpha_0), Y)$ since $\mathbb{E}\,\mathrm{qs}_{\alpha_n}(X, Y) \geq \mathbb{E}\,\mathrm{qs}_{\alpha_n}(F^{-1}(\alpha_n), Y)$ for all $n \in \mathbb{N}$. This shows that $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big)$ is an α-quantile of F for $\alpha \in (0,1)$. By construction in Section 6 of Arnold and Ziegel (2023), $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big)$ is the smallest possible minimizer of $\mathbb{E}\,\mathrm{qs}_\alpha(X, Y)$, so it coincides with $F^{-1}(\alpha)$ for all $\alpha \in (0,1)$ and, hence, $N = \emptyset$. Clearly, $\mathrm{DSC}_{\mathrm{QS}} = 0$ if $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = q_\alpha(Y)$ for $\alpha \in (0,1)$. Conversely, if $\mathrm{DSC}_{\mathrm{QS}} = 0$ then $q_\alpha\big(Y \mid \mathscr{L}(F^{-1}(\alpha))\big) = q_\alpha(Y)$ for $\alpha \in (0,1)$. □

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

Corollary 6.4.5. *Under Assumption 6.4.1, an auto-calibrated forecast yields* $\mathrm{MCB}_{\mathrm{CT}} = \mathrm{MCB}_{\mathrm{ISO}} = \mathrm{MCB}_{\mathrm{BS}} = \mathrm{MCB}_{\mathrm{QS}} = 0.$

Corollary 6.4.6. *Under Assumption 6.4.1, it holds that*

$$\mathbb{E}\,\mathrm{CRPS}(F,Y) \geq \mathrm{MCB}_{\mathrm{CT}} \geq \mathrm{MCB}_{\mathrm{ISO}} \geq \max\{\mathrm{MCB}_{\mathrm{BS}}, \mathrm{MCB}_{\mathrm{QS}}\}. \tag{6.49}$$

Proof of Corollary 6.4.6. For any $z \in \mathbb{R}$, $P_{Y|F}(\cdot,(z,\infty))$ minimizes $\mathbb{E}(p - \mathbb{1}\{Y > z\})^2$ over all $\sigma(F)$-measurable random variables p, and hence, also over all $\mathscr{L}(F)$-measurable random variables since any $\mathscr{L}(F)$-measurable random variable is also $\sigma(F)$-measurable, see Arnold and Ziegel (2023, Lemma 3.1). Thus, we apply Fubini to derive

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

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

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

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

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

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

Importantly, while formulated at the population level, the above results apply to the empirical versions of the decompositions, by identifying the joint distribution $\mathbb{P}$ of the tuple (F,Y) with the empirical law of the data at (6.7). In particular, the relations in (6.49) nest the respective inequalities (6.24) for the empirical decompositions. For the isotonicity-based decomposition, if modified CDFs $F^{(a,b)}$ are used the results apply to the latter, and we refer to (6.28) for relationships to the respective components computed on the original CDFs.

Finally, we consider the Hersbach decomposition from Proposition 6.4.1, which struggles to satisfy the desirable properties from Section 6.4.1. By definition, properties (P_1) and (P_5) hold. The miscalibration component is clearly nonnegative. However, $\mathrm{DSC}_{\mathrm{HB}}$ may be negative as in Example 6.C.3, i.e., property (P_2) is violated. Moreover, the example in the proof of Proposition 6.2.3 shows that the Hersbach decomposition fails to satisfy (P_4). Concerning (P_3), Hersbach (2000) and Candille and Talagrand (2005) argue that the Hersbach reliability component is closely related to the rank histogram and hence one might expect that $\mathrm{MCB}_{\mathrm{HB}} = 0$ if, and only if, F is probabilistically calibrated. However, the examples in Section 6.C.4 and 6.C.5 show that probabilistic calibration is neither sufficient nor necessary for $\mathrm{MCB}_{\mathrm{HB}} = 0$. The following proposition collects calibration properties in relation to the Hersbach decomposition.

Proposition 6.4.7. *Let Assumption 6.4.1 hold and consider the population version of the Hersbach decomposition at* (6.38).

(a) *If $Y \in \mathrm{supp}(F)$ almost surely, then $\mathrm{MS} = 0$, where MS is defined at* (6.39).

(b) *For an auto-calibrated forecast, it holds that $\mathrm{MS} = \mathrm{MCB}_{\mathrm{HB}} = 0$.*

(c) *Suppose that F belongs to a location family, i.e., for all $x \in \mathbb{R}$, $F(x) = F_0(x-\mu)$ for some $F_0 \in \mathcal{P}(\mathbb{R})$ and random location μ. Suppose furthermore that F_0 has no jumps and F_0^{-1} is absolutely continuous. Then $\mathrm{MCB}_{\mathrm{HB}} = 0$ if F is probabilistically calibrated.*

Proof of Proposition 6.4.7. The claim in part (a) follows from the definition of MS at (6.39). For part (b), suppose that F is auto-calibrated. Then $Y \in \mathrm{supp}(F)$ almost surely and hence $\mathrm{MS} = 0$ by part (a). The tower property implies for any $A \in \mathcal{B}(0,1)$ that

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

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

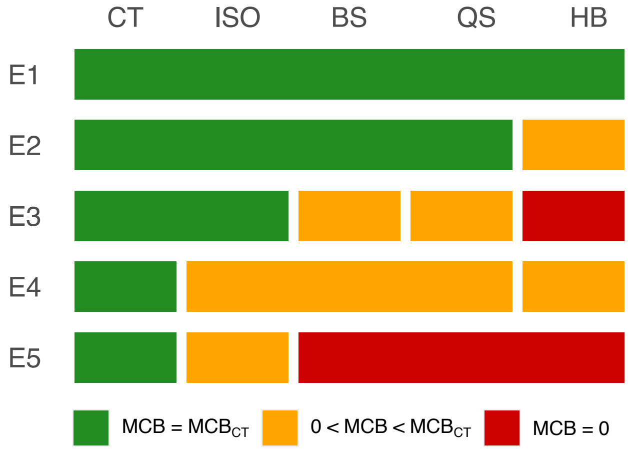

Figure 6.1: The graphic indicates for the population level examples E1, $\ldots$, E5 in Appendix 6.C whether the $\mathrm{MCB}_\bullet$ term, where $\bullet$ stands for CT, ISO, BS, QS, or HB, respectively, agrees with the theoretically preferred quantity $\mathrm{MCB}_{\mathrm{CT}}$ (green), is smaller than $\mathrm{MCB}_{\mathrm{CT}}$ but remains positive (orange), or deceptively equals zero (red). Connected segments indicate equality of corresponding terms. For analytic results, see Table 6.2.

$F^{-1}(B) = \{x \in \mathbb{R} \mid F(x) \in B\} = \emptyset$ and hence $\nu_F(B) = 0$ almost surely. That is, $\nu_F(\{p \in (0,1) : F(F^{-1}(p)) > p\} = 0$ almost surely and thus

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

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

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

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

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

In Appendix 6.C we compare the different types of decompositions in a number of analytic examples at the population level. Figure 6.1 summarizes how the respective miscalibration terms relate to the theoretically preferred $\mathrm{MCB}_{\mathrm{CT}}$ component.

6.5 Case studies

We now illustrate the use of the isotonicity-based decomposition from Section 6.3 in case studies on weather forecasts and benchmark regression tasks from machine learning, both of which are also discussed in Chapter 5. For simplicity, we use an abbreviated notation for the components of the mean score $\overline{\mathrm{CRPS}}$ throughout this section, namely, $\overline{\mathrm{MCB}} = \overline{\mathrm{MCB}}_{\mathrm{ISO}}$, $\overline{\mathrm{DSC}} = \overline{\mathrm{DSC}}_{\mathrm{ISO}}$, and $\overline{\mathrm{UNC}} = \overline{\mathrm{UNC}}_0$, respectively. Note the opposite orientation of $\overline{\mathrm{MCB}}$ and $\overline{\mathrm{DSC}}$, in that higher $\overline{\mathrm{DSC}}$ corresponds to better discrimination ability, whereas lower $\overline{\mathrm{MCB}}$ indicates better calibration.

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

6.5.1 Probabilistic quantitative precipitation forecasts

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

Following Henzi et al. (2021), we consider forecasts and observations for 24-hour accumulated precipitation from 6 January 2007 to 1 January 2017 at Brussels, Frankfurt, London, and Zurich in millimeters. The 52 member raw ensemble (ENS) forecast operated by the European Centre for Medium-Range Weather Forecasts comprises a high resolution member, a control member at lower resolution, and 50 perturbed members at the same lower resolution but with perturbed initial conditions (Molteni et al., 1996). We use data from 2007 to 2014 to train the postprocessing techniques Bayesian model averaging (BMA; Sloughter et al., 2007), ensemble model output statistics (EMOS; Scheuerer, 2014), heteroscedastic censored logistic regression (HCLR; Messner et al., 2014) and two versions, IDR$_{\text{cw}}$ and IDR$_{\text{st}}$, of isotonic distributional regression (IDR; Henzi et al., 2021), where IDR$_{\text{cw}}$ is documented in Henzi et al. (2021) and IDR$_{\text{st}}$ uses the stochastic order on the ensemble CDFs. For further implementation details we refer the reader to Henzi et al. (2021). The years 2015 and 2016 form the evaluation period.

The ENS and IDR forecast distributions have finite support and we apply the isotonicity-based decomposition of $\overline{\text{CRPS}}$ in its pure form from Section 6.3.1. For the other forecasts, which employ mixtures of a point mass at zero (for no precipitation) and a density at positive accumulations as predictive distributions, we fix $a = 0$ and use Algorithm 2 to determine the upper bound b, which generally is identical to, or very slightly higher than, the highest accumulation observed in the test data; then we compute stochastic order relations on an equidistant grid of size 5000 over $[a, b]$ and apply the isotonicity-based decomposition in its approximate form from Section 6.3.2.

The respective $\overline{\text{MCB}}$–$\overline{\text{DSC}}$ plots for Brussels, Frankfurt, London, and Zurich are shown in Figure 6.2. We note an increase of the mean score $\overline{\text{CRPS}}$ values with the prediction horizon, which is due to a decrease in discrimination ability. The raw ensemble (ENS) forecasts discriminate very well, but are poorly calibrated. The postprocessing methods yield considerable improvement in $\overline{\text{CRPS}}$, subject to a trade-off between $\overline{\text{MCB}}$ and $\overline{\text{DSC}}$. The EMOS and HCLR techniques, which employ inflexible parametric densities with fixed shape, excel in terms of discrimination, but lack in calibration. In contrast, the BMA and IDR techniques, which are much more flexible, are better calibrated, but inferior in terms of discrimination ability.

6.5.2 Benchmark regression problems from machine learning

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

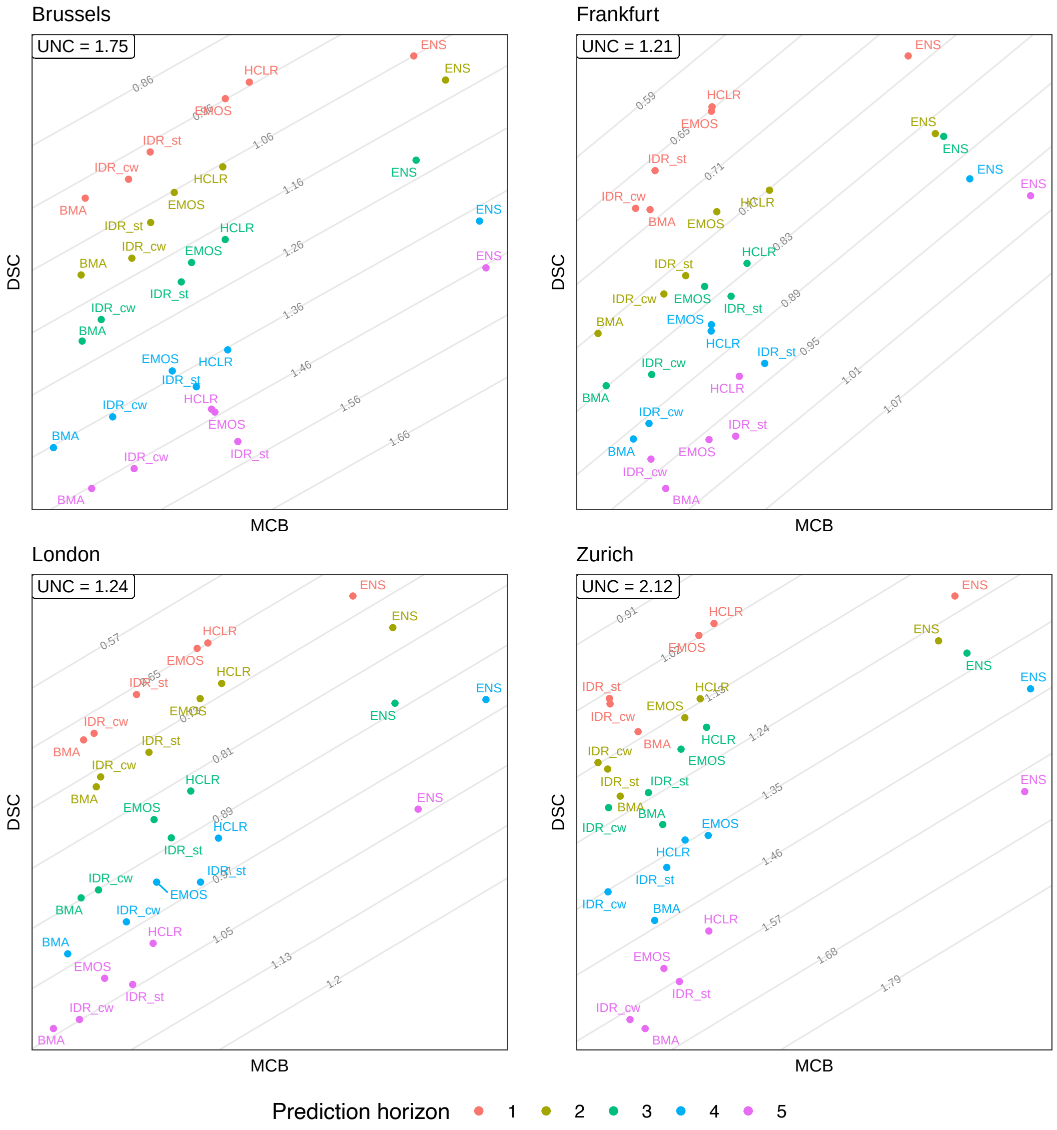

Figure 6.2: $\overline{\mathrm{MCB}}$–$\overline{\mathrm{DSC}}$ plots for forecasts of 24-hour accumulated precipitation at Brussels, Frankfurt, London, and Zurich, at prediction horizons of one to five days ahead. The mean score $\overline{\mathrm{CRPS}}$ is constant along the parallel lines and shown in the unit of millimeters. Acronyms are defined in the text, and details of the forecast methods are documented in Henzi et al. (2021, Section 5).

and its isotonicity-based decomposition into $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$ in this context.

We adopt the benchmark regression tasks setting originally proposed by Hernandéz-Lobato and Adams (2015) and consider the datasets and methods from the middle block of Table 6 in Chapter 5, except that we skip results for the Naval and Year datasets, for which there are missing entries. The experimental setting is based on single-valued output from a neural network, which learns a regression function based on a collection of covariates or features. In this setting, Walz (2023) compares competing methods for uncertainty quantification, including the popular Monte Carlo Dropout approach (MC Dropout; Gal and Ghahramani, 2016) and a scalable Laplace approximation based technique (Laplace; Immer et al., 2021; Ritter et al., 2018) that operate within the neural network learning pipeline. Their competitors include output-based methods that learn on training data of previous single-valued model output and outcomes only, without accessing feature values, namely, the Single Gaussian technique, conformal prediction (CP; Vovk et al., 2020b), and the EasyUQ technique, which is based on IDR (Henzi et al., 2021). Furthermore, we consider smoothed versions of the discrete CP and EasyUQ distributions, termed Smooth CP and Smooth EasyUQ, respectively. For implementation details, we refer the reader to Walz (2023).

The CP and EasyUQ distributions have finite support, and the Single Gaussian incurs normal distribution with a fixed variance, but varying mean. For these three methods, we use the isotonicity-based decomposition of $\overline{\mathrm{CRPS}}$ in the standard form from Section 6.3.1. The Laplace method also employs normal distributions, but with varying mean and variances. The MC Dropout technique yields mixtures of normal distributions, and the Smooth CP and Smooth EasyUQ distributions are mixtures of Student-t distributions (or normal distributions as a limit case). For these methods, we use the approximations described in Section 6.3.2.

The $\overline{\mathrm{MCB}}$–$\overline{\mathrm{DSC}}$ plots in Figure 6.3 illustrate the mean score $\overline{\mathrm{CRPS}}$ and the $\overline{\mathrm{MCB}}$, $\overline{\mathrm{DSC}}$, and $\overline{\mathrm{UNC}}$ components for the eight datasets and seven methods, respectively. The MC Dropout technique yields predictive distributions that are poorly calibrated, a finding that is well documented in the machine learning literature (Gawlikowski et al., 2023), though with high discrimination ability. The predictive distributions generated by the Laplace method trade better calibration for diminished discrimination ability. The simplistic Single Gaussian technique performs surprisingly well, typically with both the $\overline{\mathrm{MCB}}$ and the $\overline{\mathrm{DSC}}$ component being small relative to the competitors. The EasyUQ and CP distributions generally are well calibrated, with low $\overline{\mathrm{MCB}}$ components throughout, and often superior overall performance. Smoothing of the dis-

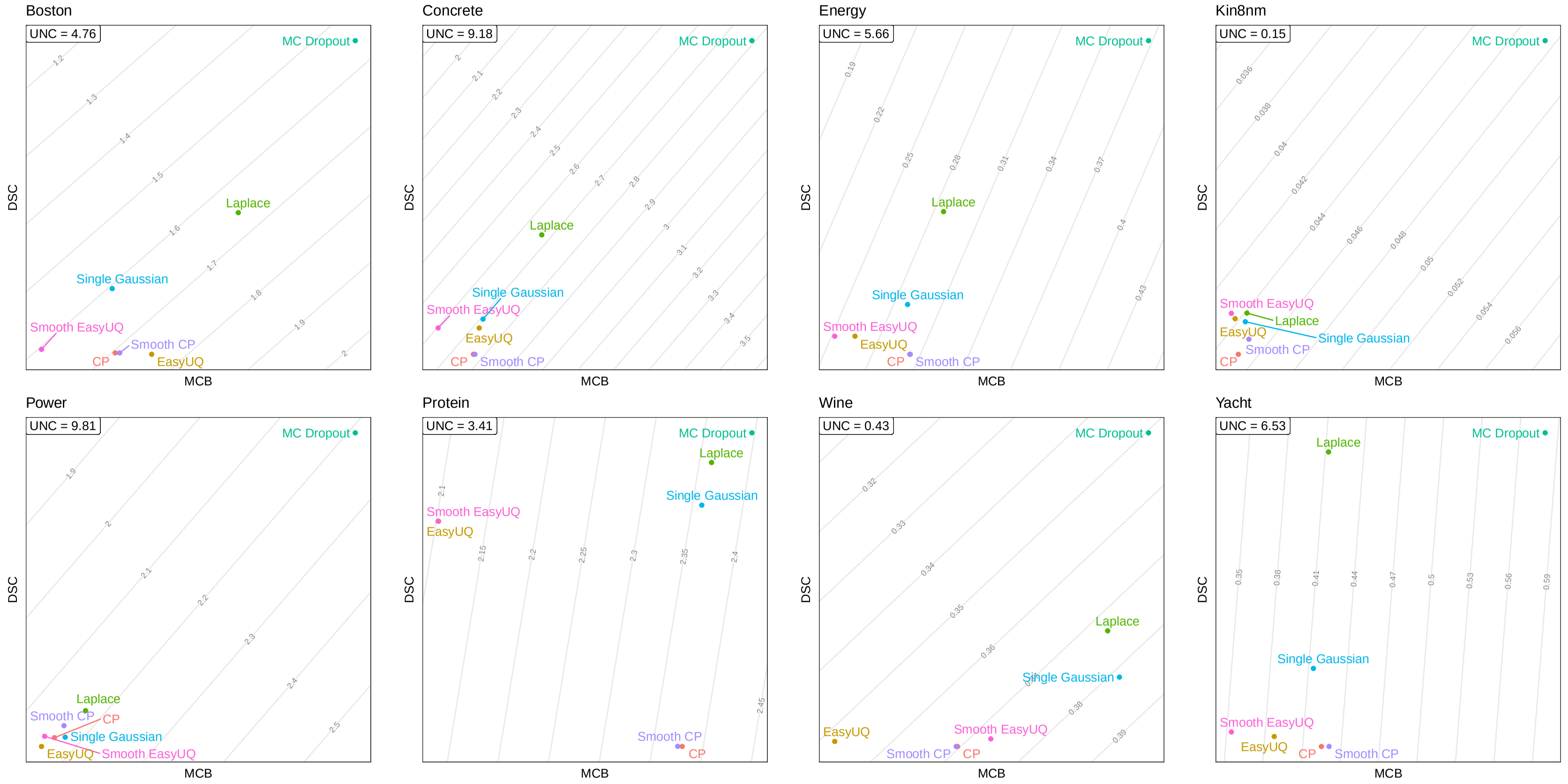

Figure 6.3: $\overline{\mathrm{MCB}}$–$\overline{\mathrm{DSC}}$ plots for methods of uncertainty quantification for neural network based regression from the middle block in Table 6 of Chapter 5. The mean score $\overline{\mathrm{CRPS}}$ is constant along the parallel lines.

crete EasyUQ and CP distributions has only small effects. The only exception is for the EasyUQ forecast for the Wine dataset, which has only ten unique outcomes that correspond to quality levels, thus favoring the discrete basic EasyUQ distributions, which place all probability mass on this small set of outcomes.

