Evaluating forecasts for high-impact events using transformed kernel scores

Sam Allen   David Ginsbourger   Johanna Ziegel

Institute of Mathematical Statistics and Actuarial Science

University of Bern

Bern, Switzerland

{sam.allen,david.ginsbourger,johanna.ziegel}@stat.unibe.ch

Abstract

It is informative to evaluate a forecaster's ability to predict outcomes that have a large impact on the forecast user. Although weighted scoring rules have become a well-established tool to achieve this, such scores have been studied almost exclusively in the univariate case, with interest typically placed on extreme events. However, a large impact may also result from events not considered to be extreme from a statistical perspective: the interaction of several moderate events could also generate a high impact. Compound weather events provide a good example of this. To assess forecasts made for high-impact events, this work extends existing results on weighted scoring rules by introducing weighted multivariate scores. To do so, we utilise kernel scores. We demonstrate that the threshold-weighted continuous ranked probability score (twCRPS), arguably the most well-known weighted scoring rule, is a kernel score. This result leads to a convenient representation of the twCRPS when the forecast is an ensemble, and also permits a generalisation that can be employed with alternative kernels, allowing us to introduce, for example, a threshold-weighted energy score and threshold-weighted variogram score. To illustrate the additional information that these weighted multivariate scoring rules provide, results are presented for a case study in which the weighted scores are used to evaluate daily precipitation accumulation forecasts, with particular interest on events that could lead to flooding.

1 Introduction

Just as important as issuing a forecast is understanding how it is expected to perform. In achieving this, forecasters gain a greater awareness of the strengths and limitations of their predictions, and, in turn, learn how they can be improved (Jolliffe and Stephenson, 2012). There are several aspects to consider when evaluating forecasts, but, intuitively, a 'good' forecaster is one whose predictions consistently agree with what materialises (Murphy, 1993). To assess to what extent this is satisfied, it is convenient to condense all information regarding forecast performance into a single numerical value, or score, thereby allowing competing forecast strategies to be objectively ranked and compared. For probabilistic forecasts, this can be achieved using scoring rules. Scoring rules are functions of the form

$$S \colon \mathcal{M} \times \mathcal{X} \to \mathbb{R} \cup \{-\infty, \infty\},$$

where $\mathcal{M}$ is a suitable class of probability measures over the measurable outcome space $(\mathcal{X}, \mathcal{A})$. Let $\mathfrak{M}$ denote the set of all probability measures on $(\mathcal{X}, \mathcal{A})$. Without loss of generality, we restrict attention to negatively oriented scoring rules, for which a lower score indicates a more accurate forecast. The score assigned to a forecast can therefore be interpreted as a loss.

It is widely accepted that scoring rules should be proper. A scoring rule S is proper with respect to $\mathcal{M} \subset \mathfrak{M}$ if

$$S(Q, Q) \leq S(P, Q) \qquad \text{for all } P, Q \in \mathcal{M}, \tag{1}$$

where $S(P,Q) = \mathbb{E}_Q[S(P,Y)]$ denotes the expectation of $S(P,Y)$ when $Y \sim Q$; it is assumed that $S(P,Q)$ exists for all $P, Q \in \mathcal{M}$, and that $S(Q,Q)$ is finite. That is, if the observations are believed to arise according to $Q \in \mathcal{M}$, then the expected value of a proper score is minimised by issuing Q as the forecast. If Q is the unique minimiser of the expected score, then the scoring rule is said to be strictly proper.

It is often of interest to evaluate a forecaster's ability to predict outcomes that have a large impact on the user, since improving the forecasts made for such outcomes may allow their impacts to be mitigated. However, Gneiting and Ranjan (2011) demonstrate that using a proper scoring rule to evaluate the predictions made when particular outcomes occur is equivalent to assessing the forecaster using an improper score, which can thus result in unreliable conclusions regarding the performance of competing forecasters. In the context of extreme events, Lerch et al. (2017) term this the forecaster's dilemma. To circumvent the forecaster's dilemma, it has become common to employ weighted scoring rules that can direct the evaluation of forecasts to certain outcomes in a theoretically sound way.

The concept of weighted scoring rules dates back at least to Matheson and Winkler (1976), though most developments in the field have occurred over the past decade. Gneiting and Ranjan (2011), for example, introduce two weighted versions of the continuous ranked probability score (CRPS), while Diks et al. (2011) propose two adaptations of the logarithmic score. Holzmann and Klar (2017) generalise the approach followed by Diks et al. (2011) and present a framework capable of constructing weighted versions of any proper scoring rule. In order to focus evaluation on particular outcomes, in this paper, a weight function is a measurable function from the sample space $\mathcal{X}$ to $[0,1]$. Alternative weight functions have also been considered in the literature (see e.g. Gneiting and Ranjan, 2011), but they are not of interest here.

Weighted scoring rules have been studied in most detail in the univariate setting, with interest often placed on extreme events, defined as instances where the outcome exceeds a chosen threshold. Often, however, a high-impact results from the interaction of several moderate events, none of which are extreme from a statistical perspective. A good example of this is a compound weather event, whereby multiple weather hazards combine and interact to generate a high-impact event, despite none of the confounding hazards themselves necessarily being 'extreme' (e.g. Zscheischler et al., 2020). The present article seeks to develop tools that permit a targeted assessment of forecasts made for high-impact events such as these. In particular, we introduce weighted multivariate scoring rules that allow emphasis to be placed on regions of a multi-dimensional outcome space when evaluating the accuracy of a forecaster.

To achieve this, we utilise kernel scores, a general class of proper scoring rules based on conditionally negative definite (c.n.d.) kernels (Gneiting and Raftery, 2007; Dawid, 2007). Here, a negative definite kernel is a symmetric function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho(x_i, x_j) \leq 0 \tag{2}$$

for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $c_1, \ldots, c_n \in \mathbb{R}$. A kernel is c.n.d. if the above criterion is satisfied for all $c_1, \ldots, c_n \in \mathbb{R}$ that sum to zero, and is strictly negative definite if equality in Equation 2 holds only when $c_1 = 0, \ldots, c_n = 0$ for distinct $x_1, \ldots, x_n$. Conversely, a kernel is said to be positive definite if the inequality in Equation 2 is reversed. Hereafter, we assume wherever necessary that the kernels are measurable.

Definition 1. Given a c.n.d. kernel $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the *kernel score corresponding to* ρ is the scoring rule

$$S_\rho(P, y) = \mathbb{E}_P[\rho(X, y)] - \frac{1}{2}\mathbb{E}_P[\rho(X, X')] - \frac{1}{2}\rho(y, y), \tag{3}$$

where $X, X' \sim P \in \mathfrak{M}$ are independent, and it is assumed that all expectations are finite.

Several familiar scoring rules fall into this kernel score framework, including the Brier score (Brier, 1950), the CRPS, and the energy score (Gneiting and Raftery, 2007). In Section 4, we demonstrate that the variogram score proposed by Scheuerer and Hamill (2015b) is also a kernel score. The final term in Equation 3 does not depend on the forecast and is not present in previous definitions of kernel scores (Gneiting et al., 2007;

Steinwart and Ziegel, 2021). Nonetheless, it is included here since it generates a scoring rule that can be interpreted as a divergence between the forecast and a Dirac measure at the outcome, even if $\rho(y, y) \neq 0$.

The deployment of c.n.d. kernels within scoring rules follows from their interpretation as generalised distance measures (e.g. Schölkopf, 2001). However, the propriety of a kernel score depends on the choice of ρ. To see this, consider the divergence function $d(P, Q) = S(P, Q) - S(Q, Q)$ associated with the scoring rule S. It is immediate from Equation 1 that a scoring rule is proper with respect to $\mathcal{M}$ if and only if its divergence function is non-negative for all $P, Q \in \mathcal{M}$. The divergence function corresponding to a kernel score is

$$d_\rho(P, Q) = \mathbb{E}_{P,Q}\left[\rho(X, Y)\right] - \frac{1}{2}\mathbb{E}_P\left[\rho(X, X')\right] - \frac{1}{2}\mathbb{E}_Q\left[\rho(Y, Y')\right], \tag{4}$$

where $X, X' \sim P$ and $Y, Y' \sim Q$ are independent. That is, the score divergence between P and Q is proportional (by a factor of one half) to the energy distance with respect to ρ (Székely and Rizzo, 2013). Sejdinovic et al. (2013) show that energy distances are special cases of squared Maximum Mean Discrepancies (MMD; Gretton et al., 2007), and kernel score divergences can be interpreted as squared MMDs under suitable integrability conditions. This provides a natural connection between the optimum score estimation theory introduced by Gneiting and Raftery (2007) and machine learning algorithms that use the MMD as a loss function (e.g. Dziugaite et al., 2015; Li et al., 2015).

The following theorem summarises existing results on the propriety of kernel scores, and is obtained by merging results in Sejdinovic et al. (2013) and Steinwart and Ziegel (2021). Previously, a special case of this result was presented in Gneiting and Raftery (2007).

Theorem 1. *Let ρ be a c.n.d. kernel on $\mathcal{X}$. If $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, then ρ is non-negative and the kernel score S_ρ is proper with respect to*

$$\mathcal{M}_\rho = \{P \in \mathfrak{M} \mid \mathbb{E}_P[\rho(X, x_0)] < \infty \text{ for some } x_0 \in \mathcal{X}\}.$$

If ρ is negative definite, then the kernel score S_ρ is proper with respect to

$$\mathcal{M}^\rho = \{P \in \mathfrak{M} \mid \mathbb{E}_P\left[\sqrt{-\rho(X, X)}\right] < \infty\}.$$

For a c.n.d. kernel ρ, we will often state the assumption that S_ρ is proper with respect to $\mathcal{M}_\rho$ or $\mathcal{M}^\rho$. For concision, it is always assumed in the former case that $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, and in the latter case that ρ is negative definite.

The strict propriety of kernel scores relates to injectivity of kernel mean embeddings (Steinwart and Ziegel, 2021). If $\rho = -k$, with k being a positive definite and bounded kernel, then this is synonymous with the kernel being characteristic (Muandet et al., 2017). On the other hand, if the c.n.d. kernel ρ is a metric, then the criterion that Equation 4 is zero if and only if $P = Q$ is exactly the definition given by Lyons (2013) for ρ to be a metric of strong negative type. We will refer to specific results where necessary throughout the paper.

Since the theory underlying kernels is well-established, results from the extant literature can be leveraged in order to choose the most appropriate kernel when evaluating forecasts in particular scenarios. The kernel can be chosen to extract the information that is most relevant for the situation at hand, allowing prior information to be incorporated directly into forecast evaluation. Bolin and Wallin (2019), for example, propose altering the kernel used within the CRPS to reduce the score's sensitivity to outliers. We study the choice of kernel in the context of weighted scoring rules, with particular interest on forecasts made for high-impact events.

In the following section, we demonstrate that the threshold-weighted continuous ranked probability score (twCRPS) introduced by Gneiting and Ranjan (2011) is a kernel score. The twCRPS is arguably the most well-known weighted scoring rule, and this result leads to a convenient representation of the score when evaluating ensemble forecasts, i.e. finite samples of point forecasts. In addition, this permits a generalisation

of the twCRPS to so-called threshold-weighted kernel scores, which we introduce in Section 3. Furthermore, established results on kernels are leveraged in order to introduce further new approaches to weighting kernel scores. Due to the flexibility of kernels, these results significantly widen the range of situations in which weighted scoring rules can be applied, and we illustrate this in Section 4 by considering outcomes in multi-dimensional Euclidean space. We introduce weighted variogram scores and weighted energy scores, and also study a new scoring rule based on a bounded kernel. The utility of these weighted multivariate scoring rules when evaluating forecasts made for high-impact events is presented in a simulation study, as well as in a case study on flood forecasts in Section 5. A discussion of the results presented herein is available in Section 6, while all proofs are deferred to the appendix.

2 Weighted versions of the CRPS

2.1 Definitions and properties

Here, we consider the case where $\mathcal{M}$ is the set of Borel probability measures on $\mathcal{X} = \mathbb{R}$ with finite first moment, and identify elements of $\mathcal{M}$ with their associated distribution functions. A popular scoring rule used to assess forecasts in this setting is the continuous ranked probability score (CRPS), defined as

$$
\begin{aligned}
\mathrm{CRPS}(F, y) &= \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z, \\
&= 2 \int_{(0,1)} (\mathbb{1}\{F^{-1}(\alpha) \geq y\} - \alpha)(F^{-1}(\alpha) - y) \, \mathrm{d}\alpha, \\
&= \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_F |X - X'|,
\end{aligned} \tag{5}
$$

where X and X' are independent random variables with distribution function $F \in \mathcal{M}$, $y \in \mathbb{R}$ is the corresponding observation, and $\mathbb{1}$ denotes the indicator function. In the second expression, F^{-1} is the lower quantile function or generalised inverse of F.

The CRPS is strictly proper with respect to $\mathcal{M}$ (Gneiting and Raftery, 2007). Its three different representations also partly explain the score's popularity. The first expression demonstrates that the CRPS is equivalent to the Brier score integrated over all possible thresholds (Matheson and Winkler, 1976), whereas the second expression highlights that it can also be written as a quantile scoring rule integrated over all quantiles (Laio and Tamea, 2007). The final representation demonstrates that the CRPS is a kernel score, where the c.n.d. kernel is $\rho(x, x') = |x - x'|$ (Gneiting and Raftery, 2007).

Due to its popularity, weighted scoring rules have been studied in most detail using the CRPS, with the most well-known version being the threshold-weighted continuous ranked probability score (twCRPS):

$$
\mathrm{twCRPS}(F, y; \nu) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}\nu(z), \tag{6}
$$

where ν is a Borel measure on $\mathbb{R}$, often chosen so that it has density equal to a particular non-negative weight function, w (Matheson and Winkler, 1976; Gneiting and Ranjan, 2011). Although analytical expressions of the twCRPS have been derived for particular families of parametric distributions (e.g. Allen et al., 2021), the integral in Equation 6 is often evaluated using numerical techniques. The following proposition provides an alternative representation of the twCRPS as a kernel score, implying a straightforward approach to computing this integral when F is an empirical distribution function.

Proposition 1. *Let ν be a Borel measure on $\mathbb{R}$. Then, there exists an increasing function v on $\mathbb{R}$ such that the threshold-weighted CRPS associated with the measure ν is the kernel score corresponding to $\rho(x, x') = |v(x) - v(x')|$. In particular, v is any function such that $v(x) - v(x') = \nu([x', x))$ for all $x, x' \in \mathbb{R}$. For*

$F \in \mathcal{M}_\rho$, it holds that

$$\text{twCRPS}(F, y; \nu) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, d\nu(z),$$

$$= 2 \int_{(0,1)} (\mathbb{1}\{F^{-1}(\alpha) \geq y\} - \alpha)(v(F^{-1}(\alpha)) - v(y)) \, d\alpha, \qquad (7)$$

$$= \mathbb{E}_F |v(X) - v(y)| - \frac{1}{2} \mathbb{E}_F |v(X) - v(X')|,$$

where $X, X' \sim F$ are independent and $y \in \mathbb{R}$.

Equation 7 generalizes the three representations of the CRPS in Equation 5, which correspond to the case where ν is the Lebesgue measure and $v(z) = z$ (up to a constant) for all $z \in \mathbb{R}$. The final equality in Proposition 1 illustrates that the twCRPS can be interpreted as the CRPS after having deformed the forecasts and observations, where the deformation is governed by the choice of measure, or weight function. We refer to v as the chaining function.

Remark 1. If the measure ν in the threshold-weighted CRPS is chosen so that it has density w, then the chaining function v is any function such that

$$v(x) - v(x') = \int_{[x', x)} w(x) \, dx.$$

In light of this remark, possible deformations corresponding to some common weight functions are displayed in Figure 1, along with the resulting kernel to be employed in the twCRPS. Knowing the chaining function permits a greater appreciation of what the weight in the twCRPS achieves. For example, if weight is placed only on values above a certain threshold, $w(z) = \mathbb{1}\{z \geq t\}$, then the forecasts and observations are projected onto $[t, \infty)$, with values lower than the threshold mapped onto t, before calculating the unweighted CRPS.

Remark 2. The kernel score representation of the twCRPS readily extends to the integrated quadratic distance (IQD), the score divergence associated with the CRPS (Thorarinsdottir et al., 2013). A threshold-weighted version of the IQD is defined analogously to the twCRPS in Equation 6, and there exists a similar representation of this weighted divergence in terms of the kernel $\rho(x, x') = |v(x) - v(x')|$.

