Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events

Sam Allen[a, b] , Jonas Bhend[c] , Olivia Martius[b, d] , and Johanna Ziegel[a, b] ,

[a] *Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland*

[b] *Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland*

[c] *Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland*

[d] *Institute of Geography, University of Bern, Bern, Switzerland*

Corresponding author: Sam Allen, sam.allen@stat.unibe.ch

ABSTRACT: To mitigate the impacts associated with adverse weather conditions, meteorological services issue weather warnings to the general public. These warnings rely heavily on forecasts issued by underlying prediction systems. When deciding which prediction system(s) to utilise when constructing warnings, it is important to compare systems in their ability to forecast the occurrence and severity of high-impact weather events. However, evaluating forecasts for particular outcomes is known to be a challenging task. This is exacerbated further by the fact that high-impact weather often manifests as a result of several confounding features, a realisation that has led to considerable research on so-called compound weather events. Both univariate and multivariate methods are therefore required to evaluate forecasts for high-impact weather. In this paper, we discuss weighted verification tools, which allow particular outcomes to be emphasised during forecast evaluation. We review and compare different approaches to construct weighted scoring rules, both in a univariate and multivariate setting, and we leverage existing results on weighted scores to introduce conditional probability integral transform (PIT) histograms, allowing forecast calibration to be assessed conditionally on particular outcomes having occurred. To illustrate the practical benefit afforded by these weighted verification tools, they are employed in a case study to evaluate probabilistic forecasts for extreme heat events issued by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss).

1. Introduction

The impacts associated with adverse weather conditions are well-documented. To mitigate these impacts, meteorological services issue weather warnings that inform the general public when hazardous conditions are expected, and outline what action should be taken to minimise the associated risks. Operational warning systems typically account not only for how likely it is that a high-impact weather event will occur, but also for other factors, such as how the public will behave in response to a warning (WMO 2015). Evaluating the quality of a warning system is thus an intrinsically difficult task. However, if a warning system has access to more accurate forecasts for high-impact weather events, then it has the potential to generate more useful warnings. Methods to evaluate forecasts for these high-impact events can therefore play an integral role when developing warning systems.

It is important to distinguish between high-impact and extreme weather events. Extreme events are rare, relative to previously observed values, and are typically defined as maxima or exceedances of a relevant threshold. Of course, extreme weather events are of practical interest because they often result in large social and economic impacts. However, not all extreme events will generate a large impact. Instead, it has become common to study high-impact weather events, defined directly as events that result in a large (usually negative) societal impact, which may or may not be extreme from a statistical perspective. For example, in 2015, the World Meteorological Organisation (WMO) launched the High-Impact Weather Project, whose goal is to improve forecasts and warnings for high-impact weather (Majumdar et al. 2021). A key component of this project is to develop methods to evaluate the quality of forecasts and warnings for high-impact weather events.

Traditionally, the evaluation of weather forecasts focuses on two aspects of forecast performance: forecast calibration and forecast accuracy. Forecast calibration considers to what extent forecasts are reliable, or trustworthy - for example, do the observed outcomes occur with the same probability with which they are predicted? This is typically assessed visually using graphical diagnostic tools, such as reliability diagrams or rank histograms (Hamill 2001; Jolliffe and Stephenson 2012; Dimitriadis et al. 2021), though statistical tests also exist to check the calibration more rigorously (e.g. Wilks 2019; Arnold et al. 2021). Forecast accuracy, on the other hand, is a measure of the agreement between a forecast and the corresponding observation, and is quantified using proper scoring rules. Scoring rules summarise forecast performance using a single numerical value, allowing competing forecasters to be ranked and compared objectively, and proper scoring rules encourage the forecaster to issue what they truly believe will occur (Gneiting and Raftery 2007).

However, when interest is on particular outcomes, such as high-impact events, classical evaluation techniques risk raising the forecaster's dilemma; in particular, Lerch et al. (2017) remark that "if forecast evaluation proceeds conditionally on a catastrophic event having been observed, always predicting calamity becomes a worthwhile strategy." Gneiting and Ranjan (2011) demonstrate that a proper scoring rule is rendered improper if it is used to evaluate only the forecasts issued when particular outcomes have occurred, and Bellier et al. (2017) note that the forecaster's dilemma also applies to checks for forecast calibration. This raises questions regarding how forecasts for high-impact events should be assessed.

If only the occurrence of a high-impact event is of interest, then forecasts for this binary outcome can be evaluated using established verification tools: contingency table-based methods assess forecasts that are themselves binary (Stephenson et al. 2008; Ferro and Stephenson 2011), whereas probabilistic forecasts for the event occurrence can be evaluated using reliability diagrams and appropriate scoring rules. However, relatively few methods exist to evaluate forecasts for the severity of a high-impact event. Over the past decade, the canonical approach to achieve this has been to employ weighted scoring rules, which emphasise particular outcomes during forecast evaluation whilst circumventing the forecaster's dilemma (Gneiting and Ranjan 2011; Diks et al. 2011; Holzmann and Klar 2017).

While weighted scoring rules have been applied almost exclusively in univariate settings, they can also be used to place more weight on certain multivariate outcomes when assessing forecast accuracy (Allen et al. 2022). The application of weighted scoring rules in a multivariate context is particularly useful when evaluating forecasts for high-impact weather events, since such events are often inherently multivariate. In particular, high-impact weather may arise not only from an extreme event, but also from the interaction of several more moderate events; this has been the catalyst for numerous recent studies on so-called compound weather events (see Zscheischler et al. 2020, for a review).

Various approaches to construct weighted scoring rules have been proposed. In this paper, we discuss and compare these approaches, and provide guidance regarding which should be employed in different circumstances, both in a univariate and multivariate setting. While weighted scoring rules provide a measure of forecast accuracy when predicting extreme events, we demonstrate that the theory underlying these weighted scores can readily be applied to checks for forecast calibration. We then introduce a novel diagnostic tool that can assess the calibration of probabilistic forecasts conditionally on particular outcomes having occurred.

To demonstrate how they can be applied in practice, these weighted verification tools are applied to forecasts of extreme heat. Extreme heat provides a salient example of a compound weather event: while short and intense periods of extreme heat can have serious implications for human health (among other things), persistent hot periods further strain the human body by inhibiting its ability to recover (Basagaña et al. 2011). Most major weather services therefore issue warnings to the public when persistently high temperatures are expected, and a greater understanding of how well such events can be predicted would allow weather services to further refine their heat warning systems.

The remainder of the paper is structured as follows. In the following section, we review the general framework for forecast evaluation and introduce relevant weighted verification tools when evaluating forecasts for high-impact weather events. These are then applied to forecasts of extreme heat events in Section 3. In particular, operational forecasts issued by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) are compared to forecasts obtained from a statistical post-processing model, also introduced in Section 3, allowing us to analyse the effect of post-processing when forecasting extreme heat. The conclusions drawn from this case study are discussed in Section 4.

2. Forecast verification

a. Forecast accuracy

Suppose we are interested in forecasting a random variable Y that takes values in a set Ω, and that our forecasts are probability distributions on Ω. Let $\mathcal{F}$ denote a set of such distributions. A scoring rule S is a function that takes a forecast $F \in \mathcal{F}$ and an observation $y \in \Omega$ as inputs, and outputs a numerical value, or score, that quantifies the forecast accuracy. All scoring rules considered herein are negatively oriented, so that a more accurate forecast receives a lower score. A scoring rule S is called proper with respect to $\mathcal{F}$ if $\mathbb{E}_G[S(G,Y)] \leq \mathbb{E}_G[S(F,Y)]$ for all $F, G \in \mathcal{F}$, where $\mathbb{E}_G$ denotes the expectation over G, and strictly proper with respect to $\mathcal{F}$ if this holds with equality if and only if $F = G$. That is, if the observations are believed to arise according to a certain distribution, then the expected score is optimised when this distribution is issued as the forecast. We assume throughout that the expectations are finite where necessary.

Proper scoring rules exist to assess forecasts for a range of different outcomes (Gneiting and Raftery 2007). When considering high-impact events, it is common to reduce the problem to a binary forecasting task, whereby we are only interested in predicting whether or not the event of interest will occur. Although forecasts for such events could themselves be binary, it is more natural to issue forecasts that are probabilistic, thereby quantifying the uncertainty inherent in the prediction. One of the most popular scoring rules to evaluate such forecasts is the Brier score (Brier 1950). Consider the case where the outcome is univariate and real-valued, i.e. $\Omega = \mathbb{R}$, and the forecast F is a cumulative distribution function over the real line. A high-impact event might then be defined as an instance where the outcome exceeds a certain threshold t, in which case the Brier score is defined as

$$\mathrm{BS}(F, y; t) = (F(t) - \mathbb{1}\{y \leq t\})^2, \tag{1}$$

where $\mathbb{1}$ denotes the indicator function.

Of course, in considering only the occurrence of a high-impact event, we cannot assess how well forecasts predict the event's severity. While the Brier score evaluates the forecast at a particular threshold, the entire forecast distribution can be evaluated by integrating the Brier score over all possible thresholds. In doing so, we obtain the continuous ranked probability score (CRPS), the most commonly used scoring rule to evaluate probabilistic forecasts. If our forecast distribution F has a finite mean, the CRPS can be expressed as

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

where X and X' are independent random variables that follow the distribution F (Matheson and Winkler 1976; Gneiting and Raftery 2007).

The CRPS is regularly employed in climate-related studies, in part because it can readily be applied to ensemble forecasts by replacing the expectations in the second expression of Equation 2 with sample means over the ensemble members. The CRPS assesses the forecast distribution over the set of all possible outcomes, providing a measure of overall forecast performance, rather than evaluating forecasts made for high-impact events. Nonetheless, several extensions of the CRPS have been proposed that can emphasise particular outcomes whilst assessing forecast accuracy.

Weighted scoring rules

In order to emphasise high-impact events during forecast evaluation, a seemingly intuitive approach would be to only evaluate the forecasts issued when such an event occurs; more generally, to assign a higher weight to outcomes corresponding to higher impacts. However, Gneiting and Ranjan (2011) demonstrate that if proper scoring rules are weighted by a function that depends on the observed outcome, then the score is generally rendered improper. For example, if evaluation is restricted to instances where high-impact events occur, then the forecaster is encouraged to always predict that such an event will occur, even though such a forecast is uninformative in practice (Lerch et al. 2017).

Instead, weighted scoring rules have been introduced to target particular outcomes during forecast evaluation in a more theoretically desirable way. Weighted scoring rules incorporate a weight function into conventional scoring rules, but do such in such a way that the resulting score remains proper. The weight function can then be chosen to emphasise particular outcomes of interest. In the following, a weight function is a function w such that $w(z) \geq 0$ for all possible outcomes z. Gneiting and Ranjan (2011) list weight functions that could be used to emphasise certain real-valued outcomes, and these are given in Table 1.

