International Journal of Forecasting 42 (2026) 1336–1356

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Enforcing tail calibration when training probabilistic forecast models

Jakob Benjamin Wessel [a,*], Maybritt Schillinger [b], Frank Kwasniok [a], Sam Allen [c]

[a] Department of Mathematics and Statistics, University of Exeter, Exeter, United Kingdom
[b] Seminar for Statistics, ETH Zurich, Zurich, Switzerland
[c] Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany

ARTICLE INFO

Article history:
Dataset link: https://github.com/jakobwes/Enforcing-tail-calibration

Keywords:
Probability forecasting
Calibration
Neural networks
Weather forecasting
Extremes

ABSTRACT

Probabilistic forecasts are typically obtained using state-of-the-art statistical and machine learning models, with model parameters estimated by optimizing a proper scoring rule over a set of training data. If the model class is not correctly specified, the learned model will not necessarily produce calibrated forecasts. Calibrated forecasts allow users to appropriately balance risks in decision-making, and it is particularly important that forecast models issue calibrated predictions for extreme events, since such outcomes often generate large socio-economic impacts. In this work, we study how the loss function used to train probabilistic forecast models can be adapted to improve the reliability of forecasts made for extreme events. We investigate loss functions based on weighted scoring rules, and additionally propose regularizing loss functions using a measure of tail miscalibration. We apply these approaches to a hierarchy of increasingly flexible forecast models for UK wind speeds, including simple parametric models, distributional regression networks, and conditional generative models. We demonstrate that state-of-the-art models do not issue calibrated forecasts for extreme wind speeds, and that the calibration of forecasts for extreme events can be improved by suitable adaptations to the loss function during model training. This introduces a trade-off between calibrated forecasts of extreme events and those of more common outcomes.

1. Introduction

Probabilistic forecasts take the form of probability distributions over the set of possible outcomes. Such forecasts comprehensively describe the uncertainty surrounding the unknown outcome, making them essential for effective risk assessment and decision-making. They have therefore become commonplace in a variety of application domains, including medicine, economics, finance, politics, and climate science.

Probabilistic forecasts are typically obtained using state-of-the-art statistical and machine learning models that learn the conditional distribution of the outcome variable given some covariates. Parameters of the models can be estimated by optimizing a *proper scoring rule* over a set of training data. Proper scoring rules quantify forecast accuracy (see e.g. Gneiting & Raftery, 2007), and this framework therefore finds the model parameters that result in the most accurate forecasts, with accuracy measured in terms of the chosen scoring rule. If the model class is well-specified (i.e., the family of parametrized candidate models contains the true condi-

* Corresponding author.
E-mail addresses: j.wessel@exeter.ac.uk (J.B. Wessel), maybritt.schillinger@stat.math.ethz.ch (M. Schillinger), f.kwasniok@exeter.ac.uk (F. Kwasniok), sam.allen@kit.edu (S. Allen).

https://doi.org/10.1016/j.ijforecast.2026.04.009

tional distribution of the outcome), then this *optimum score estimation* should return the true parameters underlying the data-generating process, provided sufficient data is available (Dawid et al., 2016; Gneiting & Raftery, 2007). This holds for any choice of proper scoring rule.

However, if the model class is not correctly specified, as is generally the case in practice, the resulting forecasts will exhibit bias relative to the true distribution underlying the data. These biases will generally change for different scoring rules. For example, it is well-known that model estimation with the continuous ranked probability score (CRPS; Matheson & Winkler, 1976), a popular proper scoring rule, tends to yield forecast distributions that are less dispersed than those obtained from maximum likelihood estimation (see e.g. Buchweitz et al., 2025; Gebetsberger et al., 2018; Gneiting et al., 2005).

In this mis-specified case, there is therefore no guarantee that the estimated model yields forecasts that are *calibrated* (or *reliable*). Probabilistic forecasts are calibrated if they align statistically with the corresponding outcomes, typically assessed by checking whether the forecast probability integral transform (PIT) values resemble a sample from a standard uniform distribution (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). Gneiting et al. (2007) remarked that the general goal of probabilistic forecasting is to issue forecasts that are as informative as possible, subject to being calibrated. In this sense, calibration is a requirement that forecasts must satisfy to be useful for decision-making and risk assessment.

This is particularly true when forecasting extreme outcomes. Extreme events typically lead to the largest impacts on forecast users, and calibrated forecasts for such outcomes are therefore particularly valuable. However, Allen, Koh et al. (2025) demonstrate that a forecast can satisfy standard notions of calibration without issuing reliable forecasts for extreme events. They therefore introduce a notion of *tail calibration* that assesses whether forecasts are calibrated when predicting extremes. This facilitates a more comprehensive understanding of how forecasts behave when interest lies in extreme outcomes, helping practitioners design forecast models that are simultaneously calibrated for predicting both extreme and non-extreme events.

One approach to obtain more accurate forecasts for extreme events is to train probabilistic forecast models by optimizing a weighted scoring rule (Wessel et al., 2025). Weighted scoring rules allow users to target particular outcomes when quantifying forecast accuracy, and are therefore commonly employed to compare forecasts for extreme outcomes (Lerch et al., 2017). However, there is again no guarantee that models trained with a weighted scoring rule will issue tail-calibrated forecasts. As Wilks (2018) remark, "calibration is not guaranteed because [proper score] minimization may be achieved as a compromise between calibration and resolution (which is related to sharpness)". The same holds for weighted scoring rules and tail calibration.

As an alternative, one could regularize the scoring rule to directly penalize (tail) miscalibration. Wilks (2018), for example, propose measuring to what extent the distribution of PIT values deviates from a standard uniform distribution, and including this measure of miscalibration as an additional term in the loss function, thereby encouraging forecasters to issue calibrated forecasts. The resulting forecasts are typically better calibrated than those obtained directly from optimum score estimation, though this may come at the expense of forecast accuracy or sharpness. If interest lies in extreme events, a similar framework could be adopted, in which one measures the tail miscalibration of the forecasts and then adds this to the scoring rule during model training.