The second equality in Proposition 1 demonstrates that, like the CRPS, the twCRPS can also be expressed as the integral of a quantile scoring rule over all possible quantiles. Any scoring function that is consistent for the α-quantile can be written in the form $(\mathbb{1}\{x \geq y\} - \alpha)(v(x) - v(y))$, with v increasing (Gneiting, 2011), and we define a quantile scoring rule by replacing x with the α-quantile of a forecast distribution. Since any increasing function v satisfies $v(x) - v(x') = \nu([x', x))$ for some Borel measure ν, Proposition 1 shows that the integral of any scoring rule that is consistent for the α-quantile over all possible values of α results in a twCRPS. Gneiting and Ranjan (2011) use the quantile score representation of the CRPS to introduce a quantile-weighted version of the CRPS. In this case, the integral is over the quantiles of the forecast distribution, and the weight function emphasises particular regions of the forecast distribution, rather than regions of the outcome space. Since the threshold-weighted CRPS can also be expressed as the integral of a quantile scoring rule, it would be straightforward to introduce a weight function into this integral in an analogous way, thereby constructing a weighted version of the CRPS that emphasises certain regions of both the outcome space and the forecast distribution.

The threshold- and quantile-weighted versions of the CRPS were introduced to circumvent the fact that scaling a proper scoring rule using a weight governed by the outcomes results in an improper scoring rule (Gneiting and Ranjan, 2011). Let S denote a scoring rule that is proper with respect to $\mathcal{M}$, let w be a weight function, and let $G \in \mathcal{M}$ be a distribution function for which $\mathbb{E}_G[w(X)] > 0$. The expectation of the score S scaled by the weight function w, $\mathbb{E}_G[w(Y)S(F, Y)]$, is minimised by issuing a weighted version of G, rather than G itself. In particular, the expected scaled score is minimised by

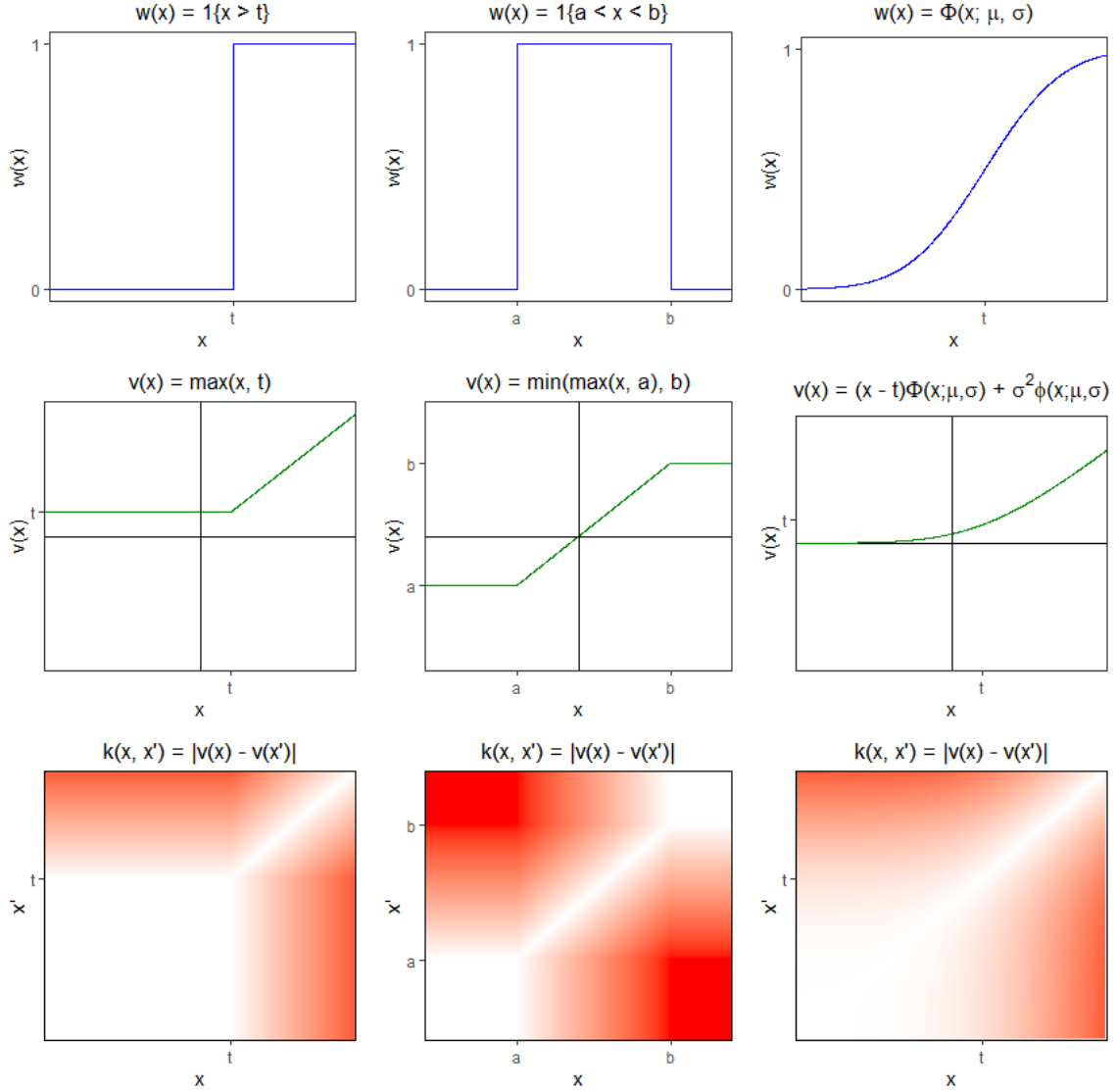$$G_w(x) = \frac{\mathbb{E}_G[\mathbb{1}\{X \leq x\}w(X)]}{\mathbb{E}_G[w(X)]}. \qquad (8)$$

Figure 1: Common weight functions (top row) with corresponding chaining functions (middle row) and the resulting kernels to be employed in the twCRPS (bottom row). $\Phi(x; \mu, \sigma)$ and $\phi(x; \mu, \sigma)$ represent the distribution and density functions, respectively, of a normal distribution with mean μ and scale σ. In the bottom row, a stronger shade of red reflects a larger value of the kernel.

Holzmann and Klar (2017) therefore suggest assessing forecast distributions through their weighted representation: if w is a weight function and S is a proper scoring rule with respect to $\{F_w | F \in \mathcal{M}, \mathbb{E}_F[w(X)] > 0\}$, with F_w defined as in Equation 8, then

$$\text{ow}S(F, y; w) = w(y)S(F_w, y) \tag{9}$$

defines a scoring rule that is proper with respect to $\{F \in \mathcal{M} | \mathbb{E}_F[w(X)] > 0\}$. Since the weighting in this case is directly dependent on the outcome y, we refer to scoring rules in this form as outcome-weighted scoring rules. For example, the outcome-weighted CRPS is defined as

$$\text{owCRPS}(F, y; w) = w(y) \int_{\mathbb{R}} (F_w(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z.$$

To appreciate how the outcome-weighted CRPS differs from the threshold-weighted CRPS, consider a weight

function of the form $w(z) = \mathbb{1}\{z \geq t\}$. The twCRPS with this weight function assesses to what extent the forecast can identify whether or not the observation will exceed the threshold t or any value larger than t. The owCRPS, on the other hand, is concerned with the forecast performance only when the observation y exceeds this threshold, in which case the forecast is assessed via its conditional distribution given that the threshold has been exceeded. Alternatively, this difference is made more explicit by noting that the outcome-weighted CRPS can be expressed in the following form.

Proposition 2. *Let w be a weight function. If $F \in \mathcal{M}$ such that $\mathbb{E}_F[w(X)] > 0$, then it holds that*

$$\text{owCRPS}(F, y; w) = \frac{1}{C_w(F)}\mathbb{E}_F\left[|X - y|w(X)w(y)\right] - \frac{1}{2C_w(F)^2}\mathbb{E}_F\left[|X - X'|w(X)w(X')w(y)\right], \quad (10)$$

where $X, X' \sim F$ are independent, $y \in \mathbb{R}$, and $C_w(F) = \mathbb{E}_F[w(X)]$.

Holzmann and Klar (2017) provide this expression for weight functions of the form $w(z) = \mathbb{1}\{z \geq t\}$. This representation of the outcome-weighted CRPS demonstrates that the weight function is applied to the output of the kernel, in contrast to the threshold-weighted version of the CRPS, which involves a prior transformation of the forecasts and outcome.

Note also that, unlike the twCRPS, the owCRPS is not a kernel score. To see this, consider a forecast that is a Dirac measure at $z \in \mathbb{R}$, $P = \delta_z$. From Equation 3, it is straightforward to verify that a kernel score S_ρ satisfies

$$S_\rho(\delta_z, y) = \rho(z, y) - \frac{1}{2}\rho(z, z) - \frac{1}{2}\rho(y, y) = S_\rho(\delta_y, z).$$

However, assuming $w(y), w(z) > 0$, $\text{owCRPS}(\delta_z, y; w) = |z - y|w(y)$, which is in general not equal to $\text{owCRPS}(\delta_y, z; w) = |z - y|w(z)$.

2.2 Localising scores

Holzmann and Klar (2017) show that outcome-weighted scoring rules exhibit some desirable properties when interest is on only a particular subset of outcomes. We recall their definitions of localising, and strictly or proportionally locally proper scoring rules.

Definition 2. Let $\mathcal{M}$ be a class of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, S a scoring rule, and w a weight function. The scoring rule S is called *localising with respect to w* if, for any $P, Q \in \mathcal{M}$, $P(\cdot \cap \{w > 0\}) = Q(\cdot \cap \{w > 0\})$ implies that $S(P, y) = S(Q, y)$ for all $y \in \mathcal{X}$. Here, $\{w > 0\} = \{x \in \mathcal{X}|w(x) > 0\}$.

That is, a weighted scoring rule is localising if it depends only on the forecast measure restricted to outcomes for which the weight function is positive. Clearly, however, if $\{w = 0\} = \{x \in \mathcal{X}|w(x) = 0\}$ is non-empty, then such a scoring rule will not be strictly proper with respect to typical choices of $\mathcal{M}$.

Definition 3. Let S be a scoring rule that is proper with respect to $\mathcal{M}$ and localising with respect to w. Then, S is called *strictly locally proper with respect to w and $\mathcal{M}$* if $S(P, Q) = S(Q, Q)$ implies $P(\cdot \cap \{w > 0\}) = Q(\cdot \cap \{w > 0\})$ for any $P, Q \in \mathcal{M}$. The scoring rule S is called *proportionally locally proper with respect to w and $\mathcal{M}$* if $S(P, Q) = S(Q, Q)$ holds if and only if $P(\cdot \cap \{w > 0\}) = cQ(\cdot \cap \{w > 0\})$ for some constant $c > 0$ that depends on P and Q, for any $P, Q \in \mathcal{M}$.

If S is a proper scoring rule with respect to $\{P_w|P \in \mathcal{M}, \mathbb{E}_P[w(X)] > 0\}$, Holzmann and Klar (2017, Theorem 3) prove that the outcome-weighted version of this score (constructed via Equation 9) will be proper and localising with respect to $\{P \in \mathcal{M}|\mathbb{E}_P[w(X)] > 0\}$, and if S is strictly proper, then the outcome-weighted score will additionally be proportionally locally proper. In order to obtain a strictly locally proper weighted scoring rule, the authors suggested complementing an outcome-weighted score S with a strictly proper scoring rule S_0 for probability forecasts of the occurrence of an arbitrary binary event, such as the logarithmic score or the Brier score:

$$\tilde{S}(P, y) = S(P, y) + \{w(y)S_0(C_w(P), 1) + (1 - w(y))S_0(C_w(P), 0)\}, \quad (11)$$

where $C_w(P)$ is as defined in Proposition 2, and is interpreted here as a probability forecast. If S is proportionally locally proper, then $\tilde{S}$ will be strictly locally proper (Holzmann and Klar, 2017, Theorem 3). To understand why this holds, note that proportionally locally proper scores only evaluate the shape of the restriction of P to $\{w > 0\}$, whereas the binary score included in Equation 11 allows $\tilde{S}$ to additionally assess the measure that is assigned to the sets $\{w = 0\}$ and $\{w > 0\}$.

While outcome-weighted scoring rules are localising by construction, the twCRPS is only localising with respect to particular weight functions. The twCRPS is not localising with respect to weights that are the indicator function of a compact interval over the real line, $w(z) = \mathbb{1}\{a \le z \le b\}$, whereas it is localising if the weight is the indicator function of a one-sided interval, e.g. $w(z) = \mathbb{1}\{z \ge t\}$. Figure 1 helps to appreciate why this is the case. If the weight is one-sided, then the chaining function maps all elements in $\{w = 0\}$ to a single point. On the other hand, if the weight is an indicator of a compact interval, then points in $\{w = 0\}$ will be mapped to one of two values, depending on where they lie in relation to the interval. As a result, the twCRPS in this case depends not only on the forecast distribution within the region of interest, but also on the measure assigned to outcomes above and below the interval bounds. Nonetheless, if the weight is an indicator function of a one-sided interval, Holzmann and Klar (2017, Theorem 4) demonstrate that not only is the twCRPS localising, but it is also strictly locally proper.

2.3   Weighted scores for ensemble forecasts

Equations 7 and 10 allow for straightforward computation of the weighted scoring rules when forecasts are in the form of an empirical distribution function, or an ensemble. In this case, we assume the predictive distribution is a sum of step functions at the M ensemble members $x_1, ..., x_M \in \mathbb{R}$, i.e. $F_{ens}(x) = \sum_{m=1}^{M} \mathbb{1}\{x_m \le x\}/M$. Gneiting et al. (2007) remark that the kernel score representation of the CRPS makes it "particularly convenient when F is represented by a sample, possibly based on Markov chain Monte Carlo output or forecast ensembles," which is ordinarily the case in weather and climate forecasting (e.g. Leutbecher and Palmer, 2008). Using the representation of the twCRPS as a kernel score, it is possible to express the threshold-weighted CRPS for an ensemble forecast in the form

$$\text{twCRPS}(F_{ens}, y; v) = \frac{1}{M} \sum_{m=1}^{M} |v(x_m) - v(y)| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{j=1}^{M} |v(x_m) - v(x_j)|, \tag{12}$$

which can be verified by substituting F_{ens} into Equation 6. Substituting the identity for v in Equation 12 recovers the corresponding well-known expression for the CRPS as a special case. From an implementation standpoint, this expression has the benefit that it involves only a (typically straightforward) transformation of the ensemble members and observations before applying existing methods and software to calculate the CRPS.

Similarly, Equation 10 permits the following representation of the outcome-weighted CRPS:

$$\text{owCRPS}(F_{ens}, y; w) = \frac{1}{M\bar{w}} \sum_{m=1}^{M} |x_m - y| w(x_m) w(y) - \frac{1}{2M^2 \bar{w}^2} \sum_{m=1}^{M} \sum_{j=1}^{M} |x_m - x_j| w(x_m) w(x_j) w(y),$$

where $\bar{w} = \sum_{m=1}^{M} w(x_m)/M$. Note, however, that this expression can result in an undefined score if the weight function is not strictly positive, since $\bar{w}$ could equal zero: for example, if the weight $w(z) = \mathbb{1}\{z \ge t\}$ is employed but all ensemble members fall below the threshold t. This is more generally the case for continuous forecast distributions that assign zero probability to $\{w > 0\}$, meaning F_w is undefined, though this will be more prevalent when dealing with finite ensembles.

3   Making new kernel scores from old

The previous section places existing weighted versions of the CRPS into the framework of kernel scores. Rasmussen and Williams (2006) list several operations under which the positive definiteness of a kernel is

conserved, and this desirable property has also been studied in the case of negative definite and c.n.d. kernels (see e.g. Berg et al., 1984; Chilès and Delfiner, 2009). For example, the sum of several c.n.d. kernels is itself c.n.d., while it is straightforward to verify that if ρ is a c.n.d. kernel, then $\rho(v(x), v(x'))$, for some $v : \mathcal{X} \to \mathcal{X}$, is also a c.n.d. kernel. If a kernel is additionally negative definite, then multiplication by a non-negative deterministic function will also yield a negative definite kernel. This permits the construction of flexible kernel scores that can be employed to assess forecasts in a range of different settings.

The threshold-weighted CRPS discussed in the previous section is one example of this. It can be generalised to construct a class of threshold-weighted kernel scores by replacing the Euclidean distance with an arbitrary c.n.d. kernel.

Definition 4. Let ρ be a c.n.d. kernel on $\mathcal{X}$ and let $v : \mathcal{X} \to \mathcal{X}$ be a measurable function. We define the *threshold-weighted kernel score* with kernel ρ and chaining function v as

$$\mathrm{tw}S_\rho(P, y; v) = \mathbb{E}_P[\rho(v(X), v(y))] - \frac{1}{2}\mathbb{E}_P[\rho(v(X), v(X'))] - \frac{1}{2}\rho(v(y), v(y)), \tag{13}$$

where $X, X' \sim P \in \mathfrak{M}$ are independent, $y \in \mathcal{X}$, and it is assumed that all expectations are finite.