Having chosen a suitable weight function for the problem at hand, several approaches have been proposed to incorporate this weight into existing scoring rules. Allen et al. (2022) list three possible methods to emphasise particular outcomes when evaluating forecasts using the CRPS. Firstly, the threshold-weighted CPRS (twCRPS) introduced by Matheson and Winkler (1976) and Gneiting and Ranjan (2011) is defined as

$$\begin{aligned} \text{twCRPS}(F, y; w) &= \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z) \mathrm{d}z \\ &= \mathbb{E}_F |v(X) - v(y)| - \frac{1}{2} \mathbb{E}_F |v(X) - v(X')|, \end{aligned} \tag{3}$$

where $X, X' \sim F$ are independent, and v is any function such that $v(z) - v(z') = \int_{z'}^{z} w(x) \mathrm{d}x$. Secondly, Holzmann and Klar (2017) proposed the outcome-weighted CRPS (owCRPS):

$$\begin{aligned} \mathrm{owCRPS}(F, y; w) &= w(y) \int_{-\infty}^{\infty} (F_w(z) - \mathbb{1}\{y \leq z\})^2 \mathrm{d}z \\ &= w(y) \mathrm{CRPS}(F_w, y), \end{aligned} \tag{4}$$

where

$$F_w(x) = \frac{\mathbb{E}_F \left[\mathbb{1}\{X \leq x\} w(X) \right]}{\mathbb{E}_F [w(X)]}, \tag{5}$$

with $X \sim F$. Lastly, Allen et al. (2022) introduced the vertically re-scaled CRPS (vrCRPS):

$$\begin{aligned} \mathrm{vrCRPS}(F, y; w, x_0) =& \mathbb{E}_F [|X - y| w(X) w(y)] - \frac{1}{2} \mathbb{E}_F [|X - X'| w(X) w(X')] \\ &+ (\mathbb{E}_F [|X - x_0| w(X)] - |y - x_0| w(y))(\mathbb{E}_F [w(X)] - w(y)), \end{aligned} \tag{6}$$

where $X, X' \sim F$ are independent, and x_0 is an arbitrary real value. Gneiting and Ranjan (2011) additionally introduced a quantile-weighted version of the CRPS, though this emphasises particular regions of the forecast distribution rather than particular outcomes, and is thus not considered here.

In all cases, the unweighted CRPS is recovered when the weight function is constant and equal to one. Of the above three approaches to weight the CRPS, the twCRPS is the most well-known, and has been applied in several studies to evaluate forecasts for extreme weather events (e.g. Lerch and Thorarinsdottir 2013; Allen et al. 2021b). The outcome-weighted CRPS, on the other hand, has been used to assess economic forecasts, but is relatively unknown within the field of weather and climate forecasting. An obvious question then is how these weighted scores differ from one another, and which (if any) should be preferred when evaluating forecasts for high-impact weather events? In this section, we seek to answer this question by providing a detailed comparison of the different approaches.

Firstly, consider how these weighted scores differ. As discussed, the CRPS is defined as an integral of the Brier score when predicting whether the observation will exceed a certain threshold, and the twCRPS simply assigns different weights to different thresholds in the integration. Note that the weight in the twCRPS depends on the variable of integration, rather than the observation. The second expression in Equation 3 demonstrates that the twCRPS can additionally be interpreted as the CRPS after having transformed the forecasts and observations, with the transformation v - which Allen et al. (2022) call the chaining function - governed by the choice of weight function.

In contrast to the twCRPS, the owCRPS employs a weight that depends on the outcome. Gneiting and Ranjan (2011) demonstrate that if the CRPS is weighted by a function that depends on the observed outcome, then the expectation of this weighted score, i.e. $\mathbb{E}_G[w(Y)\text{CRPS}(F,Y)]$ with $Y \sim G$, is minimised by issuing G_w as the forecast, rather than G, where G_w is defined analogously to F_w in Equation 5. This weighted scoring rule is therefore generally improper. To circumvent this, Holzmann and Klar (2017) suggest evaluating the forecasts via their weighted representation, providing an arguably more direct way of circumventing the forecaster's dilemma than the twCRPS.

Both the twCRPS and the owCRPS transform the forecasts and observations prior to implementing the unweighted CRPS, with the two approaches differing in the transformation they employ. Consider the common case where the weight function restricts attention to values above some threshold of interest t, i.e. $w(z) = \mathbb{1}\{z > t\}$. Figure 1 illustrates the difference between these two transformations for such a weight function. While the CRPS measures the distance between the observation and the entire forecast distribution, the twCRPS reassigns all probability assigned to values lower than the threshold to the threshold itself. This results in a left-censored distribution, with a point mass at the threshold of interest. In doing so, the score only depends on how the forecast behaves above the threshold. The owCRPS, on the other hand, truncates the distribution at the threshold, thereby evaluating the conditional distribution given that the threshold has been exceeded. This relies on the observation exceeding the threshold, and the owCRPS with this weight function is zero whenever this is not the case.

In considering this conditional distribution, the owCRPS is only sensitive to the shape of the forecast distribution above the threshold, and not to the forecast probability that the threshold will be exceeded; that is, the score cannot distinguish between two forecasts that have the same conditional distribution. It therefore only assesses the predicted severity of a high-impact event, whereas the twCRPS additionally accounts for the probability with which the event is predicted to occur. Holzmann and Klar (2017) suggest complementing the owCRPS by adding the score to a scoring rule for binary events, such as the Brier score, which can independently evaluate the

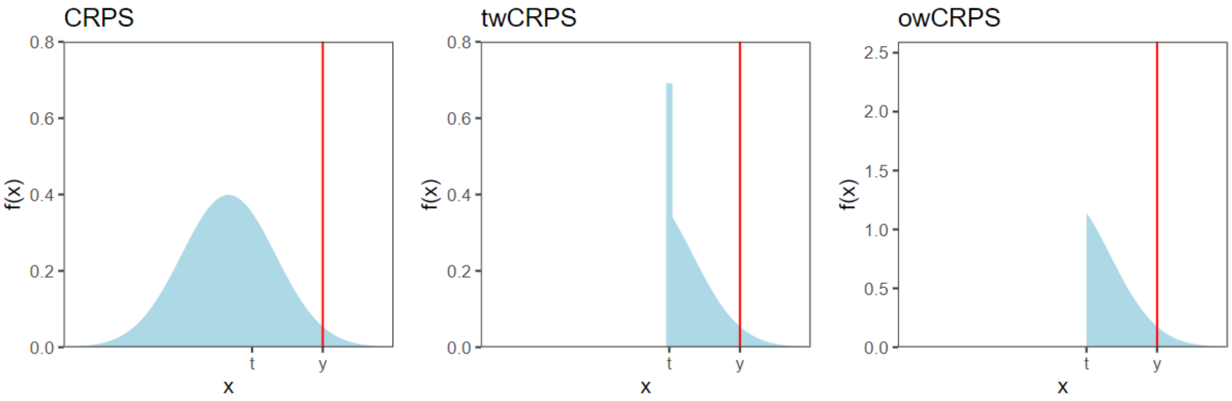

FIG. 1. Probability density functions used when calculating the CRPS, twCRPS, and owCRPS, with a weight function equal to $w(z) = \mathbb{1}\{z > t\}$. Note the contrasting scale for the owCRPS.

probability forecasts. For example,

$$\begin{aligned}\text{owCRPS}_{(\text{BS})}(F, y; w) &= \text{owCRPS}(F, y; w) + \mathbb{1}\{y \leq t\}(F(t) - \mathbb{1}\{y \leq t\})^2 \\ &= \mathbb{1}\{y > t\}\text{CRPS}(F_w, y) + \mathbb{1}\{y \leq t\}\text{BS}(F, y; t).\end{aligned} \tag{7}$$

A similar extension of the owCRPS is also possible when alternative weight functions are considered (Holzmann and Klar 2017).

However, even when complemented with such a score, the owCRPS does not consider the shape of the forecast distribution when the outcome does not exceed the threshold: in this case, if two forecasts assign the same probability to the exceedance of the threshold, then they will receive the same score, even if one predicts more severe events with a higher probability. Instead, this binary score could be replaced with a score that also accounts for the distance between the probability distribution and the threshold, penalising forecast distributions that assign higher probabilities to values much larger than the threshold. It turns out that this is in essence what the twCRPS does. In fact, if $w(z) = \mathbb{1}\{z > t\}$, it is straightforward to rewrite the twCRPS in terms of the owCRPS:

$$\begin{aligned}\text{twCRPS}(F, y; w) =& (1 - F(t))^2\text{owCRPS}(F, y; w) \\ &+ \mathbb{1}\{y > t\}\left[F(t)^2(y - t) + 2F(t)\int_t^y F(x) - F(t)\mathrm{d}x\right] \\ &+ \mathbb{1}\{y \leq t\}\int_t^\infty (F(x) - 1)^2\mathrm{d}x.\end{aligned} \tag{8}$$

For this weight function, the twCRPS thus differs from the owCRPS in two main respects. Firstly, when the outcome does not exceed the threshold of interest, the twCRPS still depends on the forecast, whereas the owCRPS is always equal to zero. While complemented versions of the owCRPS (e.g. Equation 7) address this, the twCRPS accounts not only for the probability that the threshold will be exceeded, but also the distance from the forecast distribution to this threshold. Secondly, when the threshold is exceeded by the outcome, the twCRPS is additionally comprised of two terms not present in the owCRPS, both of which penalise forecasts that issue a high probability that the outcome will not exceed the threshold.

The vrCRPS differs from the other two scores in that it does not transform the forecasts and observations, but rather weights the distance between them. In doing so, the vrCRPS depends not only on a weight function, but also on an additional parameter x_0. Although this could be construed as a practical disadvantage of the score, Allen et al. (2022) note that when $w(z) = \mathbb{1}\{z > t\}$, a canonical choice for this parameter is $x_0 = t$, in which case the vrCRPS is in fact equivalent to the twCRPS. When there is no canonical choice for x_0, it can arbitrarily be set equal to zero.

For this indicator-based weight function, a simple illustration of how the weighted scores behave is displayed in Figure 2. The forecast in this case is a standard normal distribution, and the scores are shown as a function of the observation. The CRPS clearly increases as the observed value moves away from the forecast mean, while all weighted scores are constant when the observation falls below the threshold of interest. The twCRPS and vrCRPS are proportional to the CRPS above this threshold, and the twCRPS has the desirable property that it is continuous: there is a jump in all other scores at the threshold, meaning a small difference in the observation can lead to a large change in the score. For the vrCRPS, the magnitude of this difference is controlled by x_0. When $x_0 = t$, the vrCRPS is also continuous, since the score is equivalent to the twCRPS (for comparison, the vrCRPS shown in Figure 2 employs $x_0 = 0$).

