Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Verification: assessment of calibration and accuracy

Note no SAMBA/17/17

Authors Thordis L. Thorarinsdottir
Nina Schuhen

Date 17th October 2017

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Thordis L. Thorarinsdottir is Chief Research Scientist at the Norwegian Computing Center in Oslo, Norway. Nina Schuhen is PhD Student at the Norwegian Computing Center and the University of Oslo, Norway.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Verification: assessment of calibration and accuracy**
Authors	**Thordis L. Thorarinsdottir** <thordis@nr.no> **Nina Schuhen** <nina.schuhen@nr.no>
Date	17th October 2017
Publication number	SAMBA/17/17

Abstract

In ensemble forecasting, forecast verification methods are needed to diagnose both the need for statistical post-processing, and the effectiveness of the post-processing methods in producing calibrated and accurate forecasts. This chapter discusses an array of techniques that can be used in this context, making the distinction between verification tools that are useful for ranking competing forecasters and those that are more appropriate for improving our understanding of the performance of a single method. With a focus on continuous variables, verification methods for both univariate and multivariate forecasts are discussed, including approaches that are specifically tailored to the evaluation of extreme events.

Front page photo by Aaron Burden is from https://unsplash.com.

Keywords	Probabilistic forecasting; Forecast accuracy; Forecast evaluation; Calibration; Reliability; Proper scoring rules
Target group	Forecasters
Availability	Open
Project	Postprocessing PhD
Project number	220783
Research field	Statistics, Forecasting
Number of pages	38

Introduction

In a discussion article on the application of mathematics in meteorology, Bigelow (1905) describes the fundamentals of modeling in a timeless manner:

> *"There are three processes that are generally essential for the complete development of any branch of science, and they must be accurately applied before the subject can be considered to be satisfactorily explained. The first is the discovery of a mathematical analysis, the second is the discussion of numerous observations, and the third is a correct application of the mathematics to the observations, including a demonstration that these are in agreement."*

The topic of this chapter are methods for carrying out the last item on Bigelow's list, that is, methods to demonstrate the agreement between a model and a set of observations. Ensemble prediction systems and statistically postprocessed ensemble forecasts provide probabilistic predictions of future weather. Verification methods applied to these systems should thus be equipped to handle both the verification of the best prediction derived from the ensemble and the verification of the associated prediction uncertainty.

Murphy (1993) argues that a general prediction system should strive to perform well on three types of goodness: There should be consistency between the forecaster's judgment and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user. Similarly, Gneiting et al. (2007) state that the goal of probabilistic forecasting should be to maximize the sharpness of the predictive distribution subject to calibration. Here, calibration refers to the statistical consistency between the forecast and the observation, while sharpness refers to the concentration of the forecast uncertainty; the sharper the forecast, the higher information value will it provide. The prediction goal of Gneiting et al. (2007) is thus equivalent to Murphy's second and third types of goodness.

We focus on verification methods for probabilistic predictions of continuous variables in one or more dimensions under the general framework described by Murphy (1993) and Gneiting et al. (2007). Specifically, we denote an observation in d dimensions by $y = (y_1, \ldots, y_d) \in \Omega^d$ for $d = 1, 2, \ldots$ where Ω denotes either the real axis $\mathbb{R}$, the non-negative real axis $\mathbb{R}_{\geq 0}$, the positive real axis $\mathbb{R}_{>0}$, or an interval on $\mathbb{R}$. A probabilistic forecast for y given by a distribution function with support on Ω^d is denoted by $F \in \mathcal{F}$ for some appropriate class of distributions $\mathcal{F}$, with the density denoted by f if it exists. For ensemble forecasts, we will alternatively use the notation $\mathbf{x} = \{x_1, \cdots, x_K\}$ to describe the K ensemble members or F for the associated empirical distribution function. Verification methods for deterministic predictions and other types of variables are discussed e.g. in Wilks (2011, Ch. 8) and Jolliffe and Stephenson (2012).

The chapter is organized as follows. Diagnostic tools for checking calibration are discussed in Section 2. The next Section 3 describes methods that assess the accuracy of forecasts where each forecast is issued a numerical score based on the event that materializes. Scoring rules apply to individual events while divergence functions compare the empir-

ical distribution of a series of events against a predictive distribution. The scores may focus on certain aspects of the forecast, such as the tails, and it is important to also assess the uncertainty in the scores. The properties of various univariate scores are compared in a simulation study. While the methods in Section 3 provide a decision theoretically coherent approach to model evaluation and model ranking, they may hide key information about the model performance such as the direction of bias. Additional evaluation may thus be needed to better understand the performance of a single model. Approaches for this are discussed in Section 4. The chapter then closes with a summary in Section 5.

Calibration

Calibration, or reliability, is the most fundamental aspect of forecast skill for probabilistic forecasts as it is a necessary condition for the optimal use and value of the forecast. Calibration refers to the statistical compatibility between the forecast and the observation; the forecast is calibrated if the observation cannot be distinguished from a random draw from the predictive distribution.

Univariate calibration

Several alternative notions of univariate calibration exist for both a single forecast (Gneiting et al., 2007; Tsyplakov, 2013) and a group of forecasts (Strähl and Ziegel, 2017). We focus on the so-called *probabilistic calibration* as suggested by Dawid (1984); F is probabilistically calibrated if the *probability integral transform* (PIT) $F(Y)$, the value of the predictive cumulative distribution function in the random observation Y, is uniformly distributed. If F has a discrete component, a randomized version of the PIT given by

$$\lim_{y \uparrow Y} F(y) + V\big(F(Y) - \lim_{y \uparrow Y} F(y)\big)$$

with $V \sim \mathcal{U}([0,1])$ may be used, see Gneiting and Ranjan (2013). Here, we use $y \uparrow Y$ to denote that the limit is taken as y approaches Y from below.

Assume our test set consists of n observations $y_1, \ldots, y_n$. For a forecasting method issuing continuous univariate predictive distributions $F_1, \ldots, F_n$, calibration can be assessed empirically by plotting the histogram of the PIT values

$$F_1(y_1), \ldots, F_n(y_n).$$

A forecasting method that is calibrated on average will return a uniform histogram, a $\cap$-shape indicates overdispersion and a $\cup$-shape indicates underdispersion while a systematic bias results in a triangular shape histogram. Examples of miscalibration are shown in Figure 1, including a biased forecast (subfigure a), an underdispersive forecast (subfigure b), an overdispersive forecast (subfigure c), and an example of a multiply mis-specified forecast where the left tail is too light, the main bulk of the distribution lacks mass and the right tail is too heavy (subfigure d).

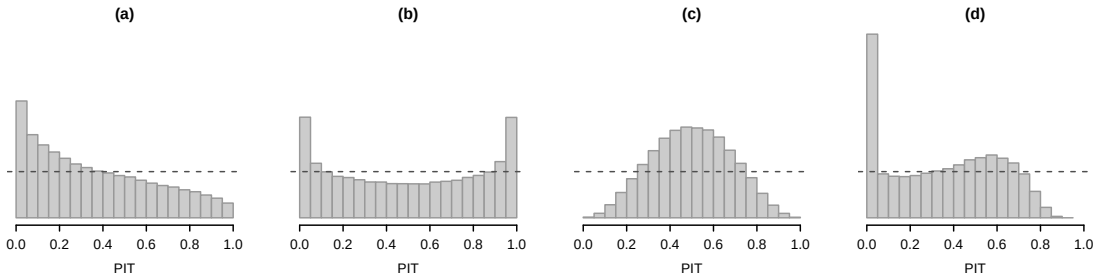

Figure 1. Probability integral transform (PIT) histograms for 100,000 simulated standard Gaussian $\mathcal{N}(0,1)$ observations and various mis-specified forecasts: (a) biased $\mathcal{N}(0.5,1)$ forecast, (b) underdispersive $\mathcal{N}(0,0.75^2)$ forecast, (c) overdispersive $\mathcal{N}(0,2^2)$ forecast, and (d) multiply mis-specified generalized extreme value GEV$(0,1,0.5)$ forecast. The theoretically optimal histograms are indicated with dashed lines.

The discrete equivalent of the PIT histogram, which applies to ensemble forecasts, is the verification rank histogram (Anderson, 1996; Hamill and Colucci, 1997). It shows the distribution of the ranks of the observations within the corresponding ensembles and has the same interpretation as the PIT histogram.

The information provided by a rank histogram may also be summarized numerically by the reliability index (RI) which is defined as

$$\mathrm{RI} = \sum_{i=1}^{I} \left| \zeta_i - \frac{1}{I} \right|,$$

where I is the number of (equally-sized) bins in the histogram and ζ_i is the observed relative frequency in bin $i = 1, \ldots, I$. The reliability index thus measures the departure of the rank histogram from uniformity (Delle Monache et al., 2006).

Multivariate calibration

For assessing the calibration of multivariate forecasts, Gneiting et al. (2008) formalized a general two-step framework. Let $S = \{x_1, \ldots x_K, y\}$ denote a set of $K+1$ points in Ω^d comprising an ensemble forecast with k members and the corresponding observation y. The rank of y in S, $\mathrm{rank}_S(y)$, is calculated in two steps,

(i) apply a pre-rank function $\rho_S : \Omega^d \to \mathbb{R}_{\geq 0}$ to calculate the pre-rank $\rho_S(u)$ of every $u \in S$ resulting in a univariate value for each u;

(ii) set the rank of the observation y equal to the rank of $\rho_S(y)$ in $\{\rho_S(x_1), \ldots, \rho_S(x_K), \rho_S(y)\}$,

$$\mathrm{rank}_S(y) = \sum_{v \in S} \mathbb{1}\{\rho_S(v) \leq \rho_S(y)\},$$

where $\mathbb{1}$ denotes the indicator function and ties are resolved at random.

Here, we focus on four different approaches that follow this general two-step framework. Further approaches are discussed in Gneiting et al. (2008), Ziegel and Gneiting (2014) and Wilks (2017). The difference between our four approaches lies in the definition of the pre-rank function ρ_S in step (i). The *multivariate ranking* of Gneiting et al. (2008) is defined

using the pre-rank function

$$\rho_S^{\mathrm{m}}(u) = \sum_{v \in S} \mathbb{1}\{v \preceq u\}, \tag{1}$$

wehre $v \preceq u$ if and only if $v_i \leq u_i$ in all components $i = 1, \ldots, d$. Gneiting et al. (2008) further consider an optional initial step in the ranking procedure in which the data is normalized in each component before the ranking. The *average ranking* proposed by Thorarinsdottir et al. (2016) provides a similar ascending rank structure and is given by the average over the univariate ranks. That is, let

$$\mathrm{rank}_S(u, i) = \sum_{v \in S} \mathbb{1}\{v_i \leq u_i\}$$

denote the standard univariate rank of the ith component of u among the values in S. The multivariate average rank is then defined using the pre-rank function

$$\rho_S^{\mathrm{a}}(u) = \frac{1}{d} \sum_{i=1}^{d} \mathrm{rank}_S(u, i). \tag{2}$$

Further two approaches assess the centrality of the observation within the ensemble. Under *minimum spanning tree ranking*, the pre-rank function $\rho_S^{\mathrm{mst}}(u)$ is given by the length of the minimum spanning tree of the set $S \setminus u$, that is, the set S without the element u (Smith and Hansen, 2004; Wilks, 2004). Here, a spanning tree of the set $S \setminus u$ is a collection of $k - 1$ edges such that all points in $S \setminus u$ are used. The spanning tree with the smallest length is then the minimum spanning tree (Kruskal, 1956); it may e.g. be calculated using the `R` package `vegan` (Oksanen et al., 2017; R Core Team, 2016).

Alternatively, the *band depth ranking* proposed by Thorarinsdottir et al. (2016) uses a pre-rank function that calculates the proportion of components of $u \in S$ inside bands defined by pairs of points from S. It can be written as

$$\rho_S^{\mathrm{bd}}(u) = \frac{1}{d} \sum_{i=1}^{d} \Big[\mathrm{rank}_S(u, i) \big[(K+1) - \mathrm{rank}_S(u, i) \big] + \big[\mathrm{rank}_S(u, i) - 1 \big] \sum_{v \in S} \mathbb{1}\{v_i = u_i\} \Big]. \tag{3}$$

If $u_i \neq v_i$ with probability 1 for all $u, v \in S$ with $u \neq v$ and $i = 1, \ldots, d$ the formula in (3) may be simplified to

$$\rho_S^{\mathrm{bd}}(u) = \frac{1}{d} \sum_{i=1}^{d} \big[(K+1) - \mathrm{rank}_S(u, i) \big] \big[\mathrm{rank}_S(u, i) - 1 \big]. \tag{4}$$

This implies that the formula in (3) should be used for forecasts with a discrete component, *e.g.* precipitation forecasts. The band depth in (3) is equivalent to the simplicial depth proposed by Liu (1990) and thus also to the simplicial depth ranking proposed by Mirzargar and Anderson (2017), see López-Pintado and Romo (2009) and Thorarinsdottir et al. (2016).