In this work, we investigate how regularizing the loss function using either weighted scoring rules or miscalibration measures affects the performance of the resulting forecasts when the focus is on extreme outcomes. These different regularization terms are implemented in a case study on the statistical post-processing of weather forecasts. We consider a hierarchy of increasingly more flexible probabilistic wind speed models, including simple baseline parametric methods (Gneiting et al., 2005), distributional regression networks (Rasp & Lerch, 2018), and conditional generative models (Chen et al., 2024). We first evaluate the performance of these methods with respect to extreme outcomes and demonstrate that all three approaches fail to issue tail-calibrated forecasts; increasing the flexibility of the model class does not resolve this. We then demonstrate how changing the loss function affects the resulting forecast performance. Penalizing tail miscalibration can improve the reliability of resulting forecasts for extreme events, though at the expense of overall forecast calibration and performance.

The remainder of the paper is structured as follows. The following section introduces probabilistic forecasts, proper scoring rules, and the theory of optimum score estimation. Section 3 defines the notions of calibration and tail calibration that we want our forecasts to satisfy, and Section 4 then proposes different regularization terms that could be employed during model training to weakly enforce (tail) calibration of the resulting probabilistic forecasts. A simple demonstration of the approach on simulated data is presented in Section 5, followed by an application to statistical weather-forecasting models in Section 6. Section 7 concludes.

2. Training probabilistic forecast models

2.1. Proper scoring rules

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the outcome variable that we wish to forecast. Here, a probabilistic forecast for Y is a probability distribution on the outcome space $\mathcal{Y}$. We use $\mathcal{F}$ to denote a class of such distributions, and represent a probabilistic forecast $F \in \mathcal{F}$ via its cumulative distribution function.

A scoring rule is a function $S : \mathcal{F} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty, \infty\}$ that takes a probabilistic forecast $F \in \mathcal{F}$ and an observation $y \in \mathcal{Y}$ as inputs, and outputs a real-valued (possibly infinite) score $S(F, y)$ that quantifies the forecast's accuracy. A lower score corresponds to a more accurate forecast, and a scoring rule is called *proper* if

$$\mathbb{E}_{Y\sim G}[S(G, Y)] \leq \mathbb{E}_{Y\sim G}[S(F, Y)] \quad \text{for all} \quad F, G \in \mathcal{F}.$$

That is, if the outcome follows the true distribution $G \in \mathcal{F}$, then the score is minimized in expectation by issuing G as

the forecast. The scoring rule is *strictly proper* if this holds with equality if and only if $F = G$.

Proper scoring rules condense forecast performance into a single value, providing a means to rank and compare competing forecasts. While all proper scoring rules are minimized in expectation by the true distribution of the outcomes, the ranking of imperfect forecasts will generally depend on the chosen scoring rule. Popular examples of proper scoring rules for real-valued outcomes include the logarithmic (log) score,

$$\mathrm{LS}(F, y) = -\log f(y), \tag{1}$$

where f is the probability density function associated with F, assuming it exists (Good, 1952); and the continuous ranked probability score (CRPS),

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(z) - \mathbb{1}\{y \le z\}]^2 \, \mathrm{d}z, \tag{2}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function (Matheson & Winkler, 1976). We refer to Gneiting and Raftery (2007) and Waghmare and Ziegel (2025) for comprehensive overviews on proper scoring rules.

2.2. Optimum score estimation

Probabilistic forecasts for Y are generally obtained using a statistical or machine learning model that estimates the conditional distribution of Y given some covariates. These probabilistic forecast models depend on parameters $\theta \in \Theta \subseteq \mathbb{R}^d$ that link the covariates to the outcome, and we denote the forecast distribution corresponding to parameter vector θ and covariate vector x by $F_{\theta,x}$. We say that a model is well-specified if there exists a parameter vector $\theta^* \in \Theta$ for which, for any covariates x, the resulting forecast distribution $F_{\theta^*,x}$ is equal to the conditional distribution of Y given the covariates.

In practice, we typically have access to a (training) set of covariates and outcomes $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from which to learn the optimal model parameters. Since forecast accuracy is assessed using proper scoring rules, it makes sense to find the parameter vector θ that optimizes the average score over the training data. This is generally referred to as optimum or minimum score estimation (Gneiting & Raftery, 2007).

Definition 1. The optimum score estimator corresponding to a proper scoring rule S is

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} S(F_{\theta,x_i}, y_i). \tag{3}$$

Optimizing the log score is equivalent to maximum likelihood estimation, while many studies have also considered using the CRPS to estimate model parameters (see e.g. Gneiting et al., 2005). The inferential properties of optimum score estimators have been studied in depth by Dawid et al. (2016).

If S is a strictly proper scoring rule, then the optimum score estimator is consistent for θ^* (Gneiting & Raftery, 2007). Hence, if the forecast model is well-specified, then optimum score estimation with any strictly proper scoring rule will asymptotically recover the parameter vector θ^* corresponding to the true conditional distribution of Y given the covariates. In this case, the choice of strictly proper scoring rule to employ in (3) is largely insignificant. However, in practice, forecasting models are generally not correctly specified – for example, it is common to assume a particular parametric distribution for the response variable – and we only have access to a finite amount of data. In most practical applications, minimizing the average score will therefore not recover the true distribution underlying the outcomes. Hence, employing different scoring rules in (3) will generally yield different parameter estimates.

3. Tail calibration

Having estimated the optimal parameters of a probabilistic forecasting model, the model fit can be assessed by checking whether the resulting forecast distributions are calibrated. Loosely speaking, forecasts are calibrated if they can be taken at face value, which is necessary for effective decision-making.

To define calibration formally, let F now denote a random cumulative distribution function, acting as a probabilistic forecaster. While F often depends on parameters and (random) covariates, as discussed above, we drop this dependence from the notation in this section for ease of presentation. The forecast-observation pairs $\{(F_1, y_1), \ldots, (F_n, y_n)\}$ that we observe in practice can be interpreted as realizations of the random pair (F, Y). We use the probability measure $\mathbb{P}$ to describe the joint distribution of F and Y. This corresponds to the prediction space setting introduced by Gneiting and Ranjan (2013), to which readers are referred for further details. Since we are interested in extremes, we assume that F is almost surely continuous. However, simple adaptations to the definitions below are available when the forecast distributions are not continuous (e.g. Gneiting & Ranjan, 2013, Definition 2.5). We study the following notion of forecast calibration (see Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007).