The twCRPS and vrCRPS additionally have the benefit that they can readily be applied to ensemble forecasts, or forecasts in the form of Monte-Carlo samples; as with the CRPS, this can be achieved by replacing the expectations in their definitions with sample means over the ensemble members (see also Allen et al. 2022). Note that for the twCRPS, the chaining function v is typically straightforward to calculate for the weights frequently employed in practice. The owCRPS, on the other hand, relies on the weighted forecast distribution F_w being well-defined (i.e. $\mathbb{E}_F[w(X)] > 0$),

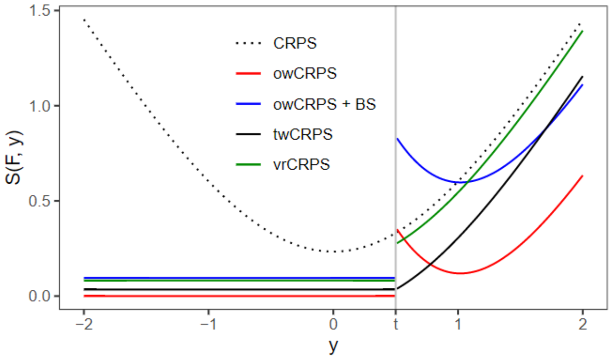

FIG. 2. CRPS and weighted versions of the CRPS for a standard normal distribution as a function of the observation y. The weight function $w(z) = \mathbb{1}\{z > t\}$ is used within the weighted scores, and a vertical grey line is shown at t.

which is often not the case if the weight function targets rare events and the forecast is an ensemble. In this case, it may be necessary to smooth the ensemble to form a continuous forecast distribution prior to assessing the forecasts. Although the use of strictly positive weight functions would ensure $\mathbb{E}_F[w(X)] > 0$ in theory, this can still lead to numerical complications in practice. Hence, when interest is on high-impact weather events, we recommend evaluating forecast accuracy using the twCRPS or vrCRPS.

MULTIVARIATE WEIGHTED SCORING RULES

Since high-impact weather often arises as a combination of weather events across multiple dimensions, forecasts for such events should be assessed using both univariate and multivariate techniques. Suppose now that $\Omega = \mathbb{R}^d$, for $d > 1$, and that the forecast F is a probability distribution on $\mathbb{R}^d$. Two of the most popular scoring rules to assess such forecasts are the energy score (ES; Gneiting and Raftery 2007) and the variogram score (VS; Scheuerer and Hamill 2015). The energy score is defined as

$$\mathrm{ES}(F, y) = \mathbb{E}_F||X - y|| - \frac{1}{2}\mathbb{E}_F||X - X'||, \tag{9}$$

where $||\cdot||$ is the Euclidean distance in $\mathbb{R}^d$, and $X, X' \sim F$ are independent (Gneiting and Raftery 2007). The energy score provides a natural generalisation of the CRPS to higher dimensions, and is commonly employed in practice since it can readily be applied to ensemble forecasts.

Values of interest	Univariate weight	Multivariate weight
All values	$w(x) = 1$	$w(x) = 1$
Central values	$w(x) = \phi_{\mu,\sigma}(x)$	$w(x) = \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x)$
Tail values	$w(x) = 1 - \phi_{\mu,\sigma}(x)/\phi_{\mu,\sigma}(\mu)$	$w(x) = 1 - \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x)/\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\mu})$
Right tail/Upper right quadrant	$w(x) = \Phi_{\mu,\sigma}(x)$	$w(x) = \boldsymbol{\Phi}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x)$
Left tail/Lower left quadrant	$w(x) = 1 - \Phi_{\mu,\sigma}(x)$	$w(x) = 1 - \boldsymbol{\Phi}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x)$

TABLE 1. Possible weight functions for univariate and multivariate weighted scores. Here, $\phi_{\mu,\sigma}$ and $\Phi_{\mu,\sigma}$ denote the density and distribution functions, respectively, of the Gaussian distribution with mean μ and standard deviation σ, while $\phi_{\mu,\boldsymbol{\Sigma}}$ and $\boldsymbol{\Phi}_{\mu,\boldsymbol{\Sigma}}$ denote the density and distribution functions, respectively, of the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

However, some studies have suggested that the energy score fails to be discriminative when comparing forecasts with different dependence structures (e.g. Pinson and Tastu 2013). Scheuerer and Hamill (2015) introduce the variogram score as an alternative scoring rule, which exploits the variogram commonly used in spatial statistics in order to directly target the forecast's multivariate dependence structure. The variogram score of order $p > 0$ is defined as

$$\mathrm{VS}_p(F, y) = \sum_{i=1}^{d} \sum_{j=1}^{d} h_{i,j} (\mathbb{E}_F |X_i - X_j|^p - |y_i - y_j|^p)^2, \tag{10}$$

where $y = (y_1, \dots, y_d) \in \mathbb{R}^d$, $X = (X_1, \dots, X_d) \sim F$, and $h_{i,j}$ are non-negative scaling parameters. In the following, p is chosen to be one half, as recommended by Scheuerer and Hamill (2015), and the scaling parameters $h_{i,j}$ are all set to one.

As in the univariate case, weighted versions of these scores exist that allow particular outcomes to be targeted during forecast evaluation. In this case, the weight functions should be defined on $\mathbb{R}^d$ rather than the real line. Gneiting and Ranjan (2011) propose several univariate weight functions based on Gaussian density and distribution functions, and weights to emphasise certain regions of the multivariate outcome space can be defined analogously in terms of multivariate Gaussian density and distribution functions. Some examples of such multivariate extensions are listed in Table 1. Of course, alternative weight functions could also be applied, and the most appropriate weight will depend on what information is to be extracted from the forecasts during evaluation.

The three approaches to generate weighted versions of the CRPS can also be applied to other scoring rules. It is possible to construct an outcome-weighted version of any proper scoring rule

(Holzmann and Klar 2017), while threshold-weighting and vertically re-scaling are applicable to the very general class of kernel scores (Gneiting and Raftery 2007; Allen et al. 2022). Since the energy score and variogram score both belong to the class of kernel scores, it is possible to introduce threshold-weighted, outcome-weighted, and vertically re-scaled versions of these multivariate scores, which can emphasise particular multivariate outcomes when evaluating forecast accuracy (Allen et al. 2022). For example, threshold-weighted energy and variogram scores can be defined as follows:

$$\mathrm{twES}(F, y; v) = \mathbb{E}_F ||v(X) - v(X')|| - \frac{1}{2}\mathbb{E}_F ||v(X) - v(X')||, \tag{11}$$

$$\mathrm{twVS}_p(F, y; v) = \sum_{i=1}^{d} \sum_{j=1}^{d} h_{i,j} (\mathbb{E}_F |v(X)_i - v(X)_j|^p - |v(y)_i - v(y)_j|^p)^2, \tag{12}$$

where $X, X' \sim F$ are independent, and $v : \mathbb{R}^d \to \mathbb{R}^d$. As with the twCRPS, these scores involve a transformation of the forecasts and observations prior to calculating the unweighted scores. Outcome-weighted and vertically re-scaled versions of these scores can similarly be introduced (see Allen et al. 2022, for details).

We can again consider how these weighted scores differ. Firstly note that, although the energy score and variogram score are arguably the most popular scoring rules to evaluate multivariate weather forecasts, other multivariate scoring rules exist, such as the logarithmic score and the Dawid-Sebastiani score (Dawid and Sebastiani 1999). While it is possible to construct outcome-weighted versions of these scores, these scores do not fit into the kernel score framework, and hence threshold-weighted and vertically re-scaled versions of these scores cannot readily be defined. This approach of outcome-weighting is therefore more general than threshold-weighting and vertically re-scaling. However, the outcome-weighted multivariate scores again rely on $\mathbb{E}_F[w(X)]$ being non-zero, and since multivariate weather forecasts almost exclusively take the form of ensembles, implementing outcome-weighted scores to evaluate forecasts for high-impact weather events becomes yet more challenging in a multivariate setting.

The threshold-weighted multivariate scores are defined in terms of a chaining function v that is used to transform the forecasts and observations. However, in contrast to the univariate case, there is no general framework with which to obtain a chaining function from a weight function on $\mathbb{R}^d$. Allen et al. (2022) show that if the weight function is always equal to either zero or one, then a canonical choice for the chaining function is

$$v(z) = \begin{cases} z & \text{if } w(z) = 1, \\ z_0 & \text{if } w(z) = 0, \end{cases} \tag{13}$$

where z_0 is an arbitrary point in $\mathbb{R}^d$. With such a weight function, the score will depend only on how the forecast distribution behaves at points z for which $w(z) = 1$. However, for more general weight functions, there is no obvious and general framework to construct a chaining function from a weight, and choosing a chaining function to emphasise the events of interest is somewhat less intuitive than selecting an appropriate weight function.

Conversely, the vertically re-scaled energy and variogram scores depend directly on a multivariate weight function. As a result, they can readily be applied with arbitrarily complex weight functions, without having to additionally define a relevant chaining function. This is a practical advantage of these weighted scores. Moreover, as in the univariate case, the vertically re-scaled scores are the same as the threshold-weighted scores for particular choices of the weight and chaining functions, and both classes of weighted scores can easily be applied to multivariate ensemble forecasts. Hence, due to the ease with which they can be implemented in practice, we generally recommend using vertically re-scaled multivariate scores to emphasise particular outcomes in multiple dimensions, though threshold-weighted scores are equally appealing if a canonical choice of the chaining function exists.

b. Forecast calibration

Although proper scoring rules allow competing prediction systems to be ranked and compared objectively, they cannot be used to determine whether a prediction system is trustworthy, in the sense that the observed outcomes are statistically consistent with the forecasts that were issued. If the forecasts do align with the observations, then the prediction system is said to reliable, or calibrated.

When the outcomes are univariate and real-valued, the most popular tool to assess forecast calibration is the rank or probability integral transform (PIT) histogram (Dawid 1984; Gneiting et al. 2007). PIT histograms rely on the result that if the outcome Y is a continuous random variable with cumulative distribution function F, then $F(Y)$ will follow the standard uniform distribution; a simple extension of this result exists for when Y is not continuous. Hence, to check for calibration, we can evaluate each forecast distribution function at the observed outcome, $F(y)$, and display these values in a histogram. If the observations are indeed draws from the corresponding forecast distributions, then the resulting histogram should be uniform. If the histogram is not uniform, then there is evidence to suggest the forecasts are miscalibrated, and the behaviour of the deviations can be used to diagnose the nature of the forecast errors (Hamill 2001).

Although forecast calibration in this setting could also be visualised using other techniques, one reason why PIT histograms are so commonly applied in practice is because there exists a discrete analogue when forecasts are in the form of an ensemble. So-called rank histograms display the relative frequency of the rank of the observation when pooled among the corresponding ensemble members (e.g. Hamill and Colucci 1997). If the prediction system is calibrated, then the observation should be equally likely to assume any rank on average, resulting in a uniform rank histogram. As with PIT histograms, the forecast miscalibration can be quantified by measuring the deviation between the observed histogram and a uniform histogram, and statistical tests for forecast calibration can then be derived by assessing whether or not this deviation is significantly large (Delle Monache et al. 2006; Wilks 2019; Arnold et al. 2021).

Conditional PIT histograms

The forecaster's dilemma also applies to diagnostic checks for calibration: if a rank or PIT histogram is constructed from only the forecasts issued when a high-impact event occurs, then the resulting histogram of a calibrated prediction system will in general not be uniform. For this reason, Bellier et al. (2017) "strongly advise against observation-based stratification when constructing rank histograms." While weighted scoring rules have been proposed to emphasise particular outcomes when calculating forecast accuracy, no similar extensions have been introduced when assessing forecast calibration. In this section, we leverage the previous discussion on weighted scoring rules to introduce conditional PIT (cPIT) histograms, which can be used to check the calibration of probabilistic forecasts conditionally on certain outcomes having occurred.