While all four methods return a uniform rank histogram for a calibrated forecast, the interpretation of the histogram shape for a misspecified forecast varies between the methods as demonstrated in the example below.

Example: Comparing multivariate ranking methods

The four multivariate ranking methods are compared in Figure 2 for several different settings where $y \in \mathbb{R}^d$ can be thought of as a temporal trajectory of a real valued variable observed at $d = 10$ equidistant time points $t = 1, \ldots, 10$. In the first two examples (row one and two), y is a realization of a zero-mean Gaussian AR(1) (autoregressive) process Y with a covariance function given by

$$\text{Cov}(Y_i, Y_j) = \exp(-|i - j|/\tau), \quad \tau > 0. \tag{5}$$

The process Y thus has standard Gaussian marginal distributions while the parameter τ controls how fast correlations decay with time lag. We set $\tau = 3$ for Y and consider ensemble forecasts with 50 members of the same type but with a different parameter value τ. That is, we set $\tau = 1.5$ in row one (too strong correlation) and $\tau = 5$ in row two (too weak correlation). It follows from this construction that a univariate calibration test at a fixed time point would not detect any miscalibration in the forecasts.

While all four methods are able to detect the misspecification in the correlation structure, the resulting histograms vary in shape. The shape of the average rank histograms and the band depth rank histograms offer a similar interpretation as that of the univariate rank histograms in Figure 1 with a $\cup$-shape when the correlation is too strong (underdispersion across components) and a $\cap$-shape when the correlation is too weak (overdispersion across components). In these 10-dimensional examples, the pre-rank ordering of the multivariate rank histograms (1) is only able to detect miscalibration related to the highest ranks, see also the discussion in Pinson and Girard (2012) and Thorarinsdottir et al. (2016). Under minimum spanning tree ranking, too many observations have high ranks when the correlation in the forecasts is too strong and the opposite holds for the example with too weak correlation in the forecasts.

In the latter two examples in Figure 2 (rows three and four), both observations and forecasts are i.i.d. variables in ten dimensions. However, the marginal distributions of the ensemble forecasts are misspecified. The observations follow a standard Gaussian distribution, the forecasts in row three have a standard deviation of 1.25 (overdispersion) and the forecasts in row four have a standard deviation of 0.85 (underdispersion). The shape of the average rank histograms is exactly that of their univariate counterparts in Figure 1, indicating that the ranking method cannot distinguish between miscalibration in the marginals and the higher order structure. For the two ranking methods based on centrality, the marginal overdispersion results in too many high ranks while the marginal underdispersion results in too many low ranks. For this dimensionality, the multivariate ranking is unable to detect the miscalibration.

Further comparison of the four ranking methods is provided in Thorarinsdottir et al. (2016) and Wilks (2017). In general, it is a challenging task to represent and compare a multi-faceted higher order structure with a single value. As the different methods vary in their strengths and weaknesses, it is recommended to apply several of these methods when assessing multivariate calibration. The multivariate ranking of Gneiting et al. (2008), for instance, does not satisfy affine invariance (Mirzargar and Anderson, 2017)

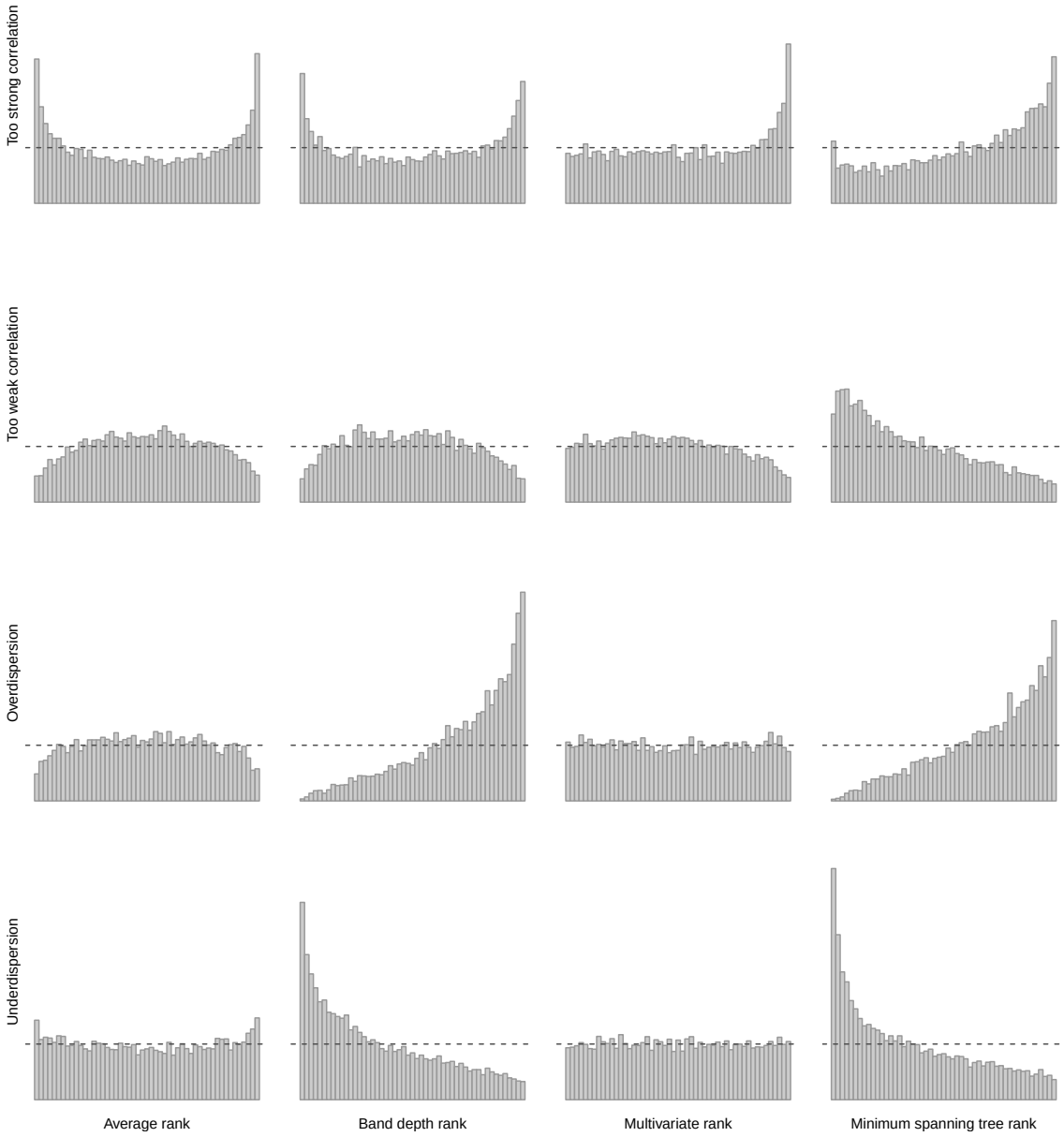

Figure 2. Rank histograms for multivariate data showing various types of miscalibration under different ranking methods: Average ranking (first column), band depth ranking (second column), multivariate ranking (third column) and minimum spanning tree ranking (forth column). 10,000 simulated observations of dimension 10 are compared against ensemble forecasts with 50 members. In the top two rows, the observations are realizations of a zero-mean Gaussian AR(1) process with the covariance function in (5) where $\tau = 3$. The forecasts follow the same model with $\tau = 1.5$ (first row) and $\tau = 5$ (second row). In the bottom two rows, the observations are i.i.d. standard Gaussian variables while the forecasts have variance 1.25^2 (third row) and 0.85^2 (forth row). The theoretically optimal histograms are indicated with dashed lines.

while lower-dimensional positive and negative biases may cancel out under average ranking (Thorarinsdottir et al., 2016).

Furthermore, a prior assessment of the marginal calibration may increase the information value in the multivariate rank histograms and ease the interpretation of the resulting shapes. As the multivariate methods perform a simultaneous assessment of the marginal and the higher-order calibration, a specific non-uniform shape may represent multiple types of mis-specifications. For example, depth-based approaches such as the band depth ranking and the minimum spanning tree ranking are not able to distinguish between underdispersive and biased forecasts (Mirzargar and Anderson, 2017).

Accuracy

In this section, we discuss methods for assessing forecast accuracy that are appropriate for ranking and comparing competing forecasting methods. Alternative assessment techniques that may provide additional insights for understanding the performance and errors of a single forecasting model while not appropriate for forecast ranking are discussed in Section 4 below.

Univariate assessment

Scoring rules

Scoring rules assess the accuracy of probabilistic forecasts by assigning a numerical penalty to each forecast-observation pair. Specifically, a scoring rule is a mapping

$$S : \mathcal{F} \times \Omega^d \to \mathbb{R} \cup \{\infty\} \tag{6}$$

where for every $F \in \mathcal{F}$ the map $y \mapsto S(F, y)$ is quasi-integrable for every $F \in \mathcal{F}$. In our notation, a smaller penalty indicates a better prediction. A scoring rule is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y) \tag{7}$$

for all probability distributions $F, G \in \mathcal{F}$, that is, if the expected score for a random observation Y is optimized if the true distribution of Y (G) is issued as the forecast. The scoring rule is *strictly proper* relative to the class $\mathcal{F}$ if (7) holds with equality only if $F = G$. Propriety will encourage honesty and prevent hedging, which coincides with Murphy's first type of goodness (Murphy, 1993). That is, the scores cannot be hedged to improve the preceived performance by a willful divergence of the forecast from the true distribution, see e.g. the discussion in Section 1 of Gneiting (2011).

Competing forecasting methods are compared based on a proper scoring rule by comparing the mean score over an out-of-sample test set, and the method with the smallest mean score is preferred. Formal tests of the null hypothesis of equal predictive performance can also be employed, see Section 3.7. While average scores are directly comparable if they refer to the same set of forecast situations, this may no longer hold for distinct sets of forecast cases, for instance, due to spatial and temporal variability in the predictability

of weather. For ease of interpretability and to address this issue, verification results are sometimes represented as a *skill score* of the form

$$S_n^{\text{skill}}(A) = \frac{\frac{1}{n}\sum_{i=1}^n S(F_i^A, y_i) - \frac{1}{n}\sum_{i=1}^n S(F_i^{\text{ref}}, y_i)}{\frac{1}{n}\sum_{i=1}^n S(F_i^{\text{perf}}, y_i) - \frac{1}{n}\sum_{i=1}^n S(F_i^{\text{ref}}, y_i)} \tag{8}$$

for the forecasting method A where F^{ref} denotes the forecast from a reference method, F^{perf} denotes the perfect forecast and n is the size of the test set. The skill score is standardized such that it takes the value 1 for an optimal forecast and the value 0 for the reference forecast. Negative values thus indicate that the forecasting method A is of a lesser quality than the reference forecast. However, it is vital to select the reference forecast with care (Murphy, 1974, 1992) as skill scores of the form (8) may be improper even if the underlying scoring rule S is proper (Gneiting and Raftery, 2007; Murphy, 1973a).

The most popular proper scoring rules for univariate real valued quantities are the *ignorance* (or *logarithmic*) *score* (IGN) and the continuous ranked probability score, see Gneiting and Raftery (2007) for a more comprehensive list. IGN is defined as

$$\text{IGN}(F, y) = -\log f(y), \tag{9}$$

where f denotes the density of F (Good, 1952). It thus applies to absolutely continuous distributions only and cannot be applied directly to ensemble forecasts. For a large enough ensemble, the density of the ensemble forecast may potentially be approximated using e.g. kernel density estimation or by fitting a parametric distribution. Alternatively, IGN may be replaced by the *Dawid-Sebastiani* (DS) *score* (Dawid and Sebastiani, 1999),

$$\text{DS}(F, y) = \log \sigma_F^2 + \frac{(y - \mu_F)^2}{\sigma_F^2}, \tag{10}$$

where μ_F denotes the mean value of F and σ_F^2 its variance. While the proper DS score equals IGN for a Gaussian predictive distribution F, it only requires the estimation of the ensemble mean and variance.

The *continuous ranked probability score* (CRPS; Matheson and Winkler, 1976) is of particular interest in that it simultaneously assesses both calibration and sharpness, and thus all three types of goodness discussed by Murphy (1993). The CRPS applies to probability distributions with finite mean and has three equivalent definitions (Gneiting and Raftery, 2007; Gneiting and Ranjan, 2011; Hersbach, 2000; Laio and Tamea, 2007),

$$\begin{aligned}
\text{CRPS}(F, y) &= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F|X - X'| && (11)\\
&= \int_{-\infty}^{+\infty} \big(F(x) - \mathbb{1}\{y \le x\}\big)^2 \mathrm{d}x && (12)\\
&= \int_0^1 \big(F^{-1}(\tau) - y\big)\big(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau\big)\mathrm{d}\tau. && (13)
\end{aligned}$$

Here, X and X' denote two independent random variables with distribution F, $\mathbb{1}\{y \le x\}$ denotes the indicator function which is equal to 1 if $y \le x$ and 0 otherwise, and $F^{-1}(\tau) = \inf\{x \in \mathbb{R} : \tau \le F(x)\}$ is the quantile function of F.