Theorem 1 implies that if ρ is c.n.d. with $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, then the threshold-weighted kernel score with kernel ρ and chaining function v is proper with respect to $\mathcal{M}_{\tilde{\rho}}$, where $\tilde{\rho}(x, x') = \rho(v(x), v(x'))$. If ρ is negative definite, then the score is proper with respect to $\mathcal{M}^{\tilde{\rho}}$. If the kernel score S_ρ is strictly proper, then it is also possible to characterise the chaining functions that preserve this strict propriety.

Proposition 3. *Let ρ be a c.n.d. kernel on $\mathcal{X}$ and let $v : \mathcal{X} \to \mathcal{X}$ be a measurable function. If S_ρ is strictly proper with respect to $\mathcal{M}_\rho$ ($\mathcal{M}^\rho$), then $\mathrm{tw}S_\rho(\cdot, \cdot; v)$ is strictly proper with respect to $\mathcal{M}_{\tilde{\rho}}$ ($\mathcal{M}^{\tilde{\rho}}$) if and only if the chaining function v is injective.*

It could be argued that the strict propriety of a weighted scoring rule is often not of primary concern, since interest is typically only on a subset of possible outcomes: the set $\{w > 0\}$, given the chosen weight function. However, threshold-weighted kernel scores require the specification of a chaining function, which may or may not be associated with a measure, or weight, and there is no canonical way to derive a chaining function that corresponds directly to a given weight. Nevertheless, if a certain weight function has been chosen, it is possible to characterise the chaining functions for which a threshold-weighted kernel score is localising and strictly locally proper.

Proposition 4. *Let ρ be a c.n.d. kernel on $\mathcal{X}$ such that $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, let w be a weight function, and let $v : \mathcal{X} \to \mathcal{X}$ be a measurable function. Then, $\mathrm{tw}S_\rho(\cdot, \cdot; v)$ is localising with respect to w if and only if $\rho(v(z), v(\cdot)) = \rho(v(z'), v(\cdot))$ for all $z, z' \in \{w = 0\}$.*

Hence, whether or not a threshold-weighted kernel score is localising with respect to a given weight will depend on the choice of chaining function.

Remark 3. The requirement $\rho(v(z), v(\cdot)) = \rho(v(z'), v(\cdot))$ for all $z, z' \in \{w = 0\}$ can easily be satisfied by choosing a chaining function such that $v(z) = v(z') = x_0$ for all $z, z' \in \{w = 0\}$, and some $x_0 \in \mathcal{X}$. If ρ is strictly c.n.d., then it is straightforward to show that this is implied by the requirement. For such a chaining function, we say that the threshold-weighted kernel score is *centred at x_0*.

Proposition 5. *Let ρ be a c.n.d. kernel on $\mathcal{X}$, let w be a weight function, and let $v : \mathcal{X} \to \mathcal{X}$ be a measurable function. If S_ρ is strictly proper with respect to $\mathcal{M}_\rho$ ($\mathcal{M}^\rho$), then $\mathrm{tw}S_\rho(\cdot, \cdot; v)$ is strictly locally proper with respect to $\mathcal{M}_{\tilde{\rho}}$ ($\mathcal{M}^{\tilde{\rho}}$) if and only if it is localising and the restriction of v to $\{w > 0\}$ is injective.*

In contrast to threshold-weighted kernel scores, outcome-weighted scores can be generalised to arbitrary c.n.d. kernels whilst maintaining a direct connection to the weight function. In particular, similarly to the outcome-weighted CRPS in Equation 10, we define outcome-weighted kernel scores as follows; they are a special case of the weighted scores proposed by Holzmann and Klar (2017).

Definition 5. Let ρ be a c.n.d. kernel on $\mathcal{X}$ and let w be a weight function. Let $P \in \mathcal{M}_\rho$ such that $\mathbb{E}_P[w(X)] > 0$. We define the *outcome-weighted kernel score* with kernel ρ and weight function w as

$$\text{ow}S_\rho(P, y; w) = \frac{1}{C_w(P)}\mathbb{E}_P[\rho(X, y)w(X)w(y)] - \frac{1}{2C_w(P)^2}\mathbb{E}_P[\rho(X, X')w(X)w(X')w(y)] - \frac{1}{2}\rho(y, y)w(y),$$
(14)

where $X, X' \sim P$ are independent, $y \in \mathcal{X}$, and $C_w(P) = \mathbb{E}_P[w(X)]$.

As mentioned for the owCRPS, outcome-weighted kernel scores provide a means of weighting existing kernel scores, but they themselves do not fit into the kernel score framework. The results of Holzmann and Klar (2017) discussed in Section 2.2 clarify when outcome-weighted kernel scores are proper, localising, proportionally locally proper, and how they can be modified to be strictly locally proper.

It is well-known that if ρ is a negative definite kernel and w is a weight function, then $\check{\rho}(x, x') = \rho(x, x')w(x)w(x')$ is also a negative definite kernel. We therefore propose constructing weighted scoring rules based on this weighted kernel, and we label such scores vertically re-scaled kernel scores.

Definition 6. Let ρ be a c.n.d. kernel on $\mathcal{X}$ and let w be a weight function.

(i) If ρ is negative definite, we define the *vertically re-scaled kernel score* with kernel ρ and weight function w as

$$\text{vr}S_\rho(P, y; w) = \mathbb{E}_P[\rho(X, y)w(X)w(y)] - \frac{1}{2}\mathbb{E}_P[\rho(X, X')w(X)w(X')] - \frac{1}{2}\rho(y, y)w(y)^2, \quad (15)$$

where $X, X' \sim P \in \mathcal{M}^\rho$ are independent and $y \in \mathcal{X}$.

(ii) If ρ satisfies $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, we define the *vertically re-scaled kernel score* with kernel ρ, weight function w, and centre $x_0 \in \mathcal{X}$ as

$$\begin{aligned}\text{vr}S_\rho(P, y; w, x_0) =& \mathbb{E}_P[\rho(X, y)w(X)w(y)] - \frac{1}{2}\mathbb{E}_P[\rho(X, X')w(X)w(X')] \\ &+ (\mathbb{E}_P[\rho(X, x_0)w(X)] - \rho(y, x_0)w(y))(\mathbb{E}_P[w(X)] - w(y)),\end{aligned}$$
(16)

where $X, X' \sim P \in \mathcal{M}_\rho$ are independent and $y \in \mathcal{X}$.

Since $\check{\rho}$ is itself a negative definite kernel and $\mathcal{M}^\rho \subset \mathcal{M}^{\check{\rho}}$, it follows that the associated kernel score in Equation 15 is proper with respect to $\mathcal{M}^\rho$. However, although multiplication by a non-negative function preserves the negative definiteness of a kernel, if ρ is only c.n.d., then it is not necessarily the case that $\check{\rho}$ will be. On the other hand, if ρ is a c.n.d. kernel with $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, then

$$\rho^*(x, x') = \rho(x, x') - \rho(x, x_0) - \rho(x', x_0) \quad (17)$$

will be negative definite, for arbitrary $x_0 \in \mathcal{X}$ (Berg et al., 1984, Lemma 2.1). Since ρ^* is negative definite, it follows that $\rho^*(x, x')w(x)w(x')$ is also negative definite, and this kernel is used in Equation 16 to construct the vertically re-scaled kernel score centred at x_0. Proposition 20 in Sejdinovic et al. (2013) can be used to show that this score is proper with respect to $\mathcal{M}_\rho$.

Regardless of whether ρ is negative definite or not, the unweighted kernel score is recovered by choosing $w(z) = 1$, and does not depend on x_0. In general, however, the vertically re-scaled kernel score will depend on x_0. Although it is not immediately obvious how the choice of x_0 will affect the score's behaviour in practice, it does not alter the theoretical properties of this weighted score, and for the applications in Section 4 there is always a canonical choice. If a vertically re-scaled kernel score is strictly proper with respect to a class of distributions that contains Dirac measures, then $\{w = 0\}$ can contain at most one element. Ensuring strict propriety of a vertically re-scaled score requires stronger assumptions.

Proposition 6. *Let ρ be a negative definite kernel on $\mathcal{X}$ and let $w > 0$ be a weight function. If $-\rho$ is strictly integrally positive definite with respect to the maximal possible set of signed measures in the sense of Steinwart and Ziegel (2021, Definition 2.1), then $\text{vr}S_\rho(\cdot, \cdot; w)$ is strictly proper with respect to $\mathcal{M}^\rho$.*

If $-\rho$ is a bounded continuous strictly positive definite function on $\mathbb{R}^d$ for any $d \geq 1$, then the requirement in Proposition 6 is satisfied. However, if ρ is a c.n.d. kernel with $\rho(x,x) = 0$, $x \in \mathcal{X}$, applying Proposition 6 to ρ^* at (17) is not always possible since the literature on kernel embedding with unbounded kernels is limited.

It can be seen immediately from their definition that vertically re-scaled kernel scores depend on the forecast P only through its restriction to the set $\{w > 0\}$, and they are therefore localising. Slightly generalizing Proposition 6, we obtain the following result.

Proposition 7. *Let ρ be a negative definite kernel on $\mathcal{X}$ and let w be a weight function. If $-\rho$ is strictly integrally positive definite with respect to the maximal possible set of signed measures on $\{w > 0\}$ in the sense of Steinwart and Ziegel (2021, Definition 2.1), then $\mathrm{vr}S_\rho(\cdot, \cdot; w)$ is strictly locally proper with respect to $\mathcal{M}^\rho$.*

Hence, vertically re-scaled kernel scores provide a direct means of obtaining a strictly locally proper scoring rule. Furthermore, they also fit into the kernel score framework. While threshold-weighted kernel scores deform the inputs of the kernel, vertically re-scaled kernel scores weight the kernel's output. However, for particular weight and chaining functions, vertically re-scaled and threshold-weighted kernel scores are equivalent.

Proposition 8. *Let ρ be a c.n.d. kernel on $\mathcal{X}$ with $\rho(x,x) = 0$ for all $x \in \mathcal{X}$, let w be a weight function such that $w(x) \in \{0,1\}$ for all $x \in \mathcal{X}$, and let $x_0 \in \mathcal{X}$. Consider the chaining function $v(x) = xw(x) + x_0(1 - w(x))$, $x \in \mathcal{X}$. Then, the threshold-weighted kernel score with kernel ρ and chaining function v equals the vertically re-scaled kernel score with kernel ρ, weight function w, and centre x_0.*

Given a kernel score, this section has described three possible approaches that can be used to weight the scoring rule in order to emphasise particular outcomes. Since these approaches apply to the entire class of kernel scores, they are applicable in a wide range of settings. As an example of this, in the following section, we investigate the application of kernel scores in a multivariate context and use results from this section to introduce weighted versions of popular multivariate scoring rules.

4 Weighted multivariate scoring rules

4.1 Energy and variogram scores

Let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{M}$ denote the set of Borel probability measures on $\mathbb{R}^d$. Forecast verification in a multivariate setting is significantly less developed than in the univariate case (Gneiting and Katzfuss, 2014). In particular, there are relatively few recognised scoring rules to quantify the accuracy of multivariate forecasts. The logarithmic score can be used to assess multivariate predictive densities, but multivariate forecasts are commonly in the form of a finite ensemble. Applying the logarithmic score to a multivariate normal density recovers the Dawid-Sebastiani score (Dawid and Sebastiani, 1999), which evaluates forecasts only through their first two moments. Although this makes the score readily applicable to ensemble forecasts, it can become uninformative when the number of dimensions under consideration is large compared with the number of ensemble members. Hence, two alternative multivariate scoring rules are commonly preferred in practice: the energy score and the variogram score.

The energy score is generally defined as

$$\mathrm{ES}_\beta(P, y) = \mathbb{E}_P ||X - y||^\beta - \frac{1}{2}\mathbb{E}_P ||X - X'||^\beta, \tag{18}$$

where $||\cdot||$ is the Euclidean norm and the exponent $\beta \in (0, 2)$ is typically set to one (Gneiting and Raftery, 2007). Here, and throughout this section, we assume that all relevant expectations are finite. Clearly, Equation 18 defines a kernel score associated with the c.n.d. kernel $\rho(x,x') = ||x - x'||^\beta$, for $x, x' \in \mathbb{R}^d$, and the energy score thus generalises the CRPS to multiple dimensions. The energy score is strictly proper with respect to $\{P \in \mathcal{M} : \mathbb{E}_P ||X||^\beta < \infty\}$.

The results presented in the previous section allow us to generate three distinct weighted energy scores. Firstly, Definition 4 can be used to construct a threshold-weighted energy score, which provides a natural extension of the threshold-weighted CRPS to higher dimensions:

$$\text{twES}_\beta(P, y; v) = \mathbb{E}_P ||v(X) - v(y)||^\beta - \frac{1}{2}\mathbb{E}_P ||v(X) - v(X')||^\beta,$$

where $v : \mathbb{R}^d \to \mathbb{R}^d$ is a chaining function. Alternatively, applying Equation 14 to the energy score recovers the outcome-weighted energy score proposed by Holzmann and Klar (2017):

$$\text{owES}_\beta(P, y; w) = \frac{1}{C_w(P)}\mathbb{E}_P[||X - y||^\beta w(X)w(y)] - \frac{1}{2C_w(P)^2}\mathbb{E}_P[||X - X'||^\beta w(X)w(X')w(y)],$$

where $w : \mathbb{R} \to [0, 1]$ is a weight function. Similarly, Definition 6 can be used to introduce a vertically re-scaled energy score. Since the kernel used in the energy score is only c.n.d., this requires choosing a point $x_0 \in \mathbb{R}^d$ at which to centre the weighted score. The natural choice is $x_0 = 0$:

$$\begin{aligned}\text{vrES}_\beta(P, y; w) =& \mathbb{E}_P\left[||X - y||^\beta w(X)w(y)\right] - \frac{1}{2}\mathbb{E}_P\left[||X - X'||^\beta w(X)w(X')\right] \\ &+ (\mathbb{E}_P\left[||X - x_0||^\beta w(X)\right] - ||y - x_0||^\beta w(y))(\mathbb{E}_P\left[w(X)\right] - w(y)).\end{aligned}$$

While the results in Section 3 provide conditions on the weight and chaining functions that ensure strict (local) propriety of the threshold-weighted and outcome-weighted energy score, we can only guarantee propriety for the vertically re-scaled energy score.

Although the energy score is possibly the most widely implemented multivariate scoring rule, several studies have found evidence to suggest that it is over-sensitive to marginal forecast performance (e.g. Pinson and Tastu, 2013). Scheuerer and Hamill (2015b) argue that since marginal performance can be assessed using univariate scoring rules, multivariate assessment should instead focus on evaluating the forecast's dependence structure. To this end, the authors introduce the variogram score as an alternative multivariate scoring rule. Given an observation $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, the variogram score of order $p > 0$ is defined as

$$\text{VS}_p(P, y) = \sum_{i=1}^d \sum_{j=1}^d h_{i,j}(\mathbb{E}_P|X_i - X_j|^p - |y_i - y_j|^p)^2, \tag{19}$$

where $X = (X_1, \ldots, X_d) \sim P$, and $h_{i,j} \in [0, 1]$ are non-negative scaling parameters. In contrast to the energy score, the variogram score is proper with respect to the set $\{P \in \mathcal{M} : \mathbb{E}_P|X_i|^{2p} < \infty \text{ for each } i = 1, \ldots, d\}$, but is not strictly proper (Scheuerer and Hamill, 2015b). Nonetheless, it is straightforward to verify that the variogram score is also a kernel score, corresponding to the c.n.d. kernel

$$\rho(x, x') = \sum_{i=1}^d \sum_{j=1}^d h_{i,j}(|x_i - x_j|^p - |x'_i - x'_j|^p)^2,$$

where $x = (x_1, \ldots, x_d), x' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$. The variogram score can thus also be expressed as

$$\text{VS}_p(P, y) = \mathbb{E}_P\left[\sum_{i=1}^d \sum_{j=1}^d h_{i,j}(|X_i - X_j|^p - |y_i - y_j|^p)^2\right] - \frac{1}{2}\mathbb{E}_P\left[\sum_{i=1}^d \sum_{j=1}^d h_{i,j}(|X_i - X_j|^p - |X'_i - X'_j|^p)^2\right],$$

where $X, X' \sim P$ are independent.

The variogram score was introduced as a multivariate scoring rule that is more sensitive to errors in the forecast's dependence structure than the energy score, and hence is itself an example of how the kernel within the kernel score framework can be chosen in order to emphasise particular features of the forecasts.

In addition, just as we have introduced weighted versions of the energy score, threshold-weighted, outcome-weighted, and vertically re-scaled versions of the variogram score can also easily be introduced. For example, the threshold-weighted variogram score of order p with chaining function v is

$$\text{twVS}_p(P, y; v) = \sum_{i=1}^{d} \sum_{j=1}^{d} h_{i,j}(\mathbb{E}_P |v(X)_i - v(X)_j|^p - |v(y)_i - v(y)_j|^p)^2.$$

Note, however, that since the variogram score is not strictly proper, the outcome-weighted version of this score is localising and proper, but not necessarily proportionally locally proper.