Definition 2. A forecaster is *probabilistically calibrated* if the forecast probability integral transform (PIT) $F(Y)$ follows a standard uniform distribution, that is,

$$\mathbb{P}(F(Y) \le u) = u \quad \text{for all} \quad u \in [0, 1]. \tag{4}$$

Probabilistic calibration can be assessed in practice by checking whether the empirical PIT values $\{F_1(y_1), \ldots, F_n(y_n)\}$ resemble a sample from a standard uniform distribution, which is typically performed by plotting a histogram or pp-plot of the PIT values. The shape of the diagnostic plot can provide useful information regarding what biases occur in the forecast.

As Gneiting and Ranjan (2013) remark, "checks for the uniformity of the probability integral transform have formed a cornerstone of density forecast evaluation". However, probabilistic calibration is a fairly weak notion of calibration, since it concerns unconditional facets of the forecasts and observations (e.g. Gneiting & Resin, 2023). Allen, Koh et al. (2025) demonstrate that a forecaster can be probabilistically calibrated without necessarily issuing reliable forecasts for extreme events, and

therefore introduce notions of forecast *tail calibration*, which assess the calibration of forecasts when interest is only in the tails of the predictive distribution.

Given a threshold $t \in \mathbb{R}$, define the forecast excess distribution as

$$F_t(x) = \frac{F(x) - F(t)}{1 - F(t)}, \quad x \geq t, \tag{5}$$

with $F_t(x) = 0$ for $x < t$, and $F_t(x) = 1$ for all $x \in \mathbb{R}$ if $F(t) = 1$. If $Y \sim G$, then the excess distribution G_t describes the distribution function of Y conditioned on Y exceeding the threshold t. The term "excess distribution" is also often used to describe the conditional distribution of $Y - t$ given $Y > t$, but we employ the definition above here since it simplifies the notation below.

Definition 3. A forecaster F is *probabilistically tail calibrated* at a threshold $t \in \mathbb{R}$ if

$$\mathbb{P}(Y > t) = \mathbb{E}[1 - F(t)], \tag{6}$$

and

$$\mathbb{P}(F_t(Y) \leq u \mid Y > t) = u \quad \text{for all } u \in [0, 1]. \tag{7}$$

Remark 4. This differs slightly from the definition of tail calibration in Allen, Koh et al. (2025), which considers the limit as t tends to the upper endpoint of the distribution of Y. Since both approaches require assessing tail calibration at a finite threshold (or set of thresholds) in practice, this difference is largely irrelevant to this study.

The first condition in Definition 3 focuses on the forecast threshold exceedance probability, thereby assessing forecasts for the occurrence of an extreme event, while the second condition concerns the calibration of the conditional forecast distribution given that the threshold t has been exceeded, thereby assessing forecasts for the severity of extreme events given their occurrence. The latter condition is equivalent to assessing whether the forecast excess distribution is probabilistically calibrated for outcomes exceeding the threshold of interest, as suggested by Allen, Ginsbourger et al. (2023) and Mitchell and Weale (2023). This can be assessed empirically via conditional PIT (CPIT) values,

$$z_{i,t} = F_{i,t}(y_i), \text{ for } i \in \mathcal{I}_t,$$

where $\mathcal{I}_t = \{i \in \{1, \ldots, n\} : y_i > t\}$ is the set of indices for which the outcome exceeds the threshold, and $F_{i,t}$ is the excess distribution of forecast F_i at threshold t. If a histogram of the empirical CPIT values appears flat, or a pp-plot lies close to the diagonal, there is evidence to suggest that (7) is satisfied.

The CPIT values assess only the shape of the forecast distribution above the threshold and ignore the probability that threshold exceedances are forecast; this probability is captured separately by (6). The two conditions in Definition 3 can be combined into a single criterion that incorporates both the shape of the excess distribution and the occurrence of threshold exceedances.

Proposition 5. *Let $O_t = \mathbb{P}(Y > t)/\mathbb{E}[1 - F(t)]$, and let H_{Z_t} denote the distribution function of $F_t(Y)$ given that $Y > t$. If H_{Z_t} is strictly increasing, with inverse $H_{Z_t}^{-1}$, then the two conditions in Definition 3 hold if and only if*

$$O_t H_{Z_t}^{-1}(u) = u \quad \text{for all } u \in [0, 1]. \tag{8}$$

Proof. Eq. (6) is recovered from (8) when $u = 1$, and (7) follows by rearranging (8). The opposite direction, that (6) and (7) imply (8), is trivial. □

Allen, Koh et al. (2025) suggest an alternative combination of (6) and (7) (namely, $O_t H_{Z_t}(u) = u$ for all $u \in [0, 1]$). Still, we argue in Appendix A that this can be less interpretable than the expression in (8), and is less suited to the methodology proposed in the following section.

To assess whether (8) holds in practice, we can estimate $O_t H_{Z_t}^{-1}$ using

$$\hat{Q}_t(u) = \underbrace{\frac{|\mathcal{I}_t|}{\sum_{i=1}^{n}[1 - F_i(t)]}}_{\hat{O}_t} \cdot \underbrace{\hat{q}_{z_t}(u)}_{\hat{H}_{Z_t}^{-1}(u)} \qquad \text{for } u \in [0, 1], \tag{9}$$

where $\hat{q}_{z_t}(u)$ is the sample u-quantile derived from the conditional PIT values $z_{i,t}$, $i \in \mathcal{I}_t$. If $\hat{Q}_t(u)$ is close to u for all $u \in [0, 1]$, then there is evidence to suggest that the forecasts are probabilistically tail calibrated. In this case, a plot of $\hat{Q}_t(u)$ against u should therefore lie close to the diagonal. In the limit as $t \to -\infty$, probabilistic tail calibration recovers probabilistic calibration: condition (6) then becomes trivial, condition (7) coincides with (4), and the estimate $\hat{Q}_t$ is equal to the empirical quantile function of the standard PIT values. Plotting $\hat{Q}_t(u)$ against $u \in [0, 1]$ is then equivalent to a qq-plot of the PIT values.

Probabilistic calibration does not imply probabilistic tail calibration (or vice versa), meaning if the standard checks for probabilistic calibration are satisfied, this does not necessarily ensure that the forecasts are calibrated in the tail (Allen, Koh et al., 2025). One consequence of this is that probabilistic forecast models may produce calibrated forecasts that are not tail-calibrated. In the following, we discuss how loss functions can be adapted to enforce tail calibration during model training.

