Isotonic Distributional Regression

Alexander Henzi[1], Johanna F. Ziegel[1], and Tilmann Gneiting[2,3]

[1]University of Bern, Switzerland
[2]Heidelberg Institute for Theoretical Studies (HITS), Germany
[3]Karlsruhe Institute of Technology (KIT), Germany

September 10, 2019

Abstract

Isotonic distributional regression (IDR) is a powerful nonparametric technique for the estimation of conditional distributions under order restrictions. In a nutshell, IDR learns conditional distributions that are calibrated, and simultaneously optimal relative to comprehensive classes of relevant loss functions, subject to isotonicity constraints in terms of a partial order on the covariate space. Nonparametric isotonic quantile regression and probabilistic classifiers for binary events emerge as special cases. For prediction, we propose an interpolation method that generalizes extant specifications under the pool adjacent violators algorithm. We recommend the use of IDR as a generic benchmark technique in probabilistic forecast problems, as it does not involve any parameter tuning nor implementation choices, except for the selection of a partial order on the covariate space. The method can be combined with subsample aggregation, with the benefits of smoother regression functions and computational efficiency under large training sets. In a case study on raw and postprocessed quantitative precipitation forecasts from a leading numerical weather prediction system, IDR is competitive with state of the art techniques.

Keywords: conditional distribution estimation; continuous ranked probability score; ensemble methods; monotonicity; probabilistic forecast; proper scoring rule; stochastic order; subagging; weather prediction

1   Introduction

There is an emerging consensus in the transdisciplinary literature that regression analysis should be distributional, with Hothorn et al. (2014) arguing forcefully that

> [t]he ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables.

Distributional regression marks a clear break from the classical view of regression, which has focused on estimating the conditional mean of the response variable in terms of one or more explanatory variable(s) or covariate(s). Later extensions have considered other functionals of the conditional distributions, such as quantiles or expectiles (Koenker, 2005; Newey and Powell, 1987; Schulze Waltrup et al., 2015). However, the reduction of a conditional distribution to a single-valued functional results in tremendous loss of information. Therefore, from the perspectives of both estimation and prediction, regression analysis ought to be distributional.

In the extant literature, both parametric and nonparametric approaches to distributional regression are available. Parametric approaches assume that the conditional distribution of the response is of a specific type (e.g., Gaussian) with an analytic relationship between the covariates and the distributional parameters. Key examples include statistically postprocessed meteorological and hydrologic forecasts, as exemplified by Gneiting et al. (2006), Schefzik et al. (2013) and Vannitsem et al. (2018). In powerful semiparametric variants, the conditional distributions remain parametric, but the influence of the covariates on the parameter values is modeled nonparametrically, e.g., by using generalized additive models (Rigby and Stasinopoulos, 2005; Klein et al., 2015; Umlauf and Kneib, 2018) or modern neural networks (Rasp and Lerch, 2018; Gasthaus et al., 2019).

Non-parametric approaches to distributional regression include nearest neighbor methods that depend on a suitable notion of distance on the covariate space. Then, the empirical distribution of the response for neighboring covariates in the training set is used for distributional regression, with possible weighting in dependence on the distance to the covariate value of interest. Kernel smoothing methods and mixture approaches allow for absolutely continuous conditional distributions (Hall et al., 1999; Dunson et al., 2007). Classification and regression trees partition the covariate space into leaves, and assign constant regression functions on each leave (Breiman et al., 1984). Linear aggregation via bootstrap (bagging) or subsample aggregation (subagging) yields random forests (Breiman, 2001), which are increasingly being used to generate conditional predictive distributions, as proposed by Hothorn et al. (2004) and Meinshausen (2006).

Isotonicity is a natural constraint in estimation and prediction problems. Consider, e.g., postprocessing techniques in weather forecasting, where the covariates stem from the output of numerical weather prediction (NWP) models, and the response variable is the respective future weather quantity. Intuitively, if the NWP model output indicates a larger precipitation accumulation, the associated regression functions ought to be larger as well. Isotonic relationships of this type hold in a plethora of applied settings.

Consequently, concerning (non-distributional) regression for a conditional functional, such as the mean or a quantile, there is a sizable literature on estimation under the constraint of isotonicity, i.e., monotonicity in the val-

ues of the covariate(s). The classical work of Brunk (1955), Ayer et al. (1955), van Eeden (1958), Bartholomew (1959a,b) and Miles (1959) is summarized in Barlow et al. (1972) and de Leeuw et al. (2009). Subsequent approaches include Bayesian and non-Bayesian smoothing techniques (e.g., Mammen, 1991; Neelon and Dunson, 2004; Dette et al., 2006; Shively et al., 2009), and reviews are available in Groeneboom and Jongbloed (2014) and Guntuboyina and Sen (2018).

In distributional regression it may not be immediately clear what is meant by isotonicity, and the literature typically considers one ordinal covariate only (e.g., Hogg, 1965; Rojo and El Barmi, 2003; El Barmi and Mukerjee, 2005; Davidov and Iliopoulos, 2012) with a notable exception being the work of Mösching and Dümbgen (2019), whose considerations allow for a real-valued covariate. In the general case of a partially ordered covariate space, which we consider here, it is unclear whether semi- or nonparametric techniques might be capable of handling monotonicity contraints, and suitable notions of isotonicity remain to be developed.

To this end, we assume that the response Y is real-valued, and equip the covariate space $\mathcal{X}$ with a partial order $\preceq$. Our aim is to estimate the conditional distribution of Y given the covariate X, for short $\mathcal{L}(Y|X)$, on training data, in a way that respects the partial order, and we desire to use this estimate for prediction. Formally, a distributional regression technique generates a mapping from $x \in \mathcal{X}$ to a probability measure F_x, which serves to model the conditional distribution $\mathcal{L}(Y|X = x)$. This mapping is isotonic if $x \preceq x'$ implies $F_x \leq_{\text{st}} F_{x'}$, where $\leq_{\text{st}}$ denotes the usual stochastic order, that is, $G \leq_{\text{st}} H$ if $G(y) \geq H(y)$ for $y \in \mathbb{R}$, where we use the same symbols for the probability measures G, H and their associated conditional cumulative distribution functions (CDFs). Equivalently, $G \leq_{\text{st}} H$ holds if $G^{-1}(\alpha) \leq H^{-1}(\alpha)$ for $\alpha \in (0, 1)$, where $G^{-1}(\alpha) = \inf\{y \in \mathbb{R} : G(y) \geq \alpha\}$ is the standard quantile function (Shaked and Shanthikumar, 2007).

Useful comparisons of predictive distributions are in terms of proper scoring rules, of which the most prominent and most relevant instance is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007). We show that there is a unique isotonic distributional regression that is optimal with respect to the CPRS (Theorem 2.1), and refer to it as the *isotonic distributional regression* (IDR). As it turns out, IDR is a universal solution, in that the estimate is optimal with respect to an extraordinarily broad class of proper scoring rules (Theorem 2.2). Classical special cases such as nonparametric isotonic quantile regression and probabilistic classifiers for threshold-defined binary events are subsumed by IDR. Simultaneously, IDR avoids pitfalls commonly associated with nonparametric distributional regression, such as suboptimal partitions of the covariate space and level crossing (Athey et al., 2019, p. 1167).

For illustration, consider the joint distribution of (X, Y), where X is
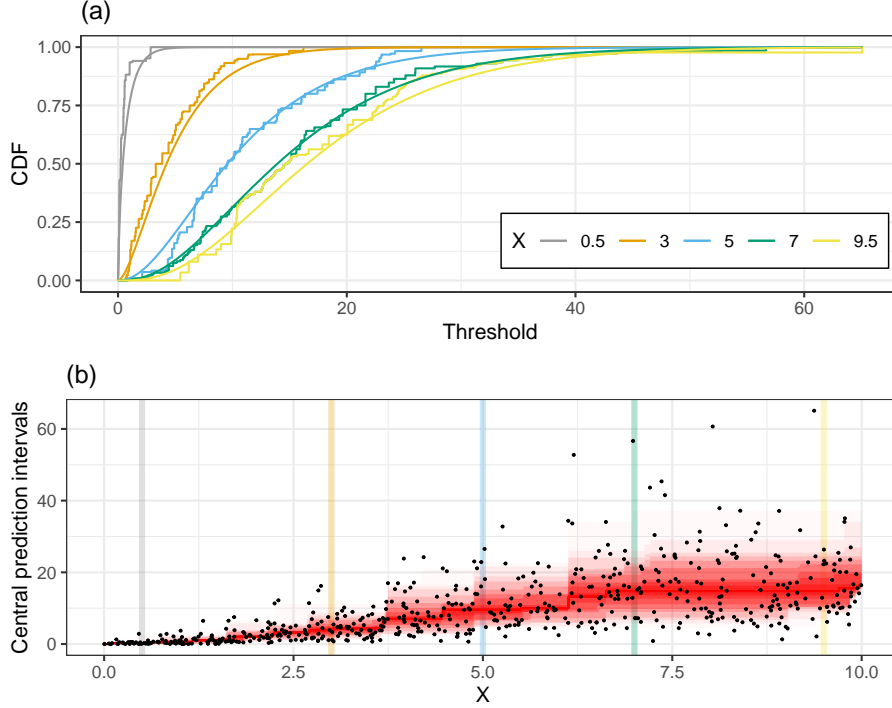
Figure 1: Simulation example for a sample of size $n = 600$ from the distribution in (1): (a) True conditional CDFs (smooth) and IDR estimates (step functions) for selected values of the covariate. (b) IDR estimated conditional distributions. The shaded bands correspond to probability mass 0.10 each, with the darkest shade marking the central interval. Vertical strips indicate the cross-sections corresponding to the values of the covariate in panel (a).

uniform on $(0, 10)$ and

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

so that $\mathcal{L}(Y \mid X = x) \leq_{\text{st}} \mathcal{L}(Y \mid X = x')$ if $x \leq x'$. Figure 1 shows IDR conditional CDFs and quantiles as estimated on a training set of size $n = 600$. IDR is capable of estimating both the strongly right-skewed conditional distributions for lower values of X and the more symmetric distributions as X increases. The CDFs are piecewise constant, and they never cross each other. The computational cost of IDR is of order at least $\mathcal{O}(n^2)$ and thus becomes prohibitive as n grows. However, IDR can usefully be combined with subsample aggregation (subbagging), much in the spirit of random forests (Breiman, 2001), with the benefits of reduced computational cost under large training samples, smoother regression functions, and (frequently) improved predictive performance.

The remainder of the paper is organized as follows. The methodological

4

core of the paper is in Section 2, where we prove existence, uniqueness and universality of the IDR solution, followed by a discussion of computational issues and strategies for prediction. In Section 3 we discuss the critical issue of the choice of a partial order on the covariate space. Section 4 is devoted to a case study on probabilistic quantitative precipitation forecasts, with covariates provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble system. Precipitation accumulations feature unfavorable properties that challenge parametric approaches to distributional regression: The conditional distributions have a point mass at zero, and they are continuous and right skewed on the positive half-axis. In a comparison to state-of-the-art methods that have been developed specifically for the purpose, namely Bayesian Model Averaging (BMA; Sloughter et al., 2007) and Ensemble Model Output Statistics (EMOS; Scheuerer, 2014), the (out-of-sample) predictive performance of IDR is competitive, despite the method being generic, and being fully automatic once a partial order on the covariate space has been chosen.

We close the paper with a discussion in Section 5, where we argue that IDR provides a very widely applicable, competitive benchmark in probabilistic forecasting problems. The use of benchmark techniques has been called for across application domains (e.g., Rossi, 2013; Pappenberger et al., 2015; Basel Committee on Banking Supervision, 2016; Vogel et al., 2018), and suitable methods should be competitive in terms of predictive performance, while avoiding implementation decisions that may vary from user to user. IDR is well suited to this purpose, as it is entirely generic, does not involve any implementation decisions, other than the choice of the partial order, applies to all types of real-valued outcomes with discrete, continuous or mixed discrete-continuous distributions, and accommodates general types of covariate spaces.

2 Isotonic distributional regression

We proceed to introduce the isotonic distributional regression (IDR) technique. To this end, we first review basic facts on proper scoring rules and notions of calibration. Then we define the IDR solution, prove existence, uniqueness and universality, and discuss its computation. Thereafter, we turn from estimation to prediction and describe how IDR can be used in out-of-sample forecasting. The section closes with a discussion of computational issues and subsample aggregation. Throughout, we identify a Borel probability measure on the real line $\mathbb{R}$ with its cumulative distribution function (CDF), and we denote the extended real line by $\bar{\mathbb{R}} = [-\infty, \infty]$.