The energy score and variogram score are established kernel scores. The kernel score framework also permits the introduction of novel kernel scores by choosing an appropriate kernel. We illustrate this here by introducing a scoring rule based on the inverse multiquadric kernel (Micchelli, 1984; Schölkopf and Smola, 2002). In particular, we define the inverse multiquadric score (IMS) as

$$\text{IMS}(P, y) = \mathbb{E}_P \left[-(1 + ||X - y||^2)^{-\frac{1}{2}} \right] - \frac{1}{2} \mathbb{E}_P \left[-(1 + ||X - X'||^2)^{-\frac{1}{2}} \right] + \frac{1}{2},$$

where $X, X' \sim P$ are independent.

The IMS can be used to assess both univariate and multivariate forecasts, and weighted versions of this score can be constructed using the results presented in the previous section. The strict propriety of the IMS with respect to the entire class $\mathcal{M}$ follows from the fact that the inverse multiquadric kernel is strictly positive definite and bounded. Boundedness of the kernel is in contrast to the kernels used within the CRPS, energy score, and variogram score. In particular, it implies strict local propriety of the vertically re-scaled IMS. By comparing the IMS to these established scores, we can see to what extent these properties of the kernel affect the behaviour of the resulting kernel score.

4.2 Weight and chaining functions

Thus far, we have argued that applying a weighted scoring rule is often synonymous with choosing a suitable kernel to employ within the kernel score framework. A natural question then arises regarding what weight or chaining function, and hence what kernel, to choose for a given problem. In this section, we consider the case where the outcome space is the d-dimensional Euclidean space, $\mathbb{R}^d$, and discuss possible weight and chaining functions that could be used within the weighted multivariate scores introduced above in order to evaluate forecasts made for high-impact events.

Firstly, consider possible weight functions. In the univariate setting, it is common to assess forecasts with a weight function that only emphasises values above (or below) a chosen threshold, e.g. $w(z) = \mathbb{1}\{z \geq t\}$. In the multivariate case, this can be extended seamlessly by considering weight functions that are one when a combination of the values along the different dimensions exceeds a threshold, and zero otherwise: $w(z) = \mathbb{1}\{\sum_{i=1}^{d} b_i z_i \geq t\}$ for constants $b_1, \ldots, b_d \in \mathbb{R}$ and a threshold $t \in \mathbb{R}$. An example of this weight function is displayed in Figure 2a.

However, it may be the case that high-impact events arise from the interaction of several moderate events. For example, moderate rainfall over consecutive days is likely to result in flooding, despite the rainfall on each day not being extreme from a statistical perspective. Hence, one could also consider using a weight function that depends on whether a threshold is exceeded in every dimension, e.g. $w(z) = \mathbb{1}\{z_1 \geq t_1, \ldots, z_d \geq t_d\}$ with $t_1, \ldots, t_d \in \mathbb{R}$. Such weight functions can be interpreted as indicator functions of orthants in Euclidean space, as illustrated for the bivariate case in Figure 2b.

Both of these weights are based on indicator functions, meaning they are not strictly positive. As such, the outcome-weighted and vertically re-scaled scoring rules constructed using these weights will not be strictly proper. The final column of Figure 1 provides an example of a univariate weight function that is strictly positive, the Gaussian distribution function. This can also readily be extended to higher dimensions by using

(a) $w(z) = \mathbb{1}\{z_1 + z_2 \geq t\}$ (b) $w(z) = \mathbb{1}\{z_1 \geq t_1, z_2 \geq t_2\}$ (c) $w(z) = \Phi(z; \mu, \Sigma)$
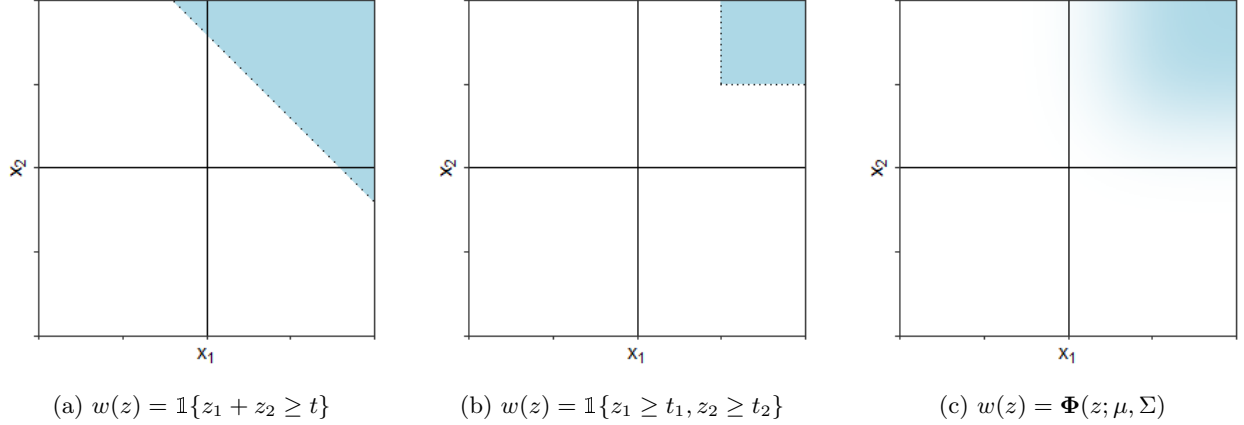
Figure 2: Possible weight functions that could be used to assess forecasts with emphasis on high-impact events in a bivariate setting. A darker shade of blue reflects a higher weight. $\Phi(z; \mu, \Sigma)$ denotes the cumulative distribution function of the bivariate Gaussian distribution with mean vector μ and covariance matrix Σ.

a weight function equal to a continuous multivariate distribution function. In this case, the weight changes smoothly over $\mathbb{R}^d$, whilst still allowing emphasis to be placed on certain outcomes, as illustrated in Figure 2c.

If a threshold-weighted kernel score is to be implemented, it is necessary to specify a chaining function, rather than a weight, which is a less trivial task. In the univariate case, a weight function can typically be converted to a suitable chaining function due to the integral representation of the threshold-weighted CRPS (Remark 1). More generally, however, there is no canonical way to derive a chaining function given a particular weight.

Proposition 4 states that the threshold-weighted kernel score will be localising if the chaining function maps all points in $\{w = 0\}$ to a single value $x_0 \in \mathbb{R}^d$, while Proposition 5 states that if the chaining function is additionally injective on $\{w > 0\}$, then the score will be strictly locally proper. If all points in $\{w > 0\}$ receive the same weight, as is the case for indicator-based weight functions, then one option is to employ the function

$$v(z) = \begin{cases} z & \text{if } z \in \{w > 0\}, \\ x_0 & \text{if } z \in \{w = 0\}, \end{cases} \tag{20}$$

for some $x_0 \in \mathbb{R}^d$. Such a chaining function maintains correspondence with the twCRPS when the weight is a one-sided interval: for example, when $d = 1$ and $w(z) = \mathbb{1}\{z \geq t\}$, choosing $x_0 = t$ recovers the chaining function $v(z) = \max(z, t)$, as presented in Figure 1.

On the other hand, if the weight is not constant on $\{w > 0\}$, then what chaining function to choose will depend strongly on what the weighting is designed to achieve. Since we are interested here in high-impact events, consider the case where the weight function is increasing along each dimension, a multivariate distribution function, for example. In the univariate case, if the weight is increasing, then the resulting chaining function is convex. One way to translate this to the multivariate setting is to use a chaining function that is convex along every dimension.

Following Remark 1, a possible chaining function that satisfies this is the integral of the weight function along each margin separately, conditional on the values along the other dimensions. For example, the final weight function considered in Figure 2 employs a multivariate Gaussian distribution function. Given a mean vector $\mu = (\mu_1, \ldots, \mu_d)$ and a diagonal covariance matrix Σ with variances $\sigma_1^2, \ldots, \sigma_d^2$, integrating the conditional Gaussian distribution along each dimension yields a chaining function of the form

$$v(z) = \left((z_1 - \mu_1)\Phi\left(\frac{z_1 - \mu_1}{\sigma_1}\right) + \sigma_1\phi\left(\frac{z_1 - \mu_1}{\sigma_1}\right), \ldots, (z_d - \mu_d)\Phi\left(\frac{z_d - \mu_d}{\sigma_d}\right) + \sigma_d\phi\left(\frac{z_d - \mu_d}{\sigma_d}\right) \right),$$
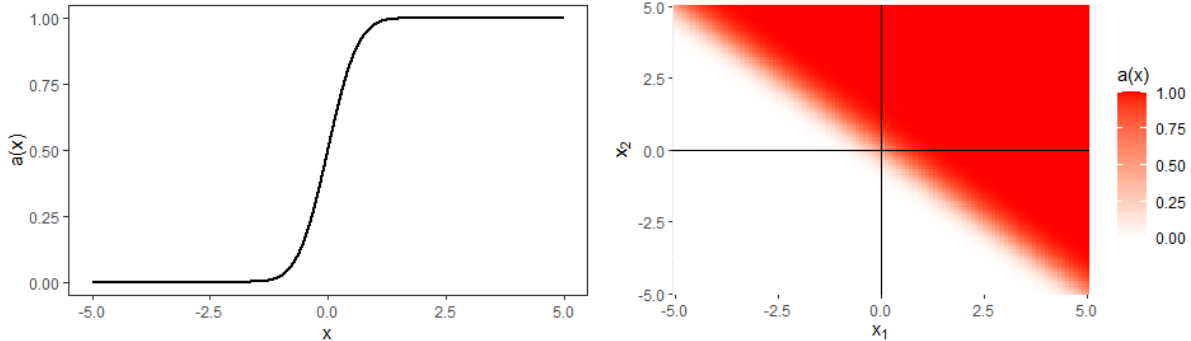
Figure 3: The mixing function $a(x)$ in the univariate case (Left) and the bivariate case (Right).

which is essentially a component-wise extension of the chaining function presented in the final column of Figure 1. Moreover, since this chaining function is injective on $\{w > 0\}$, Proposition 5 states that the resulting threshold-weighted kernel score will be strictly locally proper.

These are only a few examples of possible weight and chaining functions that could be employed when evaluating forecasts and outcomes on $\mathbb{R}^d$ whilst emphasising high-impact events. Different choices of either function will generate a scoring rule that assesses different aspects of the forecast performance, and, in general, it is a task for a domain expert to choose the appropriate weight or chaining function in order to extract the relevant information for the problem at hand. In the remainder of this section, we examine how the weights and deformations presented above can be used within weighted multivariate scoring rules to assess forecasts made for high-impact events.

4.3   Simulation study

4.3.1   Outline

In order to understand the properties of the various weighted multivariate scoring rules, we apply them to simulated forecasts and observations. The simulation study is organised as follows. Firstly, a distribution G is chosen from which to draw 100 independent observations. Secondly, two forecast distributions, F_1 and F_2, are specified, both of which are linear combinations of the true distribution G, and a mis-specified distribution H:

$$F_1(z) = a(z)G(z) + (1 - a(z))H(z),$$
$$F_2(z) = (1 - a(z))G(z) + a(z)H(z),$$

where F_1, F_2, G and H all denote distributions on $\mathbb{R}^d$, while $a : \mathbb{R}^d \to [0, 1]$ is a mixing function.

In the following, G denotes a standard multivariate Gaussian distribution, while H is a multivariate Student's t distribution with four degrees of freedom (Hothorn et al., 2001). As in Holzmann and Klar (2017), the mixing function a is a univariate Gaussian distribution function with zero mean and standard deviation equal to one half. In the multivariate case, this univariate distribution function is evaluated at $\sum_{i=1}^{d} z_i$. This mixing function is displayed for the univariate and bivariate cases in Figure 3.

In order to assess the competing forecasts, 100 ensemble members are sampled at random from F_1 and F_2, and both forecasts are then evaluated at each of the 100 observations via ensemble forecast representations of the various weighted scoring rules. A Diebold-Mariano test (Diebold and Mariano, 1995) is applied to the sample mean scores of both forecasts to assess whether or not one forecast outperforms the other when assessed using each score. This process is repeated 1000 times and the proportion of instances that the null hypothesis of equal predictive performance is rejected in favour of each forecast distribution is recorded. The rejection rates for the various scoring rules can then be examined to understand the discriminative behaviour

15

of the different scores. Such a framework has been considered in several previous studies on weighted scoring rules (Diks et al., 2011; Lerch et al., 2017; Holzmann and Klar, 2017).

In the univariate case, the forecasts are assessed using the CRPS and the IMS, while in the multivariate case, results are presented for the energy score, the variogram score and the IMS. Threshold-weighted, outcome-weighted, and vertically re-scaled versions of these kernel scores are all considered, as well as outcome-weighted scores that have been complemented with the Brier score in order to make them strictly locally proper, as discussed in Section 2.2 (Equation 11).

For simplicity, we consider only indicator-based weight functions within these weighted scores, which are commonly applied in practice. Results are presented for the univariate weight function $w(z) = \mathbb{1}\{z \geq t\}$, where interest is on values that exceed a chosen threshold, and for two multivariate, indicator-based weight functions: $w(z) = \mathbb{1}\{\sum_{i=1}^{d} z_i \geq t\}$ and $w(z) = \mathbb{1}\{z_1 \geq t, \ldots, z_d \geq t\}$. In all cases, we study the results as the threshold t is changed.

Equation 20 is used to construct a chaining function that generates localising threshold-weighted kernel scores. In the first scenario, when $w(z) = \mathbb{1}\{z_1 \geq t, \ldots, z_d \geq t\}$, the threshold-weighted scores are centred at $x_0 = (t, \ldots, t)$, while in the second case, when $w(z) = \mathbb{1}\{\sum_{i=1}^{d} z_i \geq t\}$, the scores are centred at $x_0 = (t/d, \ldots, t/d)$. Where relevant, the vertically re-scaled kernel scores are centred at $(0, \ldots, 0)$. If the kernel in the kernel is only c.n.d., this is equivalent to using a threshold-weighted kernel score centred at the origin (Proposition 8). Hence, comparing the performance of the threshold-weighted and vertically re-scaled scores allows us to assess how the scores depend on the point at which they are centred. For the threshold-weighted variogram score, however, the kernel is insensitive to whether the score is centred at $(t, \ldots, t)$ or $(0, \ldots, 0)$, and hence the results for the localising threshold-weighted variogram score are equivalent to those for the vertically re-scaled variogram score.

For the sake of comparison, these scores are compared to threshold-weighted kernel scores that are non-localising. In this case, the chaining function is chosen to resemble the chaining function in the univariate case. Firstly, consider the weight function $w(z) = \mathbb{1}\{z_1 \geq t, \ldots z_d \geq t\}$. One possible chaining function corresponding to this weight would be to take the component-wise maximum between the point and the threshold:

$$v(z) = (\max(z_1, t), \ldots, \max(z_d, t)).$$

As has been discussed, the weight function in this case can be thought of as an orthant in Euclidean space, and this chaining function simply projects any point not contained in this orthant onto the perimeter of the orthant, while leaving the remaining points unchanged. That is, points that lie in the region of interest, $\{w > 0\}$, are unchanged, whereas points outside this region are projected onto the closest point for which the weight equals one.

This approach could similarly be used when considering the weight function $w(z) = \mathbb{1}\{\sum_{i=1}^{d} z_i \geq t\}$, in which case the chaining function becomes

$$v(z) = (\max(z_1, z_1 + l), \ldots, \max(z_d, z_d + l)),$$

where $l = (t - \sum_{i=1}^{d} z_i)/d$. Here, any point in $\{w = 0\}$ is moved perpendicular to the plane defined by the weight function, until it reaches a point on the plane.

4.3.2 Results

Firstly, consider the univariate case. The forecast F_1 is correctly specified in the upper tail, but exhibits a heavy lower tail, whereas the opposite is true for F_2. As such, since evaluation is focused on the upper tail of the forecast distributions, for large values of t the weighted scoring rules should reject the null hypothesis of equal predictive performance in favour of F_1.