6.6 Discussion

In line with the general idea of the CORP approach of Dimitriadis et al. (2023) and Gneiting and Resin (2023), we have developed an isotonicity-based decomposition of the mean score $\overline{\mathrm{CRPS}}$. Both theoretically and computationally, the isotonicity-based decomposition serves as an attractive alternative to the Candille–Talagrand decomposition, which is of theoretical appeal, but yields degenerate decompositions in practice. Remarkably, Proposition 6.3.2 ensures that theoretical guarantees for the standard implementation from Section 6.3.1 very nearly carry over to the approximate implementation described in Section 6.3.2. Code in R (R Core Team, 2021) for the computation of the isotonicity-based decomposition and replication materials are available at `https://github.com/evwalz/isodisregSD` and `https://github.com/evwalz/paper_isocrpsdeco`, respectively.

Due to its linear computational complexity, the Hersbach decomposition is a viable option for decomposing $\overline{\mathrm{CRPS}}$ for ensemble forecasts with a moderate number m of members, even when the size n of the evaluation set at (6.7) is very large and the isotonicity-based approach with its quadratic complexity is not feasible. We recommend that it be used in the modified form described in Section 6.2.5, which allows for extensions beyond the case of ensemble forecasts. A useful facet of the Hersbach decomposition is that it applies to general (nonnegatively) weighted sums (rather than simple averages only) of CRPS scores (Hersbach, 2000). The isotonicity-based decomposition generalizes to weighted sums as well, as the theoretical guarantees for IDR (Henzi et al., 2021) continue to apply in weighted case, and software developed by Alexander Henzi (`https://github.com/AlexanderHenzi/isodistrreg`) handles the extension. We leave details to future work.

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

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

and

$$\mathrm{CRPS}_v(F,y) = \int_{0}^{1} v(\alpha)\,\mathrm{qs}_\alpha(F^{-1}(\alpha),y)\,\mathrm{d}\alpha,$$

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

However, the isotonicity-based approach fails if a mean of logarithmic scores (Gneiting and Raftery, 2007, Section 4.1) is sought to be decomposed, for the logarithmic score, which allows for the comparison of density forecasts only, cannot be applied to the discrete IDR distributions. While in principle isotonic recalibration by IDR, on which isotonicity-based decompositions are based, could be replaced by recalibration with other methods, it is not at all evident what type of technique ought to be used, and we are unaware of any such method that would share the optimality properties of IDR that underlie the theoretical guarantees enjoyed by the isotonicity-based approach.

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

Appendix 6.A Technical details for the Brier score and quantile score based decompositions

In this appendix we describe the Brier score (BS) and quantile score (QS) based decompositions from Sections 6.2.3 and 6.2.4 for the mean score $\overline{\mathrm{CRPS}}$ of the forecast–observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$ at (6.7). Both decompositions build on a general version of the pool-adjacent-violators (PAV) algorithm for nonparametric isotonic regression (Ayer et al., 1955). While historically work on the PAV algorithm has focused on the mean functional (Barlow et al., 1972; Robertson et al., 1988; de Leeuw et al., 2009), the algorithm yields optimal isotonic fits under any identifiable functional; see, e.g., Jordan et al. (2022) and Gneiting and Resin (2023, Section 3.1).

6.A.1 Brier score based decomposition

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

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

as proposed by Dimitriadis et al. (2021), namely,

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

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

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

Integration of the $\overline{\mathrm{MCB}}_{\mathrm{BS},z}$, $\overline{\mathrm{DSC}}_{\mathrm{BS},z}$ and $\overline{\mathrm{UNC}}_{\mathrm{BS},z}$ components over $z \in \mathbb{R}$ yields the Brier score based score components and decomposition at (6.14) and (6.15), respectively.

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

Proof of Proposition 6.2.1. We note that

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

which implies that (E_5) is satisfied. Property (E_1) is immediate. Dimitriadis et al. (2021) show that $\overline{\mathrm{MCB}}_{\mathrm{BS},z}$ and $\overline{\mathrm{DSC}}_{\mathrm{BS},z}$ are nonnegative for all $z \in \mathbb{R}$ and thus (E_2) is satisfied. Example 6.C.3 implies that the decomposition is not degenerate, so (E_3) is satisfied. Finally, suppose that $F_1 = \cdots = F_n$. Then for each $z \in \mathbb{R}$, the PAV algorithm for the mean functional on $\xi_1(z), \ldots, \xi_n(z)$ with respect to the order induced by $F_1(z) = \cdots = F_n(z)$ yields the constant calibrated forecast $\hat{F}_{\mathrm{mg}}(z)$. Hence $\overline{\mathrm{DSC}}_{\mathrm{BS}} = 0$, so that ($E_4$) is satisfied. □

Remark 6.A.1. The functions $\acute{F}_1, \ldots, \acute{F}_n$ are not necessarily increasing and hence they generally fail to be CDFs. For instance, let $n = 2$ and $z < z'$. If $F_1(z) < F_2(z)$, $F_1(z') = F_2(z')$ and $y_2 \le z < z' < y_1$, then $\acute{F}_2(z) = 1 > 1/2 = \acute{F}_2(z')$, so $\acute{F}_2$ is not increasing.

6.A.2 Quantile score based decomposition

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

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

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

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

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

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

Integration of the $\overline{\mathrm{MCB}}_{\mathrm{QS},\alpha}, \overline{\mathrm{DSC}}_{\mathrm{QS},\alpha}$ and $\overline{\mathrm{UNC}}_{\mathrm{QS},\alpha}$ components over $\alpha \in (0,1)$ yields the quantile score based decomposition at (6.16).

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

Proof of Proposition 6.2.2. In analogy to the proof of Proposition 6.2.1, we find that

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

and hence (E_5) is satisfied. Property (E_1) is clear by definition. Theorem 3.3 of Gneiting and Resin (2023) implies that $\overline{\mathrm{MCB}}_{\mathrm{QS},\alpha}$ and $\overline{\mathrm{DSC}}_{\mathrm{QS},\alpha}$ are nonnegative for all $\alpha \in (0,1)$ and thus (E_2) is satisfied. Example 6.C.3 shows that the decomposition is not degenerate, i.e., (E_3) is satisfied. Finally, suppose that $F_1 = \cdots = F_n$. Then for each $\alpha \in (0,1)$, applying the PAV algorithm on $y_1, \ldots, y_n$ with respect to the order induced by $F_1^{-1}(\alpha) = \cdots = F_n^{-1}(\alpha)$ yields the constant calibrated forecast $\grave{F}^{-1}(\alpha) = \hat{F}_{\mathrm{mg}}^{-1}(\alpha)$ and hence $\overline{\mathrm{DSC}}_{\mathrm{QS}} = 0$, i.e., ($E_4$) is satisfied. □

Remark 6.A.2. In analogy to the statements in Remark 6.A.1, the functions $\grave{F}_1^{-1}, \ldots, \grave{F}_n^{-1}$ are not necessarily increasing and hence may not be quantile functions. For example, let $n = 2$ and $\alpha < \alpha' < 1/2$, and suppose that $y_1 < y_2$, $F_1^{-1}(\alpha) < F_2^{-1}(\alpha)$, and $F_1^{-1}(\alpha') = F_2^{-1}(\alpha')$. Then $\grave{F}_2^{-1}(\alpha) = y_2 > y_1 = \grave{F}_2^{-1}(\alpha')$ whence $\grave{F}_2^{-1}$ is not increasing.

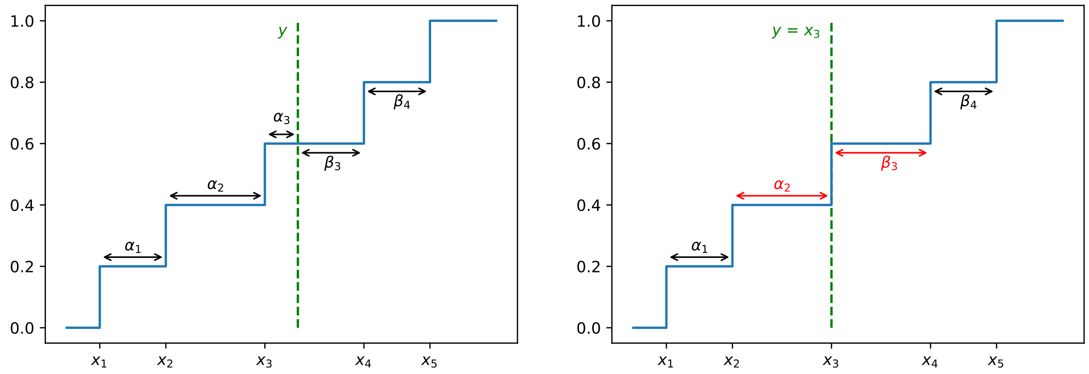

Figure 6.4: Adaptation of Figure 2 from Hersbach (2000) with the empirical CDF of $x_1 < \cdots < x_5$ and outcome y. Hersbach (2000) assumes that $y \notin \{x_1, \ldots, x_5\}$ and divides the quantity $x_{\ell+1} - x_\ell$ for $\ell = 1, \ldots, m-1$ into α_ℓ and β_ℓ, as illustrated in the left panel. When $y = x_3$ the original decomposition sets $\alpha_2 = \beta_3 = 0$. However, according to display (26) in Hersbach (2000), if $y \uparrow x_3$ then $\alpha_2 \to x_3 - x_2$, $\beta_2 \to 0$, and $\beta_3 = x_4 - x_3$, and if $y \downarrow x_3$ then $\alpha_2 = x_3 - x_2$, $\alpha_3 \to 0$, and $\beta_3 \to x_4 - x_3$. This suggests that $\alpha_2 = x_3 - x_2$, $\alpha_3 = 0$, $\beta_2 = 0$, and $\beta_3 = x_4 - x_3$ when $y = x_3$, as indicated in the right panel and in accordance with the quantity $\bar{f}_3$ in the modified Hersbach decomposition.

Appendix 6.B Technical details for the original and modified Hersbach decompositions

As in Section 6.2.5, we consider a collection of the form at (6.7) of forecast–outcome pairs $(F_1, y_1), \ldots, (F_n, y_n)$, where for $i = 1, \ldots, n$, the forecast F_i is the empirical CDF of a fixed number m of numbers $x_1^i \le \cdots \le x_m^i$. Hersbach (2000) implicitly assumes that $y_i \notin \{x_1^i, \ldots, x_m^i\}$ for $i = 1, \ldots, n$. If this condition is not satisfied, the extension of the original Hersbach decomposition at (6.19), which is implemented in the R function `crpsDecomposition` from the verification package (`https://rdrr.io/cran/verification/`), is problematic. Our suggested modified Hersbach decomposition at (6.21) resolves this issue, as illustrated graphically in Figure 6.4.

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

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

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

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

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

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

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

In contrast, we let

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

where $\bar{f}_\ell = (1/n) \sum_{i=1}^n \bar{f}_\ell^i$ with $\bar{f}_\ell^i = (1/\bar{g}_\ell)\, \mathbb{1}\{y_i < x_{\ell+1}^i\}(\alpha_\ell^i + \beta_\ell^i)$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, m-1$. In other words, Hersbach (2000) includes terms for $l = 0$ and $l = m$ in the miscalibration component and compares the nominal level p_ℓ with the quantity $\bar{o}_\ell$, which approximates the frequency of an outcome below the midpoint of bin l. In contrast, we omit the outer terms and compare p_ℓ with $\bar{f}_\ell$, which approximates the frequency of an outcome below the right endpoint of bin l.

Proof of Proposition 6.2.3. By definition, both decompositions are exact and the uncertainty component $\overline{\mathrm{UNC}}_0$ depends only on the outcomes, i.e., (E_1) and (E_5) are satisfied. Example 6.C.3 shows that (E_3) is satisfied, and that (E_2) fails to hold for the modified Hersbach decomposition. Consider the sample $(F, y_1), (F, y_2)$ with $F = (\delta_{-1/2} + \delta_{1/2})/2$, $y_1 = -1/6$ and $y_2 = 1/6$. Then $\overline{\mathrm{CRPS}} = 1/4$ and $\overline{\mathrm{UNC}}_0 = 1/12$. Moreover, $\bar{g}_1 = 1$, $\bar{g}_0 = \bar{g}_2 = 0$, $\bar{o}_1 = 1/2$, $\bar{o}_0 = \bar{o}_2 = 0$, and $\bar{f}_1 = 1$. Thus $\overline{\mathrm{MCB}}_{\mathrm{HBo}} = 0$, $\overline{\mathrm{MCB}}_{\mathrm{HB}} = 1/4$, $\overline{\mathrm{DSC}}_{\mathrm{HBo}} = -1/6$, and $\overline{\mathrm{DSC}}_{\mathrm{HB}} = 1/12$. This demonstrates that the original Hersbach decomposition does not satisfy (E_2) and (E_4) and that (E_4) fails to hold for the modified decomposition as well. Numerical examples in Hersbach (2000) show that (E_3) is satisfied for the original Hersbach decomposition. □

Table 6.2: Analytic form of the various different types of decomposition in population level examples E.1, . . . , E.5. For details and supporting calculations see the text.

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

Appendix 6.C Analytic examples at the population level

In this section we compare the population level decompositions from Section 6.4 in a number of examples in the prediction space setting. Table 6.2 collects and summarizes the analytic forms of the decomposition components in these examples. Assumption 6.4.1 is satisfied throughout.

6.C.1 Auto-calibrated Gaussian

In this example, the predictive distribution F is Gaussian with mean μ_i and standard deviation $\sigma_i > 0$ with probability w_i for $i = 1, \ldots, n$, where $w_i + \cdots + w_n = 1$. Conditionally on F, the outcome Y has distribution F, so F is auto-calibrated. We conclude that

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

Proposition 6.4.7 yields $\mathrm{MCB}_{\mathrm{HB}} = \mathrm{MS}_{\mathrm{HB}} = 0$. Finally, we apply formulas in Grimit et al. (2006) to obtain

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

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

6.C.2 Example in Candille and Talagrand (2005)

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

For $i = 1, 2$, we denote by G_i the isotonic conditional law of Y given $F = F_i$. Since $F_1 \leq_{\mathrm{st}} F_2$ and $Q_1 \leq_{\mathrm{st}} Q_2$ it follows that $Q_i = G_i$ for $i = 1, 2$ and the isotonicity-based decomposition coincides with the Candille–Talagrand decomposition. For any $z \in (-1, 1)$, $F_1(z)$ and $F_2(z)$ strictly order and hence the random variable $F(z)$ already reveals the value of F. That is, $\sigma(F(z)) = \sigma(F)$ and hence $\mathbb{P}(Y \leq z \mid F(z)) = \mathbb{P}(Y \leq z \mid F) = P_{Y|F}(z)$. Since this conditional probability is already an increasing function of $F(z)$, we may conclude by Proposition 3.2. in Arnold and Ziegel (2023) that $\mathbb{P}(Y \leq z \mid \mathscr{L}(F(z))) = P_{Y|F}(z)$ for all $z \in \mathbb{R}$ and hence the Brier score based decomposition correspond with the Candille–Talagrand decomposition. Analogously the claim can be shown for the quantile score based decomposition. Thus the isotonicity-based, Brier score based, and quantile score based decompositions coincide with the Candille–Talagrand decomposition, where $\mathbb{E}\,\mathrm{CRPS}(F, Y) = 1/6$, $\mathrm{MCB}_{\mathrm{CT}} = 1/30$, and $\mathrm{UNC}_0 = 2/5$.

The forecasts satisfy the conditions in part (c) of Proposition 6.4.7, therefore $\mathrm{MCB}_{\mathrm{HB}} = 0$. Since $Y \in \mathrm{supp}(F)$ almost surely, we have $\mathrm{MS} = 0$.

6.C.3 Example with two atoms

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

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

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

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

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

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

so $\grave{F}_2^{-1}$ fails to be increasing. Furthermore, it follows easily that $\mathrm{MCB}_{\mathrm{BS}} = 1/2 \neq 13/16 = \mathrm{MCB}_{\mathrm{QS}}$. As the conditional law of Y given F is a Dirac measure, $\mathbb{E}\,\mathrm{CRPS}(P_{Y|F}, Y) = 0$ and $\mathrm{MCB}_{\mathrm{CT}} = 1$. Similarly, $\mathrm{MCB}_{\mathrm{ISO}} = 1$ since F_1 and F_2 do not order.

According to the formulas in Section 6.2.5, $\bar{g}_1 = 2$ and $\bar{f}_1 = (\mathbb{1}\{F_1(y_1) \le \frac{1}{2}\} + 3\mathbb{1}\{F_2(y_2) \le 1/2\})/(2\bar{g}_1) = 3/4$ and thus $\mathrm{MCB}_{\mathrm{HB}} = (p_1 - \bar{f}_1)^2\,\bar{g}_1 = 1/8$, whence we conclude that $\mathrm{DSC}_{\mathrm{HB}} = \mathrm{MCB}_{\mathrm{HB}} + \mathrm{UNC}_0 - \mathbb{E}\,\mathrm{CRPS}(F, Y) = -1/8$.

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

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

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

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

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

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

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

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

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

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

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

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

Since the predictive distributions are ordered with respect to $\leq_{\mathrm{st}}$, it follows that for every threshold z, the ordering of $F_i(z)$ is the same. For $z \in (-\infty, 1]$, $F_2(z)$ and $F_3(z)$ coincide but this also holds for $G_2(z)$ and $G_3(z)$. Similarly, for $z \in [2, \infty)$, $F_1(z)$ and $F_2(z)$ coincide but this also holds for $G_1(z)$ and $G_2(z)$. This implies that the Brier score based and the isotonocity-based decompositions coincide. Since the stochastic order is equivalently characterized by pointwise orderings of lower quantile functions, the quantile score based and the isotonicity-based decompositions also coincide.

As all F_i^{-1}'s are absolutely continuous, we may apply Corollary 6.4.3 to compute $\mathrm{MCB}_{\mathrm{HB}}$.

For $p \in (0,1) \setminus \{1/4, 1/2, 3/4\}$ we find that

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

hence

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

The law of total expectation implies

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

and hence,

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

Finally, we obtain

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

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

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

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

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

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

$$G_1 = \tfrac{1}{2} \delta_{y_1} + \tfrac{4}{10} \delta_{y_2} + \tfrac{1}{10} \delta_{y_3}, \quad G_2 = \tfrac{1}{4} \delta_{y_1} + \tfrac{7}{20} \delta_{y_2} + \tfrac{4}{10} \delta_{y_3}, \quad G_3 = \tfrac{1}{4} \delta_{y_1} + \tfrac{1}{4} \delta_{y_2} + \tfrac{1}{2} \delta_{y_3}.$$

Since the image of the random vector (F, Y) is finite and ICL is the population version of IDR (Arnold and Ziegel, 2023, Proposition 4.1), one obtains the G_i's alternatively by applying IDR on the finite sample of size $n = 30$ with five occurrences of (F_1, y_1), four of (F_1, y_2), one each of $(F_1, y_3$ and (F_2, y_1), five of (F_2, y_2), four each of (F_2, y_3) and (F_3, y_1), one of (F_3, y_2), and five of (F_3, y_3). The $\mathrm{MCB}_{\mathrm{CT}}$ and $\mathrm{MCB}_{\mathrm{ISO}}$ components may be calculated in analogy to previous examples. We obtain $\mathrm{MCB}_{\mathrm{CT}} = 3t/200$ and $\mathrm{MCB}_{\mathrm{ISO}} = 3t_2/200$.