2.1 Preliminaries

Following Gneiting and Raftery (2007), we argue that distributional regression techniques should be compared and evaluated using proper scoring rules. A *proper scoring rule* is a function $S : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}}$, where $\mathcal{P}$ is a suitable class of probability measures on $\mathbb{R}$, such that $S(F, \cdot)$ is measurable for any $F \in \mathcal{P}$, the integral $\int S(G, y) \, dF(y)$ exists, and

$$\int S(F, y) \, dF(y) \leq \int S(G, y) \, dF(y)$$

for all $F, G \in \mathcal{P}$. A key example is the *continuous ranked probability score* (CRPS), which is defined for all Borel probability measures, and given as

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{y \leq z\} \right)^2 \, dz.$$

Introduced by Matheson and Winkler (1976), the CRPS has become popular across application areas and methodological communities, both for the purposes of evaluating predictive performance and as a loss function in estimation; see, e,g., Hersbach (2000), Gneiting et al. (2005), Hothorn et al. (2014), Pappenberger et al. (2015), Rasp and Lerch (2018) and Gasthaus et al. (2019). The CRPS is reported in the same unit as the response variable, and it reduces to the absolute error, $|x - y|$, if F is the point or Dirac measure in $x \in \mathbb{R}$.

Results in Laio and Tamea (2007), Ehm et al. (2016) and Ben Bouallègue et al. (2018) imply that the CRPS can be represented equivalently as

$$\text{CRPS}(F, y) = 2 \int_{(0,1)} \text{QS}_\alpha(F, y) \, d\alpha \tag{2}$$

$$= 2 \int_{(0,1)} \int_{\mathbb{R}} S^Q_{\alpha,\theta}(F, y) \, d\theta \, d\alpha \tag{3}$$

$$= \int_{\mathbb{R}} \int_{(0,1)} S^P_{z,c}(F, y) \, dc \, dz, \tag{4}$$

where the mixture representation (2) is in terms of the asymmetric piecewise linear or pinball loss,

$$\text{QS}_\alpha(F, y) = \begin{cases} (1 - \alpha) \left(F^{-1}(\alpha) - y \right), & y \leq F^{-1}(\alpha), \\ \alpha \left(y - F^{-1}(\alpha) \right), & y \geq F^{-1}(\alpha), \end{cases} \tag{5}$$

which is customarily thought of as a quantile loss function, but can be identified with a proper scoring rule (Gneiting, 2011, Theorem 3). The representations (3) and (4) express the CRPS in terms of the *elementary* or *extremal scoring functions* for the α-quantile functional, namely,

$$S^Q_{\alpha,\theta}(F, y) = \begin{cases} 1 - \alpha, & y \leq \theta < F^{-1}(\alpha), \\ \alpha, & F^{-1}(\alpha) \leq \theta < y, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\theta \in \mathbb{R}$; and for probability assessments of the binary outcome $\mathbb{1}\{y \le z\}$ at the threshold value $z \in \mathbb{R}$, namely

$$
S_{z,c}^P(F,y) = \begin{cases} 1 - c, & F(z) < c, \ y \le z, \\ c, & F(z) \ge c, \ y > z, \\ 0, & \text{otherwise,} \end{cases} \tag{7}
$$

where $c \in (0,1)$. For background information on elementary or extremal scoring functions and related concepts see Ehm et al. (2016).

Predictive distributions ought to be calibrated (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007), in the broad sense that they should be statistically compatible with the responses, and various notions of calibration have been proposed and studied. In the spirit of Gneiting and Ranjan (2013), we consider the joint distribution $\mathbb{P}$ of the response Y and the distributional regression F_X. The most widely used criterion is *probabilistic calibration*, which requires that the *probability integral transform* (PIT), namely, the random variable

$$
Z = F_X(Y-) + V\left(F_X(Y) - F_X(Y-)\right), \tag{8}
$$

be standard uniform, where $F_X(Y-) = \lim_{y \uparrow Y} F_X(y)$ and V is a standard uniform variable that is independent of F_X and Y. If F_X is continuous the PIT is simply $Z = F_X(Y)$. Here we introduce the novel notion of *threshold calibration*, requiring that

$$
\mathbb{P}(Y \le y \,|\, F_X(y)) = F_X(y) \tag{9}
$$

almost surely for $y \in \mathbb{R}$, which implies *marginal calibration*, defined as $\mathbb{P}(Y \le y) = \mathbb{E}(F_X(y))$ for $y \in \mathbb{R}$. If $F_X = \mathcal{L}(Y \,|\, X)$ then it is calibrated in any of the above senses (Gneiting and Ranjan, 2013, Theorem 2.8).

2.2 Existence, uniqueness and universality

A partial order relation $\preceq$ on a set $\mathcal{X}$ has the same properties as a total order, namely reflexivity, antisymmetry and transitivity, except that the elements need not be comparable, i.e., there might be elements $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ such that neither $x \preceq x'$ nor $x' \preceq x$ holds. A key example is the componentwise order on $\mathbb{R}^n$.

For a positive integer n and a partially ordered set $\mathcal{X}$, we define the classes

$$
\mathcal{X}_\uparrow^n = \{\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n : x_1 \preceq \cdots \preceq x_n\},
$$
$$
\mathcal{X}_\downarrow^n = \{\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n : x_1 \succeq \cdots \succeq x_n\}
$$

of the increasingly and decreasingly (totally) ordered tuples in $\mathcal{X}$, respectively. Similarly, given a further partially ordered set $\mathcal{Q}$ and a vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, the classes

$$\mathcal{Q}_{\uparrow,\boldsymbol{x}}^n = \{\boldsymbol{q} = (q_1, \ldots, q_n) \in \mathcal{Q}^n : q_i \preceq q_j \text{ if } x_i \preceq x_j\},$$
$$\mathcal{Q}_{\downarrow,\boldsymbol{x}}^n = \{\boldsymbol{q} = (q_1, \ldots, q_n) \in \mathcal{Q}^n : q_i \succeq q_j \text{ if } x_i \preceq x_j\}$$

comprise the increasingly and decreasingly (partially) ordered tuples in $\mathcal{Q}$, with the order induced by the tuple $\boldsymbol{x}$ and the partial order $\preceq$ on $\mathcal{X}$.

Let $I \subseteq \mathbb{R}$ be an interval, and let S be a proper scoring rule with respect to a class $\mathcal{P}$ of probability distributions on I that contains all distributions with finite support. Given training data in the form of a covariate vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and response vector $\boldsymbol{y} = (y_1, \ldots, y_n) \in I^n$, we may interpret any mapping from $\boldsymbol{x} \in \mathcal{X}^n$ to $\mathcal{P}^n$ as a distributional regression function. Throughout, we equip $\mathcal{P}$ with the usual stochastic order.

Definition 2.1 (S-regression). An element $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_n) \in \mathcal{P}^n$ is an *isotonic* S-*regression* of $\boldsymbol{y} \in I^n$ on $\boldsymbol{x} \in \mathcal{X}^n$, if it is a minimizer of the empirical loss

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

over all $\boldsymbol{F} = (F_1, \ldots, F_n)$ in $\mathcal{P}_{\uparrow,\boldsymbol{x}}^n$.

This definition and the subsequent results can be extended to losses of the form $\ell_{\mathrm{S}}(\boldsymbol{F}) = \sum_{i=1}^{n} w_i \mathrm{S}(F_i, y_i)$ with rational, strictly positive weights $w_1, \ldots, w_n$. The adaptations are straightforward and left to the reader.

In Appendix A we prove the following result.

Theorem 2.1 (existence and uniqueness)**.** *There exists a unique isotonic* CRPS-*regression* $\hat{\boldsymbol{F}} \in \mathcal{P}^n$ *of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$.

We refer to this unique $\hat{\boldsymbol{F}}$ as the *isotonic distributional regression* (IDR) of $\boldsymbol{y}$ on $\boldsymbol{x}$. In the particular case of a total order on the covariate space, and assuming that $x_1 < \cdots < x_n$, for each $z \in I$ the solution $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z), \ldots, \hat{F}_n(z))$ is given by

$$\hat{F}_i(z) = \min_{j=i,\ldots,n} \max_{k=1,\ldots,j} \frac{1}{j-k+1} \sum_{l=k}^{j} \mathbb{1}\{z \geq y_l\} \tag{10}$$

for $i = 1, \ldots, n$; see eqs. (1.9)–(1.13) of Barlow et al. (1972). A similar min–max formula applies under partial orders (Robertson and Wright, 1980; Jordan et al., 2019), and it follows that $\hat{F}_i$ is piecewise constant with any points of discontinuity at $y_1, \ldots, y_n$.

At first sight, the specific choice of the CRPS as a loss function may seem arbitrary. However, the subsequent result, which we prove in Appendix A,

reveals that IDR is simultaneously optimal with respect to broad classes of proper scoring rules that include all relevant choices in the extant literature. The popular logarithmic score allows for the comparison of absolutely continuous distributions with respect to a fixed dominating measure only and thus is not applicable here. Statements concerning calibration are with respect to the empirical distribution of the training data $(x_1, y_1), \ldots, (x_n, y_n)$.

Theorem 2.2 (universality). *The* IDR *solution* $\hat{\boldsymbol{F}}$ *of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$ *is threshold calibrated and has the following properties.*

i) *The* IDR *solution* $\hat{\boldsymbol{F}}$ *is an isotonic* S-*regression of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$ *under any scoring rule of the form*

$$ \mathrm{S}(F, y) = \int_{(0,1) \times \mathbb{R}} \mathrm{S}_{\alpha, \theta}^Q (F, y) \, \mathrm{d}H(\alpha, \theta) \tag{11} $$

or

$$ \mathrm{S}(F, y) = \int_{\mathbb{R} \times (0,1)} \mathrm{S}_{z, c}^P (F, y) \, \mathrm{d}M(z, c), \tag{12} $$

where $\mathrm{S}_{\alpha, \theta}^Q$ *is the elementary quantile scoring function* (6), $\mathrm{S}_{z,c}^P$ *is the elementary probability scoring rule* (7), *and* H *and* M *are locally finite Borel measures on* $(0, 1) \times \mathbb{R}$ *and* $\mathbb{R} \times (0, 1)$, *respectively.*

ii) *For every* $\alpha \in (0, 1)$ *it holds that* $\hat{\boldsymbol{F}}^{-1}(\alpha) = (\hat{F}_1^{-1}(\alpha), \ldots, \hat{F}_n^{-1}(\alpha))$ *is a minimizer of*

$$ \frac{1}{n} \sum_{i=1}^n \mathrm{s}_\alpha(\theta_i, y_i) \tag{13} $$

over all $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in I_{\uparrow, \boldsymbol{x}}^n$, *under any function* $\mathrm{s}_\alpha : I \times I \to \bar{\mathbb{R}}$ *which is left-continuous in both arguments and such that* $\mathrm{S}(F, y) = \mathrm{s}_\alpha(F^{-1}(\alpha), y)$ *is a proper scoring rule on* $\mathcal{P}$.

iii) *For every threshold value* $z \in I$, *it is true that* $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z), \ldots, \hat{F}_n(z))$ *is a minimizer of*

$$ \frac{1}{n} \sum_{i=1}^n \mathrm{s}(\eta_i, \mathbb{1}\{y_i \leq z\}) \tag{14} $$

over all ordered tuples $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n) \in [0, 1]_{\downarrow, \boldsymbol{x}}^n$, *under any function* $\mathrm{s} : [0, 1] \times \{0, 1\} \to \bar{\mathbb{R}}$ *that is a proper scoring rule for binary events, which is left-continuous in its first argument, satisfies* $\mathrm{s}(0, y) = \lim_{p \to 0} \mathrm{s}(p, y)$, *and is real-valued, except possibly* $\mathrm{s}(0, 1) = -\infty$ *or* $\mathrm{s}(1, 0) = -\infty$.

In a nutshell, isotonic CRPS-regression subsumes isotonic S-regression under all relevant loss functions S, including but not limited to the customary choices in quantile regression and binary regression.

Indeed, the classes (11) and (12) of proper scoring rules under which IDR is optimal are extraordinarily large. The quantile weighted and threshold weighted versions of the CRPS studied by Gneiting and Ranjan (2011) arise from (11) and (12) with $H = G_0 \otimes \lambda$ and $M = \lambda \otimes G_1$, where λ denotes the Lebesgue measure, and G_0 and G_1 are σ-finite Borel measures on $(0, 1)$ and $\mathbb{R}$, respectively. If G_0 and G_1 are Lebesgue measures, we recover the mixture representations (3) and (4) of the CRPS. By results of Ehm et al. (2016), if H is concentrated on $\{\alpha\} \times \mathbb{R}$ and M is concentrated on $\{z\} \times (0, 1)$, these representations cover essentially all proper scoring rules that depend on the predictive distribution F via $F^{-1}(\alpha)$ or $F(z)$ only, yielding universal optimality in statements in parts ii) and iii) of Theorem 2.2.