When interest is on particular real-valued outcomes, the outcome-weighted CRPS evaluates the conditional forecast distribution given that these outcomes have occurred. This conditional distribution can similarly be used within PIT histograms in order to assess forecast calibration for high-impact events. For example, let the outcome Y be a continuous (real-valued) random variable with distribution function F, and let $Y_{>t}$ denote the conditional outcome variable given that the outcome exceeds a threshold t; that is, $Y_{>t}$ follows the distribution G, where $G(x) = [F(x) - F(t)]/[1 - F(t)]$ for $x > t$, and $G(x) = 0$ otherwise. The probability integral transform $G(Y_{>t})$ then follows a standard uniform distribution. Hence, to evaluate the calibration of forecasts conditionally on the threshold being exceeded, we can calculate the conditional PIT values $G(y) = [F(y) - F(t)]/[1 - F(t)]$ for all observations y that exceed t, and display these values in a histogram. If the conditional distribution of Y is indeed the conditional distribution predicted by the forecasts, then the resulting histogram should be uniform, in which case the prediction system is said to be conditionally calibrated.

These cPIT histograms are not equivalent to focusing on the bins on the right-hand side of the standard PIT histogram, since an extreme observation could correspond to a low PIT value $F(y)$ if the forecast predicts more extreme events to occur with a high probability. To illustrate this, Figure 3 displays PIT and cPIT histograms for a perfect or ideal prediction system, as well as a histogram comprised of the PIT values that correspond to observations above a threshold of interest (labelled a restricted PIT histogram). The prediction system is calibrated, resulting in a uniform PIT histogram, but when interest is restricted to observations that exceed the threshold, the histogram becomes considerably skewed. The cPIT histogram, on the other hand, remains uniform, suggesting the forecasts are conditionally calibrated.

In theory, if the prediction system is calibrated, then it will additionally be conditionally calibrated, irrespective of the outcomes considered in the conditional PIT histograms. However, in practice, a forecast that appears calibrated may be significantly miscalibrated when more focus is put on particular outcomes. An example of this is presented in Section 3. Conversely, a forecast that is miscalibrated overall, leading to a non-uniform PIT histogram, may still be calibrated conditionally on the occurrence of a high-impact event.

If the cPIT histogram is not uniform, then the shape of the histogram can be used to infer what errors are present in the forecast. For example, suppose the observations are drawn from a logistic distribution with a random mean and fixed variance, and consider three competing forecasters: the first forecaster issues the normal distribution as a forecast, the second issues the logistic distribution,

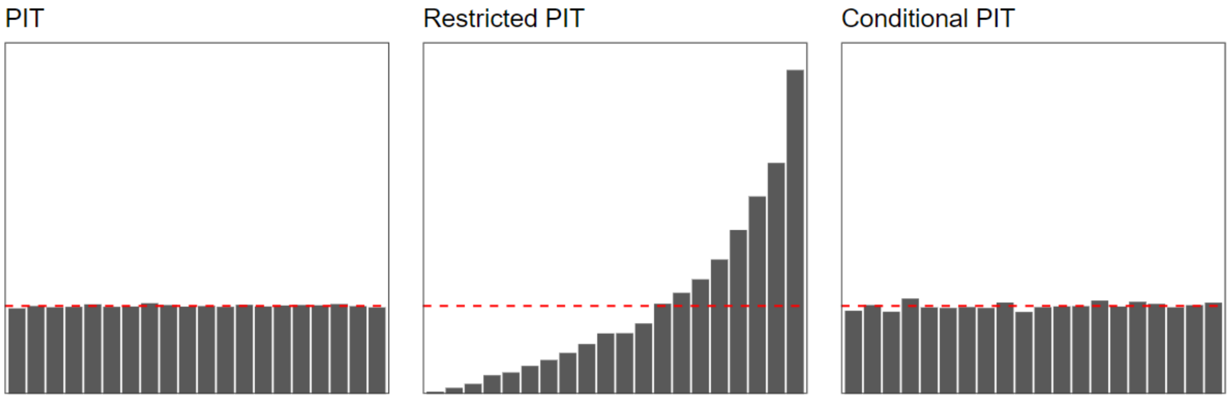

FIG. 3. PIT and conditional PIT histograms for an ideal forecaster, along with the PIT histogram constructed from only the observations that exceed a certain threshold. The histograms are comprised of 100,000 observations from a $\mathcal{N}(\mu, \sigma^2)$ distribution, where $\mu \sim \mathcal{N}(0, 1 - \sigma^2)$ and $\sigma^2 = 1/3$. A threshold of $t = 1$ is used within the weighted histograms.

and the final forecaster issues the Student's t distribution with five degrees of freedom, all of which are constructed to have the same mean and variance as the outcome distribution. Figure 4 displays the cPIT histograms for the three approaches, with a threshold equal to two. The logistic forecaster is the ideal forecaster, resulting in a uniform cPIT histogram, whereas the other two forecasters are oppositely biased: the normally distributed forecasts exhibit too light a tail, indicating a large proportion of the observations that exceed t fall in the tail of the conditional Gaussian distribution, while the Student's t distribution has a heavier tail than the logistic distribution, resulting in forecasts that over-predict the severity of extreme events.

While the number of bins to display in a PIT histogram is often chosen to equal the number of possible ranks within a reference ensemble prediction system, there is no canonical choice for the number of bins in a cPIT histogram. Hence, although histogram-based diagnostic tools are commonly employed to assess forecast calibration, we instead recommend visualising conditional calibration using PIT reliability diagrams (Gneiting and Resin 2021). PIT reliability diagrams display the empirical cumulative distribution function of the observed PIT values, and, as with standard reliability diagrams, a straight line along the graph's diagonal is indicative of a calibrated prediction system. Conditional PIT reliability diagrams analogously display the conditional PIT values, and cPIT reliability diagrams for the three forecasters in the previous example are presented in Figure 4.

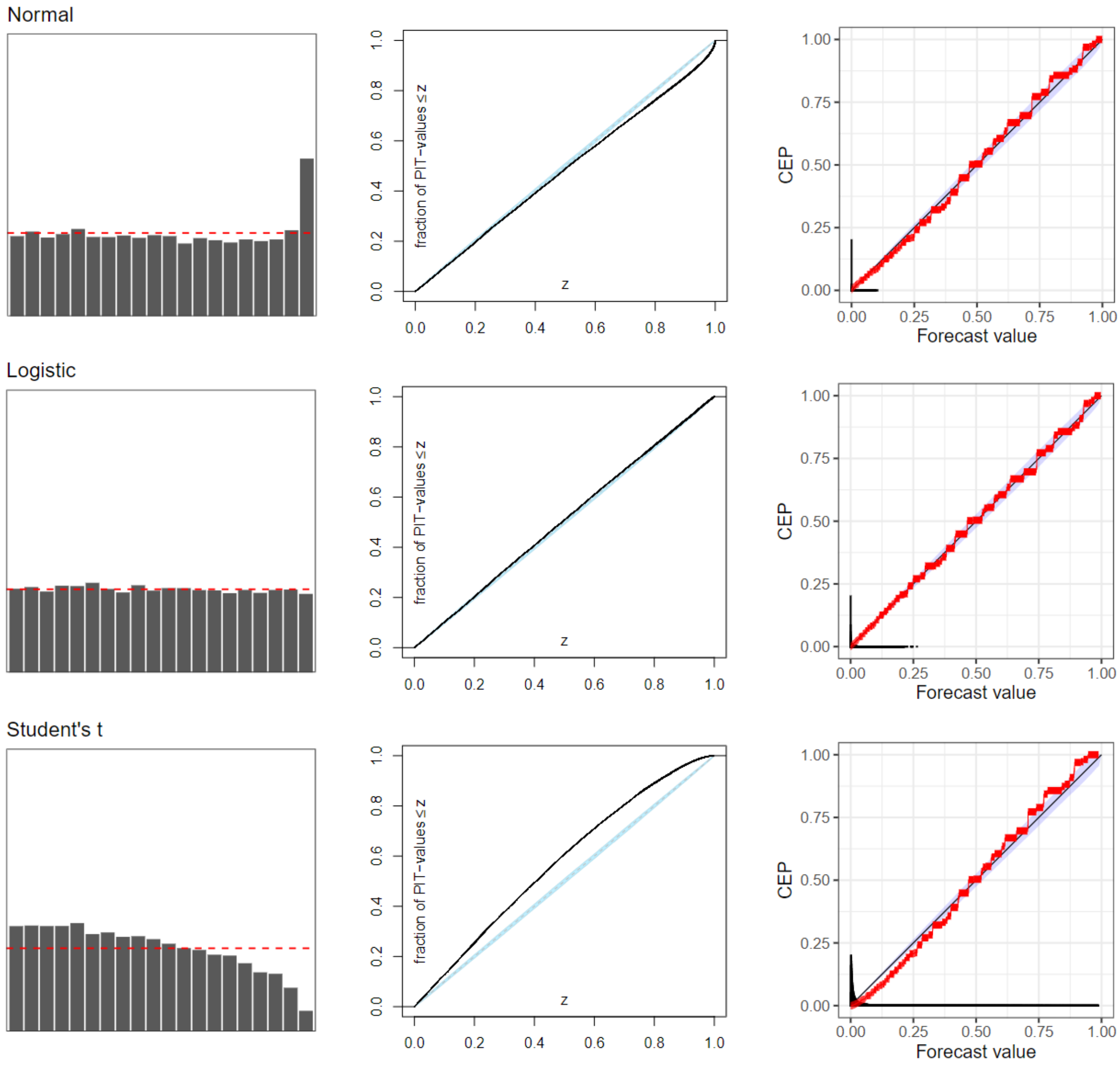

FIG. 4. Conditional PIT histograms (left) and conditional PIT reliability diagrams (middle) for forecast distributions with light (Normal), perfect (Logistic), and heavy (Student's t) tails. The histograms have been constructed using 1,000,000 observations from a logistic distribution, roughly 25,000 of which exceed the threshold $t = 2$. Standard reliability diagrams (right) also show the conditional event probabilities (CEP) given the forecast probability that the threshold will be exceeded. The blue shaded regions on the reliability diagrams are consistency intervals, constructed such that a calibrated prediction system would lie within these intervals 99% of the time.

Regardless of how the conditional calibration is visualised, by considering only the outcomes that exceed a threshold, these conditional diagnostic tools inherit some of the disadvantages associated with the outcome-weighted scores that were discussed previously. In particular, the outcome-weighted CRPS only assesses the shape of the conditional distribution, and does not consider the probability of a high-impact event occurring. This is also true for cPIT histograms and cPIT reliability diagrams, meaning they only evaluate the predicted severity of the high-impact event, and not the probability of occurrence. We therefore recommend that they are accompanied by a standard reliability diagram that separately assesses how well the forecasts can predict the occurrence of a high-impact event - akin to how Holzmann and Klar (2017) suggest complementing the owCRPS with a scoring rule for binary events. An illustration of this is presented in Figure 4 for the Gaussian, logistic, and Student's *t* forecasters. These reliability diagrams, constructed using the CORP approach proposed recently by Dimitriadis et al. (2021), highlight that, despite the differences when predicting event severity, the calibration of the three forecasters does not vary much when predicting the occurrence of a threshold exceedance.