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

$$\begin{aligned}\tau(A) &= \mathbb{E}\int_A \mathbb{1}\{F(Y) \le p\}\,\mathrm{d}\nu_F(p) \qquad (6.50)\\ &= \int_A f(p)\,\mathrm{d}\mu(p) = f_1\,\frac{2t_1}{3}\,\delta_{1/4}(A) + f_2\,\frac{t}{3}\,\delta_{1/2}(A) + f_3\,\frac{2t_2}{3}\,\delta_{3/4}(A),\end{aligned}$$

where the expectation in (6.50) may be calculated by the law of total expectation:

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

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

7 | Physics-based vs. data-driven 24-hour probabilistic precipitation forecasts for northern tropical Africa

Numerical weather prediction (NWP) models struggle to skillfully predict tropical precipitation occurrence and amount, calling for alternative approaches. For instance, it has been shown that fairly simple, purely data-driven logistic regression models for 24-hour precipitation occurrence outperform both climatological and NWP forecasts for the West African summer monsoon. More complex neural network based approaches, however, remain underdeveloped due to the non-Gaussian character of precipitation. In this study, we develop and apply a new two-stage approach, where we train an off-the-shelf convolutional neural network (CNN) on gridded rainfall data to obtain a deterministic forecast and then apply the nonparametric Easy Uncertainty Quantification (EasyUQ) approach to convert it into a probabilistic forecast. The structure of this chapter aligns with the three forecasting steps introduced in Chapter 1. Each step is successively applied using the corresponding newly developed tools from Chapters 4, 5 and 6, respectively.

7.1 Introduction

Despite the continuous improvement of numerical weather prediction (NWP) models, precipitation forecasts in the tropics remain a great challenge. Several studies (Haiden et al., 2012; Vogel et al., 2020) have shown that NWP models have difficulties in outperforming climatological forecasts. A possible explanation is the exceptional high degree of convective organization over tropical Africa (Nesbitt et al., 2006; Roca et al., 2014), a process that is difficult to capture with the convective parameterization of

NWP models (Vogel et al., 2018), although recent developments show some promise (Becker et al., 2021). Statistical postprocessing, spatial averaging, or temporal aggregation lead to improvements in the skill of raw NWP ensemble gridpoint forecasts in tropical Africa (Vogel et al., 2020; Stellingwerf et al., 2021; Gebremichael et al., 2022; Ageet et al., 2023), yet in regions of particularly poor performance of the operational forecast systems, viz. West and Central Equatorial Africa, the forecast gain over climatology is limited.

The overall poor performance of current operational systems motivates the development of alternative approaches. Vogel et al. (2020) implement a fairly simple purely data-driven logistic regression model for 24-hour precipitation occurrence, which outperforms climatology and NWP forecasts for the summer monsoon season in West Africa. The predictor variables are designed by exploiting spatial-temporal coherence patterns as developed and investigated further in Rasheeda Satheesh et al. (2023). To this end, the rainfall at each grid point is correlated with the rainfall at all other locations from 1, 2, and 3 days before using the coefficient of predictive ability (CPA) measure from Chapter 4 (Gneiting and Walz, 2022). The locations showing highest CPA for 1, 2, and 3 days before, respectively, are selected as predictor variables in the logistic regression model. The good performance of this simple logistic model, which is related to coherent, tropical wave driven spatial propagation of precipitation features in West Africa (Rasheeda Satheesh et al., 2023), motivates the development of more sophisticated data-driven models and the usage of additional weather quantities linked to rainfall occurrence and amount.

Vogel et al. (2021) and Rasheeda Satheesh et al. (2023) have only investigated the skill of probability forecasts for the binary problem of precipitation occurrence. In this chapter, the more challenging problem of producing accurate probabilistic forecasts for accumulated precipitation, a non-negative real-valued variable, is considered. Precipitation accumulation is generally considered the “most difficult weather variable to forecast” (Ebert-Uphoff and Hilburn, 2023). Indeed, precipitation accumulation follows a mixture distribution with a point mass at zero — namely, for no precipitation — and a continuous part on the positive real numbers. Therefore, despite the sweeping rise of data-driven weather prediction (Ben Bouallègue et al., 2023) and rapid progress in data-driven nowcasting of precipitation (Ayzel et al., 2020; Lagerquist et al., 2021; Ravuri et al., 2021; Schroeder de Witt et al., 2021; Espeholt et al., 2022; Zhang et al., 2023), the development of machine learning based methods for probabilistic quantitative precipitation forecasts – at least for lead times longer 12 hours – has been lagging. For example, precipitation was “not investigated” (Bi et al., 2023, p. 537) by the Pangu-

Weather team and “left out of the scope" of the GraphCast development, because “precipitation is sparse and non-Gaussian and would have required different modeling decisions than the other variables” (Lam et al., 2023, p. 6). We address these challenges by developing a novel two-stage CNN+EasyUQ approach, where we first train an off-the-shelf convolutional neural network (CNN) model to obtain a single-valued deterministic forecast, and then use the Easy Uncertainty Quantification (EasyUQ) (Walz et al., 2024) approach presented in Chapter 5 to convert the deterministic forecast into a probabilistic forecast.

The chapter is structured as followed. Section 7.2 introduces the data used in the analysis. Then, an overview of weather quantities which are known to be linked to precipitation and thus are candidates for predictor variables is provided in section 7.3. Different types of forecasting models are described in section 7.4. Importantly, we compare the CNN+EasyUQ forecasts to a comprehensive suite of state of the art methods that include physics-based raw NWP ensemble forecasts, postprocessed NWP forecasts, data-driven statistical forecasts based on logistic regression and distributional index models (DIMs), and combined statistical-dynamical (hybrid) approaches. Results from this comparison are presented in section 7.5 with the main conclusion and outlook in section 7.6.

7.2 Data

In this study, we use data from three different sources. The arguably best currently available high-resolution, gauge-calibrated, gridded precipitation product, the Integrated Multi-Satellite Retrievals for GPM (Global Precipitation Measurement) (IMERG; Huffman et al., 2020), serves as ground truth for precipitation. The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis Version 5 (ERA5; Hersbach et al., 2020) product is used to obtain estimates of other weather quantities. Finally, NWP forecasts, namely the high resolution (HRES) run and the full ECMWF ensemble prediction system (EPS) are downloaded from ECMWF’s Meteorological Archival and Retrieval System (MARS, `https://www.ecmwf.int/en/forecasts/access-forecasts/access-archive-datasets`).

The evaluation domain, visualized in Figure 7.1, is northern tropical Africa, represented by 61×19 grid boxes centered at $25°$W – $35°$E and $0°$ – $18°$N, respectively, similar to the setup in Vogel et al. (2020) and Rasheeda Satheesh et al. (2023). Five dis-

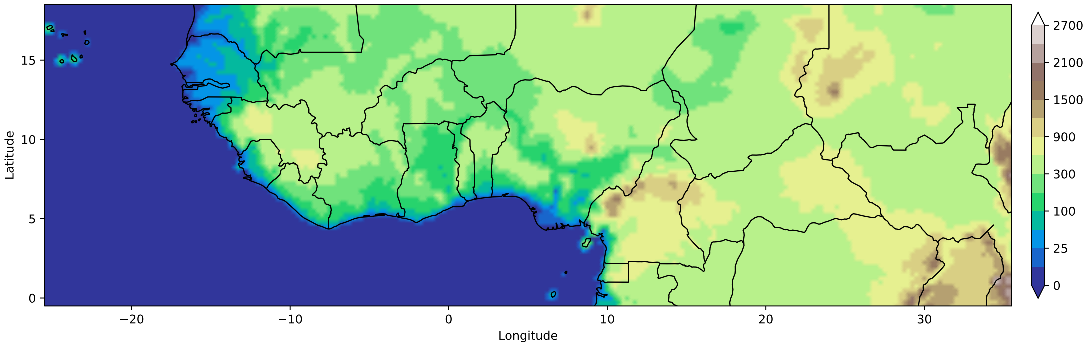

Figure 7.1: Overview of the study area. Following Rasheeda Satheesh et al. (2023), we consider an evaluation domain over northern tropical Africa that comprises 61×19 grid boxes with centers spanning from $25°$ W to $35°$ E in longitude and from $0°$ to $18°$ N in latitude, respectively. The time period considered ranges from 1 December 2000 to 30 November 2019, with 24-hour forecasts of precipitation amount and precipitation occurrence for 1 December 2010 to 30 November 2019 being evaluated. The analysis is over land only, and shading indicates altitude in meters, based on the ERA5 land–sea mask.

tinct seasons are considered as identified previously (Fink et al., 2017; Maranan et al., 2018): December–February (DJF), which is the dry season with occasional showers along the Guinea Coast; the March–April (MA) period, which features highly organized Mesoscale Convective Systems (MCSs) at the Guinea coast and the coastal hinterland; May–June (MJ), the major rainy season along most parts of the Guinea Coast; July–September (JAS), the major rainy season in the Sahel and the little dry season at the coast; and October–November (ON), the second, weaker rainy season at the Guinea Coast. To avoid cutting seasonal periods at the beginning or the end of the time period under investigation, the time period considered starts 1 December 2000 and ends 30 November 2019. Importantly, the analysis and evaluation are performed over land only, and we frequently identify a grid box with the grid point at its center. From now on, when we refer to grid boxes or grid points, we only mean boxes or points on land.

7.2.1 GPM IMERG rainfall data

We use the GPM IMERG V06B final version (Hou et al., 2014; Huffman et al., 2020) to calculate 24-hour accumulated precipitation from 06 – 06 UTC for the period under investigation. GPM IMERG has a temporal resolution of 30 minutes and a spatial resolution of $0.1° \times 0.1°$. The data were regridded to a resolution of $1° \times 1°$ using first-order conservative remapping. As we also consider 24-hour rainfall occurrence, we threshold at 0.2 mm to obtain a binary event variable representing precipitation occurrence.

The GPM IMERG algorithm uses both radar-calibrated microwave radiance from polar-orbiting satellites and infrared radiance from geostationary satellites. In the final version, the precipitation totals are calibrated with rain gauge measurements provided by the Global Precipitation Climatology Centre (GPCC; Schneider et al., 2005). The degree to which the original estimates are adjusted by the gauge calibration process within a given region is generally determined by the number of available rain gauges, which is highly variable across the Tropics.

7.2.2 Predictor variables from ERA5

Our study considers a range of meteorological variables, specified in section 7.3.2, as predictor variables for statistical models. Specifically, we use the ERA5 reanalysis (Hersbach et al., 2020), which provides a complete and consistent coverage of the study domain by combining model data with observations. For this study the resolution of the data is $1^\circ \times 1^\circ$ just like for GPM IMERG. In contrast to 24-hour accumulated precipitation, the considered ERA5 weather quantities are instantaneous values at 00 UTC, thus six hours before the 24-hour accumulation period for GPM IMERG starts. This way, observed ambient conditions well before the rainfall begins get considered. For an operational implementation of the respective statistical methods, operational analysis data would need to be used, as ERA5 is not available in near-real time, but we do not expect this to make a big difference to our results.

7.2.3 Physics-based forecasts from ECMWF

We now describe the NWP forecasts used in this study, namely, the ECMWF high resolution (HRES) model and ensemble prediction system (EPS; Molteni et al., 1996), which are also used in the case studies of Chapter 4 (see Sections 4.5 and 4.5.3), Chapter 5 (see Sections 5.2.2, 5.2.3 and 5.3.3) and Chapter 6 (see Section 6.5.1). Owing to the high resolution and the initialization with the most accurate analysis product, the HRES model is arguably the leading global deterministic NWP forecast available. As an operational product, HRES has changed considerably over time in frequent updates (`https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasti` `ng+system`). The ECMWF EPS consists of one control run and 50 perturbed members. Like the HRES model, the control run is based on the most accurate initial state of the atmosphere. The perturbed members start from slightly different initial conditions and use perturbed physics options.

The forecasts are available from MARS in a grid resolution of $0.25^\circ \times 0.25^\circ$ and first-order conservatively remapped to a resolution of $1^\circ \times 1^\circ$. HRES forecasts for total precipitation are obtained by summing forecasts for large scale precipitation and convective precipitation, which are available from April 2001 on. For the EPS, total precipitation is available from April 2006 on. To cover an equal number of seasons, we use data starting in December 2001 and December 2006, respectively. The HRES forecasts for 24-hour precipitation amounts are initialized at 00 UTC with a lead time of 24 hours. To obtain EPS forecasts for 24-hour precipitation amounts the difference between forecasts of accumulated precipitation initialized at 00 UTC with lead times of 30 and 6 hours is computed. To compute the EPS forecast probability for the occurrence of precipitation, the member forecasts are thresholded at 0.2 mm and the respective binary outcomes are averaged.

7.3 Predictor variables for data-driven forecasts

In this section we discuss and analyze potential predictor variables for data-driven forecasting methods. We distinguish predictor variables computed from IMERG data based on spatio-temporal rainfall correlation, and predictor variables based on ERA5. The initial selection of the variables stems from meteorological expertise.

7.3.1 Correlated rainfall predictors from IMERG

Vogel et al. (2021) introduced a logistic regression model to produce probability forecasts for the binary outcome of precipitation occurrence. As predictors, they used precipitation data with a lag of one and two days at locations with maximum positive and minimum negative Spearman's rank correlation coefficient. Rasheeda Satheesh et al. (2023) noted that due to propagating rainfall systems positive dependencies carry the most useful information, occasionally reaching three days backwards in time. Moreover, they suggested a replacement of Spearman's rank correlation coefficient by the coefficient of predictive ability (CPA; Gneiting and Walz, 2022) measure presented in

Chapter 4. In general, CPA is asymmetric, with the predictor variable and the outcome taking clearly identified roles, as for the classical Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) measure, to which CPA reduces when the outcomes are binary. When both the predictor variable and the outcome are continuous variables, CPA becomes symmetric and equals Spearman's rank correlation coefficient, up to a linear transformation in a continuous setting. AUC or CPA values above 0.5 correspond to positive dependencies, and values below 0.5 to negative dependencies.

Given these insights, this current study uses three correlated precipitation predictor variables, by identifying the grid points with maximum CPA at a temporal lag of one, two, and three days, respectively. Following Rasheeda Satheesh et al. (2023), correlated locations are identified within an enlarged region that comprises $68°$W – $50°$E and $0° - 20°$N, as compared to the evaluation domain depicted in Figure 7.1, which ranges from $25°$W – $35°$E and $0°$ – $18°$N.

7.3.2 Predictor variables from ERA5 reanalysis

In addition to the correlated precipitation information, various meteorological variables from ERA5 are considered as predictors (Table 7.1). For a summary of how environmental conditions affect convection, see Maranan et al. (2018). Unless noted otherwise, the variables are instantaneous quantities at 00 UTC. The first four variables in Table 7.1 are vertically integrated measures of water in different forms. TCWV has been shown to be a promising predictor for precipitation by Lafore et al. (2017b); Schroeder de Witt et al. (2021) use cloud information such as TCLW and TCC in their global statistical model. The second group comprises the three classical measures of convective instability; CAPE (the theoretical maximum of thermodynamic energy that can be converted into kinetic energy of vertical motion), CIN (the energy barrier that needs to be overcome to reach the level of free convection), and KX (based on dry static vertical stability in the 850–500 hPa layer, absolute humidity at 850 hPa, and relative humidity at 700h pa). CAPE and CIN have a complex relationship with precipitation and should be considered together and in concert with other parameters (Lafore et al., 2017a). Galvin (2010) demonstrates the usefulness of KX in assessing convective rainfall probability in relation to African Easterly Waves (AEWs).

The third group (2T, 2D, SPT) represents near-surface conditions. The former two are closely related to the equivalent potential temperature of a starting convective air

Table 7.1: Predictor variables from ERA5, all at 00 UTC.

Meteorological Variable	Acronym
Total column water vapour	TCWV
Vertically integrated moisture divergence	VIMD
Total column cloud liquid water	TCLW
Total cloud cover	TCC
Convective available potential energy	CAPE
Convective inhibition	CIN
K-index	KX
2m temperature	2T
2m dewpoint temperature	2D
24h surface pressure tendency	SPT
Temperature at 850 hPa	T850
Temperature at 500 hPa	T500
Specific humidity at 925 hPa	Q925
Specific humidity at 700 hPa	Q700
Specific humidity at 600 hPa	Q600
Specific humidity at 500 hPa	Q500
Relative humidity at 500 hPa	R500
Relative humidity at 300 hPa	R300
Shear	SHR
Streamfunction at 700 hPa	Ψ700

parcel, thereby influencing the level of cumulus condensation and free convection and thus CIN and CAPE, and have been shown to impact the intensity of convection in West Africa (Nicholls and Mohr, 2010). SPT, the tendency from 00 UTC of the day for which the prediction is made to 00 UTC of the previous day, can be related to AEW propagation and rainfall (Regula, 1936; Hubert, 1936). The fourth group characterises thermodynamic conditions in the boundary layer and free troposphere between 925 hPa and 300 hPa. For temperature, we consider 850 hPa and 500 hPa representing lower-tropospheric stability (as in KX). As moisture generally shows complex vertical structures, 925, 700, 600, and 500 hPa are chosen for specific humidity. For relative humidity, the mid- to upper-tropospheric levels of 500 hPa and 300 hPa were selected to indicate deep moistening, which facilitates cloud formation and reduces detrimental effects of entrainment on convective development. Mid-tropospheric relative humidity controls both rainfall enhancement by slow moving tropical waves (Schlueter et al., 2019) and evaporation of rainfall, and thus convective downdrafts and mesoscale organization of convection (Klein et al., 2021). The last two entries in Table 7.1 are the circulation-related variables SHR (normalized difference of horizontal wind at 600 and 925 hPa) and Ψ700 representing mid-tropospheric streamlines. SHR influences the potential for mesoscale organization and longevity through separating the areas of convective up- and downdrafts as well as the generation of cold pools (Rotunno et al., 1988; Lafore et al., 2017b). Anomalies in Ψ700 indicate variations in the African Easterly Jet (AEJ), e.g., passages of troughs and ridges of AEWs (Kiladis et al., 2006).

7.3.3 Statistical analysis of predictor variables

Thus far, the selection of predictor variables has been based on meteorological expertise and findings from other publications. Here, we use the aforementioned AUC (for rainfall occurrence) and CPA (for amount) measures (see Section 7.3.1) for a deeper analysis. In Figures 7.2 and 7.3 we show AUC and CPA values for the 20 ERA5 variables from Table 7.1. Both are computed in a co-located fashion for each grid point in the evaluation domain (Figure 7.1) and the resulting distributions are represented by boxplots.

Figure 7.2a shows AUC values for the dry season DJF. Given the overall low precipitation amounts during this period, the box plots often stretch over large ranges, indicating marked differences between grid points, and also large differences between the variables. Stable positive relations (i.e., AUC above 0.5) are found for moisture (TCWV, Q500, Q600, Q700, R500), cloud (TCLW, TCC), and instability variables (KX, CAPE), demonstrating a clear dependence on mid-tropospheric conditions, while low-level (Q925, 2D) and upper-level (R300) variables show a more ambiguous behavior. Other well-defined relations are positive with 2T, and negative with T500 and VIMD. As the variables are taken at 00 UTC, the relation to 2T may reflect warmer nights under moister and cloudier skies. CIN, SPT, and Ψ700 show weak AUC values close to 0.5. AUC values for T850 cover a wide range and stretch across 0.5, indicating that its impact depends strongly on the situation.