In particular, as a special case of (13), the IDR solution is a minimizer of the quantile loss under the asymmetric piecewise linear or pinball function (5) that lies at the heart of quantile regression (Koenker, 2005). Consequently, IDR subsumes classical nonparametric isotonic quantile regression (e.g., Casady and Cryer, 1976). Similarly, the IDR solution is a minimizer of (13) under the quadratic or Brier probability score (Gneiting and Raftery, 2007), thereby suggesting a computational implementation via quadratic programming, as discussed below.

Remark. In the special case of a binary response variable, we see from iii) that the IDR solution is an isotonic S-regression under just any applicable proper scoring rule S. Furthermore, threshold calibration is the strongest possible notion of calibration in this setting (Gneiting and Ranjan, 2013, Theorem 2.11), so the IDR solution is universal in every regard. In the further special case of a total order on the covariate space, the IDR and pool adjacent violators (PAV) algorithm solutions coincode, and under technical restrictions on the scoring rule, the optimality of this solution has been claimed, proved and discussed by Fawcett and Niculescu-Mizil (2007) and Brümmer and du Preez (2013). In particular, optimality holds under the logarithmic score (Gneiting and Raftery, 2007), which recovers a result of Lloyd (2002), in that the IDR or PAV solution yields a nonparametric maximum likelihood estimate under the constraint of isotonicity.

2.3 Prediction

As introduced, the IDR solution $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_n) \in \mathcal{P}^n_{\uparrow, \boldsymbol{x}}$ is defined at the covariate values $x_1, \ldots, x_n \in \mathcal{X}$ only. Generally, if a (not necessarily optimal) distributional regression $\boldsymbol{F} = (F_1, \ldots, F_n) \in \mathcal{P}^n_{\uparrow, \boldsymbol{x}}$ is available, a key task in practice is to make a prediction at a new covariate value $x \in \mathcal{X}$ where $x \notin \{x_1, \ldots, x_n\}$. We denote the respective predictive CDF by F.

In the specific case $\mathcal{X} = \mathbb{R}$ of a single real-valued covariate there is a simple way of doing this, as frequently implemented in concert with the PAV algorithm. For simplicity we suppose that $x_1 < \cdots < x_n$. If $x < x_1$ we may let $F = F_1$; if $x \in (x_i, x_{i+1})$ for some $i \in \{1, \ldots, n-1\}$ we may

interpolate linearly, so that

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

for $z \in \mathbb{R}$, and if $x > x_n$ we may set $F = F_n$. However, approaches that are based on interpolation do not extend to a generic covariate space, which may or may not be equipped with a metric.

In contrast, the method we describe now, which generalizes a proposal by Wilbur et al. (2005), solely uses information supplied by the partial order $\preceq$ on the covariate space $\mathcal{X}$. For a general covariate value $x \in \mathcal{X}$, we define the sets of the indices of *direct predecessors* and *direct successors* among the covariate values $x_1, \ldots, x_n$ in the training data as

$$p(x) = \{i \in \{1, \ldots, n\} : x_i \preceq x_j \preceq x \implies x_j = x_i, \, j = 1, \ldots, n\}$$

and

$$s(x) = \{i \in \{1, \ldots, n\} : x \preceq x_j \preceq x_i \implies x_j = x_i, \, j = 1, \ldots, n\},$$

respectively. If the covariate space $\mathcal{X}$ is totally ordered, these sets contain at most one element. If the order is partial but not total, $p(x)$ and $s(x)$ may, and frequently do, contain more than one element. Assuming that $p(x)$ and $s(x)$ are non-empty, any predictive CDF F that is consistent with $\boldsymbol{F}$ must satisfy

$$\max_{i \in s(x)} F_i(z) \leq F(z) \leq \min_{i \in p(x)} F_i(z) \tag{15}$$

at all threshold values $z \in \mathbb{R}$. We now let F be the pointwise arithmetic average of these bounds, i.e.,

$$F(z) = \frac{1}{2} \left(\max_{i \in s(x)} F_i(z) + \min_{i \in p(x)} F_i(z) \right) \tag{16}$$

for $z \in \mathbb{R}$. If $s(x)$ is empty while $p(x)$ is non-empty, or vice-versa, a natural choice, which we employ hereinafter, is to let F equal the available bound given by the non-empty set. If x is not comparable to any of $x_1, \ldots, x_n$ the training data lack information about the conditional distribution at x, and a natural approach, which we adopt and implement, is to set F equal to the empirical distribution of the response values $y_1, \ldots, y_n$.

The difference between the bounds (if any) in (15) might be a useful measure of estimation uncertainty and could be explored as a promising avenue towards the quantification of ambiguity and generation of second-order probabilities (Ellsberg, 1961; Seo, 2009). In the context of ensemble weather forecasts, the assessment of ambiguity has been pioneered by Allen and Eckel (2012).

2.4 Computational aspects

The key observation towards a computational implementation is the aforementioned special case of (13), according to which the IDR solution $\hat{\boldsymbol{F}} \in \mathcal{P}^n$ of $\boldsymbol{y} \in \mathbb{R}^n$ on $\boldsymbol{x} \in \mathcal{X}^n$ satisfies

$$\hat{\boldsymbol{F}}(z) = \arg \min_{\eta \in [0,1]^n_{\downarrow, \boldsymbol{x}}} \sum_{i=1}^{n} (\eta_i - \mathbb{1}\{y_i \leq z\})^2 \qquad (17)$$

at every threshold value $z \in \mathbb{R}$. In this light, the computation of the IDR CDF at any fixed threshold reduces to a quadratic programming problem. The above target function is constant inbetween the $\mathcal{O}(n)$ unique values of $y_1, \ldots, y_n$, and so it suffices to estimate the CDFs at these points only. The overall cost depends on the quadratic programming solver. In the special case of a total order, it is $\mathcal{O}(n^2 \log n)$ if the PAV algorithm is used on unsorted data, or $\mathcal{O}(n^2)$ with the implementation of Wilbur et al. (2005) for sorted data, whence direct implementations become prohibitive if n is large. Under general partial orders, the set-up costs of determining the pairwise order relations may require $\mathcal{O}(n^2)$ operations already, and active set methods for solutions to the quadratic programming problem (17) have been discussed by de Leeuw et al. (2009). In our implementation, we use the powerful quadratic programming solver OSQP (Stellato et al., 2017) implemented in the package `osqp` in the statistical programming environment R (Stellato et al., 2018; R Core Team, 2019).

These considerations pose a challenge in that IDR, like any nonparametric technique, requires large training sets to inform the estimates of the conditional distributions. However, the computational cost of estimation is at least quadratic in the size n of the training sample. As a remedy, we propose subsample aggregation, much in the spirit of random forests that rely on bootstrap aggregated (bagged) classification and regression trees (Breiman, 1996, 2001). It was observed early on that random forests generate conditional predictive distributions (Hothorn et al., 2004; Meinshausen, 2006), and recent applications include the statistical postprocessing of ensemble weather forecasts (Taillardat et al., 2016; Schlosser et al., 2019; Taillardat et al., 2019). Bühlmann and Yu (2002) and Buja and Stützle (2006) argue forcefully that subsample aggregation (subagging) tends to be equally effective as bagging, but at considerably lower computational cost.

In view of the (super) quadratic computational costs of IDR, smart uses of subsample aggregation yield major efficiency gains. Isotonicity is preserved under linear aggregation, and the aggregated conditional CDFs can be inverted to generate isotonic conditional quantile functions, with the further benefit of smoother predictive distributions. For illustration, Figure 2 returns to the simulation setting in Figure 1, but now with a much larger training sample of size $n = 10\,000$ from the distribution in (1). Linear aggregation based on 100 subsamples (drawn without replacement) of size
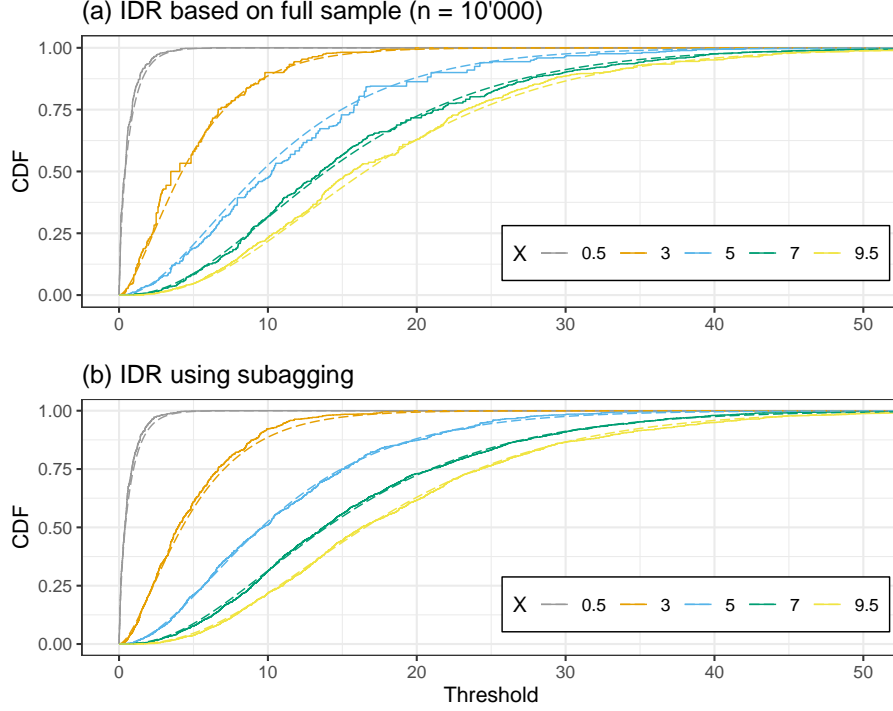
Figure 2: Simulation example for a sample of size $n = 10\,000$ from the distribution in (1). The true conditional CDFs (smooth dashed graphs) are compared to IDR estimates (step functions) based on (a) the full training sample of size $n = 10\,000$ and (b) linear aggregation of IDR estimates on 100 subsamples of size 1000 each.

$n = 1000$ each is superior to the brute force approach on the full training set in terms of both computational costs and estimation accuracy.

3 Partial orders

The choice of a sufficiently informative partial order on the covariate space is critical to any successful application of IDR. In the extreme case of distinct, totally ordered covariate values $x_1, \ldots, x_n \in \mathcal{X}$ and a perfect monotonic relationship to the response values $y_1, \ldots, y_n$, the IDR distribution associated with x_i is simply the point measure in y_i, for $i = 1, \ldots, n$. The same happens in the other extreme, when there are no order relations at all. Hence, the partial order serves to regularize the IDR solution.

Thus far, we have simply assumed that the covariate space $\mathcal{X}$ is equipped with a partial order $\preceq$, without specifying how the order might be defined. If $\mathcal{X} \subseteq \mathbb{R}^d$, the usual componentwise order will be suitable in many applications, and we investigate it in Section 3.1. For covariates that are ordinal

and admit a ranking in terms of importance, a lexicograpic order may be suitable.

If groups of covariates are exchangeable, as in our case study on quantitative precipitation forecasts, other types of order relations need to be considered. In Sections 3.2 and 3.3 we study relations that are tailored to this setting, namely, the empirical stochastic order and empirical increasing convex order. Proofs are deferred to Appendix B.

3.1 Componentwise order

Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of the covariate space $\mathbb{R}^d$. The most commonly used partial order in multivariate isotonic regression is the *componentwise order* defined by

$$x \preceq x' \iff x_i \leq x'_i \text{ for } i = 1, \ldots, d.$$

This order becomes weaker as the dimension d of the covariate space increases: If $\tilde{x} = (x_1, \ldots, x_d, x_{d+1})$ and $\tilde{x}' = (x'_1, \ldots, x'_d, x'_{d+1})$ then $x \preceq x'$ is a necessary condition for $\tilde{x} \preceq \tilde{x}'$. The following result is an immediate consequence of this observation and the structure of the optimization problem in Definition 2.1.