It follows directly from (12) and (13) that the CRPS is tightly linked to other proper scores that focus on specific parts of the predictive distribution. The form in (12) can be interpreted as the integral over the *Brier score* (Brier, 1950) which assesses the predictive probability of threshold exceedance. The Brier score is usually written on the form

$$\text{BS}(F, y \mid u) = \big(p_u - \mathbb{1}\{y \geq u\}\big)^2, \tag{14}$$

for a threshold u with $p_u = 1 - F(u)$. Similarly, the integrand in (13) equals the *quantile score* (Friederichs and Hense, 2007; Gneiting and Raftery, 2007),

$$\text{QS}(F, y \mid q) = \big(F^{-1}(q) - y\big)\big(\mathbb{1}\{y \leq F^{-1}(q)\} - q\big), \tag{15}$$

which assesses the predicted quantile $F^{-1}(q)$ for a probability level $q \in (0, 1)$.

When the predictive distribution F is given by a finite ensemble $\{x_1, \ldots, x_K\}$, the CRPS representation in (11) is equal to

$$\text{CRPS}(F, y) = \frac{1}{K}\sum_{k=1}^{K} |x_i - y| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} |x_k - x_l|, \tag{16}$$

see Grimit et al. (2006). For small ensembles, Ferro et al. (2008) propose a *fair* approximation given by

$$\text{CRPS}(F, y) \approx \frac{1}{K}\sum_{k=1}^{K} |x_i - y| - \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{l=1}^{K} |x_k - x_l|. \tag{17}$$

For large ensembles, a more computationally efficient calculation is based on the generalized quantile function (Laio and Tamea, 2007). Let $x_{(1)} \leq \cdots \leq x_{(K)}$ denote the order statistics of $x_1, \ldots, x_K$. Then

$$\text{CRPS}(F, y) = \frac{2}{K^2}\sum_{i=1}^{K} \big(x_{(i)} - y\big)\Big(K\,\mathbb{1}\{y < x_{(i)}\} - i + \frac{1}{2}\Big), \tag{18}$$

see also Murphy (1970). The formula in (18) is implemented in the `R` package `scoringRules` together with exact formulas for a large class of parametric families of distributions, see Table 1 and Jordan et al. (2017).

When the forecasting model is estimated using Bayesian analysis, the predictive distribution F is commonly given by the posterior predictive distribution under the model. Here, F is rarely know in closed form and is, instead, approximated by a large sample which is often obtained using Markov chain Monte Carlo techniques. However, such techniques may yield highly correlated samples which complicates the employment of approximation formulas as those for the CRPS above. Optimal approximations for both IGN and CRPS when the distribution F is the posterior predictive distribution from a Bayesian analysis are discussed in Krüger et al. (2016).

Table 1. Parametric families of distributions for which the CRPS is implemented in the `R` package `scoringRules` (Jordan et al., 2017). The truncated families can be defined with or without a point mass at the support boundaries.

Dist. on $\mathbb{R}$	Dist. on $\mathbb{R}_{>0}$	Dist. on intervals	Discrete dist.
Gaussian	Exponential	Generalized extreme value	Poisson
t	Gamma	Generalized Pareto	Neg. binomial
Logistic	Log-Gaussian	Trunc. Gaussian	
Laplace	Log-logistic	Trunc. t	
Two-piece Gaussian	Log-Laplace	Trunc. logistic	
Two-piece exponential		Trunc. exponential	
Mixture of Gaussians		Uniform	
		Beta	

Scoring rules derived from scoring functions

The quality of a deterministic forecast x is typically assessed by applying a *scoring function* $s(x, y)$, that assigns a numerical score based on x and the corresponding observation y. As in the case of proper scoring rules, competing forecasting methods are compared and ranked in terms of the mean score over the cases in a test set. Popular scoring functions include the squared error, $s(x, y) = (x - y)^2$, and the absolute error, $s(x, y) = |x - y|$.

A scoring function can be applied to a probabilistic prediction $F \in \mathcal{F}$ if it is *consistent* for a functional T relative to the class $\mathcal{F}$ in the sense that

$$\mathbb{E}_F s(T(F), Y) \leq \mathbb{E}_F s(x, Y) \tag{19}$$

for all $x \in \Omega$ and $F \in \mathcal{F}$. A consistent scoring function becomes a proper scoring rule if the functional T in (19) is used as the derived deterministic prediction based on F. That is, if $S(F, y) = s(T(F), y)$. The squared error proper scoring rule is given by

$$\mathrm{SE}(F, y) = (\mathrm{mean}(F) - y)^2, \tag{20}$$

where $\mathrm{mean}(F)$ denotes the mean value of F, and the absolute error proper scoring rule becomes

$$\mathrm{AE}(F, y) = |\mathrm{med}(F) - y|, \tag{21}$$

where $\mathrm{med}(F)$ denotes the median of F.

One appealing property of scoring rules that derive from scoring functions is thus the possibility to compare deterministic and probabilistic forecasts. See Gneiting (2011) for an extensive discussion of the use of scoring functions to evaluate probabilistic predictions.

Simulation study: Comparing univariate scoring rules

The purpose of this simulation study is to demonstrate a coherent approach to using proper scores and rank or PIT histograms in practice, while highlighting some of the

difficulties that might arise when working with limited data sets. In particular, we investigate how different scoring rules rank forecasts according to their skill, and how this differs with the amount of available data.

We start by generating two sets of observation data, drawn randomly from the same fixed "true" distribution. The first set consists of 100 values, which will serve as verifying observations, while the second set, the training data, consists of 300 values for each of the 100 observations. Our goal is to issue forecasts matching the observations, based on the information contained in the training data. For the first part of the simulation study, the true distribution is normal, with a random mean $\mu \sim \mathcal{N}(25, 1)$ and fixed standard deviation $\sigma = 3$. In the second part, the truth is a Gumbel distribution, with the mean following a $\mathcal{N}(25, 1)$ distribution and the scale parameter fixed to 3, see Table 2.

Table 2. Observation-generating distributions used in the simulation study. The expected values are random variables following a normal distribution, while the scale parameters are fixed. γ denotes the Euler-Mascheroni constant.

	Distribution $F(Y)$		$\mathbb{E}(Y)$	$\text{Var}(Y)$
Part 1	Normal	$\mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(25, 1)$	$\sigma^2 = 9$
Part 2	Gumbel	$\text{G}(\mu, \sigma)$	$\mu + \sigma \cdot \gamma \sim \mathcal{N}(25, 1)$	$\frac{\pi^2}{6}\sigma^2 = \frac{3\pi^2}{2}$

Using a method-of-moments approach, we estimate four competing forecast distributions per observation, which are listed in Table 3. The distribution parameters are calculated by plugging the sample mean and sample standard deviation from the training data into the equations for mean and variance. For the non-central t-distribution, the degrees of freedom are obtained numerically by a root-finding algorithm described in Brent (1973), while restricting them to $\nu \geq 3$, ensuring that both mean and variance exist. As a fifth forecaster, we use the true distribution, from which the observations are generated. An ensemble of 50 members is drawn randomly from each of the forecast distributions, and is then paired with the observations.

Table 3. Forecasters used in both parts of the simulation study, their expected values and variances as functions of the distribution parameters. γ denotes the Euler-Mascheroni constant.

Distribution $F(Y)$		$\mathbb{E}(Y)$	$\text{Var}(Y)$
Normal	$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2
Non-central t	$\text{t}(\nu, \mu)$	$\mu\sqrt{\frac{\nu}{2}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$, if $\nu > 1$	$\frac{\nu(1+\mu^2)}{\nu-2} - \frac{\mu^2\nu}{2}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2$, if $\nu > 2$
Log-normal	$\ln\mathcal{N}(\mu, \sigma^2)$	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$	$\left(\exp\left(\sigma^2\right) - 1\right)\exp\left(2\mu + \sigma^2\right)$
Gumbel	$\text{G}(\mu, \sigma)$	$\mu + \sigma \cdot \gamma$	$\frac{\pi^2}{6}\sigma^2$

The performance of the five forecasters is evaluated using the absolute error, the squared error (both Section 3.1.2), the ignorance score, the CRPS (both Section 3.1.1) and the PIT

histogram (Section 2.1). We also produced rank histograms, but they turned out to be almost identical to the PIT histograms. As we encountered variations in the scores depending on the initial random seed, the whole process is repeated 10 times with different initial seeds, so that the final number of forecast-observation pairs comes to 1000.

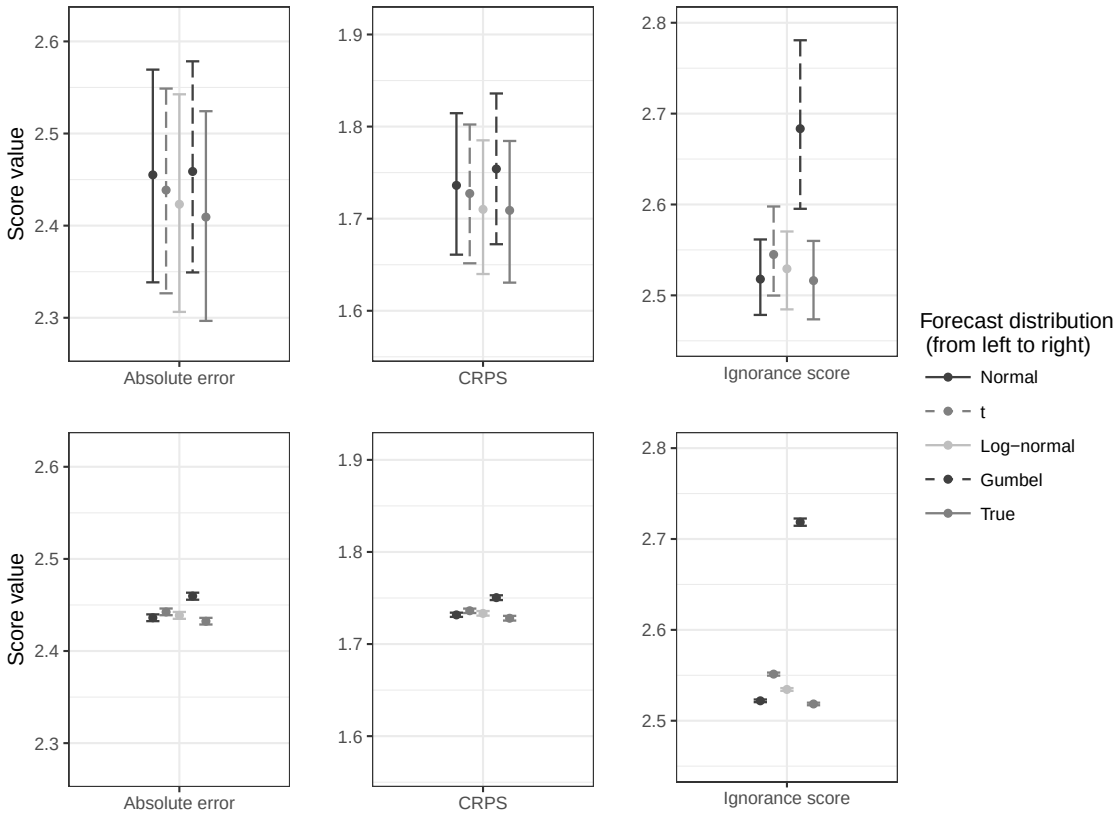

Figure 3. Top row: Mean absolute error, CRPS and ignorance score, and the 95% bootstrap confidence interval for the five forecast distributions, if the true distribution is normal. Scores are based on 1000 forecast-observation pairs. Bottom row: Same as above, but scores are based on 1 million forecast-observation pairs.

In order to understand the true ranking of the five forecasting methods in terms of skill, we reproduce the simulation study with 10 times 100 000 forecasts. For the case of a normal true distribution, Figure 3 shows the mean absolute error, mean CRPS and mean ignorance score, along with a 95% bootstrap confidence interval (see Section 3.7) computed from 1000 bootstrap samples. We have omitted the squared error from this plot, as its values are on a much larger scale than the other scores. Looking at the results for the small sample size in the top row, all scores assign the lowest mean value, and therefore the highest skill, to the normal distribution with the true parameters. However, if no knowledge about the true distribution was available, as it would be in a real forecast setting, the absolute error and the CRPS would prefer the log-normal distribution over all other forecasters, while the ignorance score judges the normal distribution with estimated parameters to be the best.

The bottom panel of Figure 3 shows the results from running the same study with the

larger sample size, and therefore the order in which we would expect the forecasters to rank. Here, all scores rightly find the Gumbel distribution, which has a completely different shape and tail behavior than the truth, to be the worst forecast, and the two forecasts based on normal distributions to be the best. This is, however, a contradiction to the results we got from the top panel, where only the ignorance score managed to rank the forecasters in the same order as we would expect.