4. Enforcing tail calibration

While proper scoring rules implicitly incorporate forecast calibration when quantifying forecast performance, training incorrectly specified forecast models by optimizing a proper scoring rule will not necessarily yield calibrated predictions. Proper scores can be decomposed into terms of discriminative ability and miscalibration (Bröcker, 2009; Murphy, 1973), and the learned forecast distributions will not necessarily result in a miscalibration term that is equal to zero since both calibration and sharpness are optimized; the scoring rule might prefer more informative forecast distributions, even if they are miscalibrated.

Moreover, even if the resulting forecasts appear probabilistically calibrated, there is no guarantee that they also yield calibrated forecasts for extreme events. In this section, we therefore discuss ways to regularize scoring-rule-based loss functions during model training to encourage

Table 1
Loss functions studied in this work. Here, $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i)$ denotes the average score corresponding to some strictly proper scoring rule S, such as the log score or CRPS, and $\bar{S}_w = \frac{1}{n}\sum_{i=1}^{n} S_w(F_i, y_i; w)$ denotes the average score corresponding to some weighted scoring rule S_w, such as the censored likelihood score or threshold-weighted CRPS.

Name	Loss function	Equation	Interpretation
Baseline	$\bar{S}$	(3) with (1) or (2)	Optimum score estimation with no penalization.
Weighted score	$\bar{S} + \gamma\bar{S}_w$	(3) with (10) or (11)	Optimum weighted score estimation.
MCB	$\bar{S} + \gamma\text{MCB}$	(13)	Penalization of probabilistic miscalibration.
TMCB	$\bar{S} + \gamma\text{TMCB}$	(15)	Penalization of probabilistic tail miscalibration.

tail calibration of the resulting forecasts. We use the terms regularization and penalization interchangeably. Table 1 provides an overview of the loss functions that we will study.

4.1. Weighted scoring rules

In the case of an incorrectly specified forecast model, Gneiting and Raftery (2007) argue that "the appeal of optimum score estimation lies in the potential adaptation of the scoring rule to the problem at hand". That is, proper scoring rules can be used to target particular aspects of forecast distributions when quantifying forecast accuracy, and by using these scoring rules as loss functions, the resulting model should generate forecasts that accurately predict these aspects of interest.

To target particular outcomes whilst quantifying forecast accuracy, scoring rules can be adapted to incorporate a weight function $w : \mathcal{Y} \to \mathbb{R}_{\geq 0}$, which can be chosen to assign greater weight to outcomes of greater interest. This recognizes that accurate forecasts for certain outcomes, such as extreme events, are often more valuable than forecasts of other outcomes.

Diks et al. (2011) introduce the censored likelihood score, a weighted version of the log score, as

$$\text{cLS}(F, y; w) = -w(y)\log f(y) - [1 - w(y)] \times \log\left[1 - \int_{-\infty}^{\infty} w(z) f(z)\,\mathrm{d}z\right]. \tag{10}$$

Here, the weight function needs to satisfy $0 \leq w(z) \leq 1$. Similarly, Gneiting and Ranjan (2011) introduce the threshold-weighted CRPS,

$$\text{twCRPS}(F, y; w) = \int_{-\infty}^{\infty} [F(z) - \mathbb{1}\{y \leq z\}]^2\, w(z)\,\mathrm{d}z. \tag{11}$$

The weight function here is not necessarily restricted to the unit interval. The log score and CRPS are recovered when $w(z) = 1$ for all $z \in \mathcal{Y}$.

To emphasize values above a threshold $t \in \mathbb{R}$, it is common to employ the weight function $w(z) = \mathbb{1}\{z \geq t\}$. In this case, the censored likelihood score and twCRPS both censor the forecast distribution and the outcome at the threshold, before applying the relevant unweighted scoring rule. This framework can readily be applied to other scoring rules, also in a multivariate context (Allen, Bhend et al., 2023; de Punder et al., 2023). Holzmann and Klar (2017) propose an alternative framework to construct weighted scoring rules that separately assess the conditional forecast distribution given that the threshold has been exceeded, and the forecast probability of a threshold exceedance; this encompasses the censored likelihood score but not the twCRPS.

Since weighted scoring rules allow events of interest to be targeted during forecast evaluation, they can similarly be used to target them during model training. The resulting forecasts should issue more accurate predictions for these events. Wessel et al. (2025) propose training statistical post-processing methods for weather forecasts by using weighted scoring rules as loss functions. They find that doing so tends to decrease overall forecast performance, as measured by an unweighted scoring rule, but improves performance in the region of interest. This illustrates the trade-off between generating accurate forecasts for more common outcomes and accurate forecasts for high-impact events.

If the weight function used in (10) and (11) is of the form $w(z) = \mathbb{1}\{z \geq t\}$, then the censored likelihood score and twCRPS are generally proper but not strictly proper; the scoring rules are insensitive to how the forecast distributions behave below the threshold. This can become problematic during model training, as it may lead to non-unique solutions to the optimization problem, depending on the class of models being trained. Instead, we consider a linear combination of weighted and unweighted scoring rules. For example, for the CRPS, we employ

$$\text{CRPS}(F, y) + \gamma\,\text{twCRPS}(F, y; w), \tag{12}$$

where $\gamma \geq 0$ represents the amount of emphasis placed on the weighted scoring rule, and the weight function in the twCRPS is $w(z) = \mathbb{1}\{z \geq t\}$. This combined score itself corresponds to the twCRPS with weight function $w(z) = 1 + \gamma\mathbb{1}\{z \geq t\}$, which has been proposed for evaluation purposes by Lerch et al. (2017) and Thorarinsdottir and Schuhen (2018). Whenever the CRPS is strictly proper, this linear combination will also be strictly proper. An analogous combination can be constructed using the log score and the censored likelihood score, or any scoring rule and its weighted counterpart. Note that the scale of the weighted and unweighted scores will generally differ, and hence γ has no direct interpretation; it must therefore be chosen using some exploratory analysis. In Sections 5 and 6, we explore how γ influences the forecast performance.