Proposition 3.1. *Let* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *and* $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$ *have components* $x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$ *and* $x_i^* = (x_{i1}, \ldots, x_{id}, x_{i,d+1}) \in \mathbb{R}^{d+1}$ *for* $i = 1, \ldots, n$, *and let* S *be a proper scoring rule.*

Then if $\mathbb{R}^d$ *and* $\mathbb{R}^{d+1}$ *are equipped with the componentwise partial order, and* $\hat{\boldsymbol{F}}$ *and* $\hat{\boldsymbol{F}}^*$ *denote isotonic* S-*regressions of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$ *and* $\boldsymbol{x}^*$, *respectively, it is true that*

$$\ell_S(\hat{\boldsymbol{F}}^*) \leq \ell_S(\hat{\boldsymbol{F}}).$$

In simple words, under the componentwise partial order, the inclusion of further covariates can only improve the in-sample fit. This behaviour resembles linear regression, where the addition of covariates can only improve the (unadjusted) R-square.

3.2 Empirical stochastic order

We now define a relation that is based on stochastic dominance and invariant under permutation.

Definition 3.1. Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$. Then x is smaller than or equal to x' in *empirical stochastic order*, for short $x \preceq_{st} x'$, if the empirical distribution of $x_1, \ldots, x_d$ is smaller than the empirical distribution of $x'_1, \ldots, x'_d$ in the usual stochastic order.

This relation is tailored to groups of exchangeable, real-valued covariates. The following results summarizes its properties and compares to the componentwise order, which we denote by $\preceq$.

Proposition 3.2. *Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$ with order statistics $x_{(1)} \leq \cdots \leq x_{(d)}$ and $x'_{(1)} \leq \cdots \leq x'_{(d)}$.*

   i) The relation $x \preceq_{\mathrm{st}} x'$ is equivalent to $x_{(i)} \leq x'_{(i)}$ for $i = 1, \ldots, d$.

   ii) If $x \preceq x'$ then $x \preceq_{\mathrm{st}} x'$.

   iii) If $x \preceq_{\mathrm{st}} x'$ and x and x' are comparable in the componentwise partial order, then $x \preceq x'$.

   iv) If $x \preceq_{\mathrm{st}} x'$ and $x' \preceq_{\mathrm{st}} x$ then x and x' are permutations of each other. Consequently, the relation $\preceq_{\mathrm{st}}$ defines a partial order on $\mathbb{R}^d_\uparrow$.

In a nutshell, the empirical stochastic order is equivalent to the componentwise order on the sorted elements, and this relation is weaker than the componentwise order. However, unlike the componentwise order, the empirical stochastic order does not degenerate as further covariates are added. To the contrary, empirical distributions of larger numbers of exchangeable variables become more informative and more easily comparable.

3.3   Empirical increasing convex order

In applications, the empirical stochastic order might be too strong, in the sense that it does not generate sufficiently informative constraints. In this light, we now define a weaker partial order on $\mathbb{R}^d_\uparrow$, which also is based on a partial order for probability measures. Specifically, let X and X' be random variables with CDFs F and F'. Then F is smaller than F' in increasing convex order if $\mathbb{E}(\phi(X)) \leq \mathbb{E}(\phi(X'))$ for all increasing convex functions ϕ such that the expectations exist (Shaked and Shanthikumar, 2007, Section 4.A.1).

Definition 3.2. *Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$. Then x is smaller than or equal to x' in empirical increasing convex order, for short $x \preceq_{\mathrm{icx}} x'$, if the empirical distribution of $x_1, \ldots, x_d$ is smaller than the empirical distribution of $x'_1, \ldots, x'_d$ in increasing convex order.*

This notion provides another meaningful relation for groups of exchangeable covariates. The following result summarizes its properties and relates to the empirical stochastic order.

Proposition 3.3. *Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$ with order statistics $x_{(1)} \leq \cdots \leq x_{(d)}$ and $x'_{(1)} \leq \cdots \leq x'_{(d)}$.*

   i) The relation $x \preceq_{\mathrm{icx}} x'$ is equivalent to

$$\sum_{i=j}^{d} x_{(i)} \leq \sum_{i=j}^{d} x'_{(i)} \;\; for \;\; j = 1, \ldots, d.$$
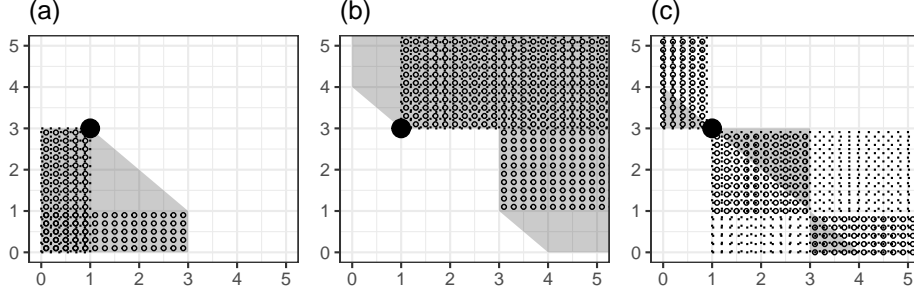
15

Figure 3: Regions of (a) smaller, (b) greater and (c) incomparable elements in the positive quadrant of $\mathbb{R}^2$, as compared to the point $(1,3)$. Shaded areas correspond to the empirical increasing convex order ($\preceq_{\mathrm{icx}}$), areas with circles to the empirical stochastic order ($\preceq_{\mathrm{st}}$), and dotted areas to the componentwise order ($\preceq$).

ii) If $x \preceq_{\mathrm{st}} x'$ then $x \preceq_{\mathrm{icx}} x'$.

iii) If $x \preceq_{\mathrm{icx}} x'$ then

$$\frac{1}{d}\sum_{i=1}^{d} x_i + \frac{d-1}{2(d+1)}g(x) \;\leq\; \frac{1}{d}\sum_{i=1}^{d} x_i' + \frac{d-1}{2(d+1)}g(x'),$$

where g is the Gini mean difference,

$$g(x) = \frac{1}{d(d-1)}\sum_{i,j=1}^{d} |x_i - x_j|. \tag{18}$$

iv) If $x \preceq_{\mathrm{icx}} x'$ and $x' \preceq_{\mathrm{icx}} x$ then x and x' are permutations of each other. Consequently, the relation $\preceq_{\mathrm{icx}}$ defines a partial order on $\mathbb{R}^d_{\uparrow}$.

Figure 3 illustrates the various types of relations for points in the positive quadrant of $\mathbb{R}^2$. As reflected by the nested character of the regions, the componentwise order is stronger than the empirical stochastic order, which in turn is stronger than the empirical increasing convex order. In fact, the latter is equivalent to *weak majorization* as studied by Marshall et al. (2011). In the special case of vectors with non-negative entries, their Corollary C.5 implies that $x \in \mathbb{R}^d$ is dominated by $x' \in \mathbb{R}^d$ in empirical increasing convex order if, and only if, it lies in the convex hull of the points of the form $(\xi_1 x'_{\pi(1)}, \ldots, \xi_d x'_{\pi(d)})$, where π is a permutation and $\xi_i \in \{0,1\}$ for $i = 1, \ldots, d$.

4 Case study: Probabilistic quantitative precipitation forecasts

The past decades have witnessed tremendous progress in the science and practice of weather prediction (Bauer et al., 2015). Arguably, the most radical innovation consists in the operational implementation of ensemble systems and an accompanying culture change from point forecasts to distributional forecasts (Leutbecher and Palmer, 2008). An ensemble system comprises multiple runs of numerical weather prediction (NWP) models, where the runs or members differ from each other in initial conditions and numerical-physical representations of atmospheric processes.

Ideally, one would like to interpret an ensemble forecast as a random sample from the conditional distribution of future states of the atmosphere. However, this is rarely advisable in practice, as ensemble forecasts are subject to biases and dispersion errors, thereby calling for some form of statistical postprocessing (Gneiting and Raftery, 2005; Vannitsem et al., 2018). This is typically done by fitting a distributional regression model, with the weather variable of interest being the response variable, and the members of the forecast ensemble constituting the covariates, and applying this model to future NWP output, to obtain conditional predictive distributions for future weather quantities. State of the art techniques include Bayesian model averaging (BMA; Raftery et al., 2005; Sloughter et al., 2007) and ensemble model output statistics (EMOS; Gneiting et al., 2005; Scheuerer, 2014).

In this case study, we apply IDR to the statistical postprocessing of ensemble forecasts of accumulated precipitation, a variable that is notoriously difficult to handle, due to its mixed discrete-continuous character, which requires both a point mass at zero and a right skewed continuous component on the positive half-axis. As competitors to IDR, we implement the BMA technique of Sloughter et al. (2007) and the EMOS method of Scheuerer (2014), which are widely used parametric approaches that have been developed specifically for the purposes of probabilistic quantitative precipitation forecasting. In contrast, IDR is a generic technique and fully automatic, once the partial order on the covariate space has been specified.

4.1 Data

The data in our case study comprise forecasts and observations of 24-hour accumulated precipitation from 06 January 2007 to 01 January 2017 at meteorological stations on airports in London, Brussels, Zurich and Frankfurt. As detailed in Table 1, data availability differs, and we remove days with missing entries station by station, so that all types of forecasts for a given station are trained and evaluated on the same data. Both forecasts and observations refer to the 24-hour period from 6:00 UTC to 6:00 UTC on the following day. The observations are in the unit of millimeter and constitute

Table 1: Meteorological stations at airports, with International Air Transport Association (IATA) airport code, World Meteorological Organization (WMO) station ID, and data availability in days (years).

	IATA Code	WMO ID	Data Availability
Brussels, Belgium	BRU	06449	3406 (9.3)
Frankfurt, Germany	FRA	10637	3617 (9.9)
London, UK	LHR	03772	2256 (6.2)
Zurich, Switzerland	ZRH	06670	3241 (8.9)

the response variable in distributional regression. They are typically, but not always, reported in integer multiples of a millimeter (mm).

As covariates, we use the 52 members of the leading NWP ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al., 1996; Buizza et al., 2005). The ECMWF ensemble system comprises a high-resolution member (x_{HRES}), a control member at lower resolution (x_{CTR}) and 50 perturbed members ($x_1, \ldots, x_{50}$) at the same lower resolution but with perturbed initial conditions, and the perturbed members can be considered exchangeable (Leutbecher, 2018). To summarize, the covariate vector in distributional regression is

$$x = (x_1, \ldots, x_{50}, x_{\mathrm{CTR}}, x_{\mathrm{HRES}}) = (x_{\mathrm{PTB}}, x_{\mathrm{CTR}}, x_{\mathrm{HRES}}) \in \mathbb{R}^{52}, \qquad (19)$$

where $x_{\mathrm{PTB}} = (x_1, \ldots, x_{50}) \in \mathbb{R}^{50}$. At each station, we use the forecasts for the corresponding latitude-longitude gridbox of size 0.25×0.25 degrees, and we consider prediction horizons of 1 to 5 days. For example, the two day forecast is initialized at 00:00 Universal Coordinated Time (UTC) and issued for 24-hour accumulated precipitation from 06:00 UTC on the next calendar day to 06:00 UTC on the day after. ECMWF forecast data are available online via the TIGGE system (Bougeault et al., 2010; Swinbank et al., 2016)

Statistical postprocessing is both a calibration and a downscaling problem: Forecasts and observations are at different spatial scales, whence the unprocessed forecasts are subject to representativeness error (Wilks, 2011, Chapter 8.9). Indeed, if we interpret the predictive distribution from the raw ensemble (19) as the empirical distribution of all 52 members — a customary approach, which we adopt hereinafter — there is a strong bias in probability of precipitation forecasts: Days with exactly zero precipitation are predicted much less often at the NWP model grid box scale than they occur at the point scale of the observations.

4.2 BMA and EMOS

Before describing our IDR implementation, we review its leading competitors, namely, state of the art parametric distributional regression approaches that have been developed specifically for accumulated precipitation.

The ensemble model output statistics (EMOS) technique of Scheuerer (2014) is based on the three-parameter family of left-censored generalized extreme value (GEV) distributions. The left-censoring generates a point mass at zero, corresponding to no precipitation, and the shape parameter allows for flexible skewness on the positive half-axis, associated with rain, hail or snow accumulations. The GEV location parameter is modeled as a linear function of x_{HRES}, x_{CTR}, $m_{\mathrm{PTB}} = \frac{1}{50} \sum_{i=1}^{50} x_i$ and

$$p_{\mathrm{ZERO}} = \frac{1}{52} \left(\mathbb{1}\{x_{\mathrm{HRES}} = 0\} + \mathbb{1}\{x_{\mathrm{CTR}} = 0\} + \sum_{i=1}^{50} \mathbb{1}\{x_i = 0\} \right),$$

and the GEV scale parameter is linear in the Gini mean difference (18) of the 52 individual forecasts in the covariate vector (19). While all parameters are estimated by minimizing the in-sample CRPS, the GEV shape parameter does not link to the covariates.