Due to assigning large penalties to outliers, the ignorance score is here able to discriminate between the shapes of the forecast distributions and shows a significant difference at the 95% level between the Gumbel and the normal, log-normal and true distributions. The relatively poor performance of the non-central t-distribution can probably be explained by the fact that, while this distribution approximates a normal distribution if the degrees of freedom are large, the asymptotic distribution will have a standard deviation of 1, which doesn't match the given standard deviation of 3 in this example.

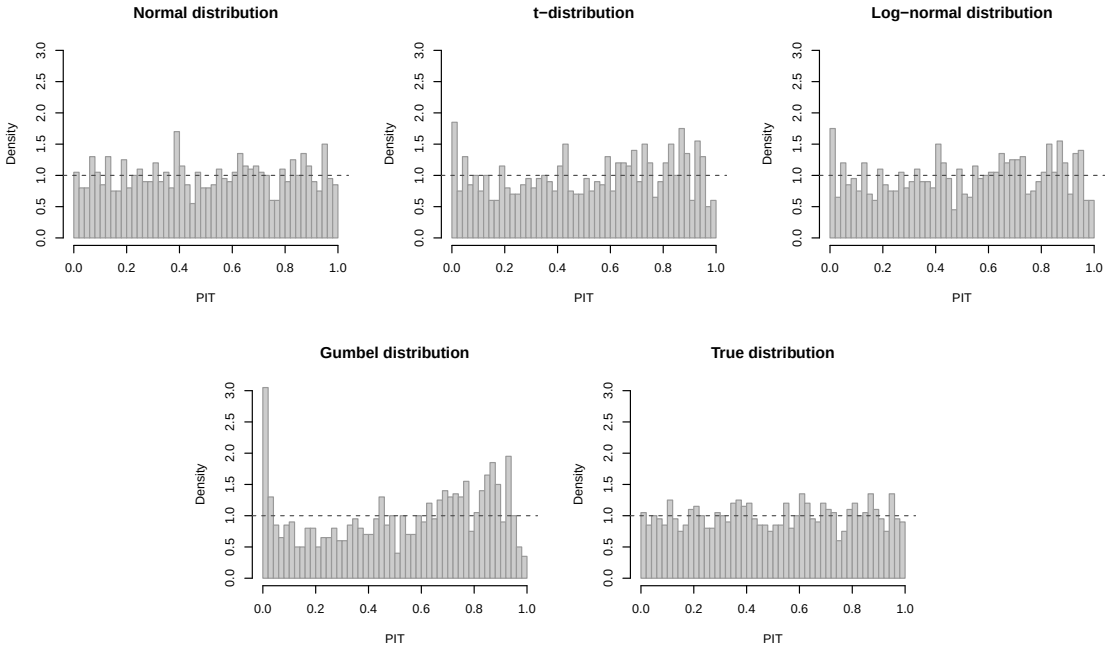

Figure 4. PIT histograms of the five forecast distributions, if the true distribution is normal, based on 1000 forecast-observation pairs.

Judging from Figure 4, which shows PIT histograms for the small-sample study with a normal true distribution, we can not make any statements about the forecast ranking, except that the Gumbel distribution forecast is clearly uncalibrated. Only when looking at the large sample equivalent in Figure 5, we see that the normal and the true forecasters are the only ones not suffering from miscalibration. A formal chi-square test (see Section 3.7) rejects the assumption of uniformity for the Gumbel distribution and even the t- and log-normal distributions (at a level of 95%) in the small sample case, and for all distributions apart from the true one in the large sample case.

Figure 6 illustrates one example forecast, for which the scores are plotted as functions of the verifying observation, in this case a sample value from a $\mathcal{N}(27.16, 9)$ distribution.

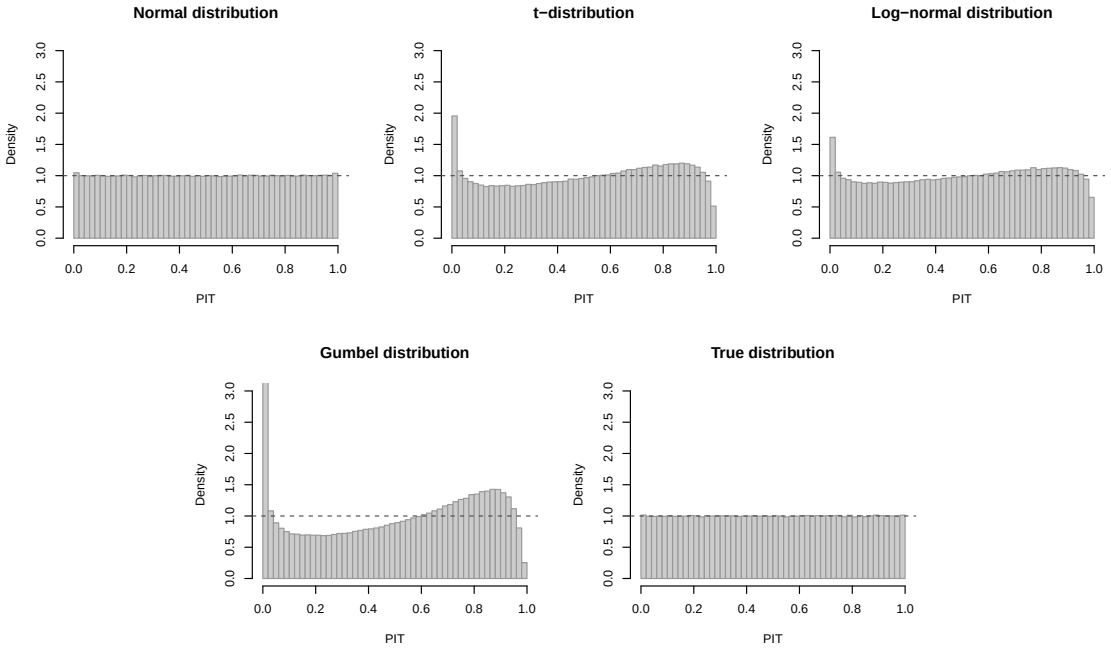

Figure 5. PIT histograms of the five forecast distributions, if the true distribution is normal, based on 1 million forecast-observation pairs

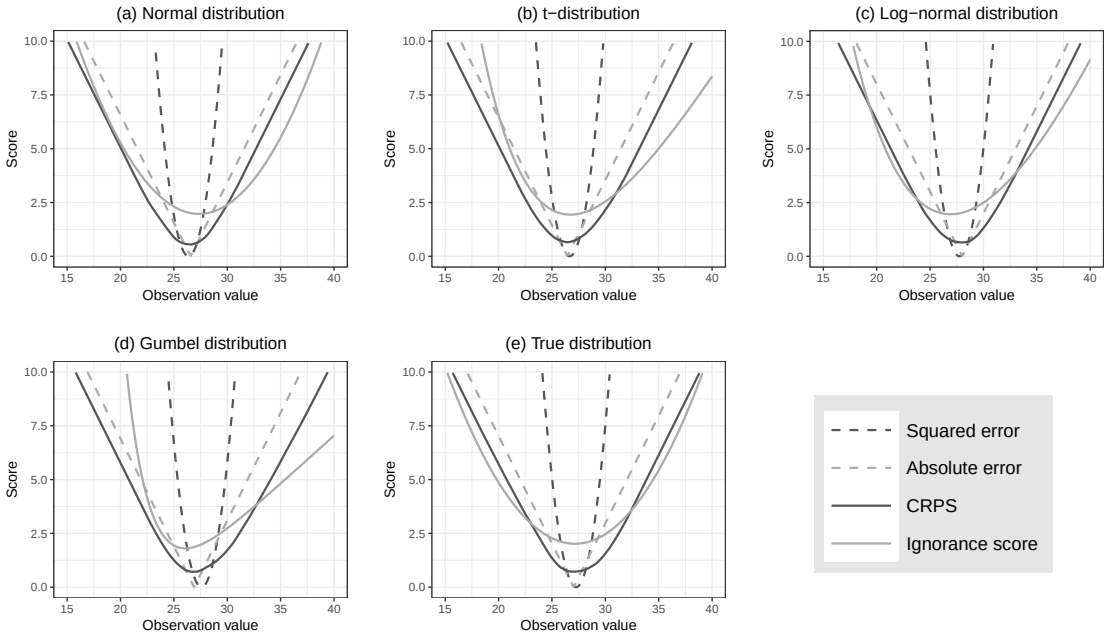

Figure 6. Squared error, absolute error, CRPS and ignorance score as a function of the verifying observation, for one forecast case in the simulation study: (a) normal distribution forecast, (b) non-central t-distribution forecast, (c) log-normal distribution forecast, (d) Gumbel distribution forecast, and (e) forecast based on the true normal distribution

While the score minima largely coincide for the true and the t-distribution, it becomes clear from the shape of the ignorance score, why it is much better at identifying the Gumbel distribution as an inferior forecaster than the other scores: because of the non-symmetry, forecasts will receive a much higher penalty, if the observation lies left of the distribution mode than if it lies on the right.

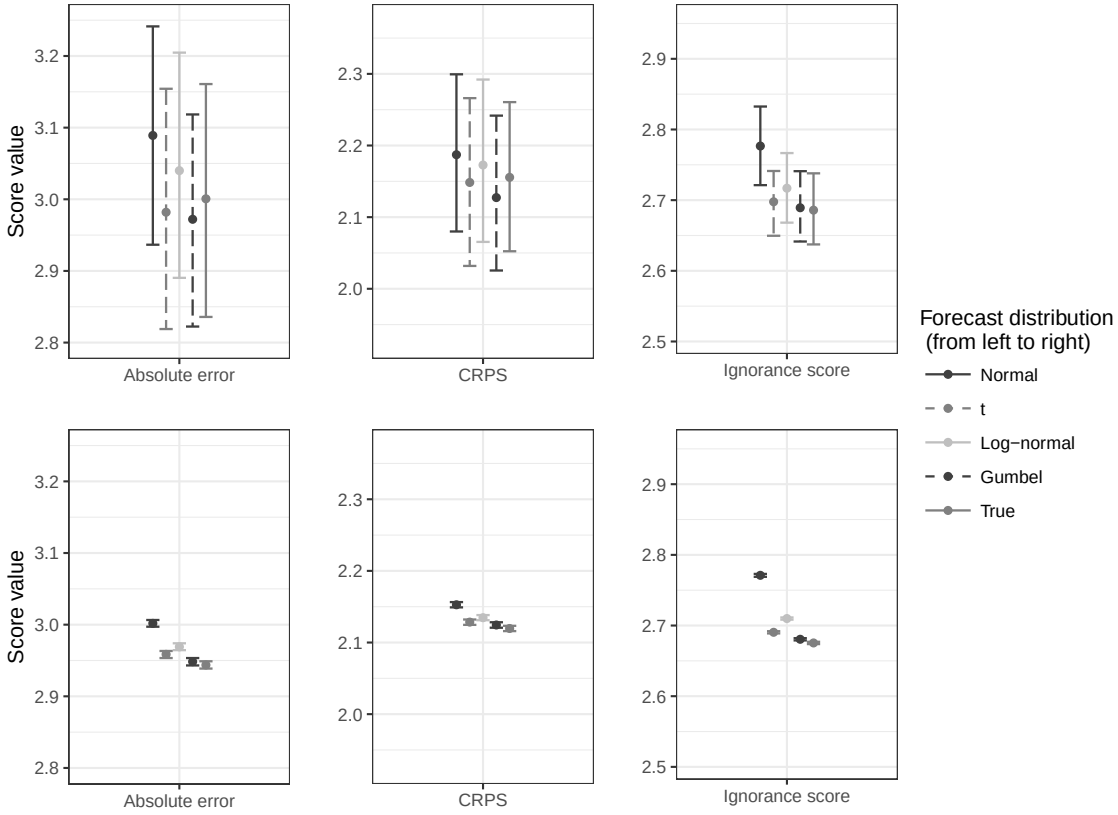

Figure 7. Top row: Mean absolute error, CRPS and ignorance score, and the 95% bootstrap confidence interval for the five forecast distributions, if the true distribution is a Gumbel distribution. Scores are based on 1000 forecast-observation pairs. Bottom row: Same as above, but scores are based on 1 million forecast-observation pairs.

For the second part of the simulation study, we used a Gumbel distribution as truth, where the mean is distributed with $\mathcal{N}(25, 1)$ and the scale parameter is 3. The same kind of forecasts are produced again: normal, non-central t, log-normal and Gumbel distributions based on the sample mean and variance of the training data. In Figure 7, the outcome of the study is shown for a small sample size (top row) and a very large sample size (bottom row). As previously, all scores agree on the forecast ranking, when the sample is large. The Gumbel distribution with estimated parameters and the true Gumbel distribution are assigned the lowest scores, while the normal forecaster now has the lowest skill.