Another disadvantage of the outcome-weighted CRPS is that it cannot easily be applied to ensemble forecasts when interest is on rare events, since the conditional forecast distribution is not always well-defined. Again, this also applies to checks for conditional calibration. As with the owCRPS, this could be addressed by smoothing the ensemble before assessing the calibration. However, generally speaking, we only advise employing checks for conditional calibration to ensemble forecasts when at least a reasonable number of ensemble members (say, 10) are expected to exceed the threshold of interest. While this limits their utility when evaluating ensemble forecasts when targeting high-impact events, cPIT histograms and cPIT reliability diagrams could still be useful when assessing the conditional calibration of ensemble forecasts relative to more moderate thresholds: for example, when interest is on precipitation accumulations that exceed zero.

Nonetheless, cPIT histograms and reliability diagrams provide a convenient and easily interpretable graphical approach to visualise calibration conditional on a high-impact event having occurred. While the interpretation of these diagnostic checks is similar to that for conventional checks for overall forecast calibration, formally testing whether a forecast is conditionally calibrated is less straightforward. In particular, the number of observations that exceed the threshold of interest is random and depends on the observed outcomes, rendering standard one-sample tests of uniformity invalid. Instead, more involved statistical tests are required that test for equality of conditional distributions, such as those commonly applied in the field of extreme value theory (Coles et al. 2001).

Throughout this section, the discussion has focused on high-impact events defined as the exceedance of a relevant threshold, i.e. corresponding to a weight function $w(z) = \mathbb{1}\{z > t\}$. In theory, cPIT histograms and reliability diagrams could be extended to more general weight functions. This would require identifying the random variable that follows the weighted distribution F_w in Equation 5. This will change for each weight function being considered, and is, in general, not a trivial task. However, we note that the weighted distribution F_w is generally not easy to interpret, and hence, even if we were able to construct the weighted PIT histogram corresponding to a general weight function, it would not be straightforward to use the resulting histogram to diagnose exactly what errors are present in the prediction system. For this reason, we restrict further attention to the cPIT histograms and reliability diagrams introduced above.

Multivariate conditional PIT histograms

Just as weighted scoring rules can be designed to target multivariate outcomes during forecast evaluation, the cPIT histograms introduced herein can also be extended to the multivariate case. However, there is no canonical definition of a multivariate rank or PIT histogram, and several contrasting approaches have been proposed to construct them (see Thorarinsdottir and Schuhen 2018). The general approach, as outlined by Ziegel (2017), is to define a pre-rank function, which condenses the multivariate forecasts and observations to univariate objects, and then to assess the calibration of these transformed forecasts using standard univariate rank or PIT histograms. The various approaches that have been proposed differ in their choice of pre-rank function.

Regardless of the chosen pre-rank function, we can straightforwardly adapt this approach to construct multivariate cPIT histograms that emphasise particular multivariate outcomes when assessing forecast calibration. In this case, consider a multivariate threshold of interest, $t \in \mathbb{R}^d$, and suppose we are interested in instances where the threshold is exceeded along all dimensions. By applying the pre-rank function to this multivariate threshold, we can obtain a univariate threshold. Note that this univariate threshold will change for each forecast case if the pre-rank function depends on the forecast, as it often does. Nonetheless, having obtained a univariate threshold, a cPIT histogram or cPIT reliability diagram can now easily be constructed as described in the previous section.

Note, however, that the challenges mentioned previously when assessing the conditional calibration of ensemble forecasts also apply here, and, since multivariate weather forecasts are more regularly in the form of finite ensembles, these issues will be yet more prevalent in the multivariate setting. In the case study presented in the following section, the multivariate forecasts are all ensemble forecasts, and hence we do not employ this approach to assess the conditional calibration of the multivariate forecasts.

3. Case study: evaluating heatwave forecasts

a. Extreme heat events

The verification techniques discussed in the previous section provide a means of evaluating forecasts with respect to high-impact events. In this section, we demonstrate the practical benefit afforded by these techniques by using them to evaluate operational weather forecasts for heatwaves and extreme heat events. The impacts associated with extreme heat events can be mitigated through effective early warning systems, and we define heat events using operational heat warning criteria adopted by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), determined following a recent study on how high temperatures affect human health in Switzerland (Ragettli et al. 2017).

MeteoSwiss issue heat warnings of three different levels, with a higher level associated with a higher impact. All heat levels are defined in terms of the daily mean temperature over a three day period, as summarised in Table 2. For completeness, Table 2 also includes a level one heat event, synonymous with the occurrence of low or moderate temperatures; the four levels therefore comprise an exhaustive set of the possible daily mean temperatures over three days. As expected, non-dangerous heat occurs on the vast majority (97%) of instances, whereas the most severe heat level occurs just 0.04% of the time. The MeteoSwiss warning levels do not change depending on the location, thereby assuming that the dangers associated with heat events do not vary substantially within the relatively small country of Switzerland.

Although these heat warning levels are specific to Switzerland, similar definitions of extreme heat are employed at other national weather centres (see e.g. McCarthy et al. 2019). This follows

Heat level	Criterion	Rel. Freq. (%)
1	T < 25°C on all three days	97.12
2	T ≥ 25°C on one or two days	2.40
3	T ≥ 25°C on all three days, T < 27°C on at least one day	0.45
4	T ≥ 27°C on all three days	0.04

TABLE 2. MeteoSwiss heat warning levels given daily mean temperatures (T) over a three day period, and the relative frequency with which each level occurs in the data under consideration.

WMO guidelines, which recommend that a heatwave be defined as “a period of marked unusual hot weather over a region persisting for at least three consecutive days during the warm period of the year based on local climatological conditions, with thermal conditions recorded above given thresholds” (WMO 2018). We highlight, however, that the weighted verification tools discussed in the previous section can readily be applied using other definitions of extreme heat (and high-impact events more generally), and they therefore provide a very flexible way to perform user-oriented forecast evaluation.

b. Data

Since these heat event definitions depend only on the daily mean temperature, we study forecasts for this weather variable. We consider daily mean temperature forecasts obtained from an operational ensemble prediction system at MeteoSwiss, which is based on a high-resolution numerical weather prediction (NWP) model from the Consortium for Small-Scale Modeling (COSMO-E). The COSMO-E model operates at a horizontal resolution of 2.2km over Switzerland and the surrounding area, and produces ensemble forecasts comprised of 21 members, all of which are initialised at 00 UTC in this study. Further details regarding COSMO-E are provided by Keller et al. (2021) and references therein.

However, even high-resolution NWP models are unable to resolve Switzerland’s complex topography, leading to large temperature biases on valley-floors and mountain-tops. To account for this, a simple lapse-rate bias correction is added to the COSMO-E forecasts, which takes into account the difference between the height of the model at each location and the true altitude; we assume a constant lapse-rate of 0.6 degrees Celsius per 100m.

The forecasts are assessed against observational temperature records at 149 weather stations across Switzerland, with the gridded COSMO-E output interpolated to individual stations using a

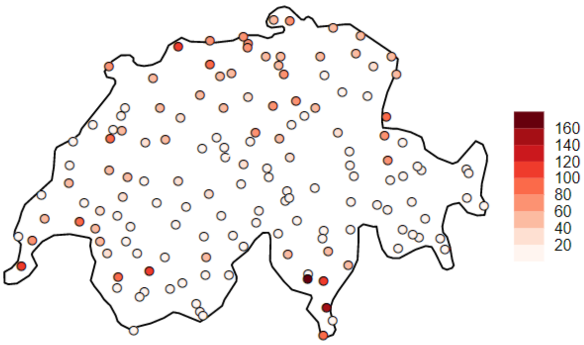

Fig. 5. Frequency of heat events of level two or greater at each station of interest, over all seven summer seasons under consideration.

nearest grid-point approach. These stations are all operated by MeteoSwiss and subject to rigorous quality control procedures. The stations are displayed in Figure 5, along with the number of extreme heat events (i.e. level two or greater) that occur at each station during the period of interest. Although there are several stations at which an extreme heat event does not occur, all stations are utilised in the subsequent analysis since forecasts should also be assessed in their ability to predict when a high-impact event will not occur.

Forecasts and observations are available for the seven year period between 2014 and 2020, and we restrict attention to extended summer seasons (May-September) in order to focus on extreme heat. This results in roughly 150,000 forecast-observation pairs to analyse at each forecast lead time. The COSMO-E forecasts extend out to five days, but since the heatwaves are defined over a three day period, forecasts are only considered over the coming three days. The forecasts are evaluated at each lead time separately using univariate verification techniques, while multivariate tools are used to assess the forecasts over the entire three day period. In doing so, forecasts can be evaluated in their ability to predict the temporal evolution of the daily mean temperature, which is key when focus is on heatwaves.

The COSMO-E ensemble forecasts are compared to two alternative forecast strategies: a climatological forecast, which always issues the local climatological temperature distribution as the prediction, and a statistically post-processed forecast, designed to remove systematic errors that occur in the COSMO-E forecasts. The post-processing method is based on an approach employed at MeteoSwiss, described in detail in the appendix.

c. Results

Overall forecast performance

The accuracy of the three prediction systems at each lead time is assessed using the CRPS. The scores for the three methods, averaged over all forecast cases and stations, are displayed in Table 3. As expected, the climatological forecast performs considerably worse than the approaches that utilise the COSMO-E output, while post-processing offers improvements upon the raw model output at all lead times.

The post-processed forecasts are consistently around 16% more accurate than the raw COSMO-E forecasts. To account for sampling uncertainty in this measurement, Table 3 additionally presents 95% confidence intervals for the relative improvement of all methods upon the COSMO-E forecasts. These confidence intervals have been obtained using non-parametric block bootstrapping, which accounts for both temporal and contemporaneous dependencies between the errors of the different forecast methods. A temporal block size of 30 days is used, though almost identical intervals are obtained using suitably smaller and larger block sizes. Further details about the block bootstrapping implemented here can be found in Wilks (2019) and Gilleland (2020).

The energy score and variogram score are also displayed in Table 3. The climatological forecasts again perform considerably worse than the two other methods, while the post-processed forecasts significantly outperform the COSMO-E ensembles when assessed using the energy score, with a relative improvement similar to that obtained from the CRPS. However, post-processing does not provide any benefit with respect to the variogram score. Since the variogram score is more sensitive to the forecast dependence structure, this suggests that the benefit of post-processing is largely due to improvements in the univariate forecast distributions, rather than in the multivariate dependence structure.

The calibration of the competing prediction systems is assessed using rank and PIT histograms, which are displayed in Figure 6. The results are shown at a lead time equal to three days, though similar conclusions are drawn at other lead times. The COSMO-E forecasts are considerably under-dispersed on average, which is commonly the case for operational weather forecasts for surface weather variables, while the post-processing model generates forecasts that are considerably better-calibrated. The climatological forecasts are also well-calibrated.