The general idea of the Bayesian model averaging (BMA; Raftery et al., 2005) approach is to employ a mixture distribution, where each mixture component is parametric and associated with an individual ensemble member forecast, with mixture weights that reflect the member's skill. Here we use the BMA implementation of Sloughter et al. (2007) for accumated precipitation in a variant that is based on x_{HRES}, x_{CTR}, $m_{\mathrm{PTB}} = \frac{1}{50} \sum_{i=1}^{50} x_i$ only, which we found to achieve more stable estimates and superior predictive scores than variants based on all members, as proposed by Fraley et al. (2010) in settings with smaller groups of exchangeable members. Hence, our BMA predictive CDF is of the form

$$F_x(y) = w_{\mathrm{HRES}} G(y|x_{\mathrm{HRES}}) + w_{\mathrm{CTR}} G(y|x_{\mathrm{CTR}}) + w_{\mathrm{PTB}} G(y|x_{\mathrm{PTB}})$$

for $y \in \mathbb{R}$, where the component CDFs $G(y|\cdot)$ are parametric, and the weights w_{HRES}, w_{CTR} and w_{PTB} are nonnegative and sum to one. Specifically, $G(y|x_{\mathrm{HRES}})$ models the logit of the point mass at zero as a linear function of $\sqrt[3]{x_{\mathrm{HRES}}}$ and $p_{\mathrm{HRES}} = \mathbb{1}\{x_{\mathrm{HRES}} = 0\}$, and the distribution for positive accumulations as a gamma density with mean and variance being linear in $\sqrt[3]{x_{\mathrm{HRES}}}$ and x_{HRES}, respectively, and analogously for $G(y|x_{\mathrm{CTR}})$ and $G(y|x_{\mathrm{PTB}})$. Estimation relies on a two-step procedure, where the (component specific) logit and mean models are fitted first, followed by maximum likelihood estimation of the weight parameters and the (joint) variance model via the EM algorithm (Sloughter et al., 2007).

Code for both BMA and EMOS is available within the `ensembleBMA` and `ensembleMOS` packages in the statistical programming environment R

([Messner, 2018](); [R Core Team, 2019]()). Unless noted differently, we use defaults options in implementation decisions.

4.3 Choice of partial order for IDR

IDR applies readily in this setting, without any need for adaptations due to the mixed-discrete continuous character of precipitation accumulation, nor requiring data transformations or other types of implementation decisions. However, the partial order on the elements (19) of the covariate space $\mathcal{X} = \mathbb{R}^{52}$, or on a suitable derived space, needs to be selected thoughtfully, considering that the perturbed members $x_1, \ldots, x_{50}$ are exchangeable.

In the sequel, we apply IDR in three variants. Our first implementation is based on x_{HRES}, x_{CTR} and $m_{\mathrm{PTB}} = \frac{1}{50} \sum_{i=1}^{50} x_i$ along with the componentwise order on $\mathbb{R}^3$, in that

$$x \preceq x' \iff m_{\mathrm{PTB}} \leq m'_{\mathrm{PTB}},\ x_{\mathrm{CTR}} \leq x'_{\mathrm{CTR}},\ x_{\mathrm{HRES}} \leq x'_{\mathrm{HRES}}. \tag{20}$$

The second implementation uses the same variables and partial order, but combined with a simple subagging approach: Before applying IDR, the training data is split into the two disjoint subsamples of training observations with odd and even indices, and we average the predictions based on these two subsamples.

Our third implementation combines the empirical increasing convex order for x_{PTB} with the usual total order on $\mathbb{R}$ for x_{HRES}, whence

$$x \preceq x' \iff x_{\mathrm{PTB}} \preceq_{\mathrm{icx}} x'_{\mathrm{PTB}},\ x_{\mathrm{HRES}} \leq x'_{\mathrm{HRES}}. \tag{21}$$

Henceforth, we refer to the implementations based on the partial orders in (20) and (21) as $\mathrm{IDR}_{\mathrm{cw}}$, $\mathrm{IDR}_{\mathrm{sbg}}$, and $\mathrm{IDR}_{\mathrm{icx}}$, respectively. As noted in Section 3, the relations (20) and (21) fail to be antisymmetric on $\mathbb{R}^{52}$, but they are partial orders on suitable derived spaces, namely $\mathbb{R}^3$ and $\mathbb{R}_{\uparrow}^{50} \times \mathbb{R}$.

We have experimented with other options as well, e.g., by incorporating the maximum $\max_{i=1,\ldots,50} x_i$ of the perturbed members in the componentwise order in (20), with the motivation that the maximum might serve as a proxy for the spread of the ensemble, or by using the empirical stochastic order $\preceq_{\mathrm{st}}$ in lieu of the empirical increasing convex order $\preceq_{\mathrm{icx}}$ in (21). IDR is robust to changes of this type, and the predictive performance remains stable, provided that the partial order honors the key substantive insights, in that the perturbed members $x_1, \ldots, x_{50}$ are exchangeable, while x_{HRES}, due to its higher native resolution, is able to capture local information that is not contained in x_{PTB} nor x_{CTR}. Hence, x_{HRES} ought to play a pivotal role in the partial order.

4.4 Selection of training periods

The selection of the training period is a crucial step in the statistical postprocessing of NWP output. Most postprocessing methods, including the

ones used in this analysis, assume that there is a stationary relationship between the forecasts and the observations. As Hamill (2018) points out, this assumption is hardly ever satisfied in practice: NWP models are updated, instruments at observation stations get replaced, and forecast biases may vary seasonally. These problems are exacerbated by the fact that quantitative precipitation forecasts require large training datasets in order to include sufficient numbers of days with non-zero precipitation and extreme precipitation events.

For BMA and EMOS, a training period over a rolling window of the latest available 720 days at the time of forecasting is (close to) optimal at all stations. This resembles choices made by Scheuerer and Hamill (2015) who used a training sample of about 900 past instances. Scheuerer (2014) took shorter temporal windows, but merged instances from nearby stations into the training sets, which is not possible here. In general, it would be preferable to select training data seasonally (e.g., data from the same month), but in our case the positive effect of using seasonal training data does not outweigh the negative effect of a smaller sample size.

As a nonparametric technique, IDR requires larger sets of training data than BMA or EMOS. As training data for IDR, we used all data available at the time of forecasting, which is about 2500 to 3000 days for the stations Frankfurt, Brussels and Zurich, and 1500 days for London Heathrow.

For evaluation, we use the years 2015 and 2016 (and 01 January 2017) for all postprocessing methods and the raw ensemble. This test dataset consists of roughly 700 instances for each station and lead time.

4.5 Results

Before comparing the BMA, EMOS, IDR_{cw}, IDR_{sbg} and IDR_{icx} techniques in terms of out-of-sample predictive performance over the test period, we exemplify them in Figure 4, where we show predictive CDFs for accumulated precipitation at Brussels on December 16, 2015, at a prediction horizon of 2 days. In panel (a) the marks at the bottom correspond to x_{HRES}, x_{CTR}, the perturbed members $x_1, \ldots, x_{50}$ and their mean m_{PTB}. The observation at 4 mm is indicated by the vertical line. Under all four techniques, the point mass at zero, which represents the probability of no precipitation, is vanishingly small. While the BMA and EMOS CDFs are smooth and supported on the positive half-axis, the IDR_{cw}, IDR_{sbg} and IDR_{icx} CDFs are piecewise constant with jump points at observed values in the training period. Panel (b) illustrates the hard and soft constraints on the IDR_{cw} CDF that arise from (15) under the order relation (20), with the thinner lines representing the IDR_{cw} CDFs of direct successors and predecessors. In this example, the constraints are mostly hard, except for threshold values between 4 and 11 mm.

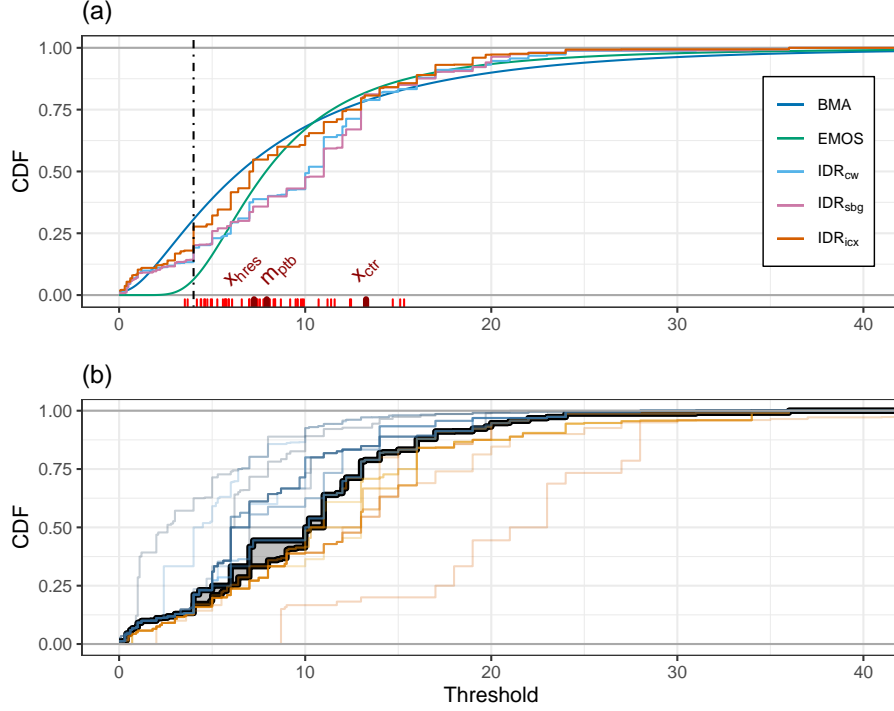We now use the mean CRPS over the test period as an overall measure

Figure 4: Distributional forecasts for accumulated precipitation at Brussels, valid 16 December 2015 at a prediction horizon of 2 days. (a) BMA, EMOS, IDR_{cw}, IDR_{sbg} and IDR_{icx} predictive CDFs. The vertical line represents the observation. (b) IDR_{cw} CDF along with the hard and soft constraints in (15) as induced by the order relation (20). The thin lines show the IDR_{cw} CDFs at direct predecessors and successors.

of out-of-sample predictive performance. Figure 5 shows the CRPS of the raw and postprocessed forecasts for all stations and lead times, with the raw forecast denoted as ENS. While EMOS performs best in terms of the CRPS, the IDR variants show scores of a similar magnitude and outperform BMA in many instances. Figure 6 shows probability integral transform (PIT; eq. (8)) histograms for distributional forecasts at a leadtime of 2 days, with the results for other prediction horizons being similar. All three IDR variants show nearly uniform PIT histograms, and so do BMA and EMOS, as opposed to the raw ensemble, which is underdispersed.