However, the rankings look different in the top panel, where the true distribution is only ranked the third best by the absolute error and the CRPS, behind the estimated Gumbel and non-central t-distributions. The ignorance score again is the only score able to repro-

duce the forecast ranking we expect from the bottom panel. This is of course concerning and hints at the fact that even for a data set of apparently sufficient size, like the 1000 50-member ensembles used here, scores don't necessarily provide robust and proper results.

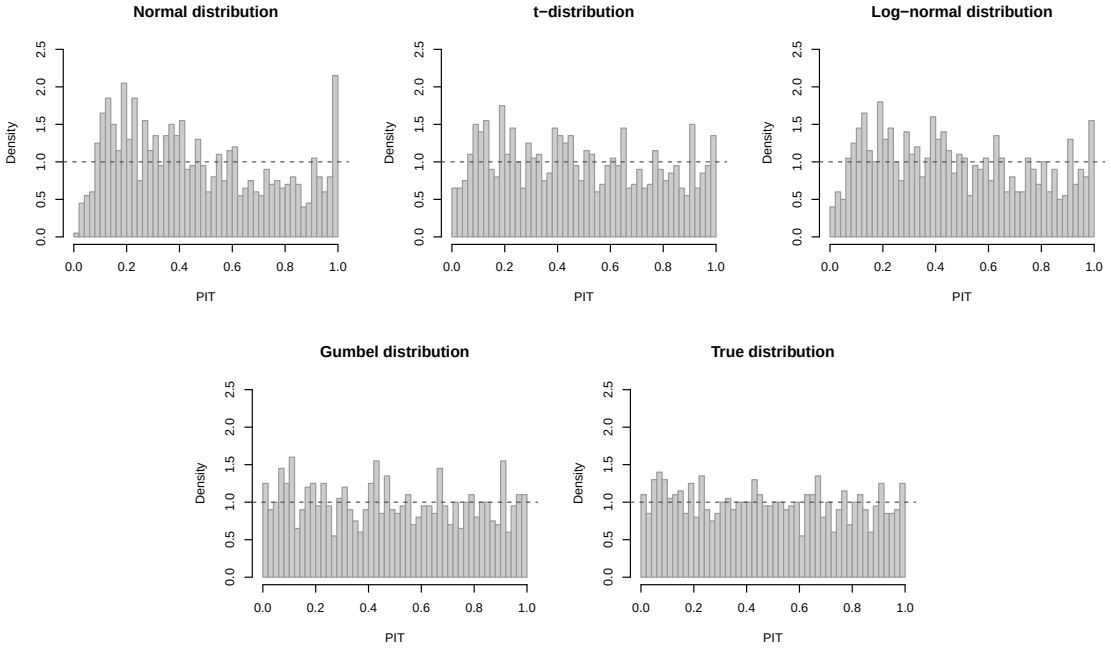

Figure 8. PIT histograms of the five forecast distributions, if the true distribution is a Gumbel distribution, based on 1000 forecast-observation pairs.

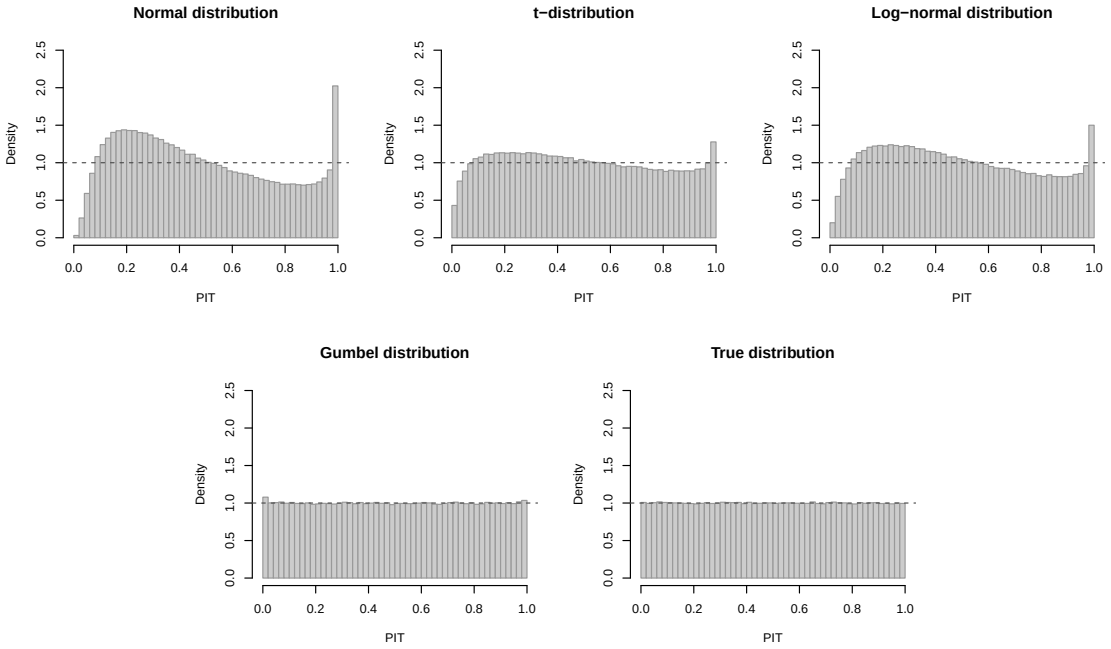

Figure 9. PIT histograms of the five forecast distributions, if the true distribution is a Gumbel distribution, based on 1 million forecast-observation pairs.

Like in the first part, we can not really judge the degree of forecast calibration by just

looking at the PIT histograms in Figure 8, except for the clearly uncalibrated normal distribution. A case could be made that the histogram for the true distribution looks slightly flatter than the other ones, but not with great certainty. It becomes clear, however, from Figure 9, that the forecasts based on non-central t and log-normal distributions also suffer from multiple types of miscalibration. These findings are confirmed by a chi-square test, which rejects the uniformity hypothesis for all except the Gumbel distributions in Figure 8 and all except the true distribution in Figure 9.

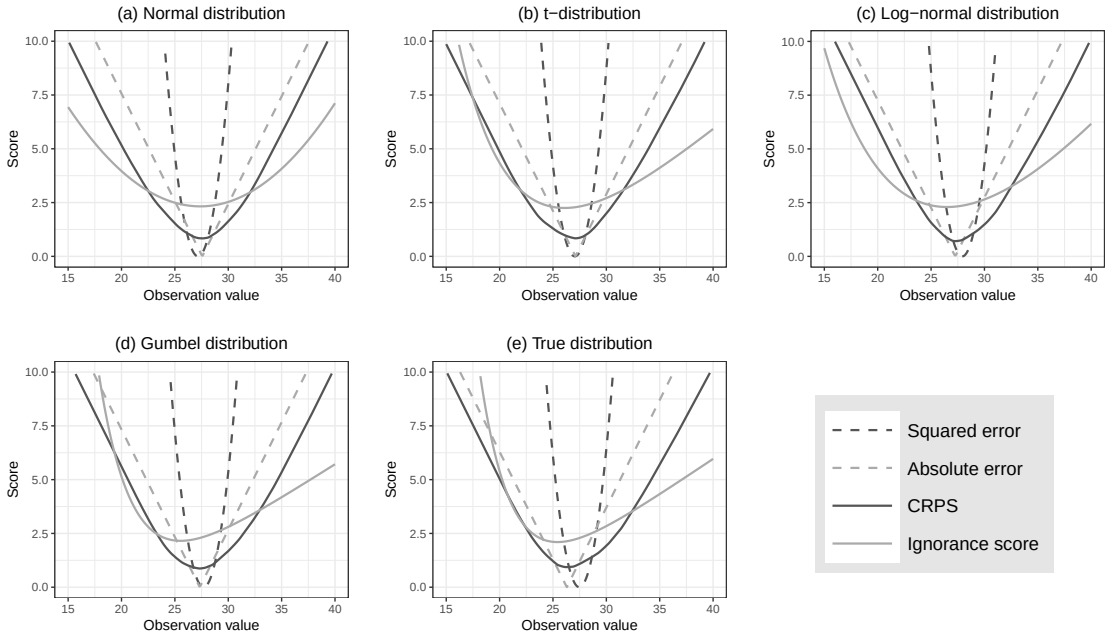

Figure 10. Squared error, absolute error, CRPS and ignorance score as a function of the verifying observation, for one forecast case in the simulation study: (a) normal distribution forecast, (b) non-central t-distribution forecast, (c) log-normal distribution forecast, (d) Gumbel distribution forecast, and (e) forecast based on the true Gumbel distribution

Picking an example forecast from the data set, as in Figure 10, shows that the ignorance score for the two Gumbel distribution forecasters is again non-symmetric, and therefore minimizes at a different value compared to the CRPS. In general, the ignorance score takes its minimum value at the mode of the distribution, and the CRPS at the median.

We can gather from this simulation study that even proper scores can behave very differently, depending on the size of the underlying data set, and aren't necessarily able to rank competing forecasters according to their actual skill. Therefore, we suggest to always use a combination of scoring rules to get a maximum amount of information about the performance of a particular model or forecaster. The ignorance score is more sensitive to the shape of a distribution and thus suitable to check if a chosen distribution actually fits the data. The CRPS is very useful for comparing models when the forecasts don't take the form of a standard probability distribution, or if for a given data set such a distribution can not be perfectly specified.

This also has implications for the ongoing discussion whether to use maximum likeli-

hood methods or minimize the CRPS to estimate model parameters (Gneiting et al., 2005), in that there might not be a definitive answer. Depending on the forecast situation and model choice, it could be preferable to switch between the two approaches. A case can be made for performing a thorough exploratory analysis of the data at hand before fitting any distributions, so as to find one that matches the data best. If it is difficult to select one distribution over the other, the simpler model should be preferred.

In all circumstances, the ranking of forecasters should not be solely based on the mean score, even if the sample size seems to be sufficiently large, but confidence intervals should be given, e.g. by applying bootstrapping techniques. We found that even for 1 million data points, differences between the forecast scores were often not significant at the 95% level.

Assessing extreme events

Forecasts specifically aimed at predicting extreme events can be assessed in a standard manner *e.g.* by using the scoring rules discussed in Sections 3.1.1 and 3.1.2 above (Friederichs and Thorarinsdottir, 2012). However, the restriction of conventional forecast evaluation to subsets of extreme observations by selecting the extreme observations after-the-fact, while discarding the non-extreme ones, and to proceed with standard evaluation tools will invalidate their theoretical properties and encourage hedging strategies (Lerch et al., 2017).

Specifically, Gneiting and Ranjan (2011) show that a proper scoring rule S is rendered improper if the product with a non-constant weight function w is formed, where w depends on the observed value y. That is, consider the weighted scoring rule

$$S_0(F, y) = w(y)S(F, y). \tag{22}$$

Then if Y has density g, the expected score $\mathbb{E}_g S_0(F, Y)$ is minimized by the predictive distribution F with density

$$f(y) = \frac{w(y)g(y)}{\int w(z)g(z)\mathrm{d}z}, \tag{23}$$

which is proportional to the product of the weight function w and the true density g. In particular, if $w(y) = \mathbb{1}\{y \geq u\}$ for some high threshold value u, then S_0 corresponds to evaluating F only on observed values exceeding u under the scoring rule S.

Instead, one can apply proper *weighted scoring rules* that are tailored to emphasize specific regions of interest. Diks et al. (2011) propose two weighted versions of the ignorance score that correct for the result in (23). The *conditional likelihood* (CL) score is given by

$$\mathrm{CL}(F, y) = -w(y) \log\Big(\frac{f(y)}{\int_\Omega w(z)f(z)\mathrm{d}z}\Big)$$

and the *censored likelihood* (CSL) score is defined as

$$\mathrm{CSL}(F, y) = -w(y) \log f(y) - \big(1 - w(y)\big) \log\Big(1 - \int_\Omega w(z)f(z)\mathrm{d}z\Big).$$

Here, w is a weight function such that $0 \leq w(y) \leq 1$ and $\int w(y)f(y)\mathrm{d}y > 0$ for all potential predictive distributions $F \in \mathcal{F}$. When $w(y) \equiv 1$, both the CL and the CSL score reduce to the unweighted ignorance score in (9).

Gneiting and Ranjan (2011) propose the *threshold-weighted continuous ranked probability score* (twCRPS), defined as

$$\text{twCRPS}(F, y) = \int_\Omega w(z)\big(F(z) - \mathbb{1}\{y \leq z\}\big)^2 \mathrm{d}z,$$

where, again, w is a non-negative weight function, see also Matheson and Winkler (1976). When $w(y) \equiv 1$, the twCRPS reduces to the unweighted CRPS in (12) while $w(y) = \mathbb{1}\{y = u\}$ equals the Brier score in (14). More generally, the twCRPS puts emphasis on a particular part of the forecast distribution F as specified by w. For focusing on the upper tail of F, Gneiting and Ranjan (2011) consider both indicator weight functions of the type $w(y) = \mathbb{1}\{y \geq u\}$ and non-vanishing weight functions such as $w(y) = \Phi(y|u, \sigma^2)$ where Φ denotes the cumulative distribution function of the Gaussian distribution with mean u and variance σ^2. Corresponding weight functions for the lower tail of F are given by $w(y) = \mathbb{1}\{y \leq u\}$ and $w(y) = 1 - \Phi(y|u, \sigma^2)$ for some low threshold value u.