The corresponding analysis for MA (Figure 7.2b) shows an overall less noisy behavior and AUC values more in line with the spatially averaged annual value of CPA that

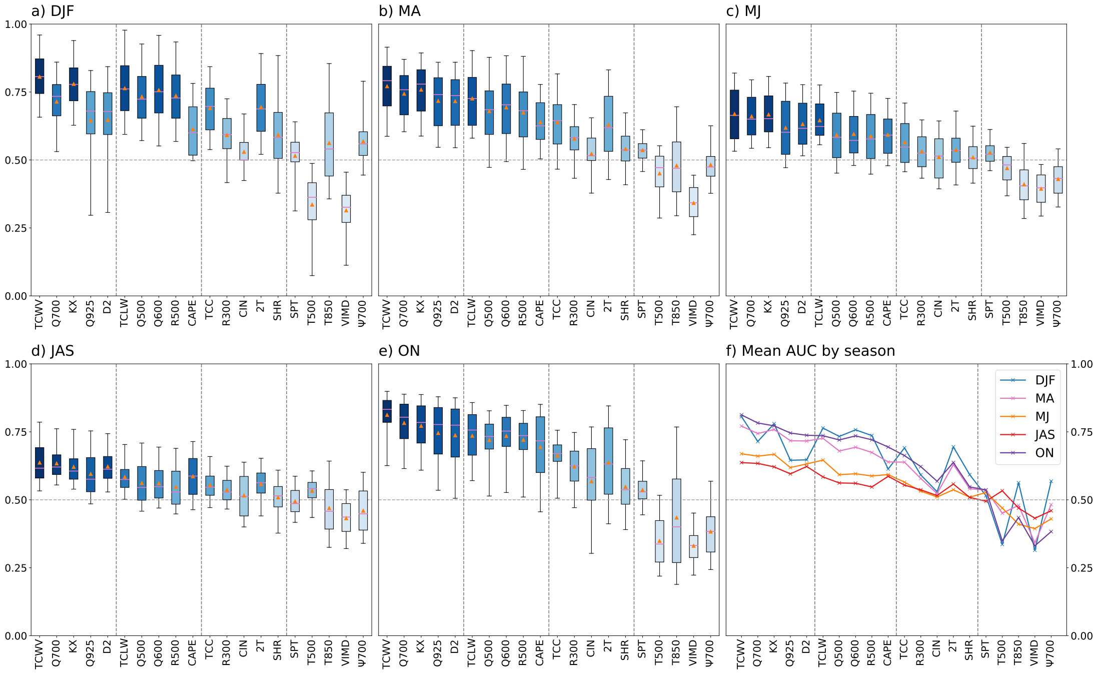

Figure 7.2: Boxplots of grid point AUC values between ERA5 variables from Table 7.1 and precipitation occurrence in season a) DJF, b) MA, c) MJ, d) JAS, and e) ON. The arrangement of the predictor variables on the horizontal axis is in the order of the spatially averaged CPA value for precipitation accumulation, when CPA is computed without splitting into seasons. The orange marks and the line plots in panel f) indicate the mean AUC value over grid points for the season at hand. The box colour from dark to light blue indicates the ranking of the seasonal mean AUC value. In combination this allows to identify differences between yearly vs. seasonal perspectives.

determines the order of the variables in all panels of Figures 7.2 and 7.3. Compared to DJF, a more stable relation to low-level moisture (Q700, Q925, 2D) is visible. There is a stronger relation to CAPE with little changes in CIN. Other remarkable changes are less dependence on cold T500, and even more ambiguous relations to T850 and Ψ700. The pre-monsoon season MJ (Figure 7.2c), when rainfalls begin to move inland, shows many similarities to MA but the point-to-point variability is smaller and AUC values tend to be closer to 0.5, while their order mostly agrees to that based on annual CPAs. Remarkable differences to MA are less dependence on 2T and clearer relations to T850 and Ψ700 (< 0.5). The latter may indicate a dependence of rainfall on the existence of cyclonic perturbations such as AEWs. The general magnitude of AUC values close to 0.5 is likely a reflection of the overall improved conditions for convection, which makes individual storms less dependent on particular circumstances, thereby creat-

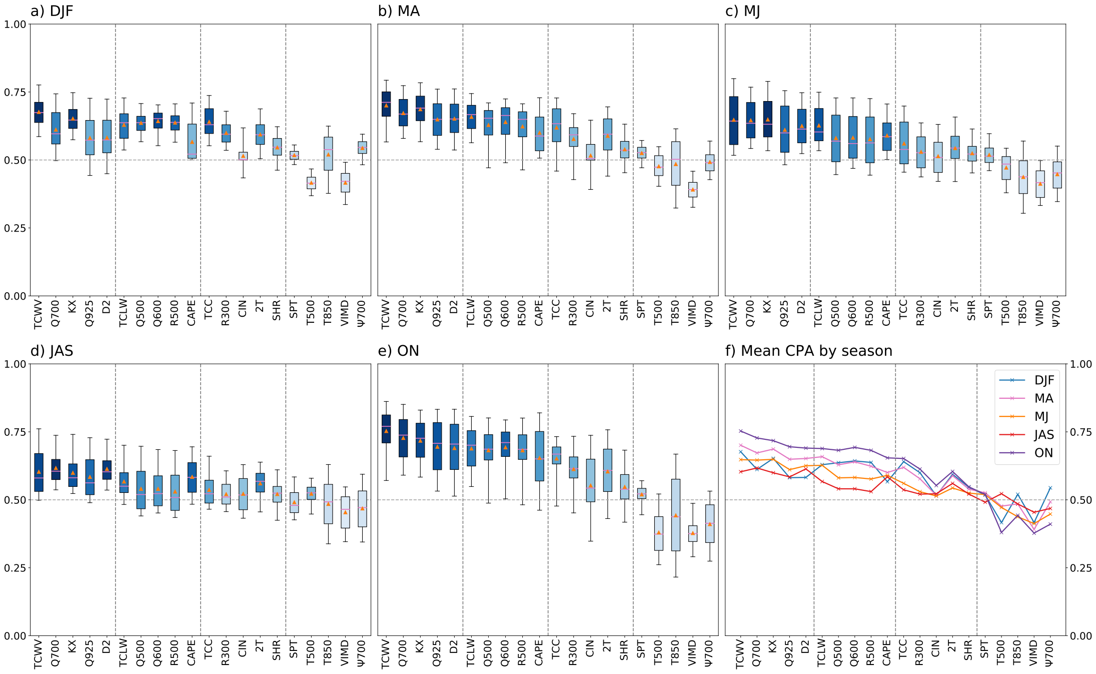

Figure 7.3: As Figure 7.2 but for CPA and precipitation amount.

ing a higher degree of stochasticity (see also discussion in Rasheeda Satheesh et al. (2023)). This trend continues going into the main monsoon season JAS (Figure 7.2d), when most variables show AUC values close to 0.5. The narrower boxplots indicate less local variability during a period when rains penetrate deeply into the continent. As expected, in the post-monsoon season ON (Figure 7.2e), conditions resemble those discussed for MA (Figure 7.2b), even with slightly larger amplitudes. Remarkable differences to MA are that rainfall occurrence depends more on CIN and 2T, possibly because in ON the solar angle is already flatter and the daytime heating is further dampened by the higher moisture availability after the rainy season. As for DJF, rain depends on cold T500 and the relation to T850 is highly variable and can take both directions, however, with a clear tendency to cooler conditions when rain occurs. ON also shows the clearest relation to cyclonic perturbations as reflected in AUC values below 0.5 for Ψ700. These may grow in importance relative to other mechanisms, as triggering by daytime heating weakens. Finally, Figure 7.2f shows a summary plot of mean AUC values for all five seasons. This plot underlines the similar behavior of MJ and JAS (with a consistently higher amplitude for MJ), as well as of MA and ON (with a consistently higher amplitude for ON). DJF often shows the highest magnitude, as rain depends strongly on unusual conditions to occur, but given the many dry days, the overall behavior appears quite noisy.

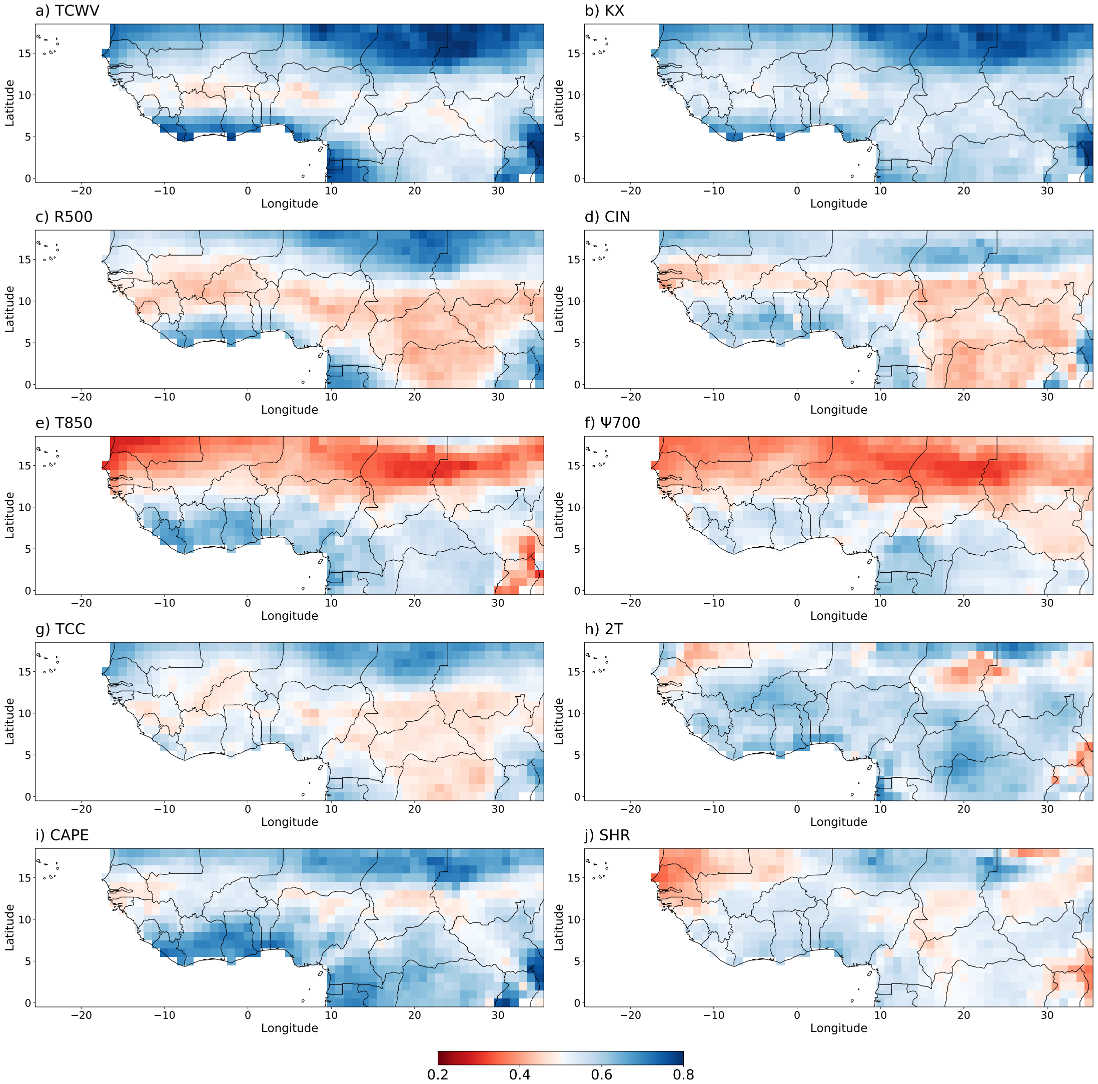

Figure 7.4: Spatial pattern of CPA between the ERA5 predictor a) TCWV, b) KX, c) R500, d) CIN, e) T850, f) Ψ700, g) TCC, h) 2T, i) CAPE, and j) SHR from Table 7.1 and precipitation amount in season JAS.

The corresponding analysis for CPA is shown in Figure 7.3. Overall there are many similarities to Figure 7.2, indicating that variables that work as predictors for occurrence also work for amount. This is particularly true for the wet part of the year (MJ, JAS and ON), where plots look largely identical (Figure 7.3c–e). For MA (Figure 7.3b), there is still large agreement across all variables but the magnitude of CPA values is smaller and the box plots are narrower than for AUC. This indicates that in this somewhat marginal rainfall season, amount is harder to predict than occurrence. This trend is even more evident for dry DJF (Figure 7.3a), when some boxplots become very narrow and magnitudes fall underneath those of ON on average, as shown by the summary plot (Figure 7.3f).

In order to better understand the ranges indicated in the boxplots in Figures 7.2 and 7.3, Figure 7.4 shows the spatial pattern of CPA of selected meteorological variables exemplary for the peak monsoon season JAS. Consistent with the leftmost boxplot in Figure 7.3d CPA values for TCWV are at or above 0.5 almost everywhere in the study region (Figure 7.4a). The spatial pattern shows an interesting three-tier structure. Over northern parts of the domain, where moisture is a general limiting factor, CPA values are high, especially over the dry eastern Sahel. Further south, along the main rain belt and stretching into the Congo Basin, CPA values are close to 0.5, indicating limitations through convective triggering or stability rather than moisture availability. To the south of the rain belt, i.e., along the Guinea Coast, and over the East African highlands, moisture appears to become a limiting factor again. A very similar pattern but with a smaller range emerges for KX (Figure 7.4b). The largest differences to TCWV are found in the Guinea Coastal area, where conditions are often close to moist neutral requiring some lifting mechanism to produce rain (see Figure 1.31 in Fink et al., 2017). Similar but slightly northward shifted structures are also found for CAPE (Figure 7.4i).

A much larger range (0.35–0.75) but with a similar three-tier structure is found for R500 and CIN (Figures 7.4c,d). One would expect that a moister mid-troposphere and less convective inhibition (recall that CIN is negatively oriented) enhances rainfall amounts and so the behavior within the rain belt is somewhat counter-intuitive. The most likely explanation is that in areas of abundant moisture and often neutral stratification, large rainfall amounts can most effectively be generated by organized convective systems that require some barrier to accumulate CAPE over the following day and a relatively dry mid-troposphere to allow rainfall evaporation and downdrafts, which in turn can trigger new convection through cold pools (cf. Table 11.2 in Lafore et al. (2017a)). It is interesting to note that TCC shows similarly low CPA values (< 0.5) in the rain belt as R500 (Figure 7.4a).

Finally, CPA values for T850 and Ψ700 are both characterised by a marked north-south division around 12°N (Figures 7.4e,f). In Figure 7.4e values over the East African highland should be largely ignored, as they are mostly extrapolated to beneath the model orography. The patterns indicate that in the north, high rainfall amounts are accompanied by lower T850, likely indicating a northward progression of the moist and cool monsoon layer, while in the south warm air at 850 hPa may indicate more instability on the following day. With respect to Ψ700 (Figure 7.4f) low values in the

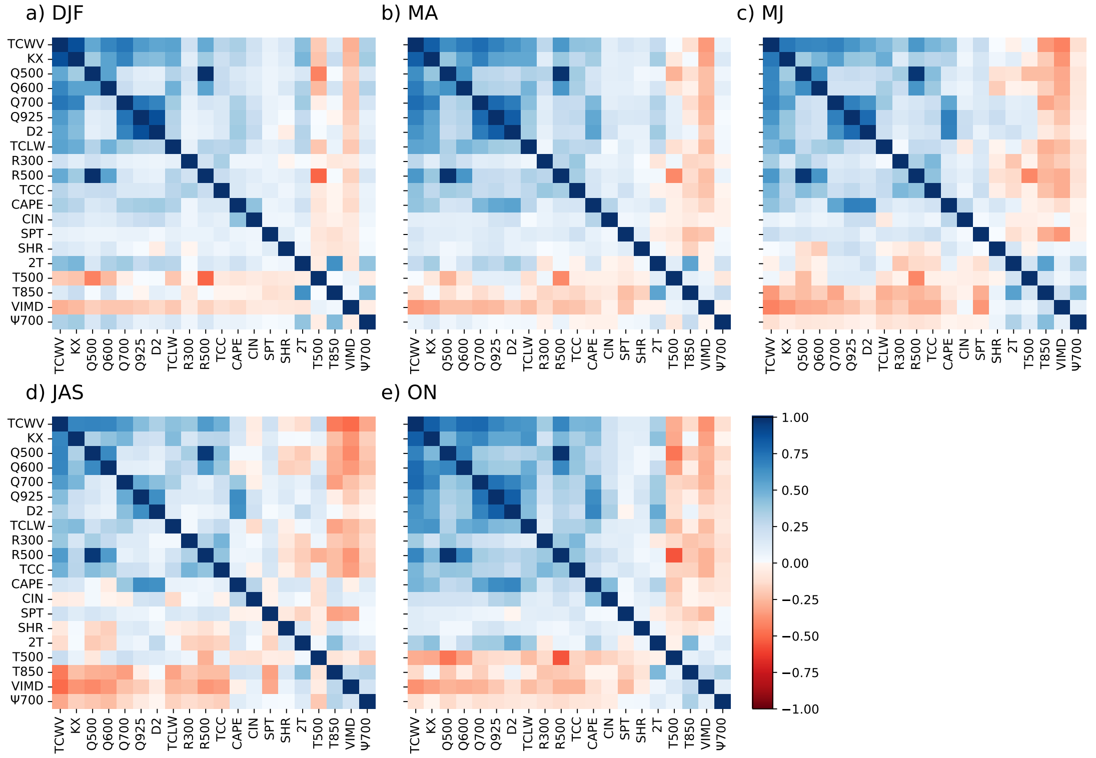

Figure 7.5: Spatially averaged Spearman's rank correlation coefficient between the ERA5 variables from Table 7.1 in season a) DJF, b) MA, c) MJ, d) JAS, and e) ON.

north indicate that rainfall is accompanied by more cyclonic conditions, likely due to the trough passage of AEWs, while in the south weak anticyclonic conditions prevail.

Most meteorological variables from Table 7.1 show spatial patterns akin to those in Figure 7.4, though some feature hard to interpret local signals that entail a wider range of CPA values (e.g., 2T and SHR, Figure 7.4h,j). It is also worth mentioning that corresponding spatial structures for AUC largely agree with CPA (not shown). Comparing Figure 7.4 with the other four seasons, we find a high consistency in the discussed patterns that largely shift north- and southward with the seasonal evolution of the West African monsoon system (not shown).

For the construction of statistical models, correlations between predictor variables matter, as they hinder interpretation and may yield unstable statistical parameters. Figure 7.5 shows Spearman's correlation coefficients for the 20 predictor variables from Table 7.1 presented in the same order as in Figures 7.2 and 7.3. We compute the correlation coefficient at each grid point, and then average over grid points. Note that here, we want the correlation coefficient to be symmetric (in contrast to the asymmetric relation between target and predictor variables). This analysis has been conducted for the same seasons as in Figures 7.2 and 7.3 and the entire year but due to the large similarities between them, we discuss JAS only (as for Figure 7.4).

Not surprisingly, there are generally high correlations between all moisture variables (TCWV, Q500, Q600, Q700, Q925, 2D, TCLW, R500, R300, and TCC) with the weakest relationship between R300 on the one hand and 2D and TCLW on the other hand. It is noteworthy that R500 is more strongly correlated to Q500 than T500. KX and CAPE show considerably different patterns with KX being highly correlated with the moisture variables but surprisingly also associated with cold T850, which to some extent counteracts the impact of moister conditions. CAPE is most sensitive to low-level moisture and associated with warm T850, as does CIN but to a smaller degree. A positive SPT is weakly associated with a moister and warmer atmosphere, consistent with the southerly flow behind an AEW trough, where the moister atmosphere suppresses longwave cooling. SHR, 2T, and T500 show overall weak and unsystematic correlations, in agreement with the difficult to interpret spatial patterns for CPA discussed above. Finally, T850, VIMD, and Ψ700 are consistently negatively correlated with the moisture variables and KX, with the exception of 2D. While the relation to VIMD is straightforward, the cooler T850 may be an indication of north-south movements of the monsoon layer, bringing overall moister or drier conditions. The negative correlation between moisture variables and Ψ700 reflects the wet conditions associated with cyclonic disturbances, e.g., AEW troughs or vortices.

7.4 Physics-based and data-driven forecast methods

Forecasts for precipitation occurrence and precipitation amount ought to be probabilistic to account for the chaotic nature of the atmosphere, thus for the former they should output a probability of precipitation (PoP) and for the latter a probability distribution. We investigate forecasts for precipitation occurrence and precipitation amount separately, which allows to connect our results to Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), where the binary setting was considered only. Furthermore, we can compare between the comparably easy task of producing PoP forecasts and the more challenging task of constructing probabilistic forecasts for precipitation amounts. To assess the skill of statistical and machine learning models it is essential to use baseline models to which to compare the forecast performance. In the following subsections, different types of forecasting models are presented that are

Table 7.2: Overview of probabilistic forecast methods for precipitation occurrence and/or accumulation, including general type, brief description, acronym, and availability of training data. Methods marked with an asterisk* yield PoP occurrence forecasts only; for methods marked** we do not present results for PoP forecasts. The final column notes from which year and month onward training data are available and used. See text for details.