In Figures 7 and 8 we evaluate probability of precipitation forecasts by means of the Brier score (Gneiting and Raftery, 2007) and reliability diagrams (Wilks, 2011, Chapter 7.4), respectively. As opposed to the raw ensemble forecast, all distributional regression methods yield reliable probability forecasts. BMA, IDR_{cw}, IDR_{sbg} and IDR_{icx} separate the estimation of the point mass at zero, and of the distribution for positive accumulations,
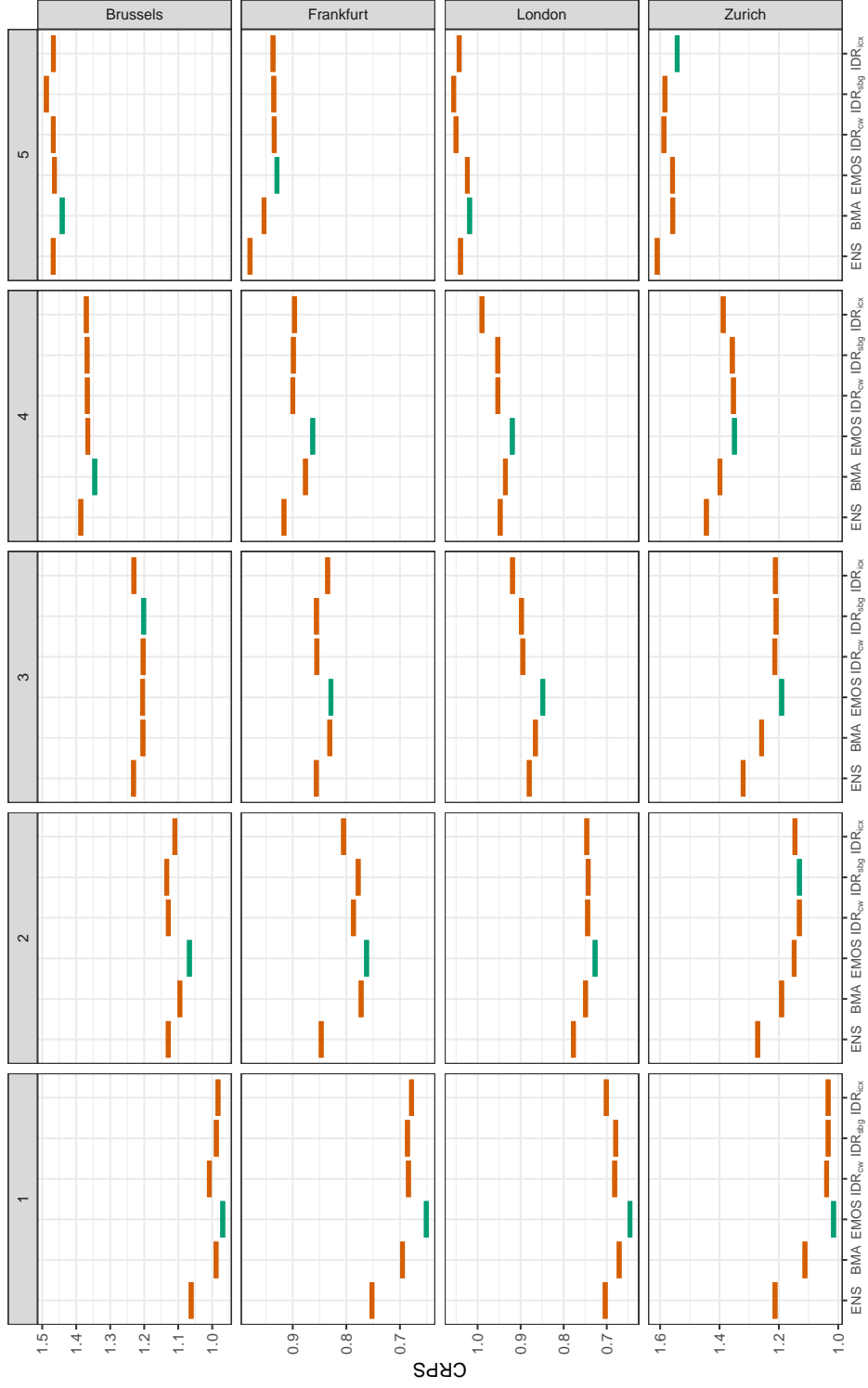
Figure 5: Mean CRPS over the test period for raw and postprocessed ensemble forecasts of 24-hour accumulated precipitation at prediction horizons of 1, 2, 3, 4 and 5 days. The lowest mean score for a given lead time and station is indicated in green.
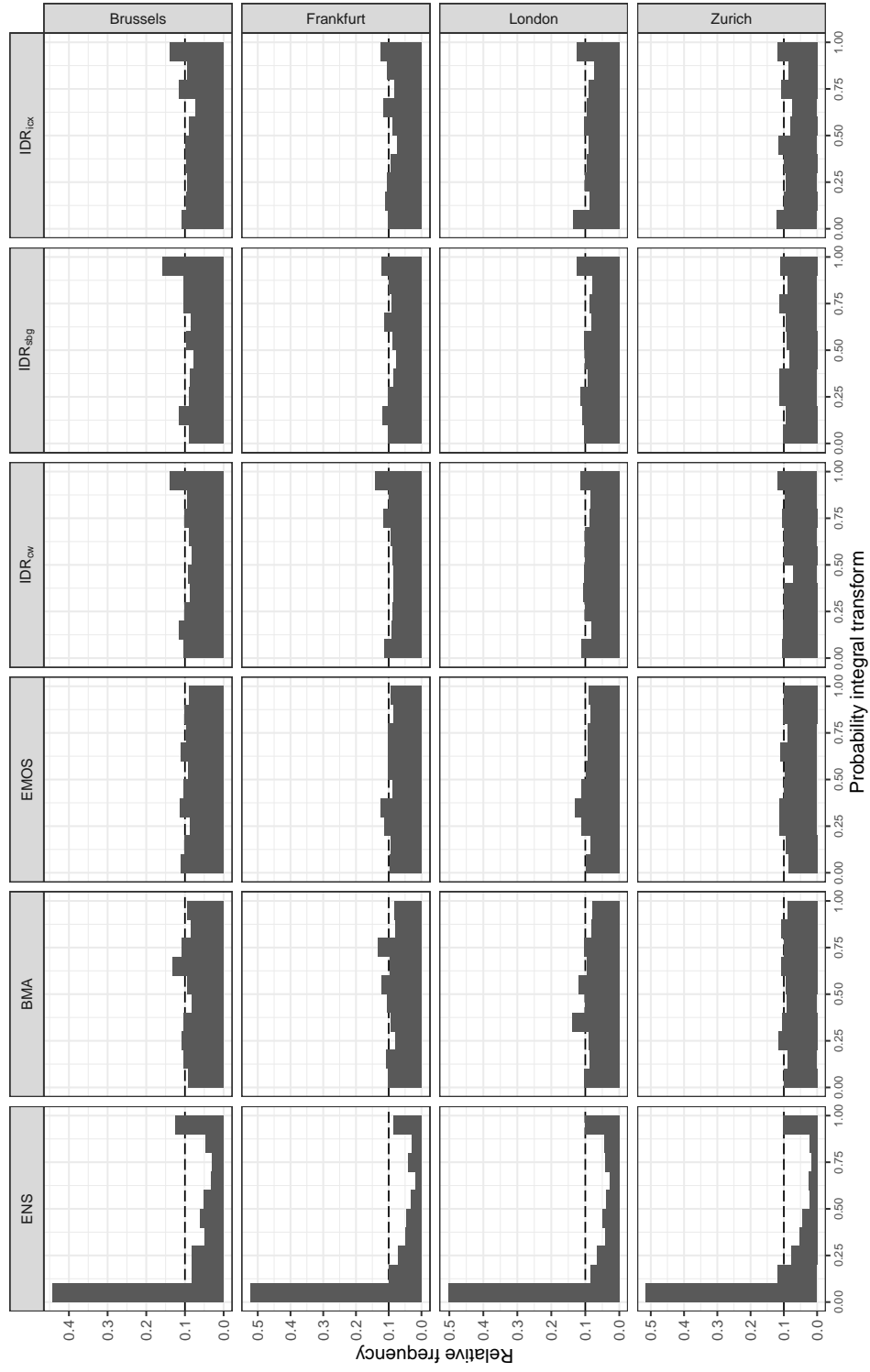
Figure 6: PIT histograms for raw and postprocessed ensemble forecasts of 24-hour accumulated precipitation at a prediction horizon of 2 days, for the test period.

and the four methods achieve about equal scores, all ahead of EMOS.

Interestingly, the IDR postprocessed predictions outperform EMOS for probability of precipitation forecasts, but not for precipitation accumulations. We attribute this to the fact that parametric techniques are capable of extrapolating beyond the range of the training responses, whereas IDR is not: The highest precipitation amount judged feasible by IDR equals the largest observation in the training set. Furthermore, unlike EMOS, IDR does not use information about the spread of the raw ensemble, which is inconsequential for probability of precipitation forecasts, but may impede distributional forecasts of precipitation accumulations.

In all comparisons, the forecast performance of IDR_{cw} and IDR_{sbg} is similar. However, in our implementation, the simple subagging method used in IDR_{sbg} reduced the computation time by up to one half.

To summarize, our results underscore the suitability of IDR as a benchmark technique in probabilistic forecasting problems. Despite being generic as well as fully automated, IDR is remarkably competitive relative to state of the art techniques that have been developed specifically for the purpose. In fact, in a wide range of applied problems that lack sophisticated, custom-made distributional regresssion solutions, IDR might well serve as a ready-to-use, top-performing method of choice.

5 Discussion

Stigler (1975) gives a lucid historical account of the 19th century transition from point estimation to distribution estimation. In regression analysis, we may be witnessing what future generations might refer to as the transition from conditional mean estimation to conditional distribution estimation, accompanied by a simultaneous transition from point forecasts to distributional forecasts (Gneiting and Katzfuss, 2014).

Isotonic distributional regression (IDR) is a nonparametric technique for estimating conditional distributions that takes advantage of partial order relations within the covariate space. It can be viewed as a far-reaching generalization of pool adjacent violators (PAV) algorithm based classical approaches to isotonic (non-distributional) regression, is entirely generic and fully automated, and provides for a unified treatment of continuous, discrete and mixed discrete-continuous real-valued response variables. Code for the implementation of IDR within R (R Core Team, 2019) is available on github (https://github.com/AlexanderHenzi/isodistrreg), with replication material for our paper and user-friendly functions for partial orders, estimation, prediction and evaluation.