4.2. Calibration and tail calibration regularization

To direct optimal score estimators towards a calibrated solution, Wilks (2018) suggest introducing a measure of miscalibration (MCB) and using it as a regularization term in the loss function. That is, to train the forecasting model by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i) + \gamma\text{MCB}, \tag{13}$$

where $\gamma \geq 0$ represents the amount of emphasis placed on the regularization term, with $\gamma = 0$ recovering the unregularized optimum score estimator. The miscalibration term MCB is implicitly a function of the training data $\{(F_1, y_1), \ldots, (F_n, y_n)\}$. Adding the regularization term should yield better-calibrated forecasts, though possibly with a worse average score. This trade-off is amplified as the parameter γ is increased.

An obvious question is what miscalibration term to use. Since the loss function will be evaluated multiple times whilst being optimized, MCB should be easy to calculate. Wilks (2018) consider a measure of miscalibration that is specific to discrete forecast distributions, but in principle, any divergence between the empirical distribution of the n PIT values $\{z_1 = F_1(y_1), \ldots, z_n = F_n(y_n)\}$ and a uniform distribution could be considered.

Here, we employ the Wasserstein-1 distance,

$$\text{MCB} = \int_0^1 \left|\hat{H}_z(u) - u\right| \, du, \tag{14}$$

where $\hat{H}_z(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{z_i \leq u\}$, $u \in [0, 1]$, is the empirical distribution function of the n PIT values. The larger the MCB, the more the distribution of PIT values deviates from a standard uniform distribution, whereas smaller MCB values indicate better calibration. The Wasserstein-1 distance can equivalently be written in terms of the quantile function associated with $\hat{H}_z$, and (14) is estimated in practice by calculating the average absolute distance between the values $\{1/n, 2/n, \ldots, 1\}$ and the sorted PIT values $\{z_{(1)}, \ldots, z_{(n)}\}$ as $\frac{1}{n}\sum_{i=1}^{n}|z_{(i)} - i/n|$.

The MCB term of (14) provides a measure of probabilistic miscalibration, and the regularized loss function of (13) therefore encourages forecast models to issue probabilistically calibrated forecasts. However, probabilistic calibration does not imply that the forecasts are calibrated in the tails. Optimizing a weighted scoring rule during model training can emphasize the tails, but it may reward discriminative ability over calibration, so the resulting forecasts will not necessarily be tail-calibrated. If we wish to enforce tail calibration, then we could adapt (13) by replacing the measure of miscalibration (MCB) with a measure of tail miscalibration (TMCB),

$$\frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i) + \gamma \text{TMCB}, \tag{15}$$

where $\gamma \geq 0$. The measure of tail miscalibration again depends on $\{(F_1, y_1), \ldots, (F_n, y_n)\}$, as well as the threshold of interest.

We employ a similar measure to quantify tail miscalibration:

$$\text{TMCB} = \int_0^1 \left|\hat{Q}_t(u) - u\right| \, du, \tag{16}$$

where $\hat{Q}_t$ is as defined in (9). TMCB tends towards MCB as the threshold $t \to -\infty$. To estimate (16), we use the average absolute difference $\frac{1}{n_t}\sum_{i=1}^{n_t}|\hat{O}_t \cdot z_{(i),t} - i/n_t|$, where $z_{(1),t} \leq \cdots \leq z_{(n_t),t}$ are the order statistics of the empirical CPIT values, with $n_t = |\mathcal{I}_t|$. This exploits the decomposition $\hat{Q}_t(u) = \hat{O}_t \cdot \hat{H}_{z_t}^{-1}(u)$ in (9).

Instead of using $\hat{Q}_t$ to define TMCB, one could also quantify tail miscalibration by penalizing the deviation of the distribution of CPIT values from a standard uniform distribution, for example, by replacing $\hat{Q}_t(u)$ in (16) with the empirical distribution $\hat{H}_{z_t}(u)$. While this approach would similarly penalize forecasts with a miscalibrated excess distribution, it would neglect the forecast probability of an extreme event. Hence, as we show in Appendix D, this yields CPIT values that are closer to uniform but severely degrades forecast performance across all other metrics considered herein. This highlights the need to assess the calibration of both threshold exceedance probabilities (forecasts for extreme event occurrence) as well as the forecast excess distribution (forecasts for extreme event severity), and is similar in essence to the forecaster's dilemma discussed by Lerch et al. (2017).

Divergences other than the Wasserstein-1 divergence could analogously be considered for the definition of MCB and TMCB, and we additionally studied the Wasserstein-2 and Kolmogorov–Smirnov distances. The results were qualitatively similar in all cases, though the Wasserstein distances penalized more homogeneously across the domain of PIT and $\hat{Q}(u)$ values, thereby yielding smoother objective functions for parameter optimization. The Cramér distance could also be employed for this purpose. Wilks (2018) demonstrated that the Cramér distance often yields more powerful calibration tests than the Wasserstein-1 distance, though we conjecture that the results will be similar to those for the Wasserstein distance when used for optimization, since the gradients of these two divergence measures point in the same direction. The Wasserstein distances have the advantage of being efficiently computed from the order statistics of the CPIT values. In the following, we therefore only present results for the Wasserstein-1 distance.

The miscalibration and tail miscalibration penalties do not correspond to average scores resulting from proper scoring rules. However, if the forecast equals the conditional distribution of the outcome given the covariates, then it will be both probabilistically calibrated and probabilistically tail-calibrated. In this case, the (tail) miscalibration penalties will be equal to zero, their minimum value. As a consequence, if the model class is well-specified and the baseline scoring rule is strictly proper, then the regularized loss functions are still minimized uniquely at the true conditional distribution. That is, the regularization terms do not affect the consistency of the optimum score estimators.

5. Simulation example

To demonstrate how the regularization terms described above affect the estimated parameters of the forecast model, consider a simple example using simulated data. Suppose $Y \mid \mu \sim N(\mu, 1)$, where $\mu \sim \text{Unif}(-3, 3)$, and define a random variable τ that is equal to 1 or -1 with equal probability, independently of (Y, μ). Suppose we are interested in outcomes Y that exceed the threshold $t = 3.23$, roughly the 95th percentile of the unconditional distribution of Y.