Type	Description	Acronym	Training
Climatological	monthly probabilistic climatology	MPC	2000 12
Physics-based	ECMWF ensemble prediction system	EPS	NA
	isotonic regression applied to EPS	EPS+ISO*	2006 12
	EMOS applied to EPS	EPS+EMOS**	2006 12
	EasyUQ applied to HRES	HRES+EasyUQ	2001 12
Statistical	logistic regresssion, baseline (5 predictors)	Logit-base*	2000 12
	same, full model (25 predictors)	Logit-full*	2000 12
	distributional index model, baseline (5 predictors)	DIM-base**	2000 12
	same, full model (25 predictors)	DIM-full**	2000 12
Machine learning	EasyUQ applied to convolutional neural network	CNN+EasyUQ	2000 12
Hybrid	mixture of HRES+EasyUQ and CNN+EasyUQ	Hybrid	NA

physics-based NWP models, purely data-driven statistical or machine learning techniques, or mixtures of both. Table 7.2 provides an overview of all considered approaches.

As discussed in section 7.2, our evaluation period for 24-hour forecasts of precipitation amount and precipitation occurrence ranges from 1 December 2010 to 30 November 2019. The DJF season runs across two subsequent calendar years and we generally assign it to the second year. Reporting a yearly seasonal or overall mean instead of a single mean score over the complete evaluation period allows for a more distinct comparison between forecasting models and provides insights into the temporal evolution of forecast skill.

Except for the ECMWF ensemble prediction system (EPS), all types of forecasting methods require training data and some form of training procedure.[1] In this study, we use annually growing, expanding training sets that resemble operational settings, where

[1] Our Hybrid model combines the CNN+EasyUQ and HRES+EasyUQ forecasts in a way that does not require additional training.

only past data are available.[2] The initial training period ranges from the first day of the month in the right most column of Table 7.2 (hereinafter, the start date) to 30 November 2010, and the thus trained methods are used to generate day-ahead 24-hour forecasts for the period from 1 December 2010 to 30 November 2011. Then, we successively add one more year to the training period, ranging now from the start date through 30 November in year $2010 + x$, and use the thus trained methods to generate forecasts for the 12-month period that begins on 1 December in year $2010 + x$, where $x \in \{1, \ldots, 8\}$. This procedure is followed until training is on data through 30 November 2018 and the thus trained methods are used to generate forecasts for 1 December 2018 through 30 November 2019. Thus, there are nine evaluation folds in total, which we associate with calendar years 2011, …, 2019, respectively.

7.4.1 Climatological forecasts

Arguably, the simplest possible type of probabilistic forecast is a climatology constructed from past observations. Here we use GPM IMERG to construct a monthly probabilistic climatology (MPC). The MPC forecast for a specific valid date is an ensemble constructed by using all past observations from the month at hand. For example, for a test date in January 2014, the MPC forecast is constructed based on data from January 2001 to January 2013, which yields an ensemble of size $31 \times 13 = 403$. To obtain the MPC PoP forecast, the relative frequency of ensemble members with rainfall exceeding 0.2 mm is computed.

7.4.2 Physics-based forecasts

Our comparison includes raw and postprocessed probabilistic forecasts from physics-based numerical weather prediction (NWP) models run by the ECMWF (section 7.2.3). The postprocessed forecasts require training, for which we use expanding training sets with start dates listed in Table 7.2 as described above. Training is performed at each grid point individually.

[2] Nonparametric statistical methods such as IDR and machine learning approaches benefit from having as much (relevant) training data available as possible. Subject to this caveat, the predictive performance generally does not depend very much on the details of the training scheme. For example, the EPS+EMOS technique using a rolling training period of the most recent 730 days yields similar results.

Operational ECMWF NWP ensemble

The operational ECMWF ensemble prediction system (EPS) comprises 51 NWP runs, namely, a control member and 50 perturbed members. Just as for the climatological MPC approach, the EPS PoP forecast is the relative frequency of members that exceed 0.2 mm.

Statistically postprocessed ECMWF NWP ensemble

Statistical postprocessing is used to correct for systematic biases in raw ensemble forecasts. Here we use Ensemble Model Output Statistics (EMOS), originally developed by Gneiting et al. (2005), to generate full predictive probability distributions by linking ensemble information to distributional parameters. The optimal coefficients are found by optimizing a performance metric on training data.

In the binary case, we recalibrate the EPS PoP by using nonparametric isotonic regression (Zadrozny and Elkan, 2002), here referred to as EPS+ISO. For precipitation amounts, we apply the EMOS technique proposed by Scheuerer (2014) which models positive rainfall accumulations with generalized extreme value distributions, to generate the EPS+EMOS forecast. While EPS+EMOS induces a PoP forecast, the predictive performance is very similar, though typically slightly inferior, to EPS+ISO. Therefore, we do not report results for the respective PoP forecasts (cf. Table 7.2).

EasyUQ on the HRES model

The high resolution (HRES) model from ECMWF generates a deterministic NWP forecast. We use the EasyUQ technique, introduced in Chapter 5, to transform this single-valued forecast into a postprocessed predictive distribution, to yield the HRES+EasyUQ forecast.

7.4.3 Statistical forecasts

Statistical approaches use training data to learn relationships between a target variable and one or more predictor variables. Here, the target variable is precipitation amount at a given grid point, which in the case of precipitation occurrence is thresholded at 0.2 mm. We use logistic regression (see Section 3.2) to obtain PoP forecasts and Distributional single Index Models (DIMs; Henzi et al., 2023) for probabilistic forecasts of precipitation amounts, based on predictor variables from section 7.3. Statistical models require training, and we use annually expanding training sets with start date in December 2000 (Table 7.2) as described above. Training is performed at each grid point individually.

The analysis in section 7.3 provides a thorough understanding of the influence of the selected variables from Table 7.1 on precipitation occurrence and amount, and enables to link them to typical seasonal weather phenomena. However, overall the effect of meteorological variables on precipitation is similar across seasons when taking into account the latitudinal shifts associated with the monsoon system. As a consequence we found little difference in model performance between fitting models on seasonal data versus the whole available training period, as temporal effects such as seasonal changes can be captured by predictor variables that encode the day of the year. Therefore, instead of fitting seasonal models, we train models that apply year-round.

We distinguish baseline models with two predictors that encode the day of the year and three correlated rainfall predictors (section 7.3.1) from full models that additionally use 20 predictor variables from ERA5 (section 7.3.2). To prevent a statistical model from overfitting, regularization techniques can be applied. However, in this experiment the performance of the statistical models, which use modest numbers of at most 25 predictor variables only, does not improve when using the regularization techniques we tested. Consequently, we refrain from performing any feature selection beyond the choices made in section 7.3, which were driven by meteorological expertise and extant literature in atmospheric physics.

Logistic regression

We use logistic regression (Logit) models, introduced in Section 3.2, to generate statistical PoP forecasts. Specifically, let m be the number of predictor variables, which we denote by $x_1, \ldots, x_m$, and let p be the PoP forecast. The logistic regression model then is of the form

$$\text{logit}(p) = \log \frac{p}{1-p} = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_j, \tag{7.1}$$

where the statistical coefficients $\alpha_0, \alpha_1, \ldots, \alpha_m$ are estimated from training data. Our baseline model (Logit-base) originates from Vogel et al. (2021) and Rasheeda Satheesh et al. (2023) and uses $m = 5$ predictor variables, namely, three correlated rainfall predictors x_1, x_2, and x_3 at temporal lags of one, two, and three days, respectively, as described in section 7.3.1, and two variables $x_4 = \sin(2\pi d/365)$ and $x_5 = \cos(2\pi d/365)$ that depend solely on the day of the year d. The full model (Logit-full) extends to $m = 25$ predictor variables in (7.1), now including the twenty ERA5 variables from Table 7.1.

Distributional index models

To produce probabilistic forecasts for accumulated precipitation we use the Distributional (single) Index Model (DIM) approach introduced by Henzi et al. (2023), which combines the classical single index model with Isotonic Distributional Regression (IDR; Henzi et al., 2021). In a nutshell, an index is learned that represents the conditional mean of the target variable (see Section 3.1), here log-transformed precipitation accumulation, and then a predictive distribution is estimated nonparametrically under a stochastic ordering constraint. As before, let m be the number of predictor variables, which we denote by $x_1, \dots, x_m$, and let y now be precipitation accumulation. The index model then assumes the relationship

$$\log\left(y + \frac{1}{100}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j, \tag{7.2}$$

where the statistical coefficients $\beta_0, \beta_1, \dots, \beta_m$ are learned from training data. Subsequent to the training of the index model, the nonparametric IDR distributions are estimated on the same training set. We distinguish a baseline model (DIM-base) and an extended model (DIM-full, $m = 25$ in (7.2)), for which we use the same sets of predictor variables as in the Logit approach from section 7.4.3.

Note that PoP forecasts can be extracted from the DIM-base and DIM-full distributions. These yield similar, though slightly inferior, results than the Logit-base and Logit-full PoP forecasts, respectively, and so we do not report the respective scores (cf. Table 7.2).

7.4.4 Machine learning based forecasts: CNN+EasyUQ

The aforementioned statistical models are applied at each grid point individually. Thus, including spatial information has to be done by manually engineering features accordingly, such as the correlated rainfall predictors from section 7.3.1. In contrast, Convolutional Neural Network (CNN) models operate directly on the two-dimensional input space and can learn spatial relations from the data without the need to extract spatial information beforehand. CNN models are most commonly used for image tasks, where the input usually is a two- or three-dimensional array of pixel values. The gridded weather data over our evaluation domain can be envisaged as two-dimensional pseudo images of size 61×19. These dimensions correspond to longitude and latitude, respectively, spanning the study domain (Figure 7.1) from 25° W to 35° E and 0° to 18° N, respectively, with a grid resolution of $1^\circ \times 1^\circ$. With a suitable architecture, a single CNN model produces a two-dimensional array with forecasts for all grid points at once, instead of training models at each grid point individually. Due to their inherent inductive bias towards local neighborhood connectivity, CNNs are well-suited for predicting precipitation on the 61×19 grids, as they effectively exploit spatial correlations and structures within a grid, recognizing patterns within local areas that may be indicative of specific weather conditions. For this reason, the three correlated rainfall predictors from section 7.3.1 are replaced by 61×19 grids of IMERG precipitation accumulations (section 7.2.1) at temporal lags of one, two, and three days, respectively.

Motivated by their successful application in related meteorological tasks (Ayzel et al., 2020; Weyn et al., 2020; Lagerquist et al., 2021; Chapman et al., 2022; Otero and Horton, 2023), we employ a CNN architecture in the form of the U-Net (Ronneberger et al., 2015). The architecture of the U-Net consists of a contracting (downsampling) path and an expansive (upsampling) path, which are symmetric in terms of individual layer properties, giving it a U-like shape. We make use of max pooling operations for downsampling and transposed convolutions for upsampling layers. A crucial feature of the U-Net is skip connections between layers of the same size in the contracting and expanding paths. Applied to the precipitation data grid, these connections allow the network to use information from multiple resolutions, combining the context from the contracting path with the localization information from the expansive path. This allows to model longer spatial range dependencies in the data. To avoid overfitting, we also make use of Dropout (Srivastava et al., 2014) throughout the network architecture.

To transform the deterministic precipitation forecasts of the CNN model into probabilistic forecasts, the EasyUQ technique introduced in Chapter 5 is applied at each grid point individually, subsequent to the training of the index model, and based on the same training data as for the neural network, augmented with the deterministic CNN output. As noted, the resulting CNN+EasyUQ forecast distributions are discrete and have mass exclusively at outcomes observed during training. Code for the implementation of the CNN+EasyUQ approach in Python (Python Software Foundation, 2021) is available under `https://github.com/evwalz/precipitation`. Once more we emphasize that, while our usage of EasyUQ in concert with the CNN model is novel, we employ standard choices, such as quadratic loss and 3x3 convolutional kernels, for the neural network architecture and neural network training.

7.4.5 Hybrid approaches

NWP models represent the physical laws of atmospheric dynamics through a set of differential equations. Statistical or machine learning based approaches, on the other hand, do not encode physical laws but learn patterns based exclusively on past data. A hybrid model is a combination of both approaches and thus can benefit from both the physical expertise embodied in NWP output and the flexibility of data-driven approaches. In this paper, we base hybrid approaches on the deterministic HRES forecast from section 7.4.2 and the deterministic CNN forecast from section 7.4.4. We consider three approaches to obtain probabilistic forecasts from the deterministic HRES and CNN forecasts. First, the NWP forecast can be used as an additional gridded feature in the CNN model, followed by grid point based application of EasyUQ. Secondly, we can apply IDR using both deterministic forecasts as input features. Lastly, a simple approach is to use a weighted or unweighted average of the predictive distributions generated by HRES+EasyUQ and CNN+EasyUQ. We found experimentally that the first two approaches do not improve predictive ability, generally showing similar forecast performance to the CNN+EasyUQ forecast. The last approach in its most basic form of an equal average between the HRES+EasyUQ and CNN+EasyUQ distributions, which does not require any additional training, shows slight forecast improvements. It is therefore selected and referred to as the Hybrid model.

7.5 Forecast evaluation

In this section we report major findings from the forecasting experiment. The discussion concentrates on the peak monsoon season JAS, but results are also provided for the other seasons. As described at the start of section 7.4, our experiment uses expanding training sets to learn the forecasting models, and we frequently report annual results from the evaluation folds for 2011, $\ldots$, 2019. As evaluation metrics, the mean Brier score (BS) and the mean continuous ranked probability score (CRPS) in-

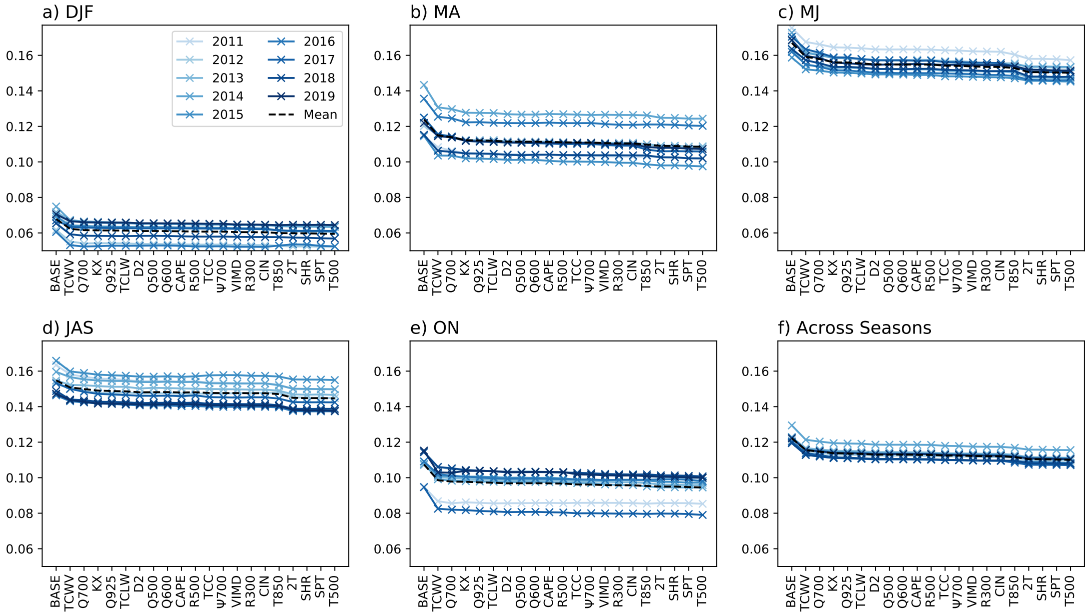

Figure 7.6: Mean Brier score (BS) for the Logit PoP forecast under successive addition of the predictor variables displayed on the horizontal axis. The base model includes three correlated rainfall predictors and two time features. The BS is averaged over space and season a) DJF, b) MA, c) MJ, d) JAS, and e) ON, and f) across seasons for evaluation folds from 2011 to 2019.

troduced in Section 2.4 are used.

7.5.1 Effects of variable selection in statistical models

To better understand the influence of the predictor variables on the forecast performance of the statistical models, namely, the Logit PoP forecast (section 7.4.3) and the DIM forecast for precipitation accumulation (section 7.4.3), a visual analysis is provided in Figure 7.6 and Figure 7.7, respectively. Starting with the mean score of the base model, which has five predictor variables, one more variable is successively added and the corresponding mean score is shown, until the full model with 25 predictor variables is reached. The variables are selected in the order of the distance between 0.5 and the mean AUC respectively CPA computed without splitting into seasons.[3]

Although the overall level of the scores varies strongly between seasons, the results are qualitatively similar. Therefore, the subsequent interpretation of the visual dis-

[3] An AUC or CPA value of 0.5 suggests a useless feature.

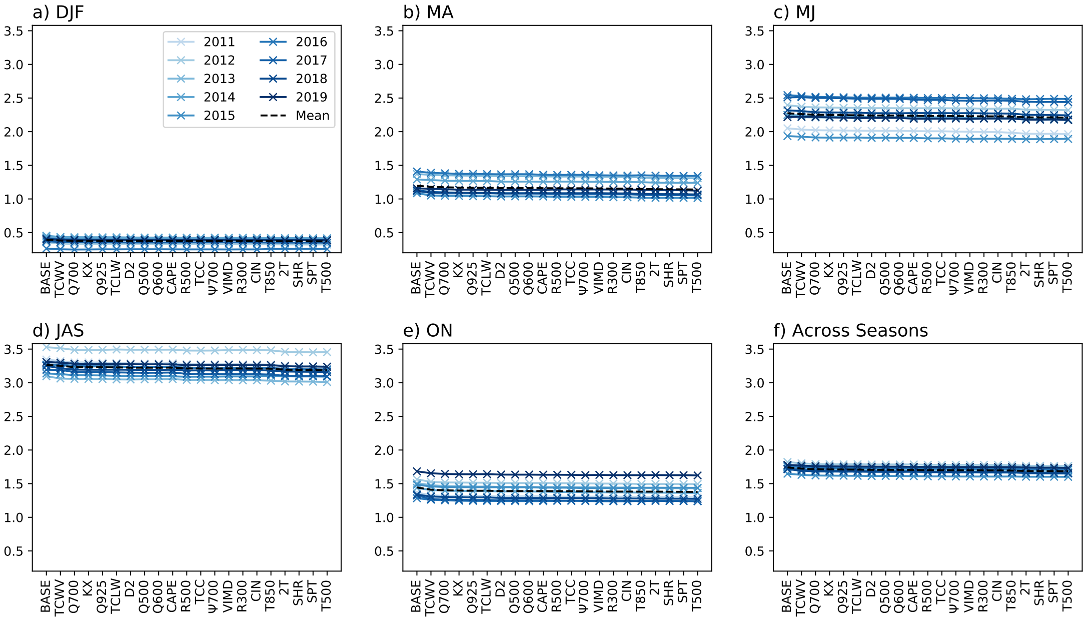

Figure 7.7: Mean continuous ranked probability score (CRPS) for probabilistic forecasts from DIM for precipitation accumulation in millimeters under successive addition of the predictor variables displayed on the horizontal axis. The base model includes three correlated rainfall predictors and two time features. The CRPS is averaged over space and season a) DJF, b) MA, c) MJ, d) JAS, and e) ON, and f) across seasons for evaluation folds from 2011 to 2019.

plays focuses on season JAS. Figure 7.6d shows that for season JAS the addition of TCWV to the Logit base model yields an improvement of the BS on the order of 5% in all years. Small further improvements of less than 1% are obtained by adding mid-level humidity (Q700) and static stability (KX). The addition of further variables yields minor improvement only, with the striking exception of 2m temperature (2T), which leads to an improvement comparable to Q700 and KX, despite AUC values barely above 0.5 (Figure 7.2d). Qualitatively, improvements in CRPS per predictor regarding precipitation amount (Figure 7.7d) show similar results, yet the percentage improvements are smaller such that adding variables other than TCWV and Q700 barely improves performance. Generally, the performance difference between years is large, and the ranking of the years differs between the BS, where the lowest values are seen for 2017, and the CRPS, where they are seen for 2013.