IDR relies on information supplied by order constraints, and the choice of the partial order on the covariate space is a critical decision prior to the analysis. Only variables that contribute to the partial order need to be re-
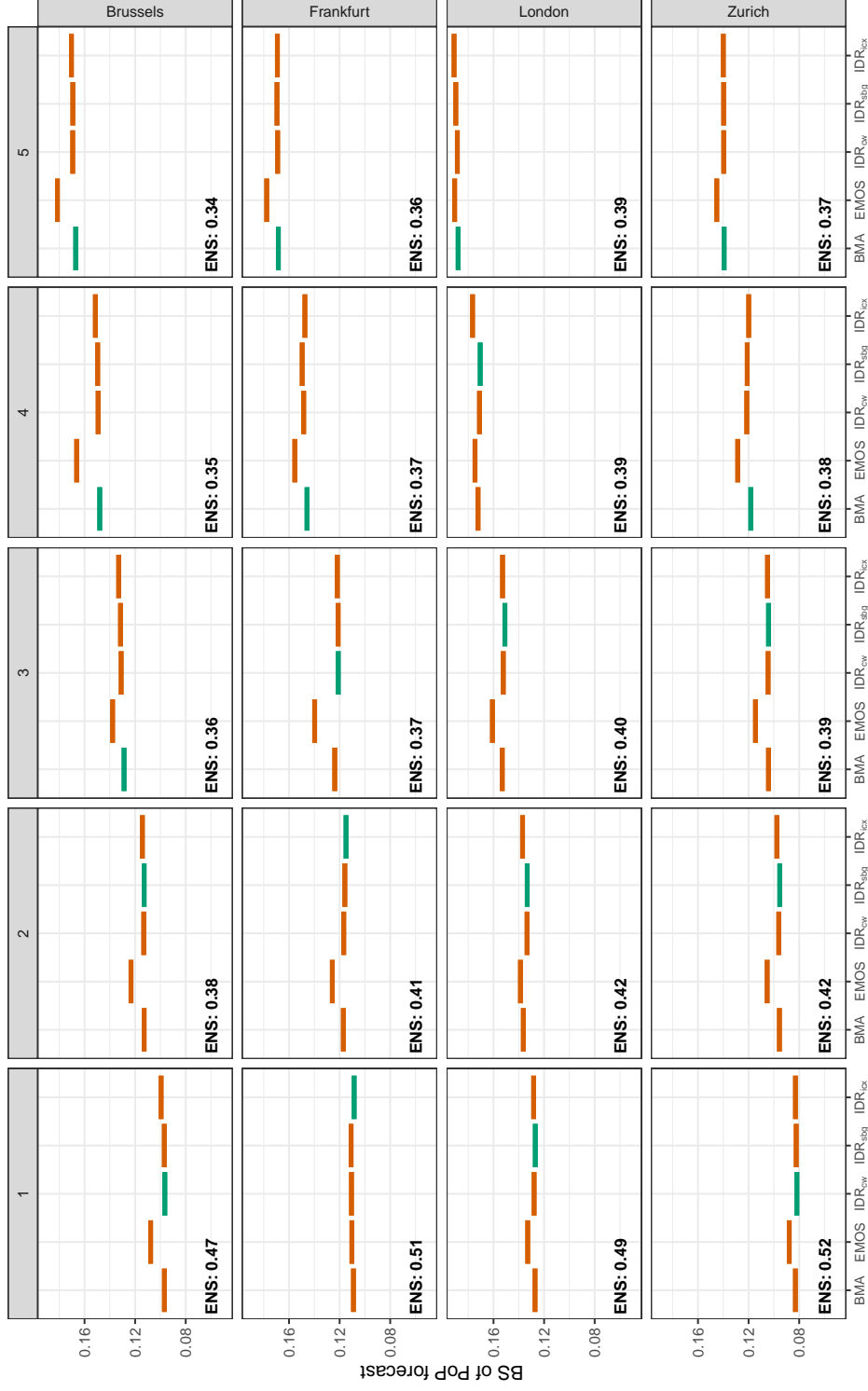
Figure 7: Mean Brier score over the test period for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days. The lowest mean score for a given lead time and station is indicated in green.
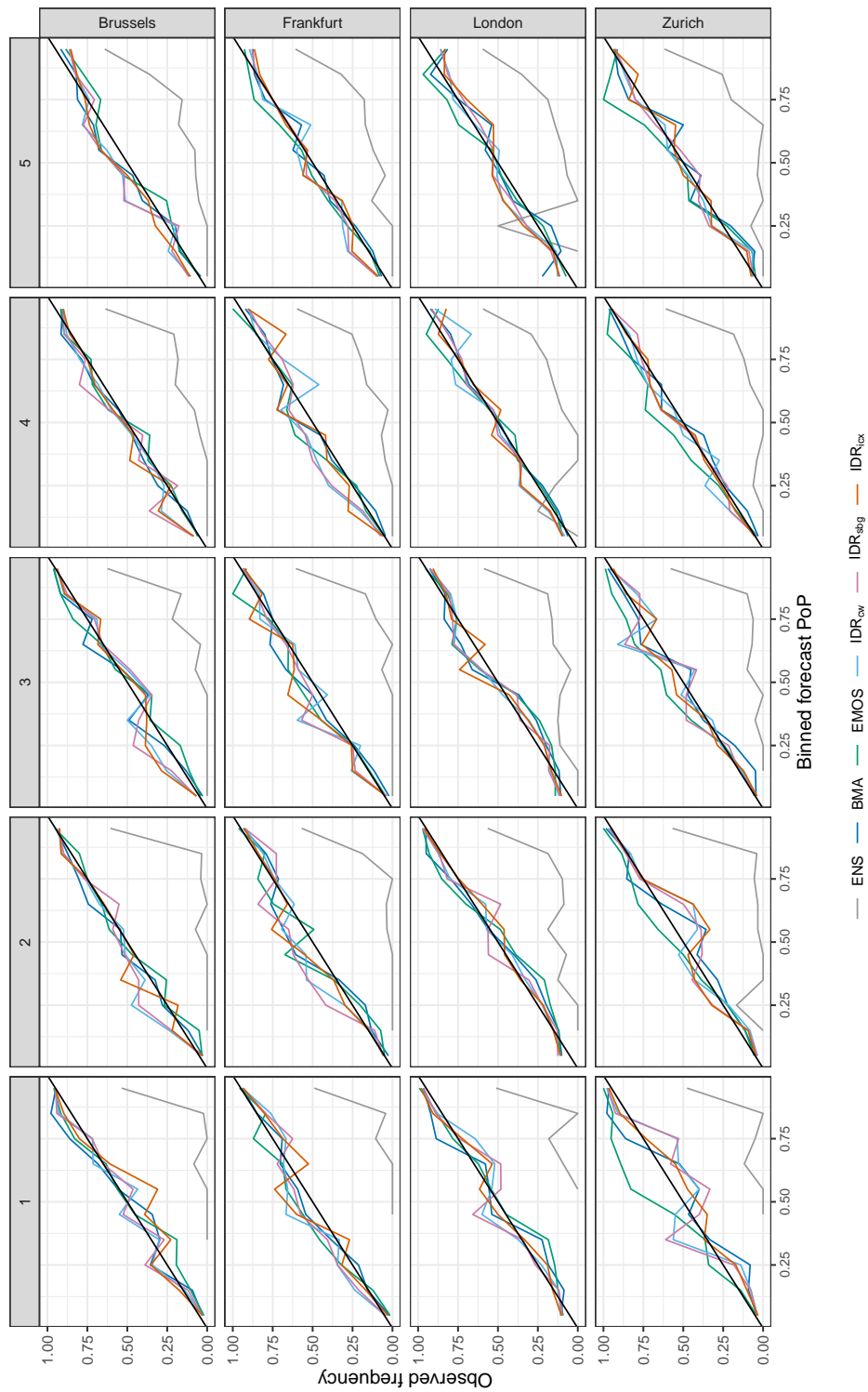
Figure 8: Reliability diagrams for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days, for the test period.

tained, and the order constraints serve to regularize the IDR solution. Weak orders lead to increased numbers of comparable pairs of training instances and predictive distributions that are more regular. The choice of the partial order is typically guided and informed by substantive expertise, as illustrated in our case study. A challenge for future research is to investigate whether the selection of the partial order can be automated.

Unlike other methods for distributional regression, which require implementation decisions, such as the specification of parametric distributions, link functions, estimation procedures and convergence criteria, to be undertaken by users, IDR is fully automatic once the partial order and the training set have been identified. In this light, we recommend that IDR be used as a benchmark technique in distributional regression and probabilistic forecasting problems. With both computational efficiency and the avoidance of overfitting in mind, IDR can be combined with subsample aggregation (subagging) in the spirit of random forests. In our case study on quantitative precipitation forecasts, we used simplistic ad hoc choices for the size and number of subsamples. Future research on computationally efficient algorithmic implementations of IDR as well as optimal and automated choices of subsampling settings is highly desirable.

A limitation of IDR in its present form is that we only consider the usual stochastic order on the space $\mathcal{P}$ of the conditional distributions. Hence, IDR is unable to distinguish situations where the conditional distributions agree in location but differ in spread, shape or other regards. This restriction is of limited concern for response variables such as precipitation accumulation or income, which are bounded below and right skewed, but may impact the application of IDR to variables with symmetric distributions. In this light, we encourage future work on ramifications of IDR, in which $\mathcal{P}$ is equipped with partial orders other than the stochastic order.

While in this paper we have studied IDR from the perspective of prediction, challenging questions also arise in estimation. In the extant literature, pointwise strong consistency of nonparametric isotonic quantile regression has been proved and associated laws of the iterated logarithm (Robertson and Wright, 1975; Casady and Cryer, 1976) and rates of convergence (Mösching and Dümbgen, 2019) have been established under suitable conditions. It will be of interest to extend these results so that they apply to IDR in general settings.

Acknowledgements

programme operated by the European Centre for Medium-Range Weather Forecasts (ECMWF). Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

References

M. S. Allen and F. A. Eckel. Value from ambiguity in ensemble forecasts. *Weather and Forecasting*, 27:70–84, 2012.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 37:1148–1178, 2019.

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

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions*. John Wiley & Sons, 1972.

D. J. Bartholomew. A test of homogeneity for ordered alternatives. *Biometrika*, 46:36–48, 1959a.

D. J. Bartholomew. A test of homogeneity for ordered alternatives II. *Biometrika*, 46:329–335, 1959b.

Basel Committee on Banking Supervision. Standard: Minimum Capital Requirements for Market Risk, 2016. URL http://www.bis.org/bcbs/publ/d352.pdf.

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

Z. Ben Bouallègue, T. Haiden, and D. S. Richardson. The diagonal score: Definition, properties, and interpretations. *Quarterly Journal of the Royal Meteorological Society*, 144:1463–1473, 2018.

P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, and S. Worley. The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072, 2010.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

N. Brümmer and J. du Preez. The PAV algorithm optimizes binary proper scoring rules, 2013. Preprint, https://arxiv.org/abs/1304.2331.

H. B. Brunk. Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics*, 26:607–616, 1955.

P. Bühlmann and B. Yu. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2002.

R. Buizza, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133:1076–1097, 2005.

A. Buja and W. Stützle. Observations on bagging. *Statistica Sinica*, 16: 323–351, 2006.

R. J. Casady and J. D. Cryer. Monotone percentile regression. *Annals of Statistics*, 4:532–541, 1976.

O. Davidov and G. Iliopoulos. Estimating a distribution function subject to a stochastic ordering restriction: A comparative study. *Journal of Nonparametric Statistics*, 24:923–933, 2012.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147:278–290, 1984.

J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–19, 2009.

H. Dette, N. Neumeyer, and K. F. Pilz. A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, 12:469–490, 2006.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883, 1998.

D. B. Dunson, N. Pillai, and J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society Series B*, 69:163–183, 2007.

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

H. El Barmi and H. Mukerjee. Inferences under a stochastic ordering constraint. *Journal of the American Statistical Association*, 100:252–261, 2005.

D. Ellsberg. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

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

C. Fraley, A. E. Raftery, and T. Gneiting. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138:190–202, 2010.

J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski. Probabilistic forecasting with spline quantile function RNNs. *Proceedings of Machine Learning Research*, 89: 1901–1910, 2019.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

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

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

T. Gneiting and R. Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29:411–422, 2011.

T. Gneiting and R. Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

T. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098–1118, 2005.

T. Gneiting, K. Larson, K. Westrick, M. G. Genton, and E. A. Aldrich. Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) method. *Journal of the American Statistical Association*, 101:968–979, 2006.

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

P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, 2014.

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *Statistical Science*, 33:563–594, 2018.

P. Hall, R. C. L. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94: 154–163, 1999.

T. M. Hamill. Practical aspects of statistical postprocessing. In S. Vannitsem, D. S. Wilks, and J. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 187–217. Elsevier, 2018.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570, 2000.

R. V. Hogg. On models and hypotheses with restricted alternatives. *Journal of the American Statistical Association*, 60:1153–1162, 1965.

T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger. Bagging survival trees. *Statistics in Medicine*, 23:77–91, 2004.

T. Hothorn, T. Kneib, and P. Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society Series B*, 76:3–27, 2014.

A. Jordan, A. Mühlemann, and J. F. Ziegel. Optimality of the PAV algorithm, 2019. Preprint, https://arxiv.org/abs/1904.04761.

N. Klein, T. Kneib, S. Lang, and A. Sohn. Bayesian structured additive distributional forecasting with an application to regional income inequality in Germany. *Annals of Applied Statistics*, 9:1024–1052, 2015.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

F. Laio and S. Tamea. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11: 1267–1277, 2007.

C.-I. C. Lee. The min-max algorithm and isotonic regression. *Annals of Statistics*, 11:467–477, 1983.

M. Leutbecher. Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, 2018. In press.

M. Leutbecher and T. N. Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227:3515–3539, 2008.

C. J. Lloyd. Estimation of a convex ROC curve. *Statistics & Probability Letters*, 59:99–111, 2002.

E. Mammen. Estimating a smooth monotone regression function. *Annals of Statistics*, 19:724–740, 1991.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and its Applications*. Springer, 2nd edition, 2011.

J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096, 1976.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

J. W. Messner. Ensemble postprocessing with R. In S. Vannitsem, D. S. Wilks, and J. Messner, editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 291–326. Elsevier, 2018.

R. Miles. The complete almagamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 46:317–327, 1959.

A. Mösching and L. Dümbgen. Monotone least squares and isotonic quantiles, 2019. Preprint, https://arxiv.org/abs/1901.02398.

F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis. The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119, 1996.

B. Neelon and D. B. Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60:398–406, 2004.

W. K. Newey and J. L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847, 1987.

F. Pappenberger, M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salomon. How do I know if my forecasts are better? Using benchmarks in hydrologic ensemble prediction. *Journal of Hydrology*, 522:697–713, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.r-project.org/.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.

S. Rasp and S. Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900, 2018.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 54:507–554, 2005.

T. Robertson and F. T. Wright. Consistency in generalized isotonic regression. *Annals of Statistics*, 3:350–362, 1975.

T. Robertson and F. T. Wright. Algorithms in order restricted statistical inference and the Cauchy mean value property. *Annals of Statistics*, 8: 645–651, 1980.

J. Rojo and H. El Barmi. Estimation of distribution functions under second order stochastic dominance. *Statistica Sinica*, 13:903–926, 2003.

B. Rossi. Exchange rate predictability. *Journal of Economic Literature*, 51: 1063–1110, 2013.

R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28:616–640, 2013.

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

M. Scheuerer. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140:1086–1096, 2014.

M. Scheuerer and T. M. Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143:4578–4596, 2015.

L. Schlosser, T. Hothorn, R. Stauffer, and A. Zeileis. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics*, 2019. In press.

L. Schulze Waltrup, F. Sobotka, T. Kneib, and G. Kauermann. Expectile and quantile regression — David and Goliath? *Statistical Modelling*, 15: 433–456, 2015.

K. Seo. Ambiguity and second-order belief. *Econometrica*, 77:1575–1605, 2009.

M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer, New York, 2007.

T. S. Shively, T. W. Sager, and S. G. Walker. A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society Series B*, 71:159–175, 2009.

J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using Baysian model averaging. *Monthly Weather Review*, 135:3209–3220, 2007.