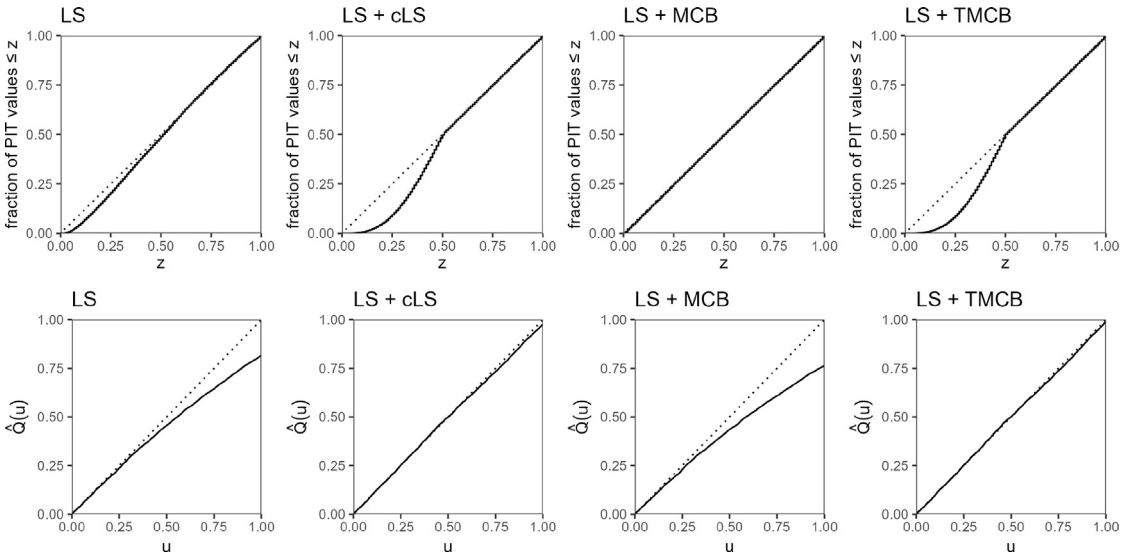

Fig. 1. Calibration plots when estimating a using standard maximum likelihood (first column), and when regularizing with the censored likelihood score (second), miscalibration (third), and tail miscalibration (fourth). Results are shown for $\gamma = 2$ when assessing standard calibration (top row), and tail calibration (bottom).

Let $F_1 = \frac{1}{2}N(\mu, 1) + \frac{1}{2}N(\mu + \tau, 1)$, and let F_2 be such that

$$F_2(x) = \begin{cases} \Phi(x - \mu) & \text{if } x \geq \mu, \\ \Phi\left(\frac{x-\mu}{2}\right) & \text{if } x < \mu. \end{cases}$$

The forecast F_1 is a slight adaptation of the unfocused forecaster of Gneiting et al. (2007, Example 3), which is a probabilistically calibrated forecast for Y, but is not probabilistically tail calibrated. The forecast F_2 is equal to the true conditional distribution of Y given μ for all $x \geq \mu$, but has the incorrect scale below μ. Since $\mu < t$ almost surely, F_2 is therefore probabilistically tail calibrated but not probabilistically calibrated.

Suppose that we wish to estimate the optimal linear combination of the two forecasts F_1 and F_2. That is, we wish to estimate the optimal mixing parameter $a \in [0, 1]$ in the combined forecast distribution $F_a = aF_1 + (1 - a)F_2$. We estimate a by minimizing the log score (maximizing the likelihood) of the combined forecast F_a over $n = 100,000$ realizations of μ, τ, and Y. The resulting parameter estimate is $\hat{a} = 0.72$, meaning F_1 is assigned a higher weight in the combination. The combined forecast $F_{\hat{a}}$ achieves an average log score of 1.52 over the n realisations, compared with 1.53 and 1.58 for F_1 and F_2, respectively. However, while $F_{\hat{a}}$ is the optimal combination of F_1 and F_2 according to the log score, the resulting forecasts are neither calibrated nor tail calibrated. Calibration plots are displayed in Fig. 1.

Consider now the parameters estimated by minimizing the log score with the regularization terms introduced in the previous section. Regularization is performed using a penalty for the censored likelihood score (cLS), overall miscalibration (MCB), and tail miscalibration (TMCB). The cLS and TMCB terms are calculated using the threshold t. The estimated parameters for all methods are shown as a function of γ in Fig. 2. When penalizing overall miscalibration during model training, the estimated mixing parameter tends to one as γ increases, recovering the unfocused forecaster F_1. The average log score of F_1 is marginally worse than that of $F_{\hat{a}}$, but F_1 is probabilistically calibrated whereas $F_{\hat{a}}$ is not (Fig. 1), confirming that the miscalibration penalty yields forecasts that are better calibrated.

By instead penalizing tail miscalibration during model training, the estimated mixing parameter tends to zero as γ increases, recovering the piecewise forecast F_2. The average log score of F_2 is also worse than that of $F_{\hat{a}}$, but F_2 is probabilistically tail calibrated whereas $F_{\hat{a}}$ is not (Fig. 1), confirming that the tail miscalibration penalty yields forecasts that are better tail calibrated.

Similar results are observed when regularizing with the censored likelihood score during model training. The censored likelihood score emphasizes outcomes exceeding the threshold t when calculating forecast accuracy, implicitly encouraging tail calibration. The estimated mixing parameter, therefore, again tends towards zero as γ increases. The parameter tends to zero more slowly than in tail miscalibration regularization. Still, since the censored likelihood score is generally on a different scale to the measures of miscalibration, the dependence on γ is not directly comparable between these methods; the γ values in Fig. 2 for this method have been rescaled by a constant factor to aid visual comparison.

6. Weather forecasting application

Consider now an application of the regularization approaches in the context of weather forecasting. We apply the approaches to a hierarchy of increasingly complex statistical post-processing models that issue probabilistic weather forecasts based on the output of numerical weather prediction models; further details on statistical post-processing can be found in Vannitsem et al.