While the rejection rate of the unweighted scores is close to 0.025 for all thresholds, Figure 4 illustrates that the frequency of rejections in favour of F_1 generally increases with the threshold when a weighted scoring rule
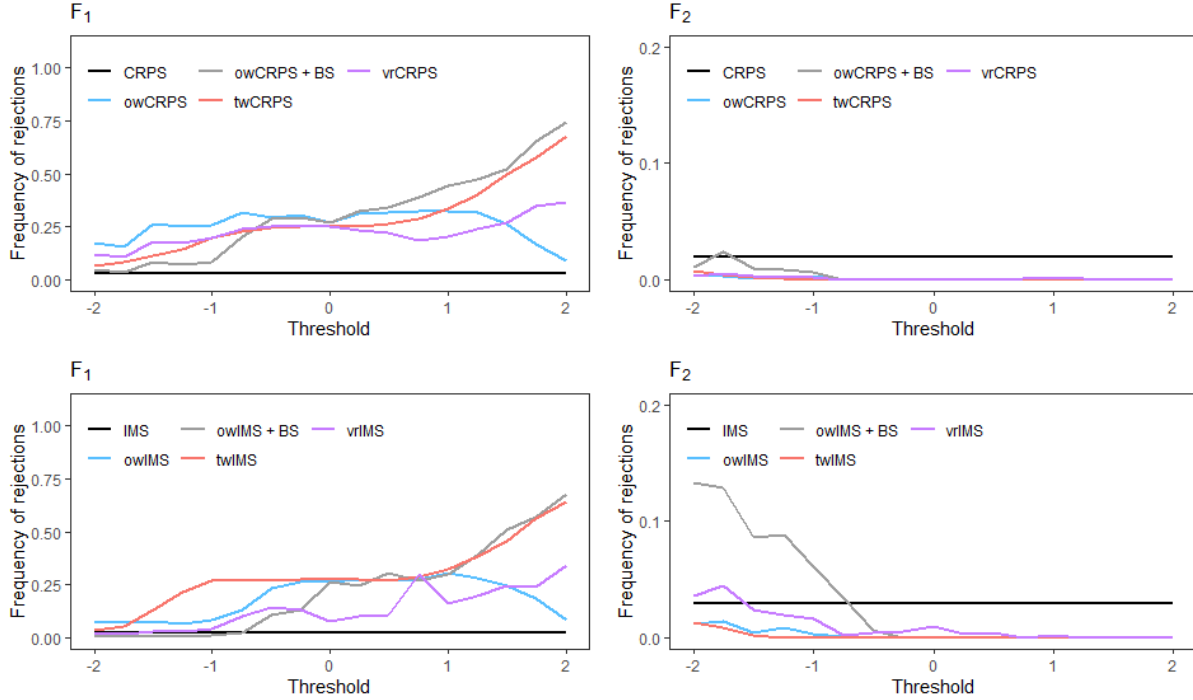
Figure 4: The proportion of instances that a Diebold-Mariano test for equal predictive performance is rejected in favour of F_1 (Left) and F_2 (Right) for each version of the CRPS (Top) and the IMS (Bottom). The rejection rate is displayed as a function of the threshold used in the weight function $w(z) = \mathbb{1}\{z \geq t\}$ when evaluating the forecasts. Note the different scales when considering F_1 and F_2.

is used to assess the forecasts. The outcome-weighted scores perform poorly for larger thresholds, since they are sensitive to the number of observations that exceed the threshold, a result also observed in Holzmann and Klar (2017). Complementing the owCRPS and owIMS with the Brier score generates scores that can better distinguish between the two forecasts. The threshold-weighted scores also perform well in this respect. Comparing the threshold-weighted CRPS to the vertically re-scaled CRPS illustrates the sensitivity of the scores to the point at which the two weighted scores are centred: in this case, centering the scores at the threshold appears more beneficial than centering them at the origin.

The right-hand panel of Figure 4 shows the proportion of rejections in favour of F_2. As would be expected, and in contrast to F_1, this rejection frequency is close to zero for large thresholds, and increases towards 0.025 as the threshold becomes smaller, mirroring the results presented in Holzmann and Klar (2017). The owIMS complemented with the Brier score, on the other hand, results in a large rejection frequency when the threshold is low.

Consider now results for the bivariate setting. In this case, F_1 is close to the true data generating process G in the upper right quadrant (when the mixing function is close to one), whereas F_2 is more similar to the true distribution in the lower left quadrant, meaning the weighted scores should again reject the hypothesis of equal predictive performance in favour of F_1 when the threshold t is large.

Figure 5 displays the rejection frequency in favour of F_1 and F_2 corresponding to each scoring rule for the weight function $w(z) = \mathbb{1}\{z_1 \geq t, z_2 \geq t\}$. Similar results are observed to those presented in the univariate case. In particular, the energy score, variogram score and the inverse multiquadric score all cannot distinguish between the two forecasts, whereas the weighted scores do, particularly when interest is on relatively large thresholds. The rejection rates corresponding to the outcome-weighted scores increase slightly with the threshold, but then tend towards zero as the number of observations that exceed the threshold decreases.
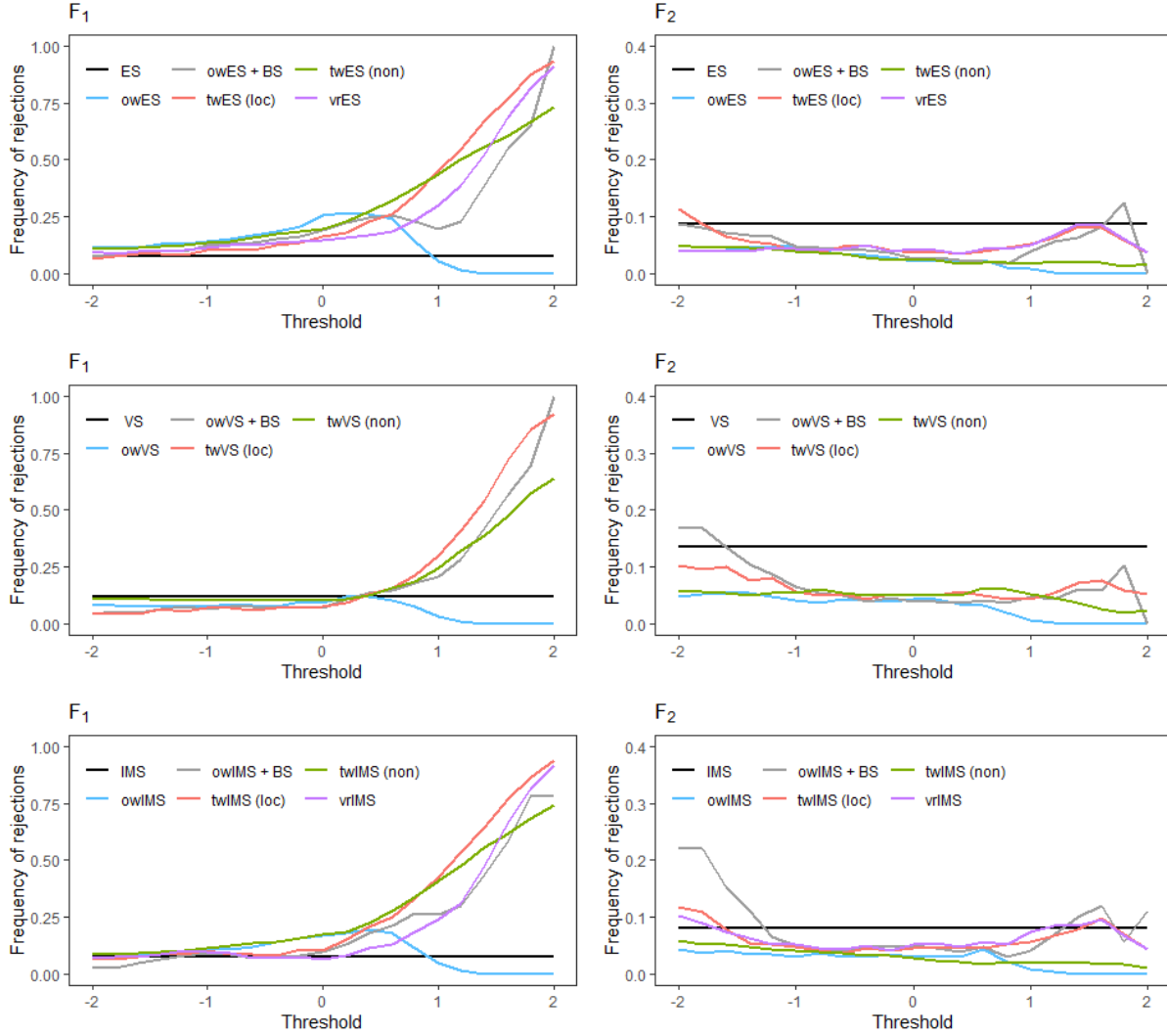
Figure 5: The proportion of instances that a Diebold-Mariano test for equal predictive performance is rejected in favour of F_1 (Left) and F_2 (Right) for each version of the energy score (Top), variogram score (Middle), and IMS (Bottom). The rejection rate is displayed as a function of the threshold used in the weight function $w(z) = \mathbb{1}\{z_1 \geq t, z_2 \geq t\}$ when evaluating the forecasts. Localising versions of the threshold-weighted scores are denoted by (loc), while non-localising variants are labelled (non). Note the different scales when considering F_1 and F_2.

Complementing these scores with the Brier score again generates scoring rules that are capable of identifying the differences between the forecasts.

The threshold-weighted scores are also adept at capturing the differences in forecast behaviour as the threshold is increased. This is true for both the localising and non-localising variants, though the localising score is generally more discriminative for large thresholds. For the energy score, the vertically re-scaled score is again slightly less informative than the localising threshold-weighted score, suggesting it is preferable to centre these scores close to the threshold of interest.

Figure 6 displays the corresponding results for the weight function $w(z) = \mathbb{1}\{z_1 + z_2 \geq t\}$. The weighted scoring rules in this case appear to be less discriminative than in the previous setting, though the conclusions are largely similar. The unweighted multivariate scores cannot distinguish between the two forecasts, whereas
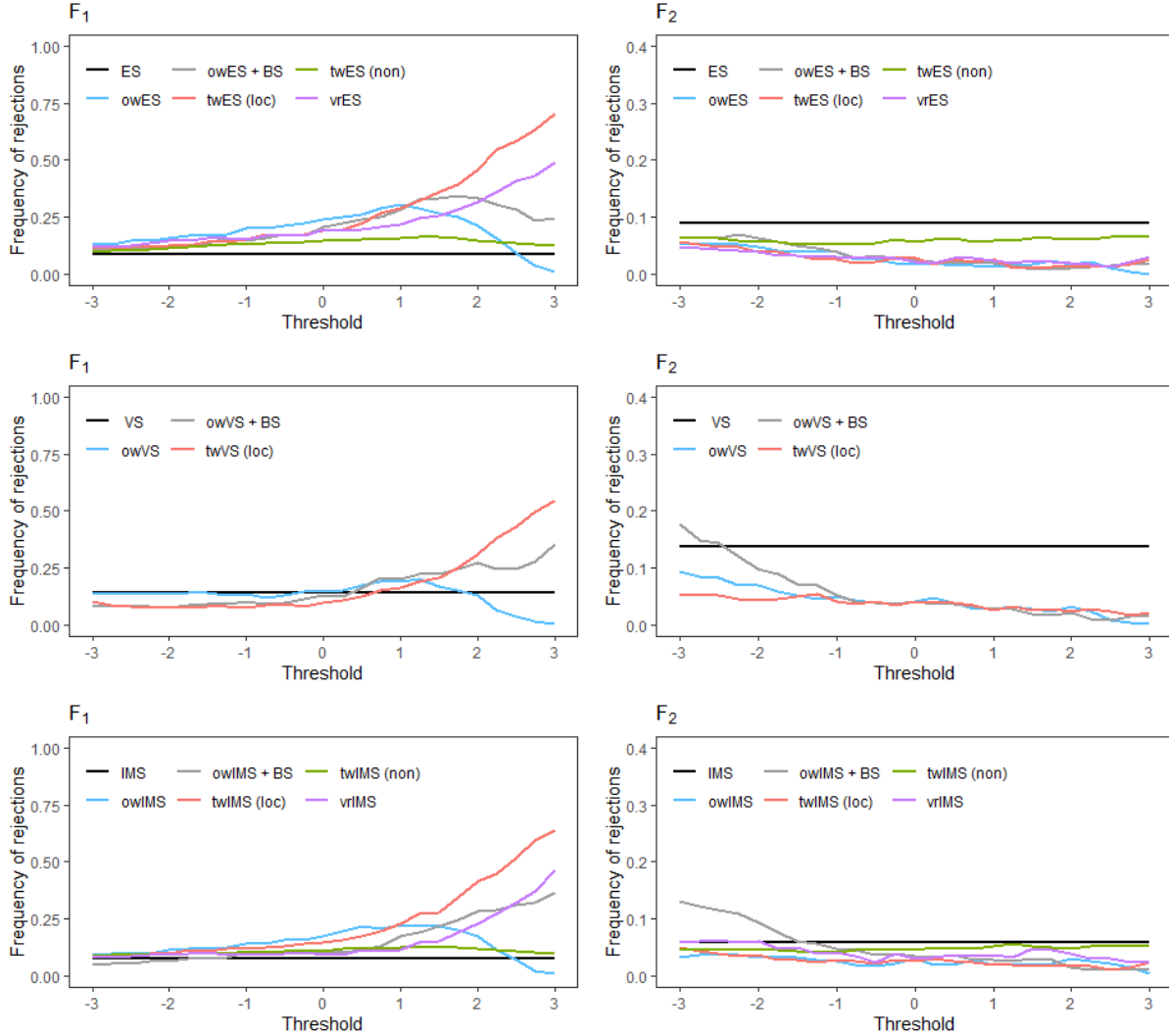
Figure 6: Same as Figure 5 for the weight function $w(z) = \mathbb{1}\{z_1 + z_2 \geq t\}$.

the weighted scores do so successfully. The exception to this is the non-localising threshold-weighted scores, which are only marginally more discriminative than the unweighted scores, regardless of the threshold used within the weight function. The reason for this is that the deformation function in this case projects points in $\{w = 0\}$ onto the line $z_1 + z_2 = t$, which still contains a lot of information. The resulting score is thus dominated by differences between points in $\{w = 0\}$.

5 Case study

5.1 Introduction

The simulation study in the previous section demonstrates the utility of weighted multivariate scoring rules when evaluating forecasts made for high-impact events. In this section, we seek to reinforce this by illustrating how these weighted scores can be applied in practice. In particular, the weighted energy, variogram and inverse multiquadric scores described previously are used to evaluate daily rainfall accumulation forecasts across several forecast lead times, with emphasis on events that could lead to flooding.

Flooding and other associated impacts could manifest, for example, from a large precipitation accumulation

19

Figure 7: The weather stations across Switzerland and the surrounding area at which precipitation forecasts are considered. The colour of each point reflects the mean daily accumulation at that station, measured in millimetres.

on a single day, or from moderate rainfall over consecutive days. Changing the weight used within the weighted multivariate scoring rules allows us to consider these different possibilities when evaluating forecasts. Whether or not an impact occurs will depend not only on the amount of rainfall, but also on other factors, such as the temperature or a location's capabilities to deal with large rainfall accumulations. These external factors are not considered in this study, though the weight or chaining function within the weighted scoring rules could possibly be adjusted to include this information. For example, a weight function could be used that employs different parameters depending on certain covariates or location-specific characteristics.

The daily precipitation accumulation forecasts considered here were issued by the Swiss Federal Office of Meteorology and Climatology's (Meteoswiss) COSMO-E ensemble prediction system. The forecasts and corresponding observations are therefore available at a large number of weather stations across Switzerland. The 245 stations are displayed in Figure 7, which also presents the average observed daily accumulation for each station over the period of interest, the four autumns seasons (September to November) between 2016 and 2019. This results in roughly 100,000 pairs of forecasts and observations. The forecasts are initialised at 00 UTC, and their performance is analysed over the three consecutive days following this initialisation time.

The COSMO-E prediction system issues forecasts in the form of a 21-member ensemble. However, operational ensemble forecasts made for surface weather variables are commonly found to be overconfident, or under-dispersed, exhibiting less spread than desired. This can be verified using rank histograms, which display the relative frequency that the observed precipitation accumulation is assigned each rank when pooled among the ensemble members. Rank histograms can thus be thought of as an empirical analogue of the probability integral transform (PIT) histogram (Dawid, 1984; Gneiting et al., 2007). If the ensemble is calibrated, then the observation should be equally likely to assume each possible rank, resulting in a uniform rank histogram. However, the rank histogram for the COSMO-E ensemble forecasts at a lead time of one day (aggregated across all stations) in Figure 8 shows that this is not the case, and the observation tends to fall above or below all ensemble members significantly more often than would be expected from a calibrated forecast.

As such, it is common for these dynamical forecasts to undergo some form of statistical post-processing. Since the physical mechanisms underlying flooding events may occur on timescales larger than one day, the COSMO-E forecasts are post-processed at individual lead times and then combined using copula approaches to generate multivariate forecasts for the precipitation accumulation for the following three days.

5.2   Statistical post-processing

The post-processing approach implemented here follows that proposed by Scheuerer and Hamill (2015a), which assumes that the daily precipitation accumulation follows a Gamma distribution that is shifted and censored below at zero; the shifting and censoring yields a zero-inflated distribution that captures the positive

probability that the precipitation is exactly zero. The mean and standard deviation of the censored, shifted Gamma distribution are assumed to depend linearly on the ensemble mean and ensemble standard deviation respectively, as is commonly assumed within the Non-homogeneous Regression, or Ensemble Model Output Statistics post-processing framework (Gneiting et al., 2005).