B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: An operator splitting solver for quadratic programs, 2017. Preprint, https://arxiv.org/abs/1711.08013.

B. Stellato, G. Banjac, P. Goulart, and S. Boyd. *osqp: Quadratic Programming Solver using the 'OSQP' Library*, 2018. URL https://CRAN.R-project.org/package=osqp. R package version 0.5.0.

S. M. Stigler. The transition from point to distribution estimation. *Bulletin of the International Statistical Institute*, 46(2):332–340, 1975.

R. Swinbank, M. Kyouda, P. Buchanan, L. Froude, T. M. Hamill, T. D. Hewson, J. H. Keller, M. Matsueda, J. Methven, F. Pappenberger, M. Scheuerer, H. A. Titley, L. Wilson, and M. Yamaguchi. The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97:49–67, 2016.

M. Taillardat, O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144:2375–2393, 2016.

M. Taillardat, A. L. Fougères, P. Naveau, and O. Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34:617–634, 2019.

N. Umlauf and T. Kneib. A primer on Bayesian distributional regression. *Statistical Modelling*, 18:219–247, 2018.

C. van Eeden. *Testing and Estimating Ordered Parameters of Probability Distributions*. PhD thesis, University of Amsterdam, 1958.

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

P. Vogel, P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting. Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, 33:369–388, 2018.

W. J. Wilbur, L. Yeganova, and W. Kim. The synergy between PAV and AdaBoost. *Machine Learning*, 61:71–103, 2005.

D. S. Wilks. *Statistical Methods in the Atmospheric Sciences.* Elsevier, 3rd edition, 2011.

A   Proofs for Section 2

Proof of Theorem 2.1. Let $\mathcal{A}$ be the lattice of all subsets of $\{1,\ldots,n\}$ that yield admissible superlevel sets for an increasing function $\{x_1,\ldots,x_n\} \to \mathbb{R}$. More precisely, a set $A \subseteq \{1,\ldots,n\}$ belongs to $\mathcal{A}$ if and only if for any $i \in A$ and any x_j with $x_i \preceq x_j$ it follows that $j \in A$.

Let $z \in \mathbb{R}$. By Jordan et al. (2019, Proposition 4.5 and Lemma 4.6), the minimizer of the criterion

$$\frac{1}{n}\sum_{i=1}^{n}(p_i - \mathbb{1}\{y_i \leq z\})^2 \tag{22}$$

over all $\boldsymbol{p} = (p_1,\ldots,p_n) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$ is uniquely determined and given by $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z),\ldots,\hat{F}_n(z)) \in \mathbb{R}^n$ with

$$\hat{F}_i(z) = \max_{A \in \mathcal{A}: i \in A} \min_{A' \in \mathcal{A}: A' \subsetneq A} \frac{1}{\#(A \setminus A')} \sum_{j \in A \setminus A'} \mathbb{1}\{y_j \leq z\}, \tag{23}$$

for $i = 1,\ldots,n$, where $\#B$ denotes the cardinality of a set B. From the definition of the CRPS it is clear that $\hat{\boldsymbol{F}}$ minimizes $\ell_{\mathrm{CRPS}}(\boldsymbol{F})$ over all tuples of functions $\boldsymbol{F} = (F_1,\ldots,F_n)$ with $F_i : \mathbb{R} \to \mathbb{R}$ such that for each $z \in \mathbb{R}$, $(F_1(z),\ldots,F_n(z)) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$. It remains to show that for each $i = 1,\ldots,n$, $\hat{F}_i$ is a valid CDF.

Let $i \in \{1,\ldots,n\}$, $z \leq z'$, $B \subseteq \{1,\ldots,n\}$. It is clear from (23) that the domain of $\hat{F}_i$ is $[0,1]$. Furthermore,

$$\frac{1}{\#B}\sum_{j \in B}\mathbb{1}\{y_j \leq z\} \leq \frac{1}{\#B}\sum_{j \in B}\mathbb{1}\{y_j \leq z'\}, \tag{24}$$

and therefore, by (23), $\hat{F}_i(z) \leq \hat{F}_i(z')$. The function $\hat{F}_i$ is also right-continuous because for $z' \downarrow z$, the right-hand side of (24) converges to the left-hand side. Finally, for $z \to \pm\infty$ the left-hand side of (24) converges to zero and one, respectively, which concludes the proof. $\qquad\square$

Proof of Theorem 2.2. First, we show threshold calibration. Let (X,Y) be a random vector with distribution $(1/n)\sum_{i=1}^{n}\delta_{(x_i,y_i)}$ where $\delta_{(x_i,y_i)}$ denotes the Dirac measure at (x_i,y_i). Let $z \in \mathbb{R}$. By Lee (1983, Theorem 6.4), there exists a partition $\{B_m\}_{m=1}^{M}$ of $\{1,\ldots,n\}$ such that

$$\hat{F}_i(z) = \hat{F}_{x_i}(z) = \sum_{m=1}^{M}\mathbb{1}\{i \in B_m\}\frac{1}{\#B_m}\sum_{j \in B_m}\mathbb{1}\{y_j \leq z\}.$$

Therefore, the σ-algebra generated by $\hat{F}_X(z)$ is contained in the σ-algebra generated by $\{\bar{B}_m\}_{m=1}^M$ with $\bar{B}_m = \{(x_i, y_i) : i \in B_m\}$. Furthermore,

$$\mathbb{E}\left(\mathbb{1}\{Y \le z\}\mathbb{1}\{(X,Y) \in \bar{B}_m\}\right) = \frac{1}{n}\sum_{j \in B_m}\mathbb{1}\{y_j \le z\}$$
$$= \mathbb{E}\left(\hat{F}_X(z)\mathbb{1}\{(X,Y) \in \bar{B}_m\}\right).$$

Part i) for the scoring rules of type (12) follows directly from the arguments in the proof of Theorem 2.1. By Jordan et al. (2019, Proposition 4.5 and Lemma 4.6) the solution $\hat{\boldsymbol{F}}(z)$ at (23) is not only the unique minimizer of the criterion (22) but also the unique solution that minimizes

$$\frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{c < p_i\} - \mathbb{1}\{y_i \le z\}\right)\left(c - \mathbb{1}\{y_i \le z\}\right) \tag{25}$$

over all $\boldsymbol{p} = (p_1, \ldots, p_n) \in \mathbb{R}_{\downarrow,\boldsymbol{x}}^n$ simultaneously for all $c \in (0,1)$. Since we know that $\hat{\boldsymbol{F}} \in \mathcal{P}_{\downarrow,\boldsymbol{x}}^n$, and $(1/n)\sum_{i=1}^n \mathrm{S}_{z,c}(F_i, y_i)$ is equal to the expression at (25) with $p_i = F_i(z)$, we obtain the claim.

Part iii) is a direct consequence of the arguments for the second part of part i) and the representation theorem of Schervish (1989) for proper scoring rules of binary events.

Let $\alpha \in (0,1)$. Concerning part ii), observe that any function s_α satisfying the requirements of the theorem can be written as $\int \tilde{\mathrm{S}}_{\alpha,\theta}^Q(q,y)\,\mathrm{d}h(\theta)$ for some Borel measure h on $\mathbb{R}$; see Ehm et al. (2016, Theorem 1). Here,

$$\tilde{\mathrm{S}}_{\alpha,\theta}^Q(q,y) = \begin{cases} 1 - \alpha, & y \le \theta < q, \\ \alpha, & q \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

By Jordan et al. (2019, Propositions 4.5 and 4.7) there exists a unique solution $\hat{\boldsymbol{q}}(\alpha) = (\hat{q}_1(\alpha), \ldots, \hat{q}_n(\alpha)) \in \mathbb{R}_{\downarrow,\boldsymbol{x}}^n$ that minimizes

$$\frac{1}{n}\sum_{i=1}^n \tilde{\mathrm{S}}_{\alpha,\theta}^Q(q_i, y_i)$$

over all $\boldsymbol{q} = (q_1, \ldots, q_n) \in \mathbb{R}_{\downarrow,\boldsymbol{x}}^n$ simultaneously over all $\theta \in \mathbb{R}$ such that for each $i \in \{1, \ldots, n\}$, $\hat{q}_i(\alpha)$ is the lower α-sample-quantile of some subset of observations $B_i \subseteq \{y_1, \ldots, y_n\}$. Indeed, the solution has a max-min representation as in (23) with the empirical mean of the indicators replaced by the lower α-sample quantile over all observations in $A \backslash A'$. The max-min representation for $\hat{q}_i(\alpha)$ yields that $\hat{q}_i(\cdot)$ is increasing and left-continuous because lower α-sample-quantiles are increasing and left-continuous as a function of α. Therefore, $\hat{q}_i(\cdot)$ is a valid quantile function for each $i =$

$1, \ldots, n$, and the generalized inverse $\hat{\boldsymbol{q}}^{-1} = (\hat{q}_1^{-1}, \ldots, \hat{q}_n^{-1})$ is a member of $\mathcal{P}_{\uparrow, \boldsymbol{x}}^n$.

Since $\mathrm{S}_{\alpha, \theta}^Q(F, y) = \tilde{\mathrm{S}}_{\alpha, \theta}^Q(F^{-1}(\alpha), y)$ for any CDF F, it follows from (3) that $\hat{\boldsymbol{q}}^{-1}$ is an isotonic CRPS-regression of $\boldsymbol{y}$ on $\boldsymbol{x}$. To conclude the proof of part ii), it remains to note that $\hat{\boldsymbol{q}}^{-1} = \hat{\boldsymbol{F}}$ due to the uniqueness of $\hat{\boldsymbol{F}}$. The first part of part i) is now also immediate. $\qquad\square$

B   Proofs for Section 3

Proof of Proposition 3.2. Denote the CDF corresponding to the empirical distribution of $x_1, \ldots, x_d$ and of $x_1', \ldots, x_d'$ by F and G, respectively. For part i), assume that $x_{(i)} \leq x_{(i)}'$ for $i = 1, \ldots, d$, and let $z \in \mathbb{R}$. Then,

$$F(z) = \frac{\#\{i : x_{(i)} \leq z\}}{d} \geq \frac{\#\{i : x_{(i)}' \leq z\}}{d} = G(y),$$

hence F is smaller than G in the usual stochastic order. Conversely, if F is smaller then G, by choosing $z = x_{(k)}'$, $k = 1, \ldots, d$, we obtain

$$\frac{\#\{i : x_{(i)} \leq x_{(k)}'\}}{d} = F(x_{(k)}') \geq G(x_{(k)}') = \frac{\#\{i : x_{(i)}' \leq x_{(k)}'\}}{d}.$$

By definition of the k-th order statistic, we know that $\#\{i : x_{(i)}' \leq x_{(k)}'\} \geq k$ (with equality if the x_i' are distinct). Therefore, $\#\{i : x_{(i)} \leq x_{(k)}'\} \geq k$. This can only be true if $x_{(k)} \leq x_{(k)}'$.

Concerning part ii), we can assume without loss of generality that $x_1 \leq \cdots \leq x_d$, otherwise we reorder the pairs (x_i, y_i). Now apply part i): We know that $x_1 \leq x_1'$ and $x_{(1)}' \geq x_j$ for some j. But the components of x are sorted, hence $x_{(1)}' \geq x_j \geq x_1 = x_{(1)}$, and also $x_1' \geq x_{(1)}' \geq x_j$. So we can think of the positions of x_1' and $x_{(1)}'$ in x' to be exchanged, without violating the condition $x \preceq x'$. Now we can ignore the pair $(x_1', x_{(1)}')$ and proceed in the same way for remaining components $(x_i)_{i=2}^d$ and $(x_i')_{i=2}^d$.

For the proof of part iii), assume the opposite, that is, $x_i \geq x_i'$ for $i = 1, \ldots, d$. By ii), we know that $x \succeq_{\mathrm{st}} x'$. But by assumption $x \preceq_{\mathrm{st}} x'$, hence x and x' are permutations of each other. But then either $x = x'$, or x and x' cannot be comparable in the componentwise order.

The last part is immediate from part i). $\qquad\square$

Proof of Proposition 3.3. Part i) is a consequence of Theorem 4.A.3 of Shaked and Shanthikumar (2007). Part ii) follows from part i) and Proposition 3.2 i). For part iii) note that the Gini mean difference has the equivalent formula

$$g(x) = \frac{2}{d(d-1)} \sum_{i=1}^d x_{(i)}(2i - d - 1),$$

which can be rewritten as

$$g(x) = \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x_{(j)} - 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i.$$

Part i) implies that

$$g(x') + 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i' = \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x_{(j)}'$$

$$\geq \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x_{(j)} = g(x) + 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i. \quad \square$$