These rank and PIT histograms have been constructed from the forecasts and observations at all stations of interest. However, due to the complex topography of Switzerland, forecast calibration will likely change depending on the location. Figure 6 additionally contains a reliability index corresponding to each of the 149 stations, as a function of the station latitude and altitude (defined as the height above sea level). The reliability index, introduced in Delle Monache et al. (2006, Equation 1), measures the absolute deviation of the bars in the histogram from the uniform red line: the index is therefore minimised at zero, with larger values indicating more severe miscalibration. The reliability index for the COSMO-E forecasts tends to be marginally smaller at higher altitudes than lower altitudes, though the improvement in calibration gained by post-processing appears to be fairly insensitive to the station's location. The climatological forecasts produce yet smaller reliability indices.

Predicting heatwave severity

In the univariate case, to evaluate how well the forecasts capture the severity of the extreme heat events, the three weighted versions of the CRPS are employed at each lead time. Figure 7 displays these scores at a lead time of three days as a function of the threshold employed in the weight function $w(z) = \mathbb{1}\{z > t\}$, which emphasises events that exceed the threshold t. The owCRPS has been complemented with the Brier score (Equation 7), and, to ensure this score is well-defined, the COSMO-E ensembles are smoothed using a normal distribution prior to calculation. The additional parameter in the vrCRPS is set to $x_0 = 0$.

	CRPS			ES	VS
	1 day	2 days	3 days		
Clim.	2.36 $[-1.40, -1.03]$	2.36 $[-1.36, -0.97]$	2.37 $[-1.40, -1.02]$	4.44 $[-1.34, -1.03]$	2.40 $[-0.88, -0.66]$
COSMO	1.05 $[0.00, 0.00]$	1.08 $[0.00, 0.00]$	1.07 $[0.00, 0.00]$	2.02 $[0.00, 0.00]$	1.36 $[0.00, 0.00]$
Post-proc.	0.88 $[0.14, 0.18]$	0.92 $[0.13, 0.18]$	0.92 $[0.11, 0.17]$	1.76 $[0.11, 0.15]$	1.37 $[-0.02, 0.02]$

Table 3. The CRPS (at each lead time), energy score, the variogram score for the climatological, COSMO-E, and post-processed forecasts. The scores have been aggregated over all years and stations. Below each score is a 95% confidence interval for the corresponding skill score, with the COSMO-E forecasts used as reference.

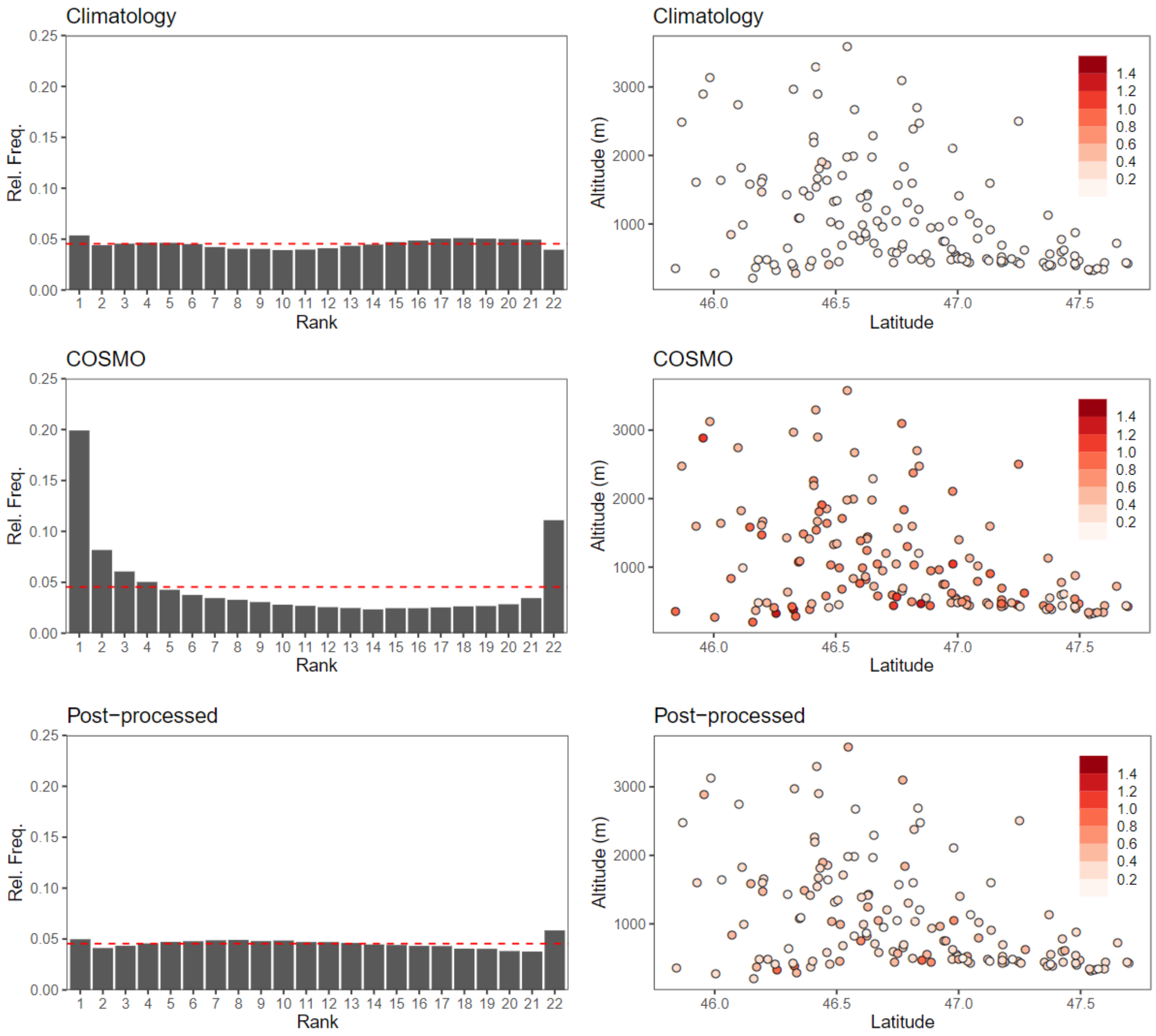

FIG. 6. Rank histogram for the COSMO-E ensemble and PIT histograms for the climatological and post-processed forecast distributions at a lead time of three days. The ranks have been aggregated over all years and stations, and the horizontal red line is indicative of perfect calibration. A measure of the miscalibration in the histogram is also shown as a function of the station latitude and altitude for all three methods.

The scores are displayed in the form of skill scores, with the raw COSMO-E ensemble forecasts as the reference. A positive skill score indicates an improvement upon the COSMO-E forecasts, whereas a negative skill score suggests the reference forecasts are more accurate. As the threshold decreases, the weight function tends to one, and all scores should therefore tend to the skill score obtained from the unweighted CRPS. As expected in this case, the climatological forecasts are significantly worse than the COSMO-E output, while the post-processed forecasts offer improvements of roughly 16%. Confidence intervals for the skill scores are also shown at each threshold, calculated using the block bootstrap approach described previously.

However, for all weighted versions of the CRPS, the skill score increases as higher thresholds are considered, suggesting the COSMO-E forecasts perform particularly poorly when predicting these more extreme events. This is true not only for the post-processed forecasts, but also for the climatological predictions. The COSMO-E forecasts, and hence also the post-processed forecasts to a lesser degree, tend to over-predict exceedances of extreme thresholds (see Figures 8 and 9). However, the extreme thresholds are rarely exceeded by the observations, meaning the average weighted scores for the climatological forecasts are very close to zero. As a result, the climatological forecasts improve even upon the post-processed forecasts at very extreme temperature thresholds, with skill scores that tend towards one.

When evaluating the three competing prediction systems with respect to multivariate high-impact events, separate weight functions are chosen to emphasise the different heat levels. For example, when interest is on level two heat events, the weight is equal to one when the level two criteria in Table 2 are satisfied, and zero otherwise. Equation 13 is then used to construct a chaining function for the threshold-weighted energy and variogram scores from this weight function, with $z_0 = (25, 25, 25)$ for heat levels one, two, and three, and $z_0 = (27, 27, 27)$ for heat level four.

For concision, only the threshold-weighted scores are presented here. The outcome-weighted scores cannot be readily applied to the multivariate ensemble forecasts without some appropriate smoothing, as discussed previously, while the vertically re-scaled scores are equivalent to the threshold-weighted scores for appropriate choices of x_0. Of course, the weighted scores could be calculated using alternative weight functions, though in this example there are fixed definitions of extreme heat events, providing obvious weight functions with which to emphasise these events when calculating multivariate forecast accuracy.

The threshold-weighted energy and variogram scores with the above weight and chaining functions are displayed in Table 4. The scores corresponding to level one heat events are similar to those obtained from the unweighted ES and VS, while the climatological forecasts appear to perform best with respect to the most extreme heat level. COSMO-E forecasts appear to be significantly

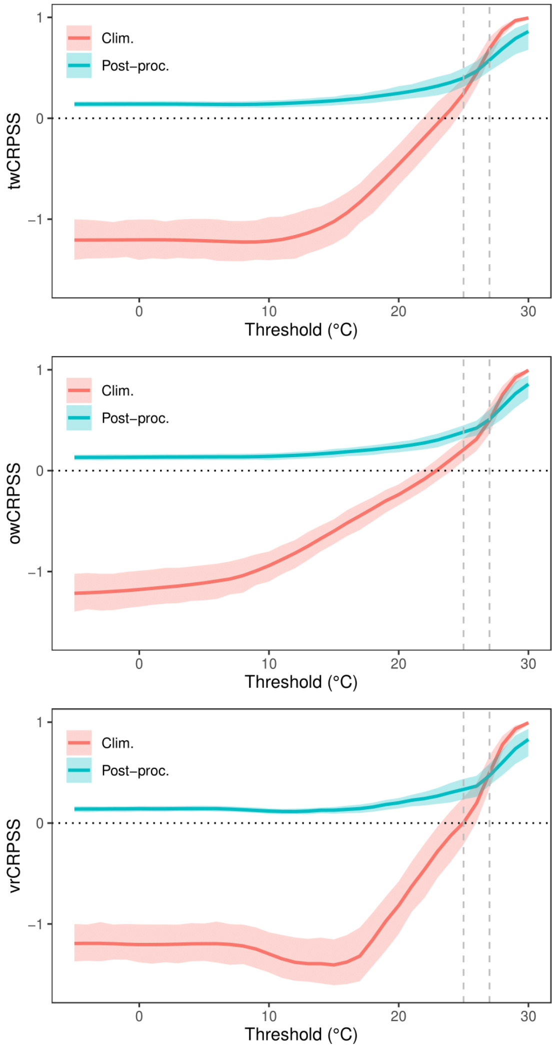

Fig. 7. Skill scores for the twCRPS, owCRPS, and vrCRPS as a function of the threshold used in the weight function $w(z) = \mathbb{1}\{z > t\}$ at a lead time of three days. The skill scores are shown for the climatological and post-processed forecast distributions, with the COSMO-E forecasts as the reference approach. The shaded regions represent pointwise 95% confidence intervals. Dashed vertical grey lines are shown at the thresholds $t = 25$ and $t = 27$.

less accurate than the alternative strategies when predicting the severity of level three and four heat events.