In particular, letting Y denote the daily precipitation accumulation at a particular station and lead time, and x the corresponding COSMO-E ensemble forecast, with mean $\bar{x}$ and standard deviation s, the post-processing model is

$$Y|x \sim \Gamma_0(\kappa, \theta, \xi), \qquad \mu = \kappa\theta = \alpha + \beta\bar{x}, \qquad \sigma = \theta\sqrt{\kappa} = \gamma + \delta s,$$

where $\Gamma_0(\kappa, \theta, \xi)$ denotes the Gamma distribution with shape κ and scale θ, shifted negatively by ξ and censored below at zero. The mean of the Gamma distribution is represented by μ, and the standard deviation by σ. These are linked to regression parameters α, β, γ and δ, which, along with ξ, are estimated using maximum likelihood estimation over a training data set. All parameters are constrained to be non-negative using a square-root link function.

The training data set consists of all forecasts issued during the autumn seasons of 2016 and 2017, while the resulting forecasts are then evaluated out-of-sample on the data available during 2018 and 2019. A separate post-processing model is fit to forecasts at each lead time and each station under consideration: the distributional assumptions are the same in each case, but separate sets of model coefficients are estimated. The right-hand panel of Figure 8 demonstrates that this post-processing method successfully re-calibrates the COSMO-E ensemble forecasts for daily precipitation accumulation on average. The calibration of the statistically post-processed forecasts does not change depending on the lead time, whereas the COSMO-E output becomes gradually less under-dispersed as forecast lead time increases.

Flooding could occur due to rainfall events that persist over several days, and in order to anticipate such an event, the forecasts should capture the temporal dependence in the precipitation observations. Hence, the post-processed forecasts for the daily precipitation accumulation at each lead time are combined into a single multivariate forecast distribution. This is achieved using a copula to model the dependence between the precipitation accumulation on consecutive days. Four different copulas are considered.

An independence copula assumes that there is no dependence between the precipitation accumulation on successive days (conditional on the ensemble output), whereas a comonotonic copula conversely assumes perfect dependence, so that heavy rainfall on one day is followed by heavy rainfall on the next day. These copulas therefore serve as convenient baseline approaches to which alternative methods can be compared. The third approach that we consider utilises an empirical copula based on the dependence structure observed in the COSMO-E ensemble forecast; it thus assumes that the numerical weather model correctly simulates
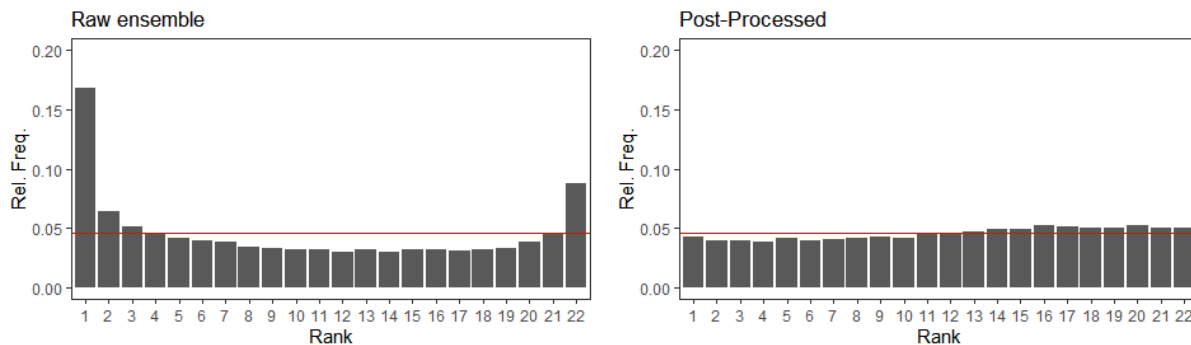

Figure 8: Rank histograms for the COSMO-E prediction system (Left) and the statistical post-processing method (Right) at a lead time of one day. A horizontal red line has been added to indicate perfect calibration. The ranks have been aggregated across all locations and all forecast instances in the test data set, and ties between ranks have been resolved at random.

the dependence structure observed in reality. This approach, called ensemble copula coupling (ECC), is well-established in the post-processing literature, and is commonly implemented in operational post-processing suites (Schefzik et al., 2013). Finally, we employ a Gaussian copula, which was found here to outperform alternative parametric copula families.

In all cases, 21 ensemble members are generated by sampling from the univariate post-processed distributions at equidistant quantiles, which are then reordered according to the four copulas. Hence, like the COSMO-E output, the resulting forecasts are in the form of a three-dimensional, 21 member ensemble forecast. This implementation of a Gaussian copula differs from previous applications in a post-processing context that draw a random sample from the copula (e.g. Möller et al., 2013; Lerch et al., 2020), which is then transformed using the quantile function of the univariate post-processed distributions. The new approach implemented and advocated here, described in detail in the Appendix, ensures that all multivariate post-processing methods exhibit the same marginal forecast performance.

The marginal forecast performance at each lead time is evaluated using the CRPS and the univariate IMS, while the multivariate forecast distributions are assessed with the energy score, the variogram score and the IMS. Since interest is predominantly on forecasts made for events that could lead to flooding, threshold-weighted and vertically re-scaled versions of these kernel scores are also employed, both in a univariate context for the individual daily accumulations, and in a multivariate context for the combined daily accumulations over three days.

Before evaluating the accuracy of the multivariate forecasts using these weighted scores, the calibration of the forecasts is assessed using multivariate rank histograms. Multivariate rank histograms provide a natural extension of the rank histograms in Figure 8, and a uniform histogram is again indicative of a calibrated forecast. Several approaches have been proposed to construct multivariate rank histograms and we implement the approach introduced by Gneiting et al. (2008), which is presented in Figure 9 for the four post-processing approaches.

In this case, the observation assumes the highest rank if it is larger than the forecast ensemble in all dimensions. Figure 9 therefore suggests that the comonotonic copula overestimates the dependence between precipitation on successive days, as expected, while the independence copula, ECC, and Gaussian copula approaches all result in forecasts that slightly underestimate this dependence. Indeed, this is known to be a disadvantage of ECC when forecasting precipitation, since several ensemble members often predict zero precipitation and are then reordered at random (Scheuerer and Hamill, 2018). For further details regarding the interpretation of these multivariate histograms, readers are diverted to Gneiting et al. (2008); Ziegel and Gneiting (2014).

5.3 Weighted scoring rules

Since the multivariate post-processed forecasts differ only in their choice of copula, all exhibit the same univariate forecast performance. Table 1 displays the CRPS and the IMS for these post-processed forecasts, as well as the COSMO-E model output. Both scores suggest that statistical post-processing is beneficial at all lead times considered here, though the improvement decreases as lead time increases. Table 1 also presents the threshold-weighted and vertically re-scaled versions of these two univariate scores. The weight in this case is an indicator function that emphasises daily precipitation accumulations that exceed 50mm, roughly corresponding to the 99th percentile of the observed accumulations in the training data set. The corresponding chaining function used within the threshold-weighted scores is $v(z) = \max(z, 50)$. Similarly to the unweighted scores, post-processing improves upon the raw ensemble forecast at all lead times, regardless of the weighted score used to assess the forecasts.

The multivariate forecasts made for the precipitation accumulation across the three lead times are evaluated using the energy score, the variogram score and the IMS, which are presented in Table 2. The comonotonic copula approach performs worst with respect to all unweighted scores, whereas the COSMO-E ensemble outperforms the post-processed forecasts when evaluated using the variogram score. Since the variogram
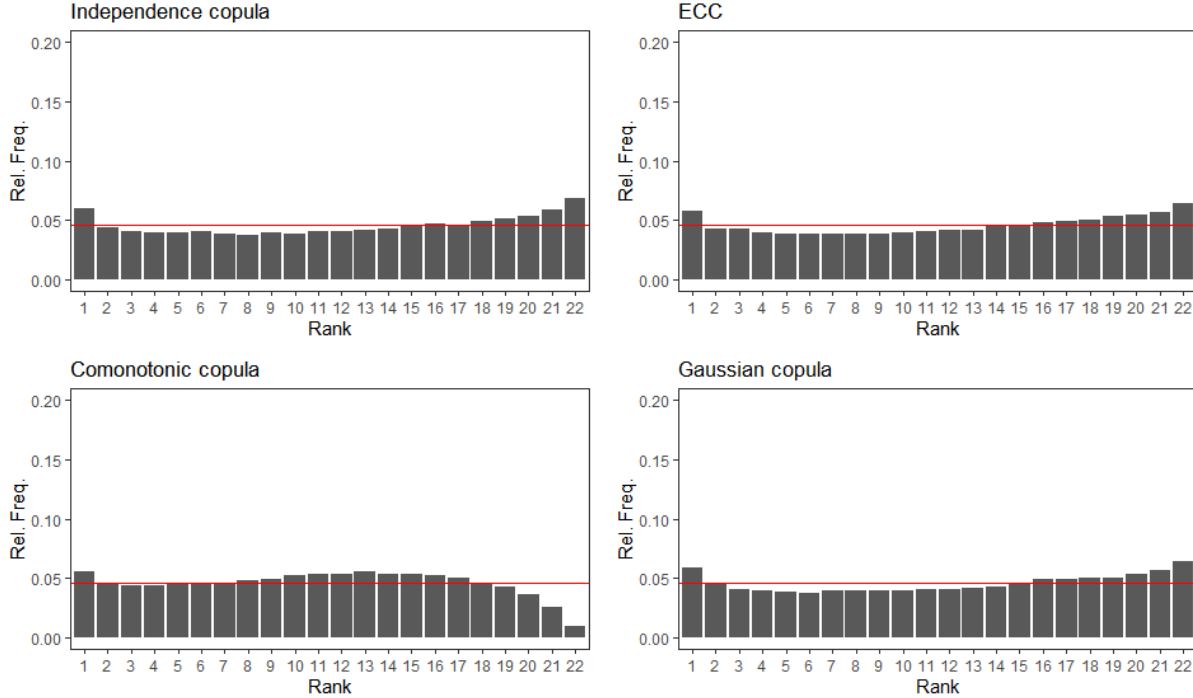
Figure 9: Multivariate rank histograms for the COSMO-E prediction system and the three copula-based multivariate post-processing methods. A horizontal red line has been added to indicate perfect calibration. The ranks have been aggregated across all locations and all forecast instances in the test data set, and ties between ranks have been resolved at random.

score is more sensitive to the forecast dependence structure than the energy score and the IMS, this result suggests that any improvements gained by post-processing are principally due to an improved univariate performance. As in Figure 9, the independence copula, ECC, and Gaussian copula approaches all perform similarly, suggesting the COSMO-E output already captures the majority of the dependence between the precipitation accumulation on successive days.

Results are also presented in Table 2 when evaluating forecasts using threshold-weighted and vertically re-scaled versions of these multivariate scores. Firstly, consider a weight function that is equal to one when the daily precipitation exceeds 25mm on the three consecutive days - this is labelled a "successive exceedance" in Table 2. As in the simulation study, two different chaining functions are used within the threshold-weighted energy and variogram scores, one of which results in a localising weighted score, while the other does not.

The conclusions drawn from the two chaining functions are not the same. The raw ensemble output and the comonotonic copula approach perform worst according to the non-localising threshold-weighted scores, regardless of whether the energy score, variogram score or IMS is considered. However, when evaluated using a localising threshold-weighted score, or a vertically re-scaled score, these two approaches typically outperform the other post-processing methods. This is particularly the case for the energy score, even though the raw ensemble and the comonotonic approaches perform worst with respect to the unweighted energy score.

This highlights that although one forecast strategy might result in the best overall forecast performance, this may not be the preferred approach when interest is on high-impact events. To reinforce this, Figure 10 displays the energy score against the (localising) threshold-weighted energy score for the five forecasting approaches. By analysing forecast performance with respect to multiple objectives, we can clearly see the trade-offs between the different approaches: the independence copula, ECC, and Gaussian copula generate

23

	1 day		2 days		3 days	
	Raw	Post proc.	Raw	Post proc.	Raw	Post proc.
CRPS	1.30 (0.02)	1.23 (0.02)	1.38 (0.02)	1.32 (0.02)	1.51 (0.02)	1.47 (0.02)
twCRPS	0.262 (0.011)	0.244 (0.010)	0.285 (0.011)	0.264 (0.010)	0.277 (0.010)	0.269 (0.010)
vrCRPS	0.642 (0.019)	0.609 (0.018)	0.676 (0.020)	0.635 (0.019)	0.680 (0.019)	0.665 (0.020)
IMS	1.41 (0.01)	1.30 (0.01)	1.40 (0.01)	1.33 (0.01)	1.47 (0.01)	1.42 (0.01)
twIMS	0.167 (0.005)	0.160 (0.005)	0.168 (0.004)	0.161 (0.005)	0.171 (0.004)	0.166 (0.005)
vrIMS	0.132 (0.004)	0.127 (0.004)	0.131 (0.004)	0.126 (0.004)	0.132 (0.004)	0.128 (0.004)

Table 1: Unweighted and weighted univariate scores for the raw COSMO-E output and the post-processed forecasts at each lead time. The scores have been averaged across all locations and all forecast instances in the test data set. Standard errors for the scores are shown in brackets, and the inverse multiquadric scores have been scaled by 10 to ease interpretation.

forecasts that outperform the COSMO-E ensemble when evaluated using the energy score, but this comes at the expense of forecast accuracy when interest is on high-impact events.

Consider now employing a weight function in these weighted scores that is equal to one if the total precipitation over all three days exceeds 75mm, and is equal to zero otherwise - this is labelled "total exceedance" in Table 2. In this case, the weight function acknowledges that flooding could result from moderate precipitation on consecutive days, or extreme precipitation on just a single day. The corresponding scores are displayed in Table 2. Since the simulation study in the previous section suggested that the non-localising variant of the threshold-weighted scores was not effective for this weight function, only the localising versions have been applied.

The independence copula, ECC, and Gaussian copula approaches generate the most accurate forecasts for these events, while the comonotonic copula and particularly the raw ensemble forecast perform comparatively poorly. This weight function depends less on the multivariate dependence structure than the previous weight function, and hence the results are more similar to those when an unweighted scoring rule is used to evaluate forecast performance.

6 Discussion

Evaluating forecasts with an emphasis on extreme events is an intrinsically challenging task (Taillardat et al., 2019; Brehmer and Strokorb, 2019). Nonetheless, weighted scoring rules have become the standard approach to do so. In this article, we have introduced and examined weighted scoring rules that can be applied to multivariate forecasts. We contend that high-impact events often result from the interaction of several
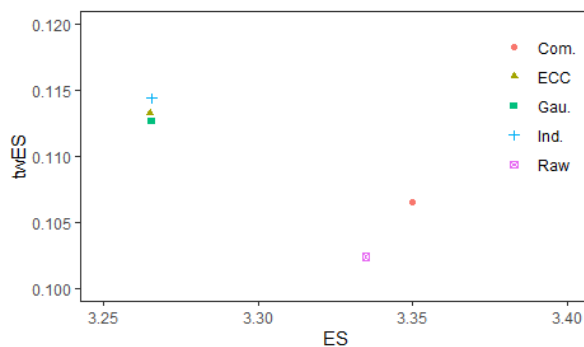

Figure 10: The energy score against the localising threshold-weighted energy score for the five multivariate forecast approaches.

Unweighted					
	Raw	Ind.	ECC	Com.	Gau.
ES	3.33 (0.03)	**3.27 (0.03)**	**3.27 (0.03)**	3.35 (0.03)	**3.27 (0.03)**
VS	**8.94 (0.08)**	9.11 (0.08)	9.12 (0.08)	9.88 (0.09)	9.11 (0.08)
IMS	26.8 (0.1)	**26.0 (0.1)**	**26.0 (0.1)**	26.9 (0.1)	**26.0 (0.1)**

Successive exceedance: $w(z) = \mathbb{1}\{z_1 \geq 25, z_2 \geq 25, z_3 \geq 25\}$					
	Raw	Ind.	ECC	Com.	Gau.
twES (non)	0.733 (0.018)	**0.681 (0.017)**	**0.681 (0.017)**	0.686 (0.018)	**0.681 (0.017)**
twES (loc)	**0.102 (0.011)**	0.114 (0.012)	0.113 (0.012)	0.107 (0.010)	0.113 (0.012)
vrES	**0.141 (0.014)**	0.162 (0.017)	0.161 (0.017)	0.149 (0.014)	0.160 (0.017)
twVS (non)	3.26 (0.08)	**3.01 (0.08)**	**3.01 (0.08)**	3.04 (0.08)	**3.01 (0.08)**
twVS (loc)	0.330 (0.041)	0.333 (0.042)	0.333 (0.042)	**0.318 (0.035)**	0.328 (0.042)
twIMS (non)	4.21 (0.07)	**4.05 (0.07)**	**4.05 (0.07)**	4.07 (0.07)	**4.05 (0.07)**
twIMS (loc)	**0.169 (0.016)**	0.179 (0.018)	0.178 (0.018)	0.175 (0.016)	0.178 (0.018)
vrIMS	0.123 (0.012)	**0.122 (0.012)**	**0.122 (0.012)**	0.125 (0.012)	**0.122 (0.012)**

Total exceedance: $w(z) = \mathbb{1}\{z_1 + z_2 + z_3 \geq 75\}$					
	Raw	Ind.	ECC	Com.	Gau.
twES (loc)	0.674 (0.028)	0.601 (0.027)	**0.600 (0.027)**	0.602 (0.026)	0.602 (0.027)
vrES	0.589 (0.025)	**0.535 (0.024)**	**0.535 (0.024)**	0.540 (0.023)	0.537 (0.024)
twVS (loc)	2.45 (0.13)	**2.11 (0.11)**	**2.11 (0.11)**	2.14 (0.11)	**2.11 (0.11)**
twIMS (loc)	0.906 (0.036)	**0.865 (0.037)**	0.866 (0.037)	0.876 (0.036)	0.867 (0.037)
vrIMS	0.621 (0.026)	**0.609 (0.026)**	**0.609 (0.026)**	0.616 (0.026)	**0.609 (0.026)**

Table 2: Unweighted and weighted multivariate scoring rules for each of the five prediction methods. The scores have been averaged across all locations and all forecast instances in the test data set. Standard errors for the scores are shown in brackets, and the best approach with regard to each score is shown in bold. For convenience, all variogram scores have been scaled by 10, and all inverse multiquadric scores by 100.

features, and the weighted multivariate scoring rules proposed herein therefore allow for a more thorough evaluation of forecasts with regards to high-impact events.