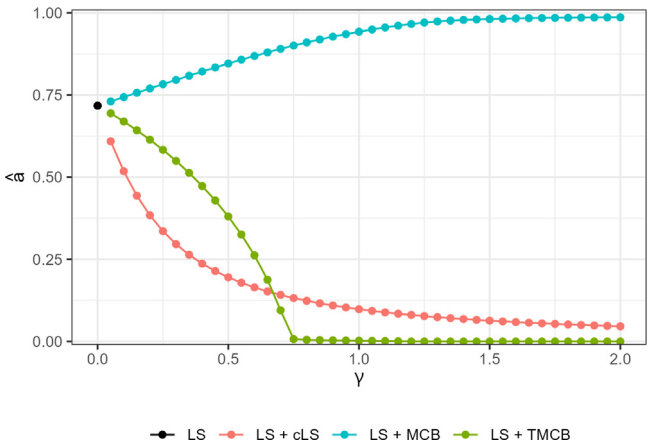

Fig. 2. Estimated parameter a in the simulation example as a function of γ when penalizing miscalibration (blue), tail miscalibration (green), and censored likelihood score (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(2018). The three statistical models we consider are: ensemble model output statistics (EMOS), a simple linear parametric approach that is well established in operational post-processing suites (Gneiting et al., 2005); distributional regression networks (DRNs), a non-linear extension of EMOS models based on neural networks (Rasp & Lerch, 2018); and conditional generative models (CGMs), a non-parametric generative model-based framework with a neural network architecture (Chen et al., 2024). These methods are state-of-the-art in the post-processing literature, and we study how adapting the loss function used to train these models can influence the behavior of the resulting forecasts, particularly when the focus is on extreme events.

6.1. Data and methods

We consider forecasts of 10 m wind speed from the UK Met Office's global ensemble prediction system (MOGREPS-G; Porson et al., 2020; Walters et al., 2017). The gridded MOGREPS-G output is bilinearly interpolated to the locations of 124 synoptic observation stations, shown in Fig. 3, at which wind speed observations are available. We consider forecasts initialized at 00UTC and issued 72 h in advance. Forecasts and observations from 1 April 2019 to 31 December 2020 are used to train the statistical post-processing methods, which are then evaluated on data from 1 January 2021 to 31 March 2022.

The three post-processing models are discussed in the following sections. All models are trained by optimizing the CRPS, the most popular scoring rule for assessing probabilistic weather forecasts. The same models are then additionally trained using the CRPS with the miscalibration, tail miscalibration, and twCRPS regularization terms discussed in Section 4, with the results then compared to the baseline CRPS-trained models with no regularization. We explore the sensitivity of the results to different values of the penalty parameter γ.

All models use the same covariates: the mean and standard deviation of the MOGREPS-G ensemble forecast at the corresponding time and location, as well as the sine and cosine of the day of the year, allowing the model to capture seasonal variations in the wind speed; these harmonic terms additionally ensure that model performance is more consistent across the year, helping to account for climatological differences in the training and test (evaluation) datasets that arise due to them containing data for differing months. Further details about the model design and implementation are provided in Appendix B.

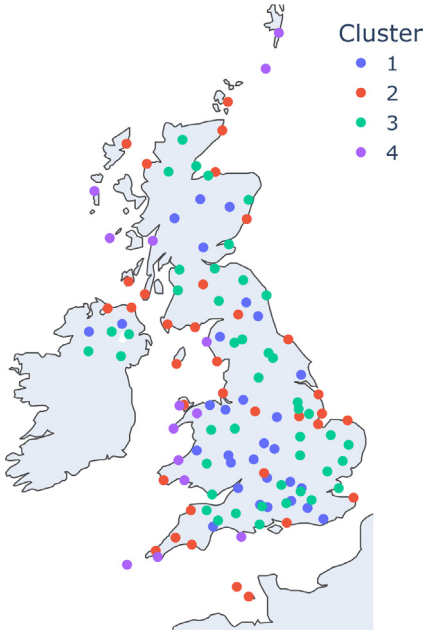

Fig. 3. Locations of the 124 weather stations considered in the application. The color of each point corresponds to the cluster it is assigned to for the semi-local parameter estimation of the EMOS model (see Section 6.2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For the definition of an extreme event, we use an absolute wind speed threshold of 12.5 m/s, which is roughly equal to the 97.5th percentile of the distribution

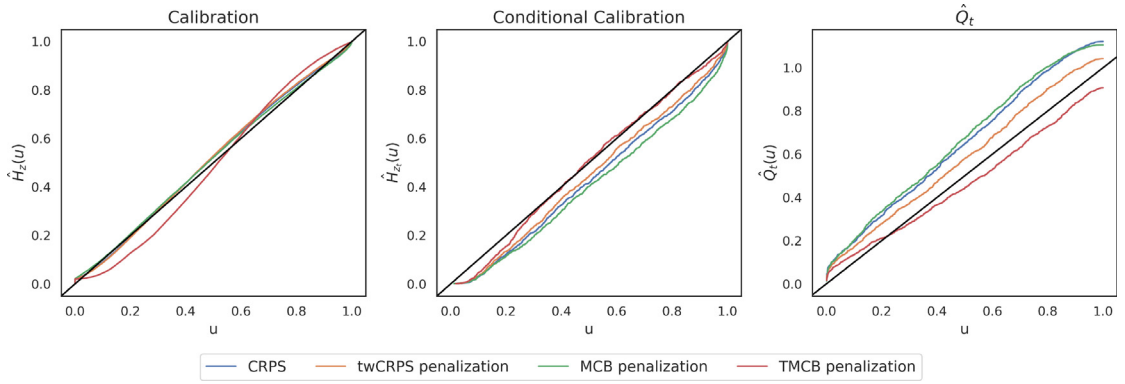

Fig. 4. Calibration plots for EMOS models trained by minimizing the CRPS with different regularization terms, for $\gamma = 5$. Left: empirical distribution function of PIT values. Middle: empirical distribution function of conditional PIT values. Right: $\hat{Q}_t(u)$ as a function of $u \in [0, 1]$. In all cases, a calibrated forecast should result in lines close to the black diagonal line. Results are pooled across all stations.

of wind speeds in the training data (across all stations). This threshold is used for both model training and evaluation. A range of thresholds was also studied, all of which yielded similar conclusions.