However, the weighted scoring rules cannot be used to infer what biases are present in the COSMO-E forecasts for these high-impact heat events. For this, conditional PIT histograms and conditional PIT reliability diagrams are displayed in Figures 8 and 9, where interest is on instances when the temperature exceeds 25°C or 27°C, respectively. These checks for conditional calibration are accompanied by standard reliability diagrams for predictions that these thresholds will be exceeded. As with the owCRPS, the COSMO-E ensembles are first smoothed using a normal distribution.

The climatological forecasts appear to issue better-calibrated forecasts for the probability of an extreme temperature event occurring, though the range of the predictions issued is much smaller than the COSMO-E and post-processed forecasts, highlighting that the climatological forecasts are less discriminative. The climatological forecast distributions also exhibit a heavy tail, suggesting parametric families other than the normal distribution may be more appropriate when modelling summer-time temperatures (Allen et al. 2021a). Figures 8 and 9 suggest that the COSMO-E ensembles over-estimate both the occurrence and severity of high temperature events, particularly for the more extreme threshold. This behaviour is also observed for the post-processed forecasts, albeit to a lesser degree. Hence, although statistical post-processing improves upon the raw COSMO-E model output, these forecasts themselves exhibit systematic biases when predicting high-impact events.

4. Conclusions

If meteorological services could accurately and reliably predict high-impact weather events, then the impacts associated with these events could be mitigated through the design of effective

	twES				twVS			
Level:	1	2	3	4	1	2	3	4
Clim.	4.44 [−1.39, −1.05]	0.83 [−0.03, 0.16]	0.12 [0.15, 0.43]	0.07 [0.71, 0.96]	2.53 [−0.77, −0.54]	3.55 [−0.22, 0.08]	0.25 [0.13, 0.48]	0.16 [0.69, 0.95]
COSMO	1.99 [0.00, 0.00]	0.87 [0.00, 0.00]	0.16 [0.00, 0.00]	0.41 [0.00, 0.00]	1.53 [0.00, 0.00]	3.25 [0.00, 0.00]	0.35 [0.00, 0.00]	0.78 [0.00, 0.00]
Post-proc.	1.75 [0.10, 0.14]	0.71 [0.13, 0.24]	0.10 [0.29, 0.47]	0.11 [0.66, 0.85]	1.47 [0.01, 0.07]	2.55 [0.15, 0.29]	0.21 [0.29, 0.50]	0.23 [0.63, 0.81]

TABLE 4. Threshold-weighted ES and VS for the three forecasting strategies with emphasis on each heat event level. The weight and chaining functions used within the scores are discussed in the text. Below each score is a 95% confidence interval for the corresponding skill score, with the COSMO-E forecasts used as reference. For readability, all scores for level two and three heat events have been scaled by 10, and those for level four by 100. The skill scores are unaffected by this scaling.

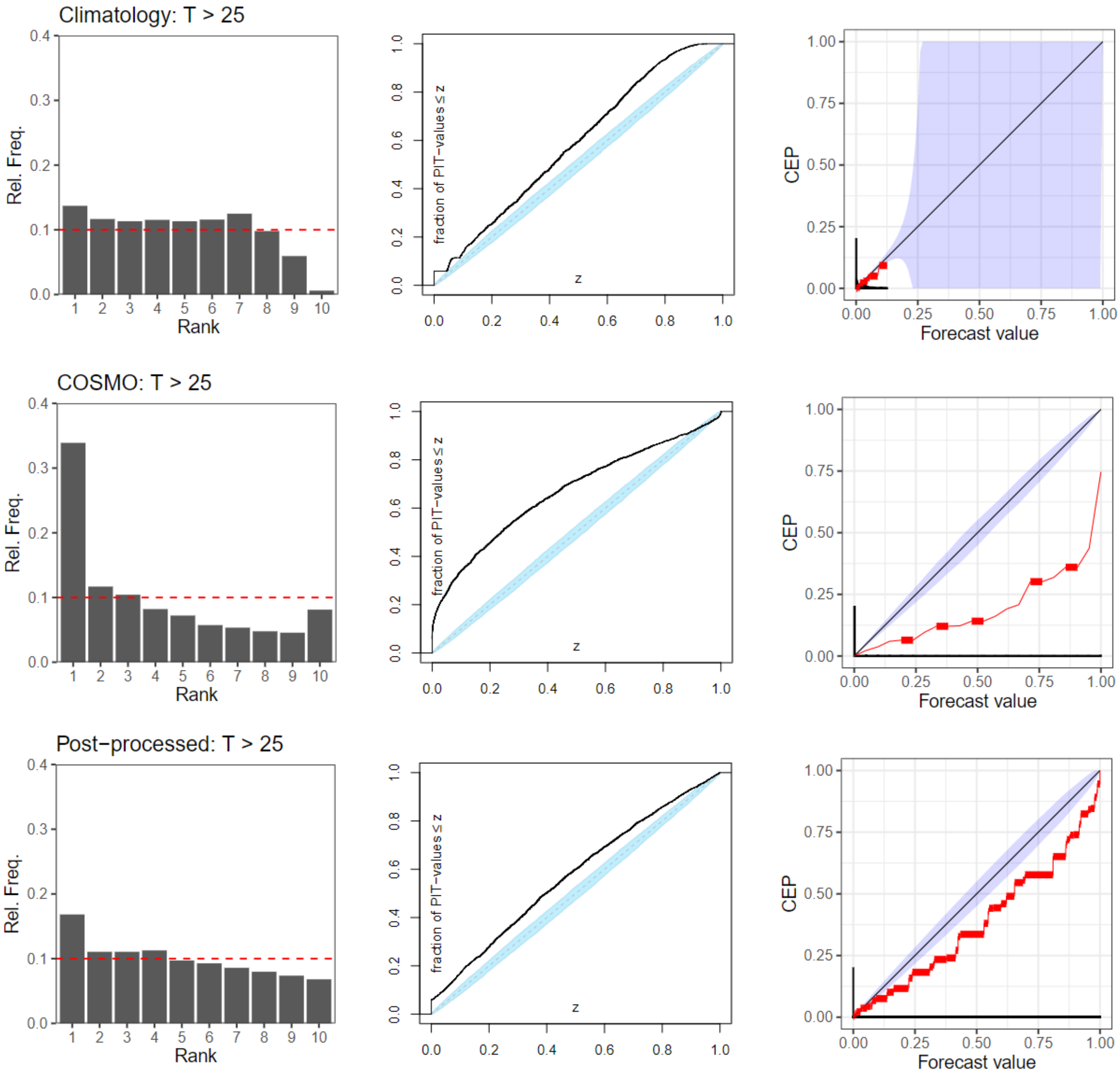

FIG. 8. Conditional PIT histograms (left) and conditional PIT reliability diagrams (right) for the three forecasting strategies at a lead time of three days. Emphasis is on daily mean temperatures that exceed 25°C. Standard reliability diagrams (right) also show the conditional event probabilities (CEP) given the forecast probability that the threshold will be exceeded.

early warning systems. Methods to evaluate forecasts made for high-impact weather are therefore crucial when developing warning systems. This paper has reviewed techniques to evaluate forecasts for high-impact events, highlighting in particular how weighted verification tools allow certain

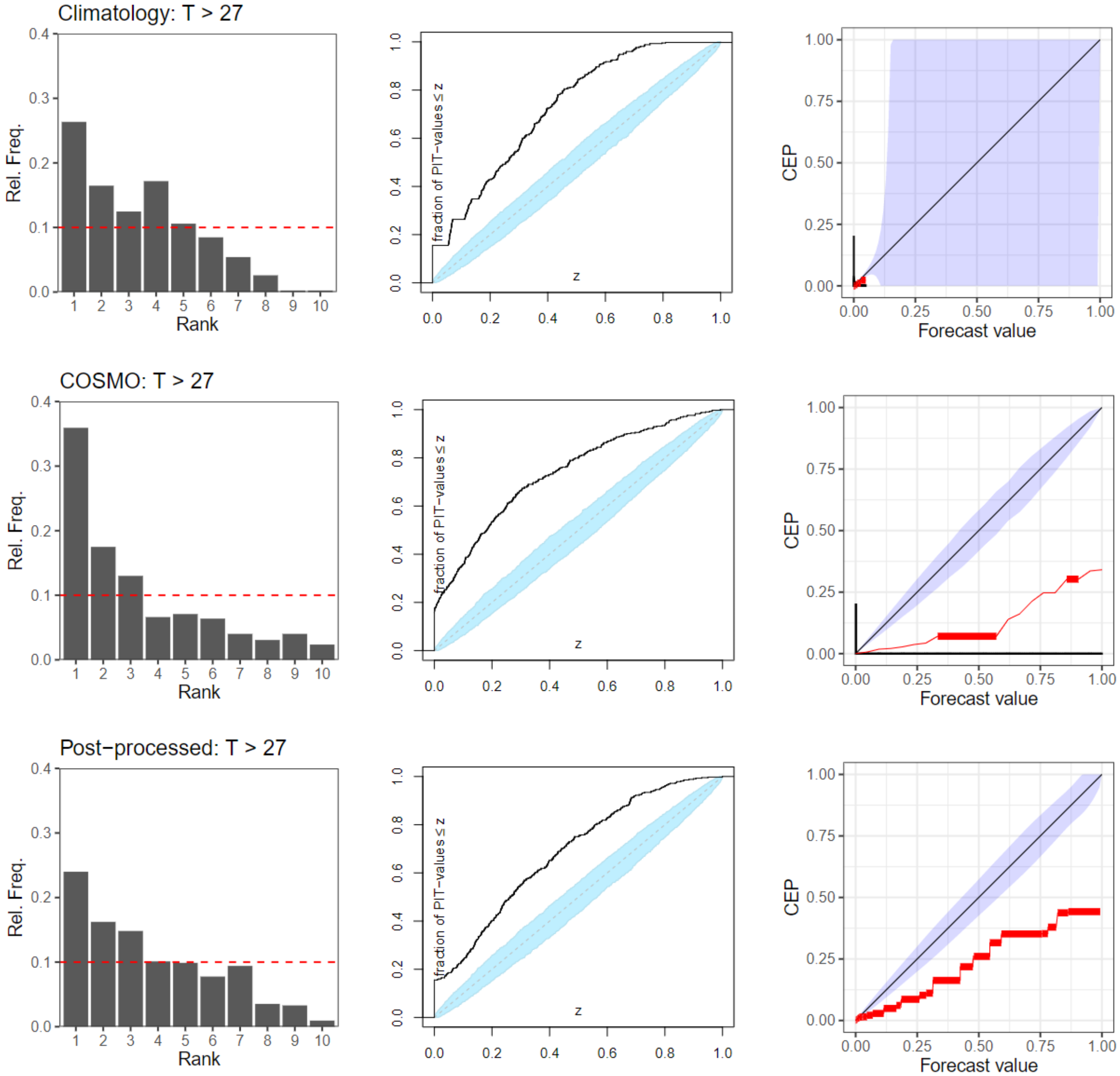

FIG. 9. As in Figure 8 but with emphasis on daily mean temperatures that exceed 27°C.

outcomes to be emphasised during forecast evaluation. We review and compare approaches to construct weighted scoring rules, both in a univariate and multivariate setting, and we then leverage the existing theory on weighted scoring rules to introduce diagnostic checks that assess forecast calibration conditionally on particular outcomes having occurred. To illustrate how these verification tools can be employed in practice, they are used to assess how well operational weather forecasts can predict dangerous heat events, defined using criteria adopted by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss).