Non-stationarity in the mean climate *e.g.* due to spatial heterogeneity may render it difficult to define a common threshold value u over a large number of forecast cases. Here, it may be more natural to define a weight function in quantile space using the CRPS representation in (13),

$$\text{twCRPS}(F, y) = \int_0^1 w(\tau)\big(F^{-1}(\tau) - y\big)\big(\mathbb{1}\{y \leq F^{-1}(\tau)\} - \tau\big)\mathrm{d}\tau,$$

where w is a non-negative weight function on the unit interval (Gneiting and Ranjan, 2011; Matheson and Winkler, 1976). Setting $w(\tau) \equiv 1$ retrieves the unweighted CRPS in (13) while this definition of twCRPS with $w(\tau) = \mathbb{1}\{\tau = q\}$ equals the quantile score in (15). Examples of more general weight functions for this setting include $w(\tau) = \mathbb{1}\{\tau \geq q\}$ and $w(\tau) = \tau^2$ for the upper tail, and $w(\tau) = \mathbb{1}\{\tau \leq q\}$ and $w(\tau) = (1 - \tau)^2$ for the lower tail with appropriate threshold values q, see also Gneiting and Ranjan (2011).

Lerch et al. (2017) find that there are limited benefits in using weighted scoring rules compared to using standard, unweighted scoring rules when testing for equal predictive performance. However, the application of weight functions as described here may facilitate interpretation of the forecast skill.

Example: proper and non-proper verification of extremes

In the following, we illustrate that the use of non-proper methods to verify and compare competing forecasts for extremes can lead to a distortion of the results and possibly false inference. Taking the same setting as the first part of the simulation study in Section 3.2, we generate sets of observation and training data from a normal distribution with standard deviation 3 and the mean a random value from a $\mathcal{N}(25, 1)$ distribution.

Four of the forecasting methods in Section 3.2 are compared: a normal distribution with estimated parameters based on the training data, a Gumbel distribution with estimated

parameters, a normal distribution with the true parameters and a Gumbel distribution with the true means as location parameter and scale parameter $\sigma = 3$. The forecasters' performance for extremes, which we consider to be values greater or equal to the 97.5% quantile of the observations u, will be measured using the threshold-weighted CRPS with three different weight functions and the unweighted CRPS, where the cases are restricted to observations above the threshold. The weight functions considered are variations on the indicator function:

$$
\begin{aligned}
w_1(y) &= \mathbb{1}\{y \geq u\} \\
w_2(y) &= 1 + \mathbb{1}\{y \geq u\} \\
w_3(y) &= 1 + \mathbb{1}\{y \geq u\} \cdot u
\end{aligned}
$$

Mean scores and 95% confidence intervals, calculated by numerical integration based on the small sample data set from Section 3.2, are shown in Figure 11 for the threshold-weighted CRPS and the CRPS with restricted observations along with the unweighted CRPS. The results for the twCRPS with weight function w_1 are omitted, as they are equal to zero for all forecasters.

However, just by adding 1 to the indicator function, we obtain meaningful scores with weight function w_2, showing the Gumbel distribution with fixed parameters to be the least skillful forecast, while the two normal distribution forecasters are of significantly better quality. The twCRPS with weight function w_3 and the unweighted CRPS lead to similar conclusions, although the differences between the scores are sometimes not significant. In contradiction to the other scores, the CRPS based on the restricted data set clearly shows the Gumbel distribution with fixed parameters to be the preferred forecaster.

Although its parameters and shape are obviously wrong, this is no surprise, as this distribution was purposely chosen because it has a heavy tail. Figure 12 pictures predictive densities for one example from the data set. If we restrict the evaluation to the area above the chosen threshold, represented by the black vertical line, the Gumbel distribution with fixed parameters is indeed the supposedly best forecast, as it assigns the highest probabilities to extreme values. The two normal distributions and the Gumbel distribution with estimated parameters, which tries to approximate the true normal distribution, have a very similar tail behavior, explaining their similar performance in terms of all scores.

We come to the same conclusion as Lerch et al. (2017), that conditioning a data set on extremal observations can result in preferring a forecaster who predicts extremes with inflated probabilities. When evaluating forecasts for a certain range of values, proper methods like the threshold-weighted CRPS should be used, where the whole data set is considered.

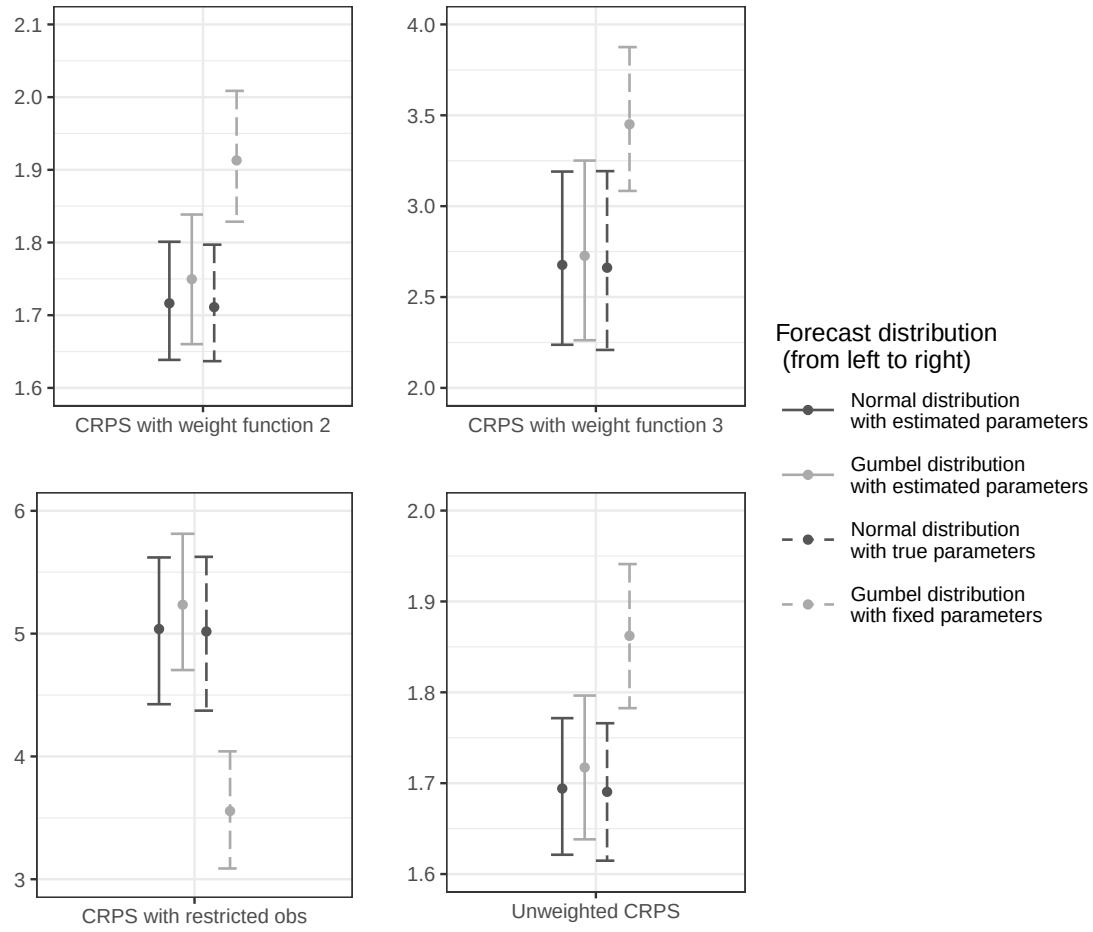

Figure 11. Mean scores and 95% bootstrap confidence interval for the four version of the CRPS. Top row: twCRPS with weight functions w_2 and w_3. Bottom row: CRPS restricted on observations above the threshold u and unweighted CRPS.

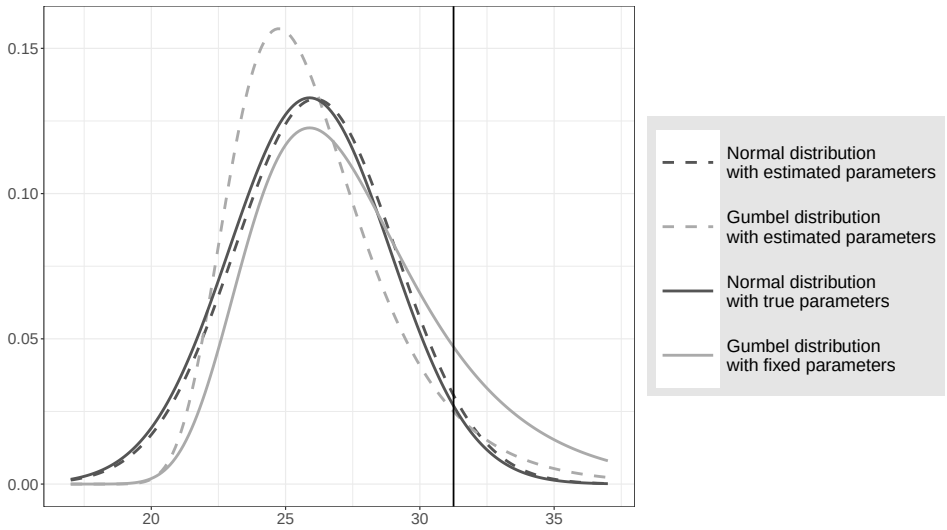

Figure 12. Example of the predictive densities given by the four competing forecasters. The black vertical line shows the threshold u, above which observations are considered to be extreme.

Multivariate assessment

Two general approaches can be employed to assess multivariate forecasts with scoring rules: Use specialized multivariate scores or reduce the multivariate forecast to a univariate quantity and, subsequently apply the univariate scores discussed above. For the latter approach, the appropriate univariate quantities depend on the context. Multivariate forecasts of single weather quantities are usually in the form of temporal trajectories, spatial fields or space-time fields. Here, it can, for instance, be useful to assess the predictive performance of derived quantities such as maxima, minima and accumulated totals, all of which depend on accurate modelling of both marginal and higher order structures, see *e.g.* Feldmann et al. (2015) for an assessment of spatial forecast fields for temperature.

Scores that directly assess multivariate forecasts are rather scarce and, as noted by Gneiting and Katzfuss (2014), there is a need to develop further decision-theoretically principled methods for multivariate assessment. The univariate Dawid-Sebastiani score in (10) can be applied in a multivariate setting with

$$\mathrm{DS}(F, y) = \log \det \Sigma_F + (y - \mu_F)^\top \Sigma_F^{-1} (y - \mu_F), \tag{24}$$

where μ_F is the mean vector and Σ_F the covariance matrix of the predictive distribution with $\det \Sigma_F$ denoting the determinant of Σ_F (Dawid and Sebastiani, 1999). However, note that unless the sample size is much larger than the dimension of the multivariate quantity, sampling errors can effect the calculation of $\det \Sigma_F$ and Σ_F^{-1} (see e.g. Table 2 in Feldmann et al., 2015). Similarly, if the multivariate predictive density is available, the ignorance score in (9) can be employed (Roulston and Smith, 2002).

Gneiting and Raftery (2007) propose the *energy score* (ES) as a multivariate generalization of the CRPS. It is given by

$$\mathrm{ES}(F, y) = \mathbb{E}_F \|X - y\| - \frac{1}{2} \mathbb{E}_F \mathbb{E}_F \|X - X'\|, \tag{25}$$

where X and X' are two independent random vectors distributed according to F and $\|\cdot\|$ is the Euclidean norm. For ensemble forecasts the natural analogue of the formulas in (16) and (17) apply. If the multivariate observation space Ω^d consists of weather variables on varying scales, the margins should be standardized before computing the joint energy score for these variables (Schefzik et al., 2013). This can be done using the marginal means and standard deviations of the observations in the test set. The energy score has been developed with low-dimensional quantities in mind and it may lose discriminatory power in higher dimensions (Pinson, 2013).

Scheuerer and Hamill (2015) propose a multivariate scoring rule that considers pairwise differences of the components of the multivariate quantity. In its general form, the *variogram score (VS) of order* p is given by

$$\mathrm{VS}_p(F, y) = \sum_{i=1}^{d} \sum_{j=1}^{d} \omega_{ij} \left(|y_i - y_j|^p - \mathbb{E}_F |X_i - X_j|^p \right)^2, \tag{26}$$

where y_i and y_j are the ith and the jth component of the observation, X_i and X_j are the ith and the jth component of a random vector X that is distributed according to F, and

ω_{ij} are nonnegative weights. Scheuerer and Hamill (2015) compare different choices of the order p and find that the best results in terms of discriminative power are obtained with $p = 0.5$. Furthermore, they recommend using weights proportional to the inverse distance between the components unless a prior knowledge regarding the correlation structure is available.