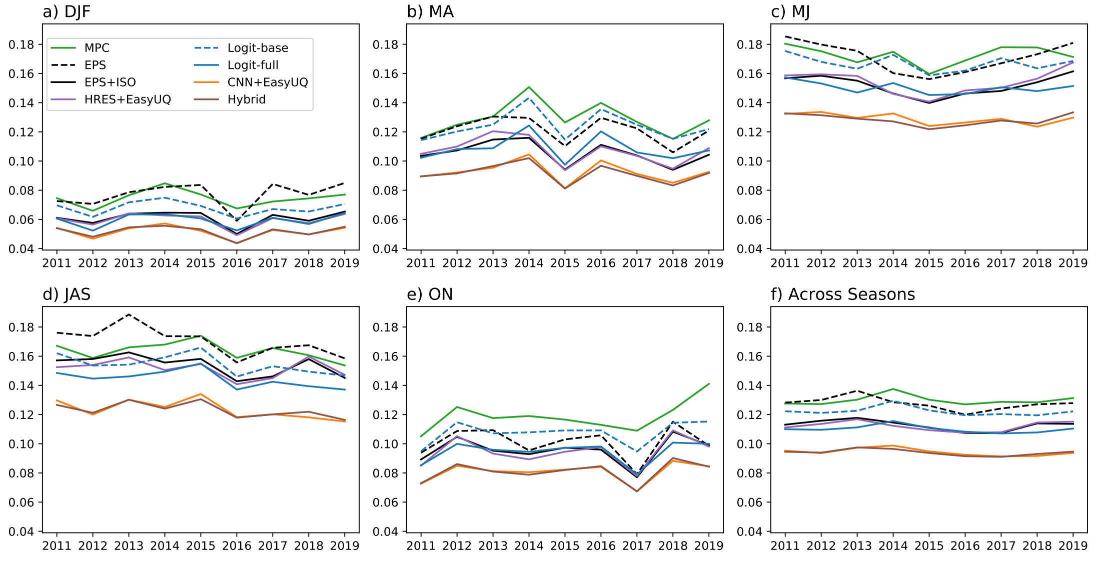

Figure 7.8: Mean Brier score (BS) for PoP forecasts from Table 7.2 in season a) DJF, b) MA, c) MJ, d) JAS, e) ON, and f) across seasons, for evaluation folds from 2011 to 2019.

7.5.2 Comparative evaluation of predictive performance

Figure 7.8 visualizes the mean Brier score (BS) for the PoP forecasting models from Table 7.2. For season JAS, the results are similar to the findings in Vogel et al. (2021). The ECMWF ensemble prediction system (EPS) shows inferior or, in later years, comparable performance to MPC, and both EPS and MPC are outperformed by a simple logistic regression approach based on correlated rainfall predictors only (Logit-base). The inclusion of ERA5 predictors into the logistic regression model (Logit-full) leads to a clear improvement even beyond the postprocessed EPS-ISO and HRES+EasyUQ PoP forecasts. Surprisingly, the HRES+EasyUQ forecast shows better performance than the ensemble-based EPS+ISO forecast. The CNN+EasyUQ forecast outperforms all other methods, except for the Hybrid forecast, which shows nearly the same performance. The ranking of the forecasting methods based on their average BS remains consistent across seasons. Throughout, the CNN+EasyUQ and Hybrid forecasts perform similarly to each other, and outperform their competitors by considerable margins.

The mean CRPS for the forecasting models for precipitation accumulation from Table 7.2 is displayed in Figure 7.9. Through 2014 in season JAS, EPS clearly shows the lowest forecast skill; thereafter, its skill improves and gets close to the performance of MPC and DIM-base. Unlike the Logit-full PoP forecasts, DIM-full does not outperform the

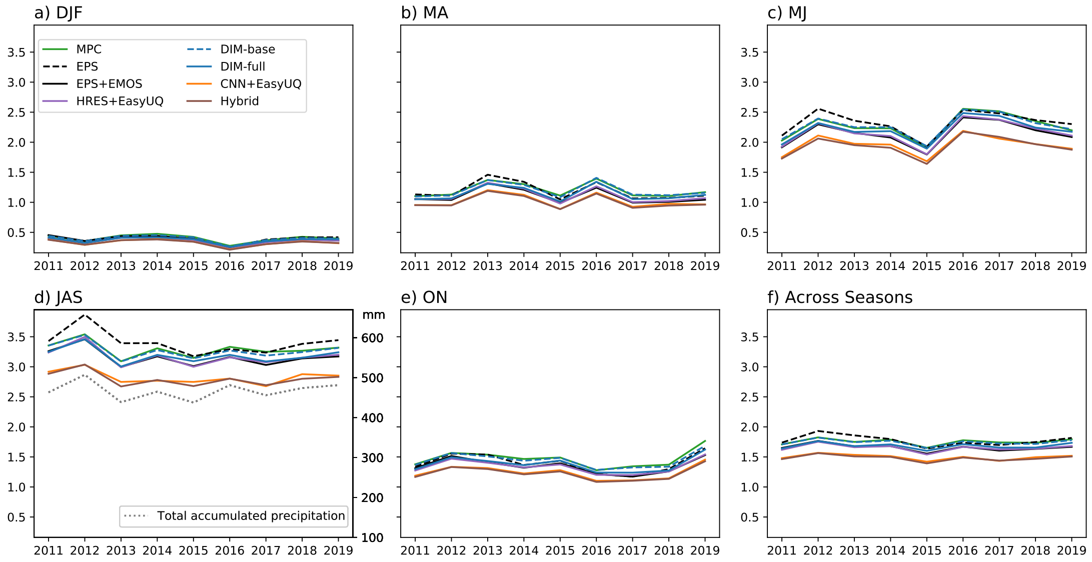

Figure 7.9: Mean continuous ranked probability score (CRPS) for probabilistic forecasts of precipitation accumulation in the millimeters from Table 7.2 in season a) DJF, b) MA, c) MJ, d) JAS, e) ON, and f) across seasons, for evaluation folds from 2011 to 2019.

postprocessed EPS+EMOS forecast. The HRES+EasyUQ approach yields better scores than EPS+EMOS, probably due to the flexibility of the EasyUQ forecast distributions. The CNN+EasyUQ approach shows a considerable forecast improvement within the evaluation period, and the Hybrid model performs similar or slightly better for some years. As can be seen by the dotted light gray line giving the JAS area-averaged rainfall in panel d), the mean CRPS co-varies with the total rainfall amount, thus the years with the best performance are usually also the driest. The ranking of the forecasting methods based on their average CRPS remains consistent across seasons. Similar to the binary setting, the CNN+EasyUQ and Hybrid forecasts show comparable performance, and outperform their competitors by considerable margins across all seasons.

7.5.3 Spatial structure of predictive performance

To facilitate the assessment of forecast performance relative to a baseline, skill scores can be used, defined as the quantity $(\overline{\mathrm{S}}_{\text{base}} - \overline{\mathrm{S}}_{\text{fcst}})/\,\overline{\mathrm{S}}_{\text{base}}$, where $\overline{\mathrm{S}}_{\text{fcst}}$ is the mean score of the forecast at hand and $\overline{\mathrm{S}}_{\text{base}}$ is the mean score of the baseline. A positive (negative) BS or CRPS skill score corresponds to predictive performance better (worse) than the baseline.

For an understanding of spatial patterns of forecast performance, skill score plots of the forecast approaches considered here with MPC as reference forecast are shown in Figure 7.10 for precipitation occurrence and in Figure 7.11 for precipitation accumulation, both for the JAS peak monsoon season and across evaluation folds.

With respect to the PoP forecasts for rainfall occurrence, EPS shows negative skill relative to MPC over the southern parts of the study domain, particularly over the relatively dry areas along the Guinea coast, over Gabon and southern Cameroon, where rainfall tends to be rather localized and short-lived such that precipitation occurrence is hard to predict (Figure 7.10a). Senegal/Mauritania and Chad/Sudan are the only areas with considerable positive skill, while the rest of the domain ranges close to zero. Applying statistical postprocessing (EPS+ISO, Figure 7.10b) removes the large negative skill along the Guinea Coast but shows remaining issues in a stretch from Nigeria to South Sudan with mostly weakly negative skill. Remarkably, postprocessing deteriorates skill around the highlands in Guinea/Sierra Leone and westernmost Ethiopia. Over the Sahel, in contrast, the postprocessing leads to an overall improvement and consistently positive skill. A possible reason is the stronger influence of predictable features such as AEWs or midlatitude perturbations here in contrast to the more stochastic rains in the south (Rasheeda Satheesh et al., 2023). The comparison between the EPS+ISO and HRES+EasyUQ (Figure 7.10c) demonstrates that for forecasts at individual sites there is no added value in running an NWP ensemble system, even after postprocessing. The structures are fairly consistent (e.g., with problematic regions in Guinea/Sierra Leone, the Central African Republic, South Sudan, and Ethiopia) but the values are consistently more positive for the HRES+EasyUQ technique, which is based on HRES model alone, as opposed to using an ensemble.

Moving to the data-based approaches (Figure 7.10d—g) we see consistent improvement over most areas of the study domain, though PoP forecasts for western Ethiopia remain a challenge, possibly related to the rough topography in this area. While in the simpler Logit-base approach (Figure 7.10d) some areas of negative skill remain, the inclusion of additional predictors in Logit-full (Figure 7.10e) leads to a consistent improvement and thus positive skill almost everywhere in the study region. It is also noteworthy that the Logit models generate overall smoother skill fields compared to the physics-based approaches. Finally, the CNN+EasyUQ and Hybrid methods (Figure 7.10f,g) outperform all other approaches to a large extent, reaching up to 40% improvement relative to the climatological benchmark MPC. The improvement relative to EPS is particular impressive over the Guinea coastal region (e.g., Ivory Coast and Ghana), where EPS performs much worse than MPC, for an illustration of the

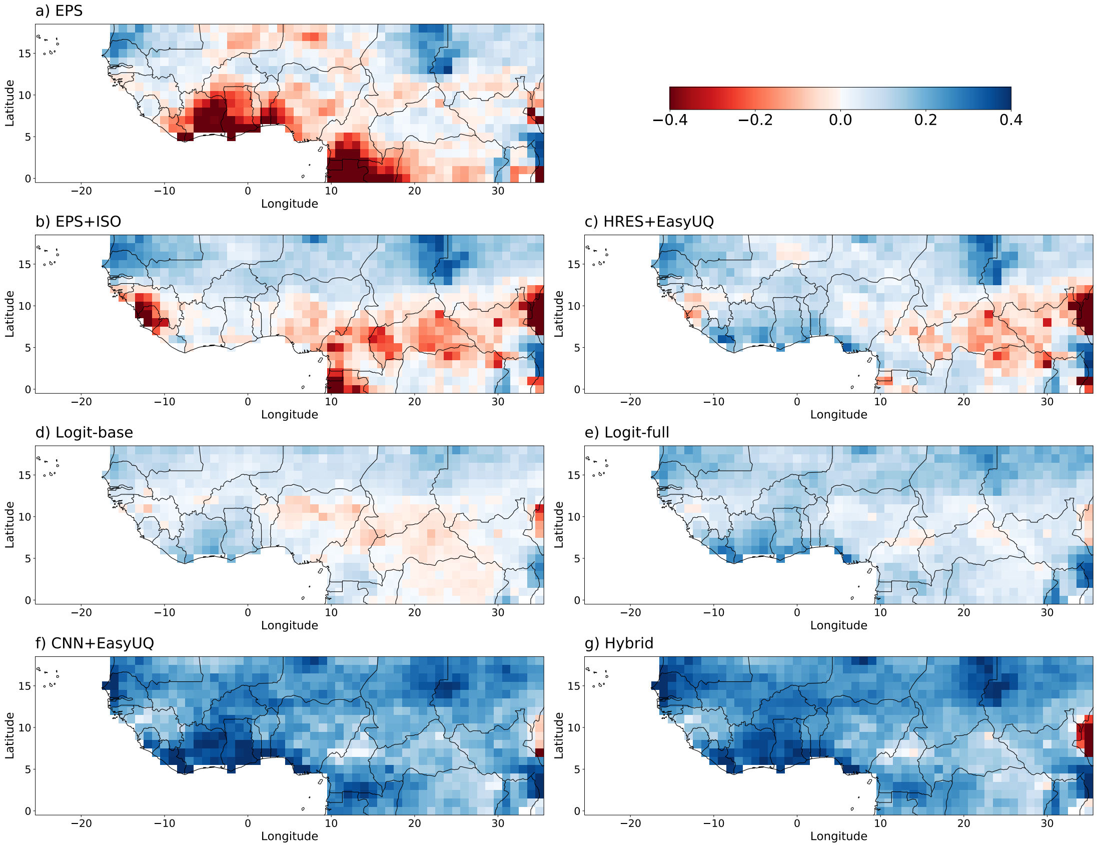

Figure 7.10: Spatial structure of the Brier skill score for probability forecasts of precipitation occurrence with a) EPS, b) EPS+ISO, c) HRES+EasyUQ, d) Logit-base, e) Logit-full, f) CNN+EasyUQ, and g) the Hybrid forecast from Table 7.2, relative to MPC as baseline, for season JAS and combined evaluation folds from 2011 to 2019.

ability of the CNN to learn complex physical relationships that determine local rainfall probability. The inclusion of NWP information from the HRES model in the Hybrid approach yields small improvements in some places but no clear advance relative to CNN+EasyUQ. This demonstrates that knowing the ambient conditions shortly before the beginning of the 24-hour forecast period is much more important than knowledge of the forecast evolution during that period.

The corresponding analysis for rainfall amount (Figure 7.11) reveals many parallels to rainfall probability. EPS (Figure 7.11a) stands out as having many areas of negative CRPS skill, with an overall similar structure to the occurrence analysis (Figure 7.10a). Postprocessing (EPS+EMOS, Figure 7.11b) cures many issues of EPS, leading to mostly weakly positive skill, but does not perform as well as the computationally much less expensive HRES+EasyUQ technique (Figure 7.11c). Here the skill fields for amount are

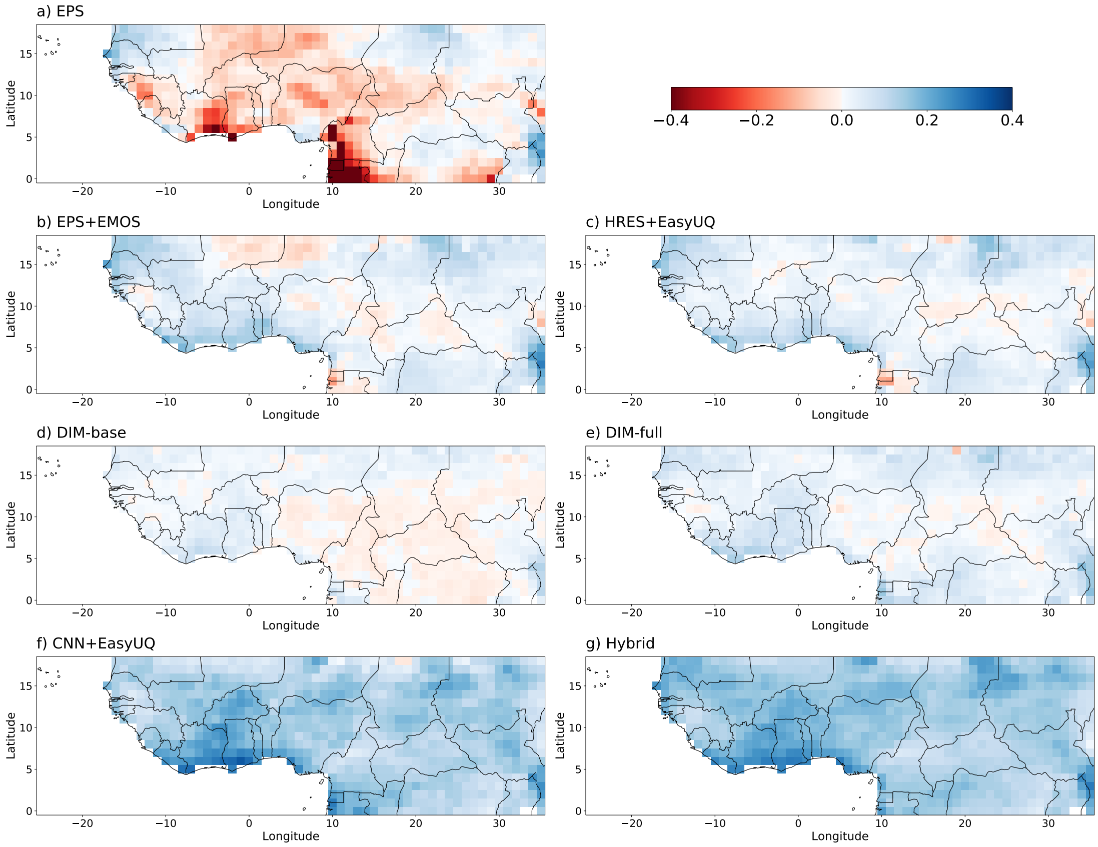

Figure 7.11: Spatial structure of the CRPS skill score for probabilistic forecasts of precipitation accumulation with a) EPS, b) EPS+EMOS, c) HRES+EasyUQ, d) DIM-base, e) DIM-full, f) CNN+EasyUQ, and g) the Hybrid forecast from Table 7.2, relative to MPC as baseline, for season JAS and combined evaluation folds from 2011 to 2019.

overall smoother than for occurrence with less contrast between the Sahel and the southern areas. The DIM models (replacing the Logit models for amount) show negligible further advance. The skill of DIM-base (Figure 7.11d) is close to zero everywhere with a negative area in the southeast and positive elsewhere, while the inclusion of additional predictors (DIM-full, Figure 7.11e) slightly improves skill over most areas. Finally, as for occurrence, the machine learning based CNN+EasyUQ and Hybrid methods (Figure 7.11f,g) outperform all other approaches to a large extent with positive CRPS skill of up to 30%. Here the Hybrid approach leads to a more considerable improvement relative to CNN+EasyUQ, yielding fairly equal skill improvement across the entire, quite heterogeneous domain. These improvements are more prominent in areas where the physics-based HRES model may better represent the time evolution of dynamical features such as AEWs and extratropical influences.

7.5.4 Calibration and discrimination ability

We now assess the calibration and discrimination ability of the forecasts. Following Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), reliability diagrams for the PoP forecasts from Table 7.2 at the grid point closest to Niamey ($13°$N, $2°$E) are presented in Figure 7.12. [4] The panels use the CORP approach of Dimitriadis et al. (2021) and show the decomposition (see Appendix 6.A) of the mean Brier score (BS) into miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components. Instead of considering each evaluation fold separately, the decomposition is computed once on forecasts in the peak monsoon season JAS from all nine evaluation years together. If the reliability curve is close to the diagonal, a PoP forecast is calibrated (reliable). Deviations from the diagonal indicate some type of miscalibration: S-shaped curves indicate underconfidence (PoP too close to center), inverse S-shaped curves correspond to overconfidence (PoP too extreme), and curves that are mostly below (above) the diagonal indicate biased PoP. The climatological MPC PoP forecast has a very limited range of forecast probabilities and lacks discrimination ability, but shows excellent calibration. The poor calibration of the raw ENS PoP is corrected by postprocessing (ENS+ISO). In agreement with the findings in Vogel et al. (2021) and Rasheeda Satheesh et al. (2023), the Logit-base PoP forecast is well calibrated and has moderate discrimination ability. In comparison, Logit-full shows a lower BS (more skillfull PoP forecasts) reflected in both better calibration and improved discrimination ability. The CNN+EasyUQ and Hybrid techniques show superior performance — they are similarly well calibrated as EPS-ISO and Logit-full but show considerably higher discrimination ability.

To assess the calibration of the probabilistic forecasts for accumulated precipitation at the grid point closest to Niamey, Figure 7.13 shows Probability Integral Transform (PIT) histograms. For the MPC and EPS ensemble forecast, a universal PIT (uPIT) histogram is shown (Vogel et al., 2018); for the other methods, the randomized version of the PIT is used (Gneiting and Resin, 2023, eq. (1)). A uniform histogram indicates calibrated forecasts while a U-shaped (hump-shaped) histogram suggests underdispersed (overdispersed) forecasts, meaning that the forecasts are overconfident (underconfident). Skewed histograms indicate biases. The ECMWF ensensemble (EPS) is underdispersed, which is corrected for in the EPS+EMOS forecast, though a bias remains. The other forecasts show PIT histograms that are nearly uniform. The associated decomposition (6.1) of the mean CRPS demonstrates the superior calibration of

[4] Python code for computation and plotting is available at `https://github.com/evwalz/corp_reldiag`.