The weighted multivariate scoring rules developed here have been constructed by exploiting existing theory on conditionally negative definite kernels and the associated kernel score framework. In particular, it is shown that the well-known threshold-weighted continuous ranked probability score is a kernel score. The effect of this is two-fold: firstly, it permits forecasts in the form of a finite ensemble to be evaluated easily using the threshold-weighted CRPS; secondly, the threshold-weighted CRPS can be generalised for use with alternative kernels, thereby producing a broader class of threshold-weighted kernel scores.

In addition to this, well-known results on negative definite kernels have been leveraged in order to introduce a novel approach to weighting scoring rules. It is shown that this is equivalent to the threshold-weighting in particular circumstances, but in general extends the existing armory of weighted scores. Like the threshold-weighted kernel scores, these vertically re-scaled kernel scores also fall into the kernel score framework, making them applicable in a range of situations.

We explore these weighted scoring rules in the context of multivariate forecast evaluation, and compare them to alternative weighted scores previously proposed by Holzmann and Klar (2017). The energy score and the variogram score are the two most popular scoring rules when assessing multivariate forecasts, and both fall into the kernel score framework. Additionally, we consider the inverse multiquadric score, a new kernel score for both univariate and multivariate outcomes based on a bounded kernel which is competitive with respect to discrimination ability and has the advantage of being strictly proper with respect to all probability measures on $\mathbb{R}^d$. We introduce various weighted energy scores, variogram scores, and inverse multiquadric scores that

can emphasise particular outcomes when evaluating multivariate forecasts. We discuss possible ways that these kernel scores could be tuned in order to achieve this, and analyse the performance of the resulting scores in an application to simulated data in Section 4. Although highly idealised, this study clearly demonstrates the utility of these weighted scores, highlighting the additional information they provide relative to the unweighted scores.

It is then demonstrated how these weighted multivariate scoring rules could be applied in practice. In particular, several weighted multivariate scores are applied to MeteoSwiss daily precipitation accumulation forecasts, with a particular focus on events that could lead to flooding. Multivariate statistical post-processing methods are compared when forecasting these events, using both weighted and unweighted scores. Importantly, the conclusions drawn from the weighted scoring rules do not always coincide with those drawn from the unweighted scores, meaning the strategy that generates the most accurate forecasts is not necessarily the best when the goal is to predict flooding events. The weighted scoring rules can discriminate well between forecast distributions that exhibit contrasting behaviour in particular regions of the outcome space, something that unweighted scoring rules are not capable of.

Choosing a kernel that is relevant for a particular problem provides an extremely flexible approach to evaluate forecast performance, as illustrated by the fact that the variogram score is itself a kernel score. To this end, we believe that kernel scores have been underappreciated in the field of forecast verification. Kernels are used widely in several areas of mathematics and machine learning, and kernel scores could be employed to evaluate predictions made in a wide range of circumstances, including those for which established forecast verification tools do not currently exist.

Results are presented here for kernel scores that employ a fairly simple weight or chaining function, and future applications of these scores could consider more elaborate weights. Furthermore, we expect that a more comprehensive analysis of forecast performance for multivariate outcomes will also inspire new developments in post-processing methodology with the goal of avoiding effects such as improving overall forecast performance at the cost of deteriorating accuracy when considering high-impact events.

Acknowledgements

References

Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F. (2021). Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, 147:1403–1418.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic analysis on semigroups*. Springer, New York.

Bolin, D. and Wallin, J. (2019). Scale dependence: Why the average CRPS often is inappropriate for ranking probabilistic forecasts. *arXiv preprint arXiv:1912.05642*.

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13:4015–4034.

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

Chilès, J.-P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, volume 497. John Wiley & Sons, New York.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society: Series A*, 147:278–292.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93.

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

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

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

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In Meila, M. and Heskes, T., editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 258–267. AUAI Press.

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

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

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

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

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

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

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:211–235.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. Cambridge, MA: MIT Press.

Holzmann, H. and Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, 11:2404–2431.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *R News*, 1:27–29.

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

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

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S., and Graeter, M. (2020). Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, 27:349–371.

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

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

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1718–1727, Lille, France. PMLR.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305.

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

Micchelli, C. A. (1984). Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation Theory and Spline Functions*, pages 143–145. Springer, New York.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139:982–991.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10:1–141.

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

Pinson, P. and Tastu, J. (2013). Discrimination ability of the energy score. Technical report, Technical University of Denmark.

Rasmussen, C. E. and Williams, C. (2006). *Gaussian processes for machine learning.* Cambridge, MA: MIT Press.

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

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

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

Scheuerer, M. and Hamill, T. M. (2018). Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *Journal of Hydrometeorology*, 19:1651–1670.

Schölkopf, B. (2001). The kernel trick for distances. In *Advances in Neural Information Processing Systems 13*, pages 301–307. Cambridge, MA: MIT Press.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels.* Cambridge, MA: MIT Press.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41:2263–2291.

Steinwart, I. and Ziegel, J. F. (2021). Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272.

Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. *arXiv preprint arXiv:1905.04022*.

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

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

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1:333–347.

Appendix

Proof of Theorem 1

If ρ is measurable and negative definite, then the result is a restatement of Steinwart and Ziegel (2021, Theorem 1.1).

If ρ is a measurable c.n.d. kernel on $\mathcal{X}$ and $x_0 \in \mathcal{X}$, then by Berg et al. (1984, Lemma 2.1), the kernel

$$k(x, x') = \rho(x, x') + \rho(x_0, x_0) - \rho(x, x_0) - \rho(x', x_0) \tag{21}$$

is negative definite, and also measurable, since ρ is measurable. Assuming all the relevant integrals exist and are finite, it is a straightforward calculation to show that $S_k(P, y) = S_\rho(P, y)$. By Steinwart and Ziegel (2021, Theorem 1.1), S_k is a proper scoring rule with respect to the class of all probability measures $P \in \mathfrak{M}$ such that $\mathbb{E}_P[\sqrt{-k(X, X)}] < \infty$. Note that $-k(x, x)$ is non-negative, since $-k$ is positive definite. In particular, S_k is a proper scoring rule with respect to the smaller class of all probability measures $P \in \mathfrak{M}$ such that $\mathbb{E}_P[-k(X, X)] < \infty$.

We will now show that when $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, this class of measures coincides with $\mathcal{M}_\rho$. We have that

$$k(x, x) = \rho(x, x) + \rho(x_0, x_0) - 2\rho(x, x_0),$$

and hence $\rho(x, x) = 0$ for all $x \in \mathcal{X}$ implies that ρ is non-negative and thus a semi-metric. In this case, Sejdinovic et al. (2013, Proposition 20) state that $P \in \mathcal{M}_\rho$ if and only if $\mathbb{E}_P[-k(X, X)] < \infty$. Although this proposition is stated in the specific situation that $\mathcal{A}$ is the Borel σ-algebra with respect to a topology on $\mathcal{X}$, this is not needed for the arguments in the proof, so the proposition holds for any measurable space $(\mathcal{X}, \mathcal{A})$.

It follows that all integrals in Equation 3 are finite for $P \in \mathcal{M}_\rho$. This is the case, since, firstly, $\mathbb{E}_P[\rho(X, x_0)] < \infty$ for some $x_0 \in \mathcal{X}$ is equivalent to $\mathbb{E}_P[\rho(X, x_0)] < \infty$ for *all* $x_0 \in \mathcal{X}$, secondly, Equation 21 holds, and finally, $-k(x, x') \leq \sqrt{-k(x, x)}\sqrt{-k(x', x')}$.

Proof of Proposition 1

Let ν be a locally finite Borel measure on $\mathbb{R}$, and let $v : \mathbb{R} \to \mathbb{R}$ be a function such that $v(x) - v(x') = \nu([x', x))$ for any points $x, x' \in \mathbb{R}$. Then,

$$\rho(x, x') = |v(x) - v(x')| = \int_{\mathbb{R}} \left(\mathbb{1}\{x' \leq u < x\} + \mathbb{1}\{x \leq u < x'\} \right) \, \mathrm{d}\nu(u).$$

If X, X' are independent random variables distributed according to $P, Q \in \mathcal{M}_\rho$ respectively, with associated distribution functions F and G, then by the Fubini-Tonelli theorem,

$$\mathbb{E}_{F,G}|v(X) - v(X')| = \int_{\mathbb{R}} \left(\mathbb{E}_{F,G}\left[\mathbb{1}\{X' \leq u < X\}\right] + \mathbb{E}_{F,G}\left[\mathbb{1}\{X \leq u < X'\}\right]\right) \, \mathrm{d}\nu(u)$$

$$= \int_{\mathbb{R}} G(u)\left[1 - F(u)\right] + F(u)\left[1 - G(u)\right] \, \mathrm{d}\nu(u).$$

From this, it is straightforward to verify that if $X, X' \sim F$ are independent, then

$$\mathbb{E}_F|v(X) - v(y)| - \frac{1}{2}\mathbb{E}_F|v(X) - v(X')| = \int_{\mathbb{R}} \left(F(u) - \mathbb{1}\{y \leq u\}\right)^2 \, \mathrm{d}\nu(u) = \text{twCRPS}(F, y; \nu).$$

Since $\rho(x, x')$ is a c.n.d. kernel, this proves that the threshold-weighted CRPS is a kernel score.

To complete the proof of Proposition 1, it remains to show that this representation of the threshold-weighted CRPS is equivalent to a quantile scoring rule integrated over all $\alpha \in (0, 1)$. The general form of a quantile scoring rule is

$$QS_{v,\alpha}(F, y) = (\mathbb{1}\{F^{-1}(\alpha) \geq y\} - \alpha)(v(F^{-1}(\alpha)) - v(y))$$

for some increasing function v, where F^{-1} is the generalised inverse of F. Integrating this with respect to α gives

$$2\int_{(0,1)} (\mathbb{1}\{F^{-1}(\alpha) \geq y\} - \alpha)(v(F^{-1}(\alpha)) - v(y)) \, \mathrm{d}\alpha$$

$$= \int_{(0,1)} (2\mathbb{1}\{F^{-1}(\alpha) \geq y\} - 1)(v(F^{-1}(\alpha)) - v(y)) \, \mathrm{d}\alpha - \int_{(0,1)} (2\alpha - 1)(v(F^{-1}(\alpha)) - v(y)) \, \mathrm{d}\alpha$$

$$= \int_{(0,1)} |v(F^{-1}(\alpha)) - v(y)| \, \mathrm{d}\alpha - \int_{\mathbb{R}} (2F(x) - 1)(v(x) - v(y)) \, \mathrm{d}F(x),$$

$$= \int_{\mathbb{R}} |v(x) - v(y)| \, \mathrm{d}F(x) - \int_{\mathbb{R}}\int_{\mathbb{R}} (2\mathbb{1}\{x' \leq x\} - 1)(v(x) - v(y)) \, \mathrm{d}F(x') \, \mathrm{d}F(x).$$

The first term is equivalent to $\mathbb{E}_F|v(X) - v(y)|$ with $X \sim F$, while the latter can be rewritten as

$$I_2 = \mathbb{E}_F\left[(2\mathbb{1}\{X' \leq X\} - 1)(v(X) - v(y))\right]$$

$$= \mathbb{E}_F\left[(2\mathbb{1}\{X' \leq X\} - 1)(v(X) - v(X'))\right] + \mathbb{E}_F\left[(2\mathbb{1}\{X' \leq X\} - 1)(v(X') - v(y))\right]$$

$$= \mathbb{E}_F|v(X) - v(X')| - \mathbb{E}_F\left[(2\mathbb{1}\{X < X'\} - 1)(v(X') - v(y))\right]$$

$$= \mathbb{E}_F|v(X) - v(X')| - I_2,$$

where $X, X' \sim F$ are independent. Rearranging gives $I_2 = \mathbb{E}_F|v(X) - v(X')|/2$, which in turn yields

$$2\int_{(0,1)} (\mathbb{1}\{F^{-1}(\alpha) \geq y\} - \alpha)(v(F^{-1}(\alpha)) - v(y)) \, \mathrm{d}\alpha = \mathbb{E}_F|v(X) - v(y)| - \frac{1}{2}\mathbb{E}_F|v(X) - v(X')|,$$

as desired.

Proof of Proposition 2

Let $\mathcal{M}$ denote the set of Borel probability measures on $\mathcal{X} = \mathbb{R}$ with finite first moment. Let $F \in \mathcal{M}$ such that $\mathbb{E}_F[w(X)] > 0$, and define F_w as in Equation 8. Since $F \in \mathcal{M}$, we also have that $F_w \in \mathcal{M}$. Then, the outcome-weighted CRPS with weight function w can be written as

$$\text{owCRPS}(F, y; w) = w(y)\text{CRPS}(F_w, y)$$

$$= \mathbb{E}_{F_w}|Z - y|w(y) - \frac{1}{2}\mathbb{E}_{F_w}|Z - Z'|w(y),$$

$$= \frac{1}{C_w(F)}\mathbb{E}_F\left[|X - y|w(X)w(y)\right] - \frac{1}{2C_w(F)^2}\mathbb{E}_F\left[|X - X'|w(X)w(X')w(y)\right],$$

where $Z, Z' \sim F_w$ and $X, X' \sim F$ are independent.

Proof of Proposition 3

Let ρ be a c.n.d. kernel on $\mathcal{X}$, let $v : \mathcal{X} \to \mathcal{X}$ be a measurable function, and let $\tilde{\rho}(x, x') = \rho(v(x), v(x'))$. Suppose that S_ρ is a strictly proper scoring rule with respect to $\mathcal{M}_\rho$.

Firstly, assume that v is injective. Let $\tilde{P}, \tilde{Q} \in \mathcal{M}_{\tilde{\rho}}$. We wish to show that $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$ implies $\tilde{P} = \tilde{Q}$. To do so, define P and Q as the push-forward of v under $\tilde{P}$ and $\tilde{Q}$, respectively, i.e. $P(A) = \tilde{P}(v^{-1}(A))$ and $Q(A) = \tilde{Q}(v^{-1}(A))$ for all $A \in \mathcal{A}$. Note that, since v is injective, it is also bi-measurable. Additionally, for $\tilde{P} \in \mathcal{M}_{\tilde{\rho}}$, we have $\mathbb{E}_{\tilde{P}}[\rho(v(X), v(x_0)] < \infty$ for some $x_0 \in \mathcal{X}$. Letting $x_1 = v(x_0)$,

$$\mathbb{E}_{\tilde{P}}[\rho(v(X), v(x_0)] = \mathbb{E}_P[\rho(X, x_1)] < \infty,$$

and hence $P, Q \in \mathcal{M}_\rho$.

The divergence function of the kernel score associated with $\tilde{\rho}$ can then be written as

$$d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = \mathbb{E}_{P,Q}\left[\rho(X, Y)\right] - \frac{1}{2}\mathbb{E}_P\left[\rho(X, X')\right] - \frac{1}{2}\mathbb{E}_Q\left[\rho(Y, Y')\right]$$
$$= d_\rho(P, Q),$$

where $X, X' \sim P$ and $Y, Y' \sim Q$ are independent. Since S_ρ is strictly proper with respect to $\mathcal{M}_\rho$, $d_\rho(P, Q) = 0$ implies $P = Q$. Then, for any $A \in \mathcal{A}$, we have that

$$\tilde{P}(A) = \tilde{P}(v^{-1}(v(A))) = P(B) = Q(B) = \tilde{Q}(v^{-1}(v(A))) = \tilde{Q}(A),$$

where $B = v(A) \in \mathcal{A}$, and $v^{-1}(v(A)) = A$ holds for all $A \in \mathcal{A}$ since v is injective. Hence, $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$ implies that $\tilde{P} = \tilde{Q}$.