A comparison of the three multivariate scores in (24)-(26) is provided in Scheuerer and Hamill (2015). The authors conclude by recommending the use of multiple scores as they complement each other in their strengths and weaknesses. The variogram score is generally able to distinquish between correct and misspecified correlation structures, it has certain limitations resulting from the fact that it is proper but not strictly proper. Some of these limitations can be addressed by also using the energy score which is more sensitive to misspecifications in the predictive mean and less affected by finite representations of the predictive distribution. While the latter is an issue for the Dawid-Sebastiani score, it performs well for continuous predictive distributions, in particular for multivariate Gaussian models (Wei et al., 2017).

Divergence functions

In some cases, in particular in climate modelling, it is of interest to compare the predictive distribution F against the true distribution of the observations which is commonly approximated by the *empirical distribution function* of the available observations $y_1, \dots, y_n$,

$$\widehat{G}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i \leq x\}. \tag{27}$$

The two distributions, F and $\widehat{G}_n$, can be compared using a *divergence*

$$D : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0} \tag{28}$$

where $D(F, F) = 0$.

Assume that the observations $y_1, \dots, y_n$ forming the empirical distribution function $\widehat{G}_n$ are independent with distribution $G \in \mathcal{F}$. A propriety condition for divergences corresponding to that for scoring rules (7) states that the divergence D is *n-proper* for a positive integer n if

$$\mathbb{E}_G D(G, \widehat{G}_n) \leq \mathbb{E}_G D(F, \widehat{G}_n) \tag{29}$$

and *asymptotically proper* if

$$\lim_{n\to\infty} \mathbb{E}_G D(G, \widehat{G}_n) \leq \lim_{n\to\infty} \mathbb{E}_G D(F, \widehat{G}_n) \tag{30}$$

for all probability distributions $F, G \in \mathcal{F}$ (Thorarinsdottir et al., 2013). While the condition in (30) is fulfilled by a large class of divergences, only score divergences have been shown to fulfill (29) for all integers n. A divergence D is a *score divergence* if there exists a proper scoring rule S such that $D(F, G) = \mathbb{E}_G S(F, Y) - \mathbb{E}_G S(G, Y)$.

A score divergence that assesses the full distributions is the *integrated quadratic divergence* (IQD)

$$\mathrm{IQD}(F, G) = \int_{-\infty}^{+\infty} \big(F(x) - G(x)\big)^2 \mathrm{d}x, \tag{31}$$

the score divergence of the continuous ranked probability score (12). Alternative score divergences that assess specific properties of the predictive distribution include the *mean value divergence* (MVD),

$$\text{MVD}(F, G) = \big(\text{mean}(F) - \text{mean}(G)\big)^2, \tag{32}$$

the divergence associated with the squared error scoring rule (20), and the *Brier divergence* (BD) associated with the Brier score (14),

$$\text{BD}(F, G \mid u) = \big(G(u) - F(u)\big)^2, \tag{33}$$

for some threshold u.

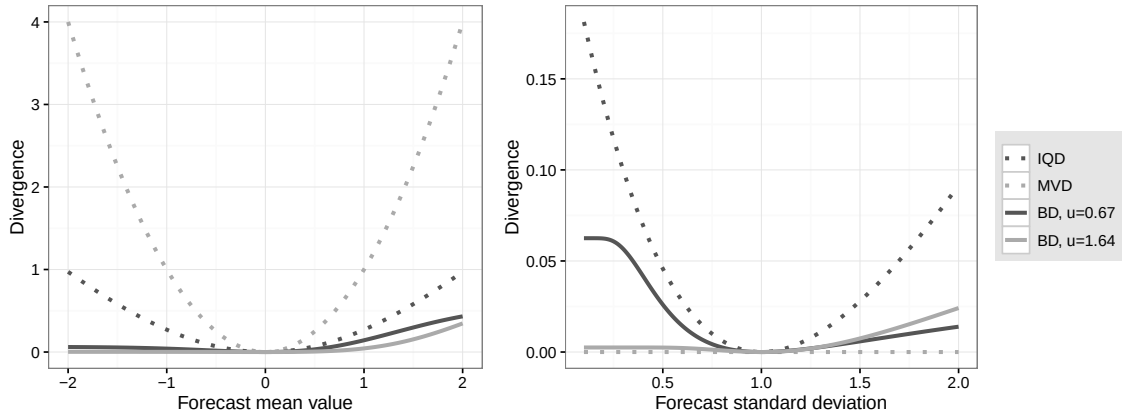

Figure 13. Comparison of expected score divergence values for a standard normal observation distribution and normal forecast distributions with varying mean values (left) or standard deviations (right).

Figure 13 provides a comparison of the score divergences in (31)-(33) for two simple settings where the observation distribution is given by a standard normal distribution and all the forecast distributions are also normal distributions but with varying parameters. In the left plot, the variance is correctly specified while the forecast mean value varies. In the right plot, the forecast mean values equal that of the observation distribution while the standard deviation varies. We compare the IQD, the MVD and the BD with thresholds $u = 0.67$ and $u = 1.64$ which equal the 75% and the 95% quantile of the observation distribution, respectively. The divergences are more sensitive to forecast errors in the mean than the spread. In particular, the MVD is, naturally, not able to detect errors in the forecast spread. Furthermore, integrating over the BD for all possible thresholds u and obtaining the IQD yields a better discrimination than investigating the differences for individual quantiles. The right plot also shows that the model ranking obtained under the BD highly depends on the threshold u.

While every proper scoring rule is associated with a score divergence, not all score divergences are practical for use in the setting where the empirical distribution function $\widehat{G}_n$ is used. One example is the Kullback-Leibler divergence, the score divergence of the ignorance score in (9). The Kullback-Leibler divergence becomes ill-defined if the forecast distribution F has positive mass anywhere where the observation distribution G has

mass zero. When G is replaced by $\widehat{G}_n$ and, especially, if the sample size n is relatively small, such issues might occur. One option to circumvent the issue is to treat the data as categorical and bin it in b bins prior to the evalution. That is, identify the probability distribution F with a probability vector $(f_1, \ldots, f_b)$ and, similarly, G with a probability vector $(g_1, \ldots, g_b)$. The *Kullback-Leibler divergence* is then given by

$$\mathrm{KLD}(F, G) = \sum_{i=1}^{b} f_i \log \frac{f_i}{g_i},$$

see also the discussion in Thorarinsdottir et al. (2013).

Historically, much of the forecast evaluation literature has focused on the evaluation of probabilistic forecasts against deterministic observations and an in-depth discussion of optimal theoretical and/or practical properties of divergences is lacking. Applied studies commonly employ divergences that are asymptotically proper rather than n-proper for all positive integer n, see e.g. Palmer (2012) and Perkins et al. (2007).

Testing equal predictive performance

As demonstrated in the simulation study in Section 3.2, the estimation of the mean score over a test set may be associated with a large uncertainty. A simple bootstrapping procedure over the individual scores may be used to assess the uncertainty in the mean score, see e.g. Friederichs and Thorarinsdottir (2012). Assume we have n score values $S(F_1, y_1), \ldots, S(F_n, y_n)$. By repeatedly resampling vectors of length n (with replacement) and calculating the mean of each sample, we obtain an estimate of the variability in the mean score. Note that some care is needed if the forecast errors, and thus the resulting scores, are correlated. A comprehensive overview over bootstrapping methods for dependent data is given in Lahiri (2003).

Formal statistical tests can be applied to test equal predictive performance of two competing methods under a proper scoring rule. The most commonly applied test is the *Diebold-Mariano test* (Diebold and Mariano, 1995) which applies in the time series setting. Consider two competing forecasting methods F and G that for each time step $t = 1, \ldots, n$ issue forecasts F_t and G_t, respectively, for an observation y_{t+k} that lies k time steps ahead. The mean scores under a scoring rule S are given by

$$\bar{S}_n^F = \frac{1}{n} \sum_{t=1}^{n} S(F_t, y_{t+k}) \qquad \text{and} \qquad \bar{S}_n^G = \frac{1}{n} \sum_{t=1}^{n} S(G_t, y_{t+k}).$$

The Diebold-Mariano test uses the test statistic

$$t_n = \sqrt{n} \frac{\bar{S}_n^F - \bar{S}_n^G}{\hat{\sigma}_n}, \tag{34}$$

where $\hat{\sigma}_n^2$ is an estimator of the asymptotic variance of the score difference. Under the null hypothesis of equal predictive performance and standard regularity conditions, the test statistic t_n in (34) is asymptotically standard normal (Diebold and Mariano, 1995). When the null hypothesis is rejected in a two-sided test, F is preferred if t_n is negative and G is preferred if t_n is positive.

Diebold and Mariano (1995) note that for ideal k-step-ahead forecasts, the forecast errors are at most $(k-1)$-dependent. An estimator for the asymptotic variance $\hat{\sigma}_n^2$ based on this assumption is given by

$$\hat{\sigma}_n^2 = \begin{cases} \hat{\gamma}_0 & \text{if } k = 1, \\ \hat{\gamma}_0 + 2\sum_{j=1}^{k-1} \hat{\gamma}_j, & \text{if } k \geq 2, \end{cases} \tag{35}$$

where $\hat{\gamma}_j$ denotes the lag j sample autocorrelation of the sequence $\{S(F_i, y_{i+k}) - S(G_i, y_{i+k})\}_{i=1}^n$ for $j = 0, 1, 2, \ldots$ (Gneiting and Ranjan, 2011). Alternative estimators are discussed in Diks et al. (2011) and Lerch et al. (2017).

In the spatial setting, Hering and Genton (2011) propose the *spatial prediction comparison test* which accounts for spatial correlation in the score values without imposing assumptions on the underlying data or the resulting score differential field. This test is implemented in the `R` package `SpatialVx` (Gilleland, 2017). Weighted scoring rules and their connection to hypothesis testing are discussed in Holzmann and Klar (2017).

A simple test for the uniformity of a rank or PIT histogram is the chi-square test. It tests if the histogram values can be considered samples from a uniform distribution and therefore if any deviations of uniformity are random or systematic (Wilks, 2004, 2011). The chi-square statistic based on n cases and K ensemble members is

$$\chi^2 = \sum_{i=1}^{K+1} \frac{(m_i - f)^2}{f}, \tag{36}$$

with m_i denoting the actual number of counts for bin i and $f = \frac{n}{K+1}$ the expected number of counts for a uniform distribution. We can reject the null hypothesis of the histogram being uniform if this statistic exceeds the quantile of the chi-squared distribution with K degrees of freedom at the chosen level of significance.

In its general form, however, the chi-square test only applies to independent data, which is not the case in many forecast settings due to e.g. temporal or spatial correlation between forecast data points. Some methods to address this effect are proposed in Wilks (2004). If the goal is to not only test for uniformity, but also for the other deficiencies in calibration shown in Section 2.1, Elmore (2005) and Jolliffe and Primo (2008) present alternatives which are more flexible and appropriate. Wei et al. (2017) propose calibration tests for multivariate Gaussian forecasts based on the Dawid-Sebastiani score in (24).

Understanding model performance

When assessing the performance of an individual model, e.g. to identify weaknesses and test potential improvements, it might be useful to look at tools which don't necessarily follow the principles of propriety described in Section 3. For instance, it can be useful to investigate the forecast bias to better understand the potential sources of forecast errors even if competing forecasting models should not be ranked based on mean bias as it is not

a proper score (Gneiting and Raftery, 2007). Here, we discuss a few tools which may be used to provide a better understanding of the performance of an individual forecasting model while it is not recommended to rank competing forecasters based on these tools.

One of the most popular measures used by national weather services is the anomaly correlation coefficient (ACC) , a valuable tool to track the gain in forecast skill over time (Jolliffe and Stephenson, 2012). The ACC quantifies the correlation between forecast anomalies and the anomalies of the observation, typically an analysis. Anomalies are defined as the difference between the forecast or analysis and the climatology for a given time and location. Usually, the climatology is based on the model climate, calculated from the range of values predicted by the NWP model over a long time period.