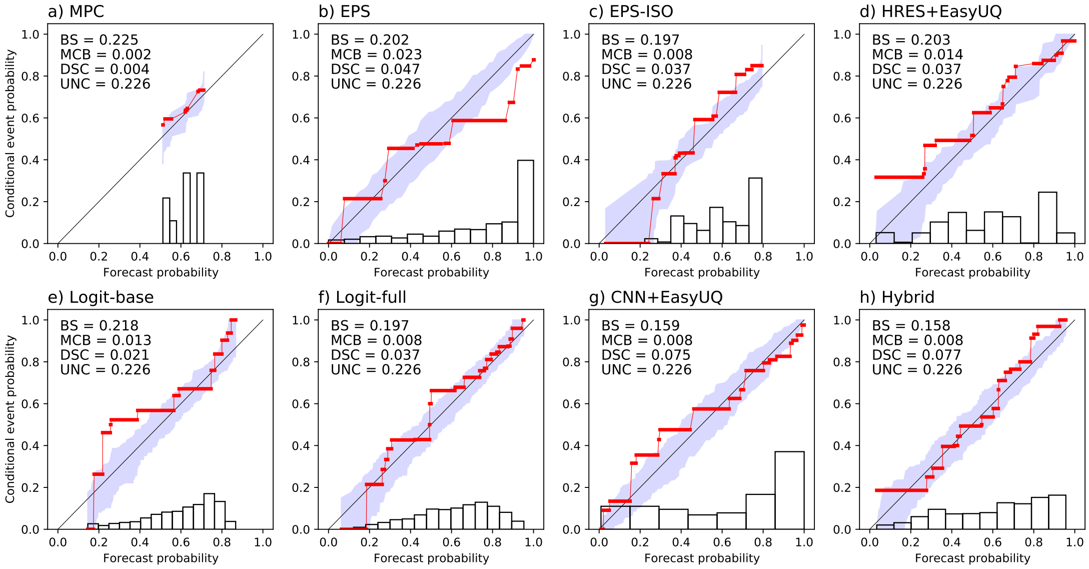

Figure 7.12: Reliability diagrams for PoP forecasts at the grid point closest to Niamey ($13°$N, $2°$E) with a) MPC, b) EPS, c) EPS+ISO, d) HRES+EasyUQ, e) Logit-base, f) Logit-full, g) CNN+EasyUQ, and h) the Hybrid approach from Table 7.2, for season JAS and combined evaluation folds from 2011 to 2019, with 90% consistency bands under the assumption of calibration (Dimitriadis et al., 2021). The panels also show the mean Brier score (BS) and its miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components from (6.1). The histograms along the horizontal axis show the distribution of the forecast probabilities.

the climatological MPC forecast and the outstanding discrimination ability and overall predictive performance of the CNN+EasyUQ and Hybrid approaches.

Finally, we use the decomposition of the mean Brier score (BS) (see Appendix 6.A) or mean continuous ranked probability score (CRPS) (see Section 6.3) into miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components for a spatially aggregated quantitative assessment. We compute the decomposition at each grid point based on forecasts for all five seasons from all nine evaluation years, and the score components are then averaged across grid points. The miscalibration–discrimination (MCB–DSC) plots for the mean BS (Figure 7.14) and mean CRPS (Figure 7.15) provide a spatially consolidated comparison of the forecast methods. In all panels, the climatological MPC forecast shows the lowest MCB and the lowest DSC component. The ECMWF raw ensemble (EPS) has higher MCB than all other methods, and the miscalibration is taken care of by postprocessing (EPS+ISO, EPS+EMOS). Regarding the statistical forecasts, the inclusion of the ERA5 predictors (Logit-full, DIM-full) models

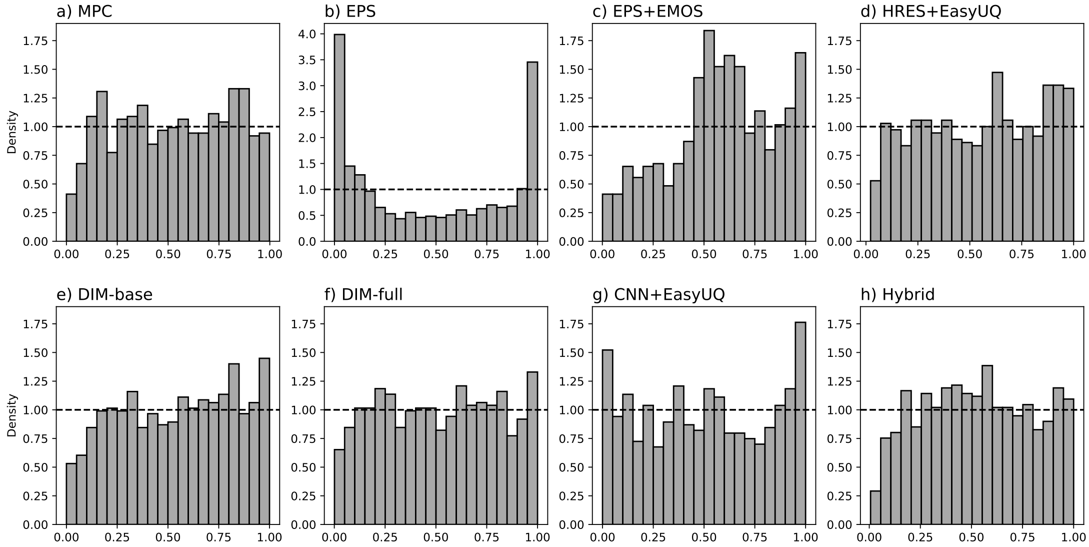

Figure 7.13: Probability Integral Transform (PIT) histograms for probabilistic forecasts of precipitation accumulation at the grid point closest to Niamey ($13°$N, $2°$E) with a) MPC, b) EPS, c) EPS+EMOS, d) HRES+EasyUQ, e) DIM-base, f) DIM-full, g) CNN+EasyUQ, and h) the Hybrid approach from Table 7.2, for season JAS and combined evaluation folds from 2011 to 2019. The vertical scale of the histograms is shared across forecasts, except for EPS. The panels also show the mean continuous ranked probability score (CRPS) and its miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components from (6.1). The vertical scale of the histograms is shared across forecasts, except for EPS.

in addition to the correlated rainfall predictors (Logit-basic, DIM-basic) improves DSC while MCB remains similar. The superiority of the CNN+EasyUQ forecast stems from its elevated discrimination ability. The Hybrid forecast shows slightly improved skill relative to CNN+EasyUQ, and trades better calibration for even higher discrimination ability.

7.6 Discussion

In this chapter the predictability of one-day ahead, 24-hour precipitation occurrence and amount over northern tropical Africa is investigated. Our study builds on previous papers with focus on forecasting rainfall occurrence for the summer season JAS, which compared the performance of climatological, raw and postprocessed ECMWF ensemble forecasts, and a simple logistic regression model based on correlated rainfall

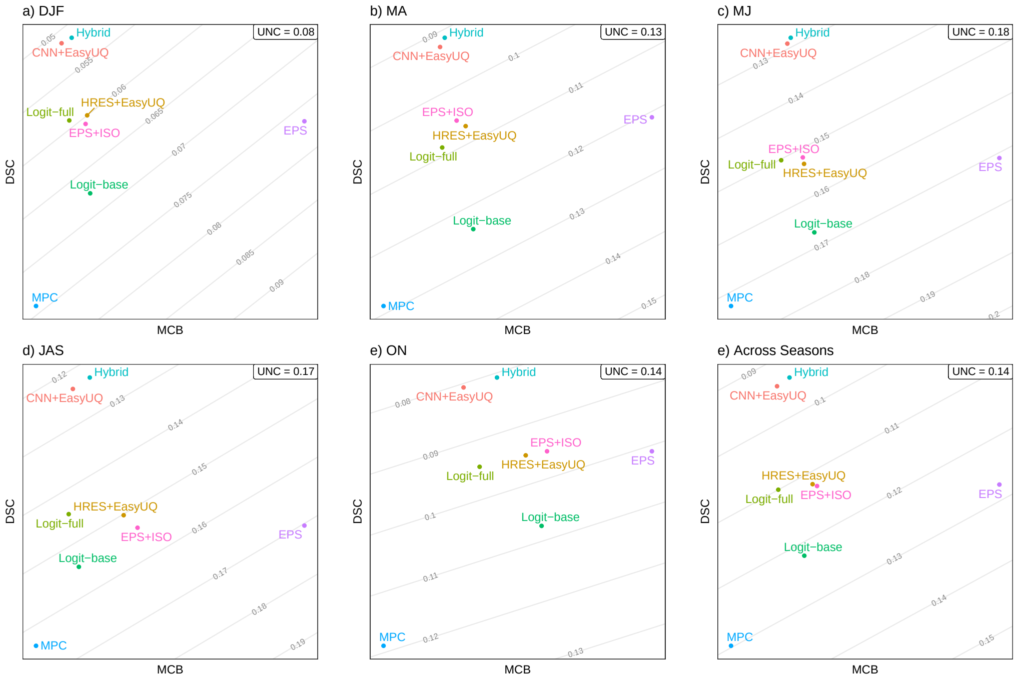

Figure 7.14: Miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) component of the mean Brier score (BS) in season a) DJF, b) MA, c) MJ, d) JAS, e) ON, and f) across seasons for PoP forecasts from Table 7.2. The CORP decomposition of Dimitriadis et al. (2021) is applied at each grid point, based on the combined evaluation folds from 2011 to 2019, and the mean score and score components are then averaged over grid points. Parallel lines correspond to equal mean scores.

predictors. This binary forecasting problem is revisited in this chapter with major adaptions. Instead of TRMM, GPM IMERG is used as ground truth data source. Forecasts are produced for the entire year instead of just the summer season (JAS) and ERA5 predictor variables are used to augment the logistic regression model. To this end, an extensive analysis of weather variables from ERA5 is performed to investigate and understand their relation to and their influence on precipitation. The meteorological interpretation of these dependencies is obtained by combining previously conducted research and results from statistical analysis performed in this work.

A key contribution of our work is that we additionally investigate the more challenging problem of producing probabilistic forecasts for accumulated precipitation. Since the climatology and the NWP model output in this chapter are in the form of ensembles, they can be readily used as probabilistic forecasts for precipitation amount. To

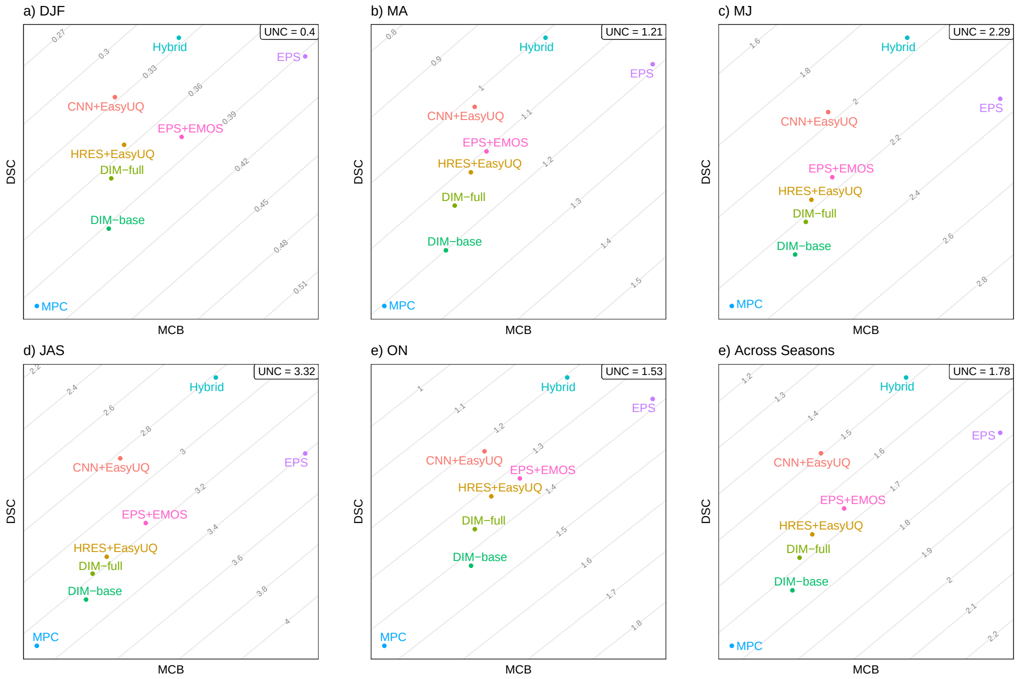

Figure 7.15: Miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) component of the mean continuous ranked probability score (CRPS) in season a) DJF, b) MA, c) MJ, d) JAS, e) ON, and f) across seasons for probabilistic forecasts of precipitation accumulation in millimeters from Table 7.2. The isotonicity-based decomposition from Section 6.3 is applied at each grid point, based on the combined evaluation folds from 2011 to 2019, and the mean score and the score components are then averaged over grid points. Parallel lines correspond to equal mean scores.

produce data-driven statistical forecasts, the Distributional Index Model (DIM) is introduced, which is simple but very effective and thus can serve as a persuasive baseline. To account for the recent rise of machine learning in weather forecasting, a CNN model is presented which has the additional benefit of inherently exploiting spatial relations. To obtain a probabilistic output, we couple the CNN model with the recently introduced EasyUQ approach, to yield the CNN+EasyUQ technique. In summary, these different forecasting approaches provide a detailed forecasting benchmark covering the range of simple to sophisticated models and ideas from NWP, statistics, and machine learning.

The CNN+EasyUQ technique outperforms its competitors by a large margin, except for the Hybrid forecast, which is a simple arithmetic average of the HRES+EasyUQ and

CNN+EasyUQ forecast distributions that does not require any additional training and yields minor only (if any) further improvement. It is interesting to place our results for one-day ahead, 24-hour forecasts in the context of recent advances in data-based precipitation forecasts. For nowcasts at prediction horizons up to 12 hours, progress has been persuasive (Ayzel et al., 2020; Lagerquist et al., 2021; Ravuri et al., 2021; Espeholt et al., 2022; Zhang et al., 2023). In stark contrast, recent developments in neural network based weather forecasts at prediction horizons of days ahead have provided sparse attention to rainfall (Bi et al., 2023; Rasp et al., 2023), arguably due to the recognition that "precipitation is sparse and non-Gaussian" (Lam et al., 2023, p. 6). The CNN+EasyUQ technique provides an elegant and computationally highly efficient way of addressing the non-Gaussianity of precipitation accumulation. In very recent work, Andrychowicz et al. (2023) find that the data-driven MetNet-3 approach outperforms the ECMWF and NOAA raw ensembles in terms of CRPS for hourly precipitation accumulation over the continental United States at lead times up to 20 hours, but not beyond. However, unlike our study, which compares the CNN+EasyUQ forecast with state of the art competitors, Andrychowicz et al. (2023) do not compare MetNet-3 to postprocessed NWP ensemble forecasts, nor to statistical forecasts of the type considered here.

The reproduction of the results in this chapter requires access to GPM IMERG precipitation data, predictor variables from ERA5, and ECMWF NWP forecasts. The first two sources are freely accessible, which makes results for MPC, the statistical approaches (Logit and DIM), and our key innovation, the CNN+EasyUQ technique, readily reproducible. For the more elaborate CNN+EasyUQ approach, code in Python (Python Software Foundation, 2021) is publicly available at `https://github.com/evwalz/precipitation`. The raw ECMWF EPS, the postprocessed versions thereof, the HRES+EasyUQ forecast, and the Hybrid model require access to ECMWF NWP forecasts which are freely available using the TIGGE (The International Grand Global Ensemble) archive (Bougeault et al., 2010) instead of MARS from ECMWF.

In view of its outstanding performance in this study, the CNN+EasyUQ approach can likely improve operational probabilistic forecasts of day ahead, 24-hour rainfall in northern tropical Africa. To make real-time forecasts feasible, one would need to use the IMERG Early Run (`https://gpm.nasa.gov/taxonomy/term/1357`) in lieu of IMERG, which is an option that remains to be tested. To obtain ensemble forecasts of entire, spatio-temporally coherent precipitation fields, rather than forecasts at individual locations and fixed prediction horizons, the HRES+EasyUQ and CNN+EasyUQ approaches can be coupled with empirical copula techniques (Clark et al., 2004; Schefzik et al., 2013), for which we encourage follow-up studies. While our study is limited in geographic scope, we feel that data-driven approaches of this type have potential throughout the tropics. Furthermore, the results of comparative studies by Little et al. (2009) for the United Kingdom and Andrychowicz et al. (2023) for the continental United States admit the speculation that the CNN+EasyUQ technique can improve probabilistic forecasts of 24-hour precipitation in the extratropics as well. Finally, a very interesting and relevant research question is whether similar advances in predictive performance are feasible at prediction horizons larger than a day ahead.

8 | Conclusion

The research in this thesis focused on statistical forecasting and evaluation within the context of real-valued outcomes, with a particular emphasis on the development and assessment of probabilistic forecasts. Therefore, this work started with an introduction section, in which the forecasting cycle was elucidated to provide the reader with a better understanding and to distinguish the purpose of the newly developed statistical methods by connecting them to relevant forecasting steps. Subsequently, this work investigated the CPA measure in Chapter 4, developed EasyUQ and Smooth EasyUQ in Chapter 5 and introduced the isotonicity-based CRPS decomposition in Chapter 6. Each of these methods was tailored to address a distinct challenge in real-valued forecasting problems. Consequently, each individual chapter outlined the motivation, provided arguments for the specific construction and offered a detailed investigation based on case studies. While these developments were designed as stand-alone concepts, Chapter 7 demonstrated how the three tools nicely complement each other and emphasized that this collection of methods offers a persuasive approach to successfully apply the forecasting cycle in the context of real-valued outcomes with corresponding probabilistic forecasts. More specifically, this work contributed three methods.

Firstly, the CPA measure, defined in Chapter 4, extends the classical AUC value to general real-valued data while maintaining desirable properties of ROC analysis. It equals the AUC measure for binary outcomes and is linearly related to Spearman's rank correlation coefficient if feature and outcome are continuous. Since the CPA is asymmetric it is particularly well suited to the purpose of feature screening and variable selection. The data examples in Subsection 4.5, highlight the usage of CPA and relate it to other rank based measures like the C index, and differentiate it from RMSE which is commonly used as evaluation measure. In addition, Chapter 7 demonstrated how the CPA measure can be used to apply the statistical data analysis step for real-valued forecasting problems. Future work may focus on a deeper theoretical investigation of the CPA measure. Furthermore, suitable adaptions of ROCM, UROC curve and CPA to specific types of data could be developed, e.g., modifications to properly handle censoring in survival analysis.

Secondly, the EasyUQ approach, introduced in Chapter 5, is a simple yet effective method to transform deterministic forecasts into calibrated predictive distributions based on a training set of model output–outcome pairs and a natural assumption of isotonicity. The method is fully automated and readily adapts to the underlying true outcome distribution without the need to specify a suitable parametric distribution beforehand. As shown in the case studies, EasyUQ is competitive to state of the art approaches. For "nice" distributions, it performs similar to conformal prediction (CP) but shows clear advantages for more "difficult" distributions. The more elaborate Smooth EasyUQ approach generates predictive distributions with Lebesgue densities, based on a kernel smoothing of the original IDR distributions, while preserving the key properties of the basic approach. Future work could focus on investigating a coupling of EasyUQ and Mondrian CP (Boström et al., 2021), improving the computational runtime of Smooth EasyUQ, comparing it to newly developed approaches based on diffusion models (Han et al., 2022) and adapting EasyUQ for specific domain problems, e.g., combining EasyUQ with empirical copula techniques in weather forecasting (Schefzik et al., 2013).

Finally, the isotonicity-based CRPS decomposition in Chapter 6 provides a persuasive technique to decompose a mean CRPS value into more informative components, namely MCB, DSC and UNC. Both theoretically and computationally, the isotonicity-based decomposition serves as an attractive alternative to the Candille–Talagrand decomposition, which is theoretically appealing, but yields degenerate decompositions in practice. Future work might focus on providing details of a generalization of the isotonicity-based decomposition to other proper scoring rules, such as the weighted CRPS. However, such a generalization fails if a mean of logarithmic scores is sought to be decomposed, as the logarithmic score can not be applied to the discrete IDR distributions.

For each method, corresponding code is available in R, Python, or both to facilitate usage for practitioner and ensure easy access to the newly developed tools across various research disciplines.