Conversely, assume that $\mathrm{tw}S_\rho$ is strictly proper with respect to $\mathcal{M}_{\tilde{\rho}}$. If v is not injective, then there exist distinct $z, z' \in \mathcal{X}$ such that $v(z) = v(z')$. Letting $\tilde{P} = \delta_z$ and $\tilde{Q} = \delta_{z'}$ be Dirac measures at z and z', respectively, we have

$$d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = \rho(v(z), v(z')) - \frac{1}{2}\rho(v(z), v(z)) - \frac{1}{2}\rho(v(z'), v(z')),$$
$$= \rho(v(z), v(z)) - \frac{1}{2}\rho(v(z), v(z)) - \frac{1}{2}\rho(v(z), v(z)) = 0.$$

That is, there exist distinct $\tilde{P}, \tilde{Q} \in \mathcal{M}_{\tilde{\rho}}$ such that $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$, which contradicts the assumption that $\mathrm{tw}S_\rho$ is strictly proper with respect to $\mathcal{M}_{\tilde{\rho}}$. Hence, if the kernel score associated with ρ is strictly proper with respect to $\mathcal{M}_\rho$, then the kernel score associated with $\tilde{\rho}$ is strictly proper with respect to $\mathcal{M}_{\tilde{\rho}}$ if and only if the chaining function v is injective. Similar arguments can be used in the case that S_ρ is strictly proper with respect to $\mathcal{M}^\rho$.

Proof of Proposition 4

Let ρ be a c.n.d. kernel on $\mathcal{X}$ such that $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, let w be a weight function, and let $v : \mathcal{X} \to \mathcal{X}$ be measurable. Set $\tilde{\rho}(x, x') = \rho(v(x), v(x'))$. For $P \in \mathcal{M}_{\tilde{\rho}}$, define $P_0 = P(\cdot \cap \{w = 0\})$ and $P_+ = P(\cdot \cap \{w > 0\})$, where $\{w > 0\} = \{x \in \mathcal{X}|w(x) > 0\}$ and $\{w = 0\} = \{x \in \mathcal{X}|w(x) = 0\}$. Since $\mathcal{X}$ can be partitioned into the union of $\{w > 0\}$ and $\{w = 0\}$, the threshold-weighted kernel score with kernel ρ and chaining function v can be decomposed as

$$\mathrm{tw}S_\rho(P, y; v) = \int_\mathcal{X} \rho(v(x), v(y)) \, \mathrm{d}P_0(x) - \frac{1}{2} \int_\mathcal{X} \int_\mathcal{X} \rho(v(x), v(x')) \, \mathrm{d}P_0(x) \, \mathrm{d}P_0(x')$$
$$+ \int_\mathcal{X} \rho(v(x), v(y)) \, \mathrm{d}P_+(x) - \frac{1}{2} \int_\mathcal{X} \int_\mathcal{X} \rho(v(x), v(x')) \, \mathrm{d}P_+(x) \, \mathrm{d}P_+(x') \quad (22)$$
$$- \int_\mathcal{X} \int_\mathcal{X} \rho(v(x), v(x')) \, \mathrm{d}P_0(x) \, \mathrm{d}P_+(x').$$

Suppose that $\rho(v(z), v(\cdot)) = \rho(v(z'), v(\cdot))$ for all $z, z' \in \{w = 0\}$. Note that this implies that $\rho(v(z), v(z')) = 0$ for all $z, z' \in \{w = 0\}$. In this case, $\mathrm{tw}S_\rho(P, y; v)$ simplifies to

$$
\begin{aligned}
\mathrm{tw}S_\rho(P, y; v) = &\, \rho(v(x_0), v(y))P(\{w = 0\}) \\
&+ \int_{\mathcal{X}} \rho(v(x), v(y)) \, \mathrm{d}P_+(x) - \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} \rho(v(x), v(x')) \, \mathrm{d}P_+(x) \, \mathrm{d}P_+(x') \\
&- P(\{w = 0\}) \int_{\mathcal{X}} \rho(v(x), v(x_0)) \, \mathrm{d}P_+(x),
\end{aligned}
$$

where x_0 is an arbitrary point in $\{w = 0\}$. Since any $P, Q \in \mathcal{M}_{\tilde{\rho}}$ are probability measures, if $P_+ = Q_+$, then $P(\{w = 0\}) = Q(\{w = 0\})$ and it becomes clear that $\mathrm{tw}S_\rho(P, y; v) = \mathrm{tw}S_\rho(Q, y; v)$ for all $y \in \mathcal{X}$. Hence, the threshold-weighted kernel score with such a kernel and chaining function is localising with respect to w.

Suppose now that $\mathrm{tw}S_\rho$ is localising with respect to w. Consider

$$
P = \frac{1}{2}\delta_z + \frac{1}{2}\delta_x, \qquad\qquad Q = \frac{1}{2}\delta_{z'} + \frac{1}{2}\delta_x,
$$

where $x \in \{w > 0\}$ and $z, z' \in \{w = 0\}$. Note that $P, Q \in \mathcal{M}_{\tilde{\rho}}$ and $P(\cdot \cap \{w > 0\}) = Q(\cdot \cap \{w > 0\})$. Since $\mathrm{tw}S_\rho$ is localising with respect to w, we have that $\mathrm{tw}S_\rho(P, y; v) - \mathrm{tw}S_\rho(Q, y; v) = 0$ for all $y \in \mathcal{X}$. From Equation 22, this means that

$$
\begin{aligned}
\mathrm{tw}S_\rho(P, y; v) - \mathrm{tw}S_\rho(Q, y; v) &= \frac{1}{2}\rho(v(z), v(y)) - \frac{1}{4}\rho(v(z), v(x)) - \frac{1}{2}\rho(v(z'), v(y)) + \frac{1}{4}\rho(v(z'), v(x)), \\
&= 0,
\end{aligned}
\tag{23}
$$

for all $z, z' \in \{w = 0\}$, $x \in \{w > 0\}$, and $y \in \mathcal{X}$.

Since this holds for all $y \in \mathcal{X}$, we can substitute $y = x$ to yield $\rho(v(z), v(x)) = \rho(v(z'), v(x))$ for all $x \in \{w > 0\}$, $z, z' \in \{w = 0\}$. Using this, and substituting $y = z'$ into Equation 23, we additionally have $\rho(v(z), v(z')) = 0$ for all $z, z' \in \{w = 0\}$. Hence, if $\mathrm{tw}S_\rho$ is localising with respect to w, then $\rho(v(z), v(x)) = \rho(v(z'), v(x))$ for all $z, z' \in \{w = 0\}$ and $x \in \mathcal{X}$.

Proof of Proposition 5

Let ρ be a c.n.d. kernel, let $v : \mathcal{X} \to \mathcal{X}$ be measurable, and define $\tilde{\rho}(x, x') = \rho(v(x), v(x'))$. Suppose that S_ρ is strictly proper with respect to $\mathcal{M}_\rho$.

Firstly, assume that the restriction of v to $\{w > 0\}$ is injective; that is, $v(z) \neq v(z')$ for all $z, z' \in \{w > 0\}$. Let $\tilde{P}, \tilde{Q} \in \mathcal{M}_{\tilde{\rho}}$. We wish to show that $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$ implies $\tilde{P}(\cdot \cap \{w > 0\}) = \tilde{Q}(\cdot \cap \{w > 0\})$. Define P and Q as the push-forward of v under $\tilde{P}$ and $\tilde{Q}$, respectively, i.e. $P(A) = \tilde{P}(v^{-1}(A))$ and $Q(A) = \tilde{Q}(v^{-1}(A))$ for all $A \in \mathcal{A}$. From the proof of Proposition 3, we know that $P, Q \in \mathcal{M}_\rho$. Also from this proof, we have that $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$ implies that $P = Q$. Then, for any $A \in \mathcal{A}$, we have that

$$
\tilde{P}(A \cap \{w > 0\}) = \tilde{P}(v^{-1}(v(A \cap \{w > 0\}))) = P(B) = Q(B) = \tilde{Q}(v^{-1}(v(A \cap \{w > 0\}))) = \tilde{Q}(A),
$$

where $B = v(A \cap \{w > 0\}) \in \mathcal{A}$, and $v^{-1}(v(A \cap \{w > 0\})) = A \cap \{w > 0\}$ for all $A \in \mathcal{A}$ since the restriction of v to $\{w > 0\}$ is injective. Hence, $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$ implies that $P = Q$, which in turn implies that $\tilde{P}(\cdot \cap \{w > 0\}) = \tilde{Q}(\cdot \cap \{w > 0\})$. The threshold-weighted kernel score with this loss function is therefore strictly locally proper with respect to w and $\mathcal{M}_{\tilde{\rho}}$.

Conversely, assume that $\mathrm{tw}S_\rho$ is strictly locally proper with respect to w and $\mathcal{M}_{\tilde{\rho}}$. If the restriction of v to $\{w > 0\}$ is not injective, then there exist distinct $z, z' \in \{w > 0\}$ such that $v(z) = v(z')$. If $\tilde{P} = \delta_z$ and $\tilde{Q} = \delta_{z'}$ are Dirac measures at z and z', respectively, then $d_{\tilde{\rho}}(\tilde{P}, \tilde{Q}) = 0$, even though $\tilde{P}(\cdot \cap \{w > 0\}) = \tilde{Q}(\cdot \cap \{w > 0\})$. This is a contradiction, and hence if $\mathrm{tw}S_\rho$ is strictly locally proper with respect to w and $\mathcal{M}_{\tilde{\rho}}$, then the restriction of v to $\{w > 0\}$ must be injective. Similar arguments can be used in the case that $\mathrm{tw}S_\rho$ is strictly proper with respect to $\mathcal{M}^{\tilde{\rho}}$.

Proof of Propositions 6 and 7

Let $-\rho$ be strictly integrally positive definite with respect to the maximal possible set of signed measures on $\{w > 0\}$ in the sense of Steinwart and Ziegel (2021, Definition 2.1), let $w > 0$ be a weight function, and let $\check{\rho}(x, x') = \rho(x, x')w(x)w(x')$.

Let $P, Q \in \mathcal{M}^\rho$ and let $\tilde{P}, \tilde{Q}$ be the measures on $\{w > 0\}$ that are absolutely continuous with respect to $P(\cdot \cap \{w > 0\})$ and $Q(\cdot \cap \{w > 0\})$, respectively, and density w. Note that $\tilde{P}, \tilde{Q}$ are finite measures on $\{w > 0\}$ but not necessarily probability measures. For the divergence function corresponding to $\mathrm{vr}S_\rho(\cdot, \cdot; w)$, we obtain that

$$0 = d_{\check{\rho}}(P, Q) = \int_{\{w>0\}} \int_{\{w>0\}} \rho(x, y)\, \mathrm{d}\tilde{P}(x)\, \mathrm{d}\tilde{Q}(y) - \frac{1}{2} \int_{\{w>0\}} \int_{\{w>0\}} \rho(x, x')\, \mathrm{d}\tilde{P}(x)\, \mathrm{d}\tilde{P}(x')$$
$$- \frac{1}{2} \int_{\{w>0\}} \int_{\{w>0\}} \rho(y, y')\, \mathrm{d}\tilde{Q}(y)\, \mathrm{d}\tilde{Q}(y')$$

implies that $\tilde{P} = \tilde{Q}$ due to $-\rho$ being integrally strictly positive definite on $\{w > 0\}$. This yields that $P(\cdot \cap \{w > 0\}) = Q(\cdot \cap \{w > 0\})$.

Proof of Proposition 8

Let ρ be a c.n.d. kernel on $\mathcal{X}$ with $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, and let w be a weight function such that $w(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$. Consider the chaining function $v(x) = xw(x) + x_0(1 - w(x))$, for $x, x_0 \in \mathcal{X}$.

For this chaining function, we have that

$$\rho(v(x), v(x')) = \rho(x, x')w(x)w(x') + \rho(x, x_0)w(x)(1 - w(x')) + \rho(x', x_0)w(x')(1 - w(x)).$$

Let $P \in \mathcal{M}_\rho$. Assuming all expectations are finite, substituting the above kernel into Equation 13 and rearranging gives

$$\mathrm{tw}S_\rho(P, y; v) = \mathbb{E}_P\left[\rho(X, y)w(X)w(y)\right] - \frac{1}{2}\mathbb{E}_P\left[\rho(X, X')w(X)w(X')\right]$$
$$+ (\mathbb{E}_P\left[\rho(X, x_0)w(X)\right] - \rho(y, x_0)w(y))(\mathbb{E}_P[w(X)] - w(y)),$$

which is the vertically re-scaled kernel score with kernel ρ, weight w, and centre x_0, as given in Equation 16.

Gaussian copula implementation

Copula analysis involves modelling a multivariate variable by first fitting distributions to each of the marginal variables, and then modelling the dependence between these marginal variables. In practice, this often requires sampling from the copula. If the copula is parametric, then this is typically performed randomly. However, it is sometimes beneficial to restrict the values of the copula that can be sampled. For example, by ensuring that the sampled points along each dimension must be equal to particular quantiles of the marginal distributions. That is, by restricting the domain of the copula from $[0, 1]^d$ to $\{\tau_1, \ldots, \tau_M\}^d$ for quantile levels $\tau_1, \ldots, \tau_M \in [0, 1]$, the canonical choice being $\tau_i = i/(M + 1)$ for $i = 1, \ldots, M$, which we have employed.

For example, in the field of statistical post-processing, using equidistant quantile levels typically results in a forecast ensemble that is more accurate than an ensemble generated at random from the copula (Schefzik et al., 2013; Lerch et al., 2020). As such, we introduce an approach that allows us to sample particular quantiles from the marginal distributions, and reorder them such that the dependence structure of the fitted copula is well-represented.

In theory, this could be achieved by computing the likelihood implied by the copula for all possible permutations of these M quantiles in each dimension, and choosing the permutation associated with the highest
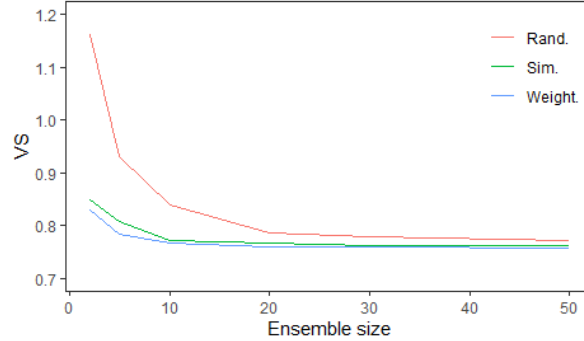
Figure 11: Variogram score against the ensemble size for three Gaussian copula-based post-processing methods applied to the MeteoSwiss COSMO-E ensemble forecasts of Section 5. The three approaches either sample randomly from the copula (Rand.), simulate combinations of particular quantile levels with probabilities proportional to the copula likelihood (Sim.), or weight each possible combination using this likelihood (Weight.).

likelihood. However, the number of such permutations is $(M!)^{d-1}$, making this approach computationally infeasible unless both M and the dimension d are very small.

Each permutation in the approach described in the previous paragraph is comprised of M combinations of the quantiles/quantile levels in the different dimensions. Instead, we calculate the likelihood of each individual combination of the quantile levels, reducing the number of evaluations to M^d. That is, we evaluate the copula density at the M^d points on the grid $\{\tau_1, \ldots, \tau_M\}^d$. A combination is sampled at random with probabilities that are proportional to the likelihood derived from the copula: combinations of the quantile levels that receive a high likelihood are more likely to be selected. Combinations that share the same quantile along any of the dimensions as the chosen combination are removed from consideration. This process is then repeated until we have M combinations. The independence copula and the comonotonic copula can both be interpreted within this general framework: for the independence copula, the probabilities corresponding to all combinations are the same, while for the comonotonic copula, the probability is one for the comonotonic combinations and zero otherwise.

Although this approach does not consider every possible permutation of the quantiles, it provides a computationally achievable alternative. When applied to the Gaussian copula-based post-processing method in the case study of Section 5, it is found to generate ensemble forecasts that are significantly more accurate than those constructed by sampling from the copula at random. Figure 11 shows the variogram score for these two approaches as a function of the ensemble size. The approach is particularly beneficial when the ensemble size is small (less than 20 members). For the results in Section 5, we choose M to be the number of members of the COSMO-E ensemble forecast, but we could easily have applied this approach with a larger number of ensemble members, which should result in better forecast performance.

Although this appears to be more beneficial than simulating randomly from the copula, sampling particular combinations of the quantiles still neglects information provided by the copula. As an alternative, rather than sampling from the set of combinations with probabilities proportional to the likelihood, these likelihoods could be used to weight each possible combination. In the post-processing set up, this would result in an ensemble forecast of M^d members, each of which is assigned a weight that is proportional to the likelihood. Figure 11 also shows the variogram score for this approach, which offers slight improvements upon the previously described approach.

34