For a deterministic forecast f_i, valid at time i, with a corresponding analysis a_i and climate statistic c_i, there are two equivalent definitions for the anomaly correlation coefficient (e.g. Miyakoda et al., 1972):

$$\begin{aligned}\mathrm{ACC} &= \frac{\sum_{i=1}^{N}(f_i - c_i)\cdot(a_i - c_i) - \sum_{i=1}^{N}(f_i - c_i)\cdot\sum_{i=1}^{N}(a_i - c_i)}{\sqrt{\sum_{i=1}^{N}(f_i - c_i)^2 - \left(\sum_{i=1}^{N}(f_i - c_i)\right)^2}\cdot\sqrt{\sum_{i=1}^{N}(a_i - c_i)^2 - \left(\sum_{i=1}^{N}(a_i - c_i)\right)^2}} \\ &= \frac{\sum_{i=1}^{N}\left(f_i' - \bar{f}'\right)\left(a_i' - \bar{a}'\right)}{\sqrt{\sum_{i=1}^{N}\left(f_i' - \bar{f}'\right)^2\sum_{i=1}^{N}\left(a_i' - \bar{a}'\right)^2}}\end{aligned}$$

Here, $f_i' = f_i - c_i$ is the forecast anomaly and $a_i' = a_i - c_i$ the anomaly of the analysis, with respective sums $\bar{f}' = \sum_{i=1}^{N}(f_i - c_i)$ and $\bar{a}' = \sum_{i=1}^{N}(a_i - c_i)$. The ACC is a preferred evaluation tool for gridded forecasts and spatial fields, as these are usually compared against an analysis or a similar gridded observation product.

However, there are certain limits and pitfalls one has to be aware of when using this measure. Due to it being a correlation coefficient, the ACC doesn't give any information about forecast biases and errors in scale, so that it can overestimate the forecast skill (Murphy and Epstein, 1989). As such, it should always be used in conjunction with an estimate of the actual bias, or applied to previously bias-corrected data.

It has been established empirically that an anomaly correlation of 0.6 corresponds to a limit in usefulness for a medium-range forecast. Murphy and Epstein (1989) warn, however, that this is rather an upper limit of the actual skill and that the ACC should be seen as a measure of potential skill. Naturally, the ACC relies to a large extent on the underlying climatology used to compute the anomalies.

When evaluating forecast skill with proper scores, it is often useful to get separate indicators for the degree of calibration and the sharpness of the forecast. The well-known and widely used decomposition of the Brier score by Murphy (1973b) separates the score value in three parts, quantifying reliability, resolution and uncertainty.

Consider a forecast sample of size N, where probability forecasts $p_u = 1 - F(u)$ are computed for exceeding a threshold u and binary observations take the form $o = \mathbb{1}\{y \geq u\}$. If the forecasts take K unique values, with n_k denoting the number of forecasts within the category k and $p_{u,k}$ the probability forecast associated with category k, then the Brier score can be written as

$$\mathrm{BS}(F, y|u) = \frac{1}{N}\sum_{k=1}^{K} n_k \left(p_{u,k} - \bar{o}_k\right)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k \left(\bar{o}_k - \bar{o}\right)^2 + \bar{o}\left(1 - \bar{o}\right), \tag{37}$$

where $\bar{o}_k$ is the event frequency for each of the forecast values and $\bar{o} = \frac{1}{N}\sum_{i=1}^{N} o_i$ the climatological event frequency, computed from the sample. The first part of the sum in (37) relates to the reliability or calibration, the second, having a negative effect on the total score, to the resolution or sharpness, and the last part is the climatological uncertainty of the event.

This representation of the Brier score relies on the number of discrete forecast values K being relatively small. If p_u takes continuous values, care must be taken when binning the forecast into categories, so as not to introduce biases (Bröcker, 2008; Stephenson et al., 2008). Several analogue decompositions have been proposed for other scores, such as the CRPS (Hersbach, 2000), the quantile score (Bentzien and Friederichs, 2014) and the ignorance score (Weijs et al., 2010). Recently, Siegert (2017) formulated a general framework allowing for the decomposition of arbitrary scores.

While it is common and advisable to look at a model's performance in certain weather situations or for certain periods of time, it is important to be aware of Simpson's paradox (Simpson, 1951). It describes the phenomenon that a certain effect appearing in several subsamples can not be found in a combination of these samples, or that the larger sample may even show the complete opposite effect.

For example, a forecast model can have superior skill over all four seasons, compared to another model, but still be worse when assessed over the whole year. Hamill and Juras (2006) showed this to be true for a synthetic data set of temperature forecasts on two islands. In this case, the climatologies of the two islands were so different, that the values of performance measures were misleadingly improved. Fricker et al. (2013) found that this spurious skill doesn't affect proper scores derived from scoring rules, but care should be taken when using scores derived from a contingency table which are not proper, and skill scores in general.

In general, it is recommended to use statistical significance testing in order to evaluate potential model improvements. Differences in scores are often very small and it is hard to judge if they are caused by genuine improvement or chaotic error growth. Geer (2016) investigate a version of the Student's *t*-test modified for multiple models and taking account of autocorrelation in the scores. They also found that in order to detect an improvement of 0.5%, at least 400 forecast fields on a global grid would be required. This confirms our findings from Section 3.2 that it is essential to make careful considerations

of the experiment sample size in order to generate meaningful and robust results.

Summary

In this chapter, a variety of methods to assess different aspects of forecast goodness were presented and discussed. Calibration errors can be diagnosed with the help of histograms, in both univariate and multivariate settings. It is recommended to use multiple such diagnostics, especially in the multivariate case, as different tools highlight different types of miscalibration.

Scoring rules provide information about the accuracy of a forecast and are valuable tools for comparing forecasting methods. In this context, only proper scores should be used, as they ensure that the forecast based on the best knowledge will receive the best score. There are many such scores available, with the CRPS and the ignorance score being amongst the most popular. However, we found that just looking at the mean of one such score can be misleading, even if the underlying sample seems to be of sufficient size. Therefore it is crucial to also provide information about the error of the mean score, and to base decisions about model preference on the evaulation of multiple scoring rules, if possible. If we don't want to compare models, but rather understand the behaviour of a model, it can be helpful to use measures which are not necessarily proper. Especially skill scores and the anomaly correlation coefficient are widely used.

By adding appropriate weight functions to the CRPS and the ignorance score, it is possible to evaluate extreme event forecasts in a proper way. These weight functions can be designed to emphasize e.g. different parts of the climatological distribution. Scores for multivariate quantities not only give information about the calibration and sharpness of the forecast, but also assess the correct representation of the covariance structure between locations, forecast times or variables. However, some of them have limitations and don't work well if the number of dimensions is large.

Given the multitude of available evaluation tools and scores, constantly growing due to new research and applications, it is essential to be aware of their properties and how to choose a suitable measure. To make sure that all aspects of a forecast's performance are addressed, a number of scores should be calculated and a quantification of the associated uncertainty given.

References

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate 9*, 1518–1530. 6

Bentzien, S. and P. Friederichs (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society 140*(683), 1924–1934.

31

Bigelow, F. H. (1905). Application of mathematics in meteorology. *Monthly Weather Review 33*(3), 90–90. 4

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, New Jersey. 14

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3. 12

Bröcker, J. (2008). Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review 136*(11), 4488–4502. 31

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A 147*, 278–292. 5

Dawid, A. P. and P. Sebastiani (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics 27*, 65–81. 11, 25

Delle Monache, L., J. P. Hacker, Z. Y., D. X., and S. R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysicl Research: Atmospheres 111*, D24307. 6

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263. 28

Diks, C., V. Panchenko, and D. Van Dijk (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics 163*(2), 215–230. 21, 29

Elmore, K. L. (2005). Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and Forecasting 20*(5), 789–795. 29

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Monthly Weather Review 143*(3), 955–971. 25

Ferro, C. A., D. S. Richardson, and A. P. Weigel (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications 15*(1), 19–24. 12

Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications 20*(2), 246–255. 31

Friederichs, P. and A. Hense (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review 135*(6), 2365–2378. 12

Friederichs, P. and T. L. Thorarinsdottir (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics 23*(7), 579–594. 21, 28

Geer, A. J. (2016). Significance of changes in medium-range forecast scores. *Tellus Ser. A 68*(1), 30229. 31

Gilleland, E. (2017). *SpatialVx: Spatial Forecast Verification*. R package version 0.6-1. 29

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association 106*(494), 746–762. 10, 13

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B 69*, 243–268. 4, 5

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*, 125–151. 25

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378. 11, 12, 25, 30

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*, 1098–1118. 21

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*(3), 411–422. 11, 21, 22, 29

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782. 5

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008). Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test 17*, 211–264. 6, 7, 8

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Ser. B 14*, 107–114. 11

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson (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. 12

Hamill, T. M. and S. J. Colucci (1997). Verification of Eta-RSM Short-Range Ensemble Forecasts. *Monthly Weather Review 125*(6), 1312–1327. 6

Hamill, T. M. and J. Juras (2006). Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society 132*(621C), 2905–2923. 31

Hering, A. S. and M. G. Genton (2011). Comparing spatial predictions. *Technometrics 53*(4), 414–425. 29

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*, 559–570. 11, 31

Holzmann, H. and B. Klar (2017). Weighted scoring rules and hypothesis testing. arXiv:1611.07345v2. 29

Jolliffe, I. T. and C. Primo (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review 136*(6), 2133–2139. 29

Jolliffe, I. T. and D. B. Stephenson (Eds.) (2012). *Forecast verification: A practitioner's guide in atmospheric science*. Chichester, UK: John Wiley & Sons. 4, 30

Jordan, A., F. Krüger, and S. Lerch (2017). Evaluating probabilistic forecasts with the R package scoringRules. arXiv:1709.04743. 12, 13

Krüger, F., S. Lerch, T. L. Thorarinsdottir, and T. Gneiting (2016). Probabilistic forecasting and comparative model assessment based on markov chain monte carlo output. arXiv:1608.06802. 12

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society 7*, 48–50. 7

Lahiri, S. N. (2003). *Resampling methods for dependent data*. New York, NY, USA: Springer. 28

Laio, F. and S. Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences Discussions 11*(4), 1267–1277. 11, 12

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science 32*(1), 106–127. 21, 22, 23, 29

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics 18*(1), 405–414. 7

López-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association 104*(486), 718–734. 7

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*, 1087–1096. 11, 22

Mirzargar, M. and J. L. Anderson (2017). On evaluation of ensemble forecast calibration using the concept of data depth. *Monthly Weather Review 145*(5), 1679–1690. 7, 8, 10

Miyakoda, K., G. D. Hembree, R. F. Strickler, and I. Shulman (1972). Cumulative results of extended forecast experiments I. Model perfomance for winter cases. *Monthly Weather Review 100*(12), 836–855. 30

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review 98*(12), 917–924. 12

Murphy, A. H. (1973a). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology 12*(1), 215–223. 11

Murphy, A. H. (1973b). A new vector partition of the probability score. *Journal of Applied Meteorology 12*(4), 595–600. 30

Murphy, A. H. (1974). A sample skill score for probability forecasts. *Monthly Weather Review 102*(1), 48–55. 11

Murphy, A. H. (1992). Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and forecasting 7*(4), 692–698. 11

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and forecasting 8*(2), 281–293. 4, 10, 11

Murphy, A. H. and E. S. Epstein (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review 117*(3), 572–582. 30

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner (2017). *vegan: Community Ecology Package*. R package version 2.4-2. 7

Palmer, T. N. (2012). Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society 138*, 841–861. 28

Perkins, S., A. Pitman, N. Holbrook, and J. McAneney (2007). Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate 20*(17), 4356–4376. 28

Pinson, P. (2013). Wind energy: Forecasting challenges for its operational management. *Statistical Science 28*(4), 564–585. 25

Pinson, P. and R. Girard (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy 96*, 12–20. 8

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. 7

Roulston, M. S. and L. A. Smith (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review 130*(6), 1653–1660. 25

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science 28*(4), 616–640. 25

Scheuerer, M. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review 143*(4), 1321–1334. 25, 26

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society 143*(703), 1178–1183. 31

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Ser. B 13*(2), 238–241. 31

Smith, L. A. and J. A. Hansen (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree. *Monthly Weather Review 132*, 1522–1528. 7

Stephenson, D. B., C. A. S. Coelho, and I. T. Jolliffe (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting 23*(4), 752–757. 31

Strähl, C. and J. Ziegel (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics 11*(1), 608–639. 5

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification 1*(1), 522–534. 26, 28

Thorarinsdottir, T. L., M. Scheuerer, and C. Heinz (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics 25*(1), 105–122. 7, 8, 10

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. Available at http://ssrn.com/abstract=2236605. 5

Wei, W., F. Balabdaoui, and L. Held (2017). Calibration tests for multivariate Gaussian forecasts. *Journal of Multivariate Analysis 154*, 216–233. 26, 29

Weijs, S. V., R. van Nooijen, and N. van de Giesen (2010). Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review 138*(9), 3387–3399. 31

Wilks, D. S. (2004). The minimum spanning tree histogram as verification tool for multidimensional ensemble forecasts. *Montly Weather Review 132*, 1329–1340. 7, 29

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Oxford, UK: Elsevier Academic Press. 4, 29

Wilks, D. S. (2017). On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society 143*(702), 164–172. 6, 8

Ziegel, J. F. and T. Gneiting (2014). Copula calibration. *Electronic Journal of Statistics 8*(2), 2619–2638. 6

Index