All in all, this thesis introduced three advanced methodologies that not only address current challenges effectively but also suggest future research endeavors. While the primary emphasis of this work is evident in the comprehensive exploration of precipitation forecasting problem, the broader implications extend far beyond this specific domain. The forecasting steps and associated tools demonstrated here offer a versatile framework that can readily be applied to general real-valued outcomes, particularly emphasizing probabilistic forecasting.

Bibliography

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Adams, N. M. and Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32:1139–1147.

Ageet, S., Fink, A. H., Maranan, M., and Schulz, B. (2023). Predictability of rainfall over equatorial East Africa in the ECMWF ensemble reforecasts on short to medium-range time scales. *Weather and Forecasting*. In press, `https://doi.org/10.1175/WAF-D-23-0093.1`.

Alley, R. B., Emanuel, K. A., and Zhang, F. (2019). Advances in weather prediction. *Science*, 363:342–344.

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *British Medical Journal*, 332.

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., Agrawal, S., and Kalchbrenner, N. (2023). Deep learning for day forecasts from sparse observations. Preprint, `arXiv:2306.06079`.

Arnold, S., Henzi, A., and Ziegel, J. (2023a). Sequentially valid tests for forecast calibration. *Annals of Applied Statistics*, 17:1909–1935.

Arnold, S., Walz, E.-M., Ziegel, J., and Gneiting, T. (2023b). Decompositions of the mean continuous ranked probability score. Preprint, `arXiv:2311.14122`.

Arnold, S. and Ziegel, J. (2023). Isotonic conditional laws. Preprint, `arXiv:2307.09032`.

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

Ayzel, G., Scheffer, T., and Heistermann, M. (2020). RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13:2631–2644.

Baker, E., Barbillon, P., Fadikar, A., Gramacy, R. B., Herbei, R., Higdon, D., Huang, J., Johnson, L. R., Ma, P., Mondal, A., Pires, B., Sacks, J., and Sokolov, V. (2022). Analyzing stochastic computer models: A review with opportunities. *Statistical Science*, 37:64–89.

Barlow, Bartholomew, Bremer, and Brunk (1972). *Statistical Inference under Order Restrictions*. Wiley.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 535:47–55.

Becker, T., Bechtold, P., and Sandu, I. (2021). Characteristics of convective precipitation over tropical Africa in storm-resolving global simulations. *Quarterly Journal of the Royal Meteorological Society*, 147:4388–4407.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F. (2023). The rise of data-driven weather forecasting: A first statistical assessment of machine-learning based weather forecasts in an operational-like context. Preprint, `arXiv:2307.10128`.

Ben Bouallègue, Z., Magnusson, L., Haiden, T., and Richardson, D. S. (2019). Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, 145:1741–1755.

Ben Bouallègue, Z., Pinson, P., and Friederichs, P. (2015). Quantile forecast discrimination ability and value. *Quarterly Journal of the Royal Meteorological Society*, 141:3415–3424.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140:1924–1934.

Berger, J. O. and Smith, L. A. (2019). On the statistical formalism of uncertainty quantification. *Annual Review of Statistics and Its Application*, 6:433–460.

Bi, J. and Bennett, K. (2003). Regression error characteristic curves. In *20th International Conference on Machine Learning*, volume 1 of *Proceedings of Machine Learning Research*, pages 43–50.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538.

Blanche, P., Dartigues, J. F., and Jacqmin-Gatta, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Quarterly Journal of the Royal Meteorological Society*, 55:687–704.

Boström, H., Johansson, U., and Löfström, T. (2021). Mondrian conformal predictive distributions. In *Proceedings of the 10th Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 24–38.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., and Worley, S. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91:1059–1072.

Bowman, A., Hall, P., and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85:799–808.

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

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13:4015–4034.

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall.

Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, 60:663–678.

Brunk, H. D. (1965). Conditional Expectation given a σ-lattice and Applications. *Annuals of Mathematical Statistics*, 36:1339–1350.

Camporeale, E. and Caré, A. (2021). ACCRUE: Accurate and reliable uncertainty estimate in deterministic models. *International Journal for Uncertainty Quantification*, 11:81–94.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131:2131–2150.

Capéraà, P. and Genest, C. (1993). Spearman's ρ is larger than Kendall's τ for positively dependent random variables. *Journal of Nonparametric Statistics*, 2:183–194.

Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S., Lerch, S., and Hayatbini, N. (2022). Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, 150:215–234.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W. (2023). FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. Preprint, `arXiv:2304.02948`.

Christensen, D. (2005). Fast algorithms for the calculation of Kendall's τ. *Computational Statistics*, 20:51–62.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In *Advances in Neural Information Processing Systems 34*, pages 10971–10984. Curran Associates, Inc.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. (2004). The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5:243–262.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 147:278–290.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. (2021). Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems 34*, pages 20089–20103. Curran Associates, Inc.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. L. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10:1–7.

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

Dimitriadis, T., Gneiting, T., and Jordan, A. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences of the United States of America*, 118:e2016191118.

Dimitriadis, T., Gneiting, T., Jordan, A., and Vogel, P. (2023). Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*. In press.

D'Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning: Generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111.

Doubleday, K., Hernandez, V. V. S., and Hodge, B.-M. (2020). Benchmark probabilistic solar forecasts: Characteristics and recommendations. *Solar Energy*, 206:52–67.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. In *37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2690–2700.

Ebert-Uphoff, I. and Hilburn, K. (2023). The outlook for AI weather prediction. *Nature*, 619:473–474.

ECMWF Directorate (2012). Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, 133:11–13.

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

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic constraint: The k-sample case. *Journal of the American Statistical Association*, 100:252–261.

Embrechts, P. and Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, 77:423–432.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Jonathan, H., Carla, B., Cenk, G., CarverRob, AndrychowiczMarcin, Jason, H., Aaron, B., and Nal, K. (2022). Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13:5145.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*, 19:242–251.

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

Fernandez, C. and Steel, M. F. J. (2009). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, 86:153–167.

Ferri, C., Hernández-Orallo, J., and Salido, M. (2003). Volume under the ROC surface for multi-class problems. In *Proceedings of 14th European Conference on Machine Learning*, pages 108–120.

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

Fink, A. H., Engel, T., Ermert, V., van der Linden, R., Schneidewind, M., Redl, R., Afiesimama, E., Thiaw, W. M., Yorke, C., Evans, M., and Janicot, S. (2017). Mean climate and seasonal cycle. In Parker, D. J. and Diop-Kane, M., editors, *Meteorology of Tropical West Africa: The Forecasters' Handbook*, pages 1–39. Wiley, Chichester.

Flach, P. A. (2016). ROC analysis. In *Encyclopedia of Machine Learning and Data Mining*. Springer.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059.

Galvin, J. (2010). Two easterly waves in West Africa in summer 2009. *Weather*, 65:219–227.

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

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

Gebremichael, M., Yue, H., and Nourani, V. (2022). The accuracy of precipitation forecasts at timescales of 1–15 days in the Volta river basin. *Remote Sensing*, 14:937.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *Journal of the American Statistical Association*, 99:575–583.

Ghanem, R., Higdon, D., and Owhadi, H., editors (2017). *Handbook of Uncertainty Quantification*. Springer.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

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

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

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

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

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

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

Gneiting, T. and Walz, E.-M. (2022). Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Machine Learning*, 111:2769–2797.

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

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132:2925–2942.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33:563–594.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Haiden, T., Janousek, M., Vitart, F., Ben Bouallègue, Z., Ferranti, L., Prates, F., and Richardson, D. (2021). Evaluation of ECMWF forecasts, including the 2020 upgrade, `https://www.ecmwf.int/sites/default/files/elibrary/2021/19879-evaluation-ecmwf-forecasts-including-2020-upgrade.pdf`.

Haiden, T., Rodwell, M., Richardson, D., Okagaki, A., Robinson, T., and Hewson, T. (2012). Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Monthly Weather Review*, 140:2720–2733.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560.

Han, X., Zheng, H., and Zhou, M. (2022). CARD: Classification and regression diffusion models. Preprint, `arXiv:2206.07275`.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186.

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

Harrell, F. E., Jr. Lee, K. L., and Mark, D. B. (1996). Tutorials in biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Heagerty, P., Lumley, T., and Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–44.

Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105.

Henzi, A. (2023). Consistent estimation of distribution functions under increasing concave and convex stochastic ordering. *Journal of Business & Economic Statistics*, 41:1203–1214.

Henzi, A., Kleger, G.-R., and Ziegel, J. F. (2023). Distributional (single) index models. *Journal of the American Statistical Association*, 118:489–503.

Henzi, A., Mösching, A., and Dümbgen, L. (2022). Accelerating the pool-adjavent-violators algorithm for isotonic distributional regression. *Methodology and Computing in Applied Probability*, 24:2633–2645.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:963–969.

Herbrich, R., Graepel, T., and Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In Bartlett, P. J., Schölkopf, B., Schuurmans, D., and Smola, A. J., editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press.

Hernandéz-Lobato, J. M. and Adams, R. P. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1861–1869.

Hernández-Orallo, J. (2013). ROC curves for regression. *Pattern Recognition*, 46:3395–3411.

Hernández-Orallo, K., Flach, P., and Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification. *Journal of Machine Learning Research*, 13:2813–2859.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C.and Radu, R., Rozum, I., Schepers, D.and Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2018). Era5 hourly data on single levels from 1979 to present. copernicus climate change service (c3s) climate data store (CDS), `https://doi.org/10.24381/cds.adbb2d47`.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S.and Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloy-aux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146:1999–2049.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Methodology)*, 76:3–27.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., and Iguchi, T. (2014). The Global Precipitation Measurement mission. *Bulletin of the American Meteorological Society*, 97:701–722.

Huang, J. and Ling, C. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17:299–310.

Hubert, H. (1936). Origine Africaine d'un cyclone tropical Atlantique. *Annales de physique du globe de la France d'outre-mer*, 6:97–115.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K.-L., Joyce, R. J., Kidd, C., Nelkin, E. J., Sorooshian, S., Stocker, E. F., Tan, J., Wolff, D. B., and Xie, P. (2020). Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) mission (IMERG). In Levizzani, V., Kidd, C., Kirschbaum, D. B., Kummerow, C. D., Nakamura, K., and Turk, F. J., editors, *Satellite Precipitation Measurement: Volume 1*, pages 343–353. Springer, Cham.

Hyndman, R. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Australia, 3rd edition.

Immer, A., Bauer, M., Fortuin, V., Rätsch, G., and Khan, M. E. (2021). Scalable marginal likelihood estimation for model selection in deep learning. In *38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4563–4573.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12):1–37.

Jordan, A. I., Mühlemann, A., and 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.

Kiladis, G. N., Thorncroft, C. D., and Hall, N. M. (2006). Three-dimensional structure and dynamics of African easterly waves. Part I: Observations. *Journal of Atmospheric Sciences*, 63:2212–2230.

Klein, C., Nkrumah, F., Taylor, C. M., and Adefisan, E. A. (2021). Seasonality and trends of drivers of mesoscale convective systems in southern West Africa. *Journal of Climate*, 34:71–87.

Kniffka, A., Knippertz, P., Fink, A., Bendetti, A., Brooks, M. E., Hill, P., Maranan, M., Pante, G., and Vogel, B. (2020). An evaluation of operational and research weather forecasts for southern west africa using observations from the dacciwa field campaign in june - july. *Quarterly Journal of the Royal Meteorological Society*, 146:1121–1148.

Knight, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61:436–439.

Köhler, M., Schindler, A., and Sperlich, S. (2014). A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review*, 82:243–274.

Kohonen, J. and Suomela, J. (2006). Lessons learned in the challenge: Making predictions and scoring them. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, pages 95–116. Springer.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53:814–861.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804.

Lafore, J. P., Chapelon, N., Beucher, F., Diop Kane, M., Gaymard, A., Kassimou, A., Lepape, S., Mumba, Z., Orji, B., Osika, D., Parker, D. J., Poan, E., Razafindrakoto, L. G., and Vincendon, J. C. (2017a). West African synthetic analysis and forecast: WASA/F. In Parker, D. J. and Diop-Kane, M., editors, *Meteorology of Tropical West Africa: The Forecasters' Handbook*, pages 423–451. Wiley, Chichester.

Lafore, J. P., Chapelon, N., Diop, M., Gueye, B., Largeron, Y., Lepape, S., Ndiaye, O., Parker, D. J., Poan, E., Roca, R., Roehrig, R., Taylor, C., and Moncrieff, M. (2017b). Deep convection. In Parker, D. J. and Diop-Kane, M., editors, *Meteorology of Tropical West Africa: The Forecasters' Handbook*, pages 90–129. Wiley, Chichester.

Lagerquist, R., Stewart, J. Q., Ebert-Uphoff, I., and Kumler, C. (2021). Using deep learning to nowcast the spatial coverage of convection from *Himawari-8* satellite data. *Monthly Weather Review*, 149:3897–3921.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11:1267–1277.

Lakshminarayanan, B., A., P., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *31st Conference on Neural Information Processing Systems (NIPS)*, Neural Information Processing Systems Foundation, pages 6405–6416.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*.

Lauret, P., David, M., and Pinson, P. (2019). Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254–271.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Leshno, M. and Levy, H. (2002). Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science*, 48:1074–1085.

Leutbecher, M. and Haiden, T. (2021). Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Quarterly Journal of the Royal Meteorological Society*, 147:425–442.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, 227:3515–3539.

Li, Q., Lin, J., and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, 31:57–65.

Little, M. A., McSharry, P. E., and Taylor, J. W. (2009). Generalized linear models for site-specific density forecasting of U.K. daily rainfall. *Monthly Weather Review*, 137:1029–1045.

Maranan, M., Fink, A., and Knippertz, P. (2018). Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment. *Quarterly Journal of the Royal Meteorological Society*, 144:1628–1648.

Marx, C., Zhou, S., Neiswanger, W., and Ermon, S. (2022). Modular conformal calibration. In *39th International Conference on Machine Learning*, volume 62 of *Proceedings of Machine Learning Research*, pages 15180–15195.

Mason, S. J. and Weigel, A. P. (2009). A generic forecast verification framework for administrative purposes. *Monthly Weather Review*, 137:331–349.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142:3003–3014.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

Müller, A., Scarsini, M., Tsetlin, I., and Winkler, R. L. (2017). Between first- and second-order stochastic dominance. *Management Science*, 63:2933–2947.

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

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Nesbitt, S. W., Cifelli, R., and Rutledge, S. A. (2006). Storm morphology and rainfall characteristics of TRMM precipitation features. *Monthly Weather Review*, 134:2702–2721.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98:544–567.

Nicholls, S. and Mohr, K. (2010). An analysis of the environments of intense convective systems in West Africa. *Monthly Weather Review*, 138:3721–3739.

Nowotarski, J. and Weron, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30:791–803.

Otero, N. and Horton, P. (2023). Intercomparison of deep learning architectures for the prediction of precipitation fields with a focus on extremes. *Water Resources Research*, 59:e2023WR035088.

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63:71–116.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salomon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrologic ensemble prediction. *Journal of Hydrology*, 522:697–713.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 22:2109–2123.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Sciences Series.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38:705–871.

Python Software Foundation (2021). Python language reference, `http://www.python.org`.

Quiñonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., and Schölkopf, B. (2006). Evaluating Predictive Uncertainty Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment*, pages 1–27. Springer.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Rasheeda Satheesh, A., Knippertz, P., Fink, A., Walz, E.-M., and Gneiting, T. (2023). Sources of predictability of synoptic-scale rainfall variability during the West African summer monsoon. *Quarterly Journal of the Royal Meteorological Society*, 149:3721–3737.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. MIT Press.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. (2020). WeatherBench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12:e2020MS002203.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallègue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. (2023). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. Preprint, `arXiv:2308.15560`.

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

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at `https://www.R-project.org/`.

Regula, H. (1936). Druckschwankungen und Tornados an der Westküste von Afrika. *Annalen der Hydrographie und maritimen Meteorologie*, 64:107–111.

Ren, Y., Suganthan, P., and Srikanth, N. (2015). Ensemble methods for wind and solar power forecasting—a state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 50:82–91.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334:1518–1524.

Ritter, H., Botev, A., and Barber, D. (2018). A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, pages 1–15.

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

Roca, R., Aublanc, J., Chambon, P., Fiolleau, T., and Viltard, N. (2014). Robust observational quantification of the contribution of mesoscale convective systems to rainfall in the Tropics. *Journal of Climate*, 27:4952–4958.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015, Part III*, pages 234–241. Springer.

Rosset, S., Perlich, C., and Zadrozny, B. (2005). Ranking-based evaluation of regression models. In *Proceedings. 5th IEEE International Conference on Data Mining*, pages 370–377.

Rotunno, R., Klemp, J., and Weismann, M. (1988). A theory for strong, long-lived squall lines. *Journal of the Atmospheric Sciences*, 45:463–485.

Roy, C. J. and Oberkampf, W. L. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computational Methods in Applied Mechanics and Engineering*, 200:2131–2144.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28:616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140:1086–1096.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics*, 13:1564–1589.

Schlueter, A., Fink, A. H., Knippertz, P., and Vogel, P. (2019). A systematic comparison of tropical waves over Northern Africa. Part I: Influence on rainfall. *Journal of Climate*, 32:1501–1523.

Schneider, U., Fuchs, T., Meyer-Christoffer, A., and Rudolf, B. (2005). Global precipitation analysis products of the GPCC. URL `https://opendata.dwd.de/climate_environment/GPCC/PDF/GPCC_intro_products_v2015.pdf`.

Schreyer, M., Paulin, R., and Trutschnig, W. (2017). On the exact region determined by Kendall's tau and Spearman's rho. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:613–633.

Schroeder de Witt, C., Tong, C., Zantedeschi, V., De Martini, D., Kalaitzis, A., Chantry, M., Watson-Parris, D., and Bilinski, P. (2021). RainBench: Towards data-driven global precipitation forecasting from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14902–14910.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379:20200097.

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

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

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

Smith, R. C. (2014). *Uncertainty Quantification: Theory, Implementation, and Applications*. SIAM.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799–811.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location, scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46.

Stellingwerf, S., Riddle, E., Hopson, T., Knievel, J., Brown, B., and Gebremichael, M. (2021). Optimizing precipitation forecasts for hydrological catchments in Ethiopia using statistical bias correction and multi-modeling. *Earth and Space Science*, 8:2019EA000933.

Strähl, C. and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639.

Sullivan, T. J. (2015). *Introduction to Uncertainty Quantification*. Springer.

Swets, J. A. (1998). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1289.

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

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

Trefethen, N. (2012). Discrete or continuous? *SIAM News*, 45(4):1.

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

Van Gestel, T., Martens, D., Baesens, B., Feremans, D., Huysmans, J., and Vanthienen, J. (2007). Forecasting and analyzing insurance companies' ratings. *International Journal of Forecasting*, 23:513–529.

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

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, 33:369–388.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T. (2020). Skill of global raw and postprocessed ensemble predictions of rainfall in the Tropics. *Weather and Forecasting*, 35:2367–2385.

Vogel, P., Knippertz, P., Gneiting, T., Fink, A., Klar, M., and Schlueter, A. (2021). Statistical forecasts for the occurrence of precipitation outperform global models over northern tropical Africa. *Geophysical Research Letters*, 48:2020GL091022.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Springer, second edition.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 37–51.

Vovk, V., Petej, I., Nouretdinov, I., Manokhin, V., and Gammerman, A. (2020a). Computationally efficient versions of conformal predictive distributions. *Neurocomputing*, 397:292–308.

Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E., and Carlsson, L. (2020b). Conformal calibration. In *Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 84–99.

Waegeman, W., De Baets, B., and Boullart, L. (2008). ROC analysis in ordinal regression learning. *Pattern Recognition Letters*, 29:1–9.

Walz, E.-M. (2023). *Replication material for "Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output"*. Available at `https://github.com/evwalz/easyuq`.

Walz, E.-M., Henzi, A., Ziegel, J., and Gneiting, T. (2024). Easy Uncertainty Quantification (EasyUQ): Generating predictive distributions from single-valued model output. *SIAM Review*. In press, `arXiv:2212.08376`.

Weihs, L., Drton, M., and Meinshausen, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika*, 105:547–562.

Weyn, J. A., Durran, D. R., and Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12:e2020MS002109.

Wilks, D. S. (2019). *Statistical Methods in the Atmospheric Sciences*. Elsevier Academic Press.

Woodbury, M. A. (1940). Rank correlation when there are equal variates. *Annals of Mathematical Statistics*, 11:358–362.

Xie, Y. (2013). animation: An R package for creating animations and demonstrating statistical methods. *Journal of Statistical Software*, 53(1):1–27.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619:526–532.