Three alternative methods to construct weighted scoring rules are compared: threshold-weighted, outcome-weighted, and vertically re-scaled scores. Outcome-weighted scores provide a direct and intuitive way to circumvent the forecaster's dilemma, allowing forecasts to be evaluated only when high-impact events occur. However, when forecasts are in the form of an ensemble, and interest is on rare events, these scores are not always well-defined, making it difficult to implement them in practice. Hence, when interest is on high-impact weather events, we instead recommend evaluating forecast accuracy using threshold-weighted and vertically re-scaled scoring rules. To aid their implementation in practice, these weighted scoring rules have recently been made available in the widely-used `scoringRules` package in `R` (Jordan et al. 2019).

In Section 3, we use these weighted scoring rules to evaluate three competing prediction systems whilst emphasising extreme heat events in Switzerland. In particular, we compare an operational, high-resolution ensemble prediction system to climatological and statistically post-processed forecasts. Although recent studies have suggested that statistical post-processing methods could deteriorate the accuracy of forecasts issued by numerical weather models when interest is on high-impact weather events, our results indicate that even simple post-processing methods can significantly improve upon the raw model output when predicting extreme heat in Switzerland. This suggests that forecasters should utilise statistically post-processed forecasts when constructing weather warnings, in addition to the raw output from numerical weather models.

However, the checks for conditional calibration introduced here - namely conditional PIT histograms and conditional PIT reliability diagrams - indicated that even the post-processed forecasts were not calibrated conditionally on extreme temperatures having occurred, despite the overall forecast distributions being reasonably well-calibrated. Future work might therefore look to remedy this, by considering how statistical post-processing methods can be developed that are tailored to the generation of weather warnings.

The conditional calibration of the three prediction systems was only evaluated in the univariate setting. In Section 2b, we also describe how multivariate cPIT histograms and cPIT reliability diagrams could be constructed to check for multivariate calibration given that a high-impact event has occurred. However, as with outcome-weighted scoring rules, the approach has practical limitations when interest is on rare events and the forecast is an ensemble, which is frequently the case for multivariate weather forecasts. It would therefore be useful design appropriate methods to convert multivariate ensemble forecasts to continuous forecast distributions, thereby allowing multivariate cPIT histograms and reliability diagrams to be applied.

Lastly, we reiterate that the methods discussed herein do not evaluate warning systems, but rather the ability of weather forecasts to predict potentially impactful events. Weather warnings rely not only on these forecasts, but also on several other factors: for example, the economic costs associated with a warning, the expected behaviour in response to the warning, and the effectiveness with which the warnings are relayed to those at risk. Although this renders the evaluation of weather warnings a multifaceted and thus complex task, methods to objectively identify effective warning systems would be highly valuable to operational forecasters. Future work might therefore look at developing methods to evaluate the quality of weather warnings, potentially building on the approaches presented herein to do so.

Acknowledgments. This work was funded by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) and the Oeschger Centre for Climate Change Research. We are also grateful for the helpful input provided by David Ginsbourger, Pascal Horton, Lionel Moret, Mark Liniger, Jon Koh, José Carlos Araujo Acuña, and three anonymous reviewers.

Data availability statement. The code used in this study is available on GitHub at `https://github.com/sallen12/WeightedForecastVerification`. For research purposes, the temperature observations used herein are freely available from MeteoSwiss' IDAweb platform (`https://www.meteoswiss.admin.ch/services-and-publications/service/weather-and-climate-products/data-portal-for-teaching-and-research.html`) and the COSMO-E forecasts are made available upon request for a data processing fee.

APPENDIX

Statistical post-processing

State-of-the-art ensemble prediction systems typically exhibit systematic biases when forecasting surface weather variables. To remove these biases and re-calibrate the ensemble output, statistical post-processing is applied to the forecasts (see Vannitsem et al. 2018, for a review). We re-calibrate the COSMO-E daily mean temperature forecasts using the ensemble model output statistics (EMOS) framework proposed by Gneiting et al. (2005). EMOS assumes that the variable to be forecast follows a certain parametric distribution, whose moments depend linearly on those of the corresponding ensemble forecast. We assume here that the daily mean temperature at a given time and location is normally distributed.

To account for local structures within the COSMO-E forecast biases, two additional predictors are incorporated into the post-processing model: a topographic position index (TPI) that reflects the change in elevation between a station and those in a local neighbourhood of 2km radius, and a measure of the height difference between the COSMO-E model and reality (MHD). The inclusion of these two spatial covariates follows other recent studies on the post-processing of COSMO-E temperature forecasts in Switzerland (e.g. Keller et al. 2021). These additional predictors allow the model to account for local features in the forecast biases despite fitting a single post-processing model simultaneously to forecasts at all stations.

The post-processing model can be formalised as follows. Let Y denote the daily mean temperature at a given station, time, and lead time, and let $\bar{x}$ and v denote the mean and variance of the corresponding COSMO-E ensemble members, respectively. Then, the model assumes that

$$Y = \beta_0 + \beta_1 \bar{x} + \beta_2 \mathrm{MHD} + \beta_3 \mathrm{TPI} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_0 + \sigma_1 v), \tag{A1}$$

where $\mathcal{N}(\mu, \gamma)$ denotes the normal distribution with mean μ and variance γ. Note that the TPI and MHD depend on the station under consideration, but not on the time or lead time. The variance of this model could similarly be set up to depend on the MHD and TPI, but this was not found to provide much benefit.

The post-processing model parameters $\beta_0, \beta_1, \beta_2, \beta_3, \sigma_0, \sigma_1$ link the predictors to the observations. A separate set of parameters is estimated for each forecast lead time, thereby acknowledging that the relationship between the forecast and the observation will change as the forecast horizon increases. As in Keller et al. (2021), the parameters are estimated by minimising the CRPS over a rolling training window containing the previous 45 forecast-observation pairs, allowing the model to also account for recent patterns in the forecast biases.

Post-processing is applied to the daily mean temperature forecast at each lead time separately. Since dangerous heat events are often a multivariate phenomenon, we use copulas to convert these individual forecast distributions into a temporally coherent multivariate forecast over the three day period. To do so, we employ ensemble copula coupling (ECC; Schefzik et al. 2013), an empirical copula-based approach. ECC works by converting the univariate post-processed forecast distributions at each lead time to an ensemble forecast, by selecting 21 evenly-spaced quantiles from each distribution, before reordering the resulting ensemble members so that the rankings of the ensemble members at each lead time are the same as in the corresponding COSMO-E ensemble.

By comparing the performance of this post-processing model to the raw COSMO-E output, we can investigate how post-processing affects predictions of high-impact events: Pantillon et al. (2018), among others, have recently postulated that post-processing can hinder forecasts of extreme events due to a regression-to-the-mean type effect. We additionally compare the COSMO-E and post-processed forecasts to a climatological prediction. The climatological forecast again assumes that the temperature is normally distributed, but no predictors are employed within this distribution. The mean and variance of this climatological distribution are estimated over a 45-day rolling window, similarly to the post-processing model, though a separate climatology is estimated for each station separately to incorporate local information. An empirical copula is then applied to the climatological forecasts to generate a coherent multivariate forecast.

References

Allen, S., G. R. Evans, P. Buchanan, and F. Kwasniok, 2021a: Accounting for skew when postprocessing MOGREPS-UK temperature forecast fields. *Monthly Weather Review*, **149**, 2835–2852.

Allen, S., G. R. Evans, P. Buchanan, and F. Kwasniok, 2021b: 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.

Allen, S., D. Ginsbourger, and J. Ziegel, 2022: Evaluating forecasts for high-impact events using transformed kernel scores. *arXiv preprint arXiv:2202.12732*.

Arnold, S., A. Henzi, and J. F. Ziegel, 2021: Sequentially valid tests for forecast calibration. *arXiv preprint arXiv:2109.11761*.

Basagaña, X., C. Sartini, J. Barrera-Gómez, P. Dadvand, J. Cunillera, B. Ostro, J. Sunyer, and M. Medina-Ramón, 2011: Heat waves and cause-specific mortality at all ages. *Epidemiology*, **22**, 765–772.

Bellier, J., I. Zin, and G. Bontron, 2017: Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Monthly Weather Review*, **145**, 3529–3544.

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

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An introduction to statistical modeling of extreme values*, Vol. 208. London: Springer.

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

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

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, **111**.

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

Dimitriadis, T., T. Gneiting, and A. I. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, **118**.

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, **26**, 699–713.

Gilleland, E., 2020: Bootstrap methods for statistical inference. Part I: Comparative forecast verification for continuous variables. *Journal of Atmospheric and Oceanic Technology*, **37**, 2117–2134.

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

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

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

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

Gneiting, T., and J. Resin, 2021: Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *arXiv preprint arXiv:2108.03210*.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312–1327.

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

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

Jordan, A., F. Krüger, and S. Lerch, 2019: Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, **90 (12)**, 1–37, https://doi.org/10.18637/jss.v090.i12.

Keller, R., J. Rajczak, J. Bhend, C. Spirig, S. Hemri, M. A. Liniger, and H. Wernli, 2021: Seamless multimodel postprocessing for air temperature forecasts in complex topography. *Weather and Forecasting*, **36**, 1031–1042.

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, **65**, 21 206.

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

Majumdar, S. J., and Coauthors, 2021: Multiscale forecasting of high-impact weather: current status and future challenges. *Bulletin of the American Meteorological Society*, **102**, E635–E659.

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

McCarthy, M., L. Armstrong, and N. Armstrong, 2019: A new heatwave definition for the UK. *Weather*, **74**, 382–387.

Pantillon, F., S. Lerch, P. Knippertz, and U. Corsmeier, 2018: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1864–1881.

Pinson, P., and J. Tastu, 2013: Discrimination ability of the energy score. Tech. rep., Technical University of Denmark.

Ragettli, M. S., A. M. Vicedo-Cabrera, C. Schindler, and M. Röösli, 2017: Exploring the association between heat and mortality in Switzerland between 1995 and 2013. *Environmental Research*, **158**, 703–709.

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

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

Stephenson, D. B., B. Casati, C. Ferro, and C. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications*, **15**, 41–50.

Thorarinsdottir, T. L., and N. Schuhen, 2018: Verification: Assessment of calibration and accuracy. *Statistical postprocessing of ensemble forecasts*, Elsevier, 155–186.

Vannitsem, S., D. S. Wilks, and J. Messner, 2018: *Statistical postprocessing of ensemble forecasts*. Elsevier.

Wilks, D. S., 2019: *Statistical methods in the atmospheric sciences*. Amsterdam: Elsevier.

WMO, 2015: WMO guidelines on multi-hazard impact-based forecast and warning services. World Meteorological Organization.

WMO, 2018: Guidelines on the definition and monitoring of extreme weather and climate events. World Meteorological Organization.

Ziegel, J., 2017: Copula calibration. *Copulae: On the Crossroads of Mathematics and Economics*, Mathematisches Forschungsinstitut Oberwolfach. Report No. 20/2015, 7–10.

Zscheischler, J., and Coauthors, 2020: A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, **1**, 333–347.