6.2. Ensemble model output statistics

Ensemble model output statistics (EMOS; Gneiting et al., 2005; Jewson et al., 2004) is a popular post-processing technique that models the target variable using a parametric distribution whose location and scale parameters depend linearly on covariates. This can be thought of as a particular Generalized Additive Model for Location, Scale and Shape (GAMLSS; Rigby & Stasinopoulos, 2005). Following Thorarinsdottir and Gneiting (2010), we model wind speed using a truncated normal distribution. For parameter optimization, we use the closed-form expressions of the CRPS and twCRPS for a truncated normal distribution given by Thorarinsdottir and Gneiting (2010) and Wessel et al. (2024), respectively.

EMOS parameters are estimated separately for four clusters of the 124 weather stations. The distribution of wind speed observations in the training dataset for each location determines the clusters. This corresponds to a semi-local post-processing approach, which has been proposed as a means to augment training data (see e.g. Lerch & Baran, 2016), making it particularly useful when interest is in extreme events. The resulting clusters are shown in Fig. 3. Further details about the model design, optimization framework, and the semi-local estimation approach are given in Appendix B.1. To ensure consistency with the evaluation of the other models in the following sections, the tail calibration of the EMOS models in each cluster is assessed using the same threshold of 12.5 m/s. However, we additionally analyzed the results for cluster-wise thresholds, and the findings were analogous (not shown).

Fig. 4 presents calibration plots for the forecasts issued by EMOS models trained using the CRPS with different regularization terms (with $\gamma = 5.0$). All models issue reasonably well-calibrated forecasts, with the distribution function of empirical PIT values lying close to the diagonal, though penalizing tail miscalibration results in slightly overdispersed forecasts. As expected, the CRPS-trained model yields forecasts that appear calibrated according to the standard notion of probabilistic calibration. However, these forecasts appear miscalibrated when predicting more extreme wind speeds: the conditional PIT values suggest that the estimated tail is too light. This reinforces the idea that standard checks for probabilistic calibration are insufficient to infer whether forecasts are tail-calibrated.

Penalizing overall miscalibration during model training has little impact on the resulting forecasts in this case, and these forecasts are therefore similarly miscalibrated in the tails. In contrast, regularizing the loss function with the twCRPS or a measure of tail miscalibration yields more reliable forecasts for extreme events. This can be seen both in improvements in the uniformity of CPIT values and in $\hat{Q}_t$ values that are closer to the diagonal. However, this comes at the expense of standard calibration.

Table 2 quantifies the improvement of the various regularization methods (again with $\gamma = 5.0$) relative to the CRPS-trained model when assessed using the CRPS, twCRPS, and the different measures of calibration. Penalizing tail miscalibration during model training can improve the tail calibration of the resulting forecasts by more than 60%, though the overall calibration becomes 187% worse; it should be noted that since the baseline CRPS-trained model is reasonably well calibrated overall but not in the tails, the 60% improvement in tail calibration corresponds to a larger absolute difference than the 187% deterioration in overall calibration (see Fig. 5). Adding the twCRPS to the loss function yields similar, but more moderate, changes in the considered metrics, with a minor improvement in twCRPS relative to the baseline.

Fig. 5 presents the absolute difference in overall miscalibration and tail miscalibration from the baseline CRPS-trained model as a function of γ. Since the twCRPS is generally on a smaller scale than measures of miscalibration, it is difficult to directly compare results across methods for a given value of γ. When penalizing the twCRPS or tail miscalibration, increasing γ results in a larger improvement with respect to tail miscalibration (before plateauing for larger γ), and a larger deterioration in overall calibration. Penalizing overall miscalibration

Table 2
Relative improvement in evaluation metrics for the regularized EMOS models compared to the baseline CRPS-trained EMOS model. Results are shown for $\gamma = 5.0$. The largest improvement for each metric is displayed in bold. The average CRPS of the baseline model is 1.01.

		Regularization term		
		MCB	TMCB	twCRPS
Proper scores	CRPS skill (%)	−0.28	−4.77	−0.09
	twCRPS skill (%)	−1.04	−1.16	**0.12**
Calibration	MCB skill (%)	**17.10**	−187.19	−13.55
	TMCB skill (%)	−10.46	**65.36**	44.81

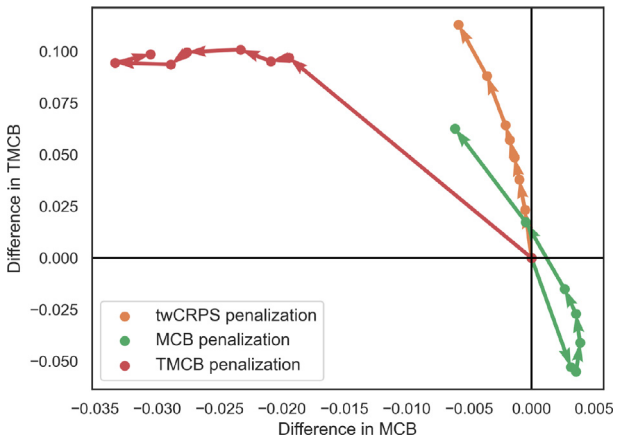

Fig. 5. Absolute difference in MCB versus the absolute difference in TMCB for the EMOS models trained using the different regularization terms, relative to the CRPS-trained EMOS model with no regularization. Positive values indicate a better (T)MCB for the regularized model than for the CRPS-trained baseline. Results are shown for $\gamma = 0$ (the point at (0, 0)) and $\gamma \in \{1, 2, 3, 4, 5, 10, 20\}$, with arrows denoting the transitions between consecutive γ values. Note the different scales of the x- and y-axes.

generally has small effects on both overall and tail calibration. For larger values of γ, no improvement in overall calibration can be seen on the test dataset, though such an improvement exists on the training dataset. These results should be interpreted with caution, as the estimation of EMOS parameters with (T)MCB regularization can become somewhat unstable for large values of γ.

6.3. Distributional regression networks

Distributional regression networks (DRNs, Rasp & Lerch, 2018) can be seen as a non-linear extension of EMOS. Wind speed is again assumed to follow a truncated normal distribution. Still, the location and scale of the distribution are now modeled as outputs of a feed-forward neural network, with the covariates as inputs. Following Rasp and Lerch (2018), the station index is also incorporated as a covariate via a learned embedding layer, allowing one spatially-aware DRN model to be fit using the data from all stations. More details on training the DRNs are given in Appendix B.2.

Since DRNs output a truncated-normal predictive distribution, the network parameters can again be trained using the available closed-form expressions for the CRPS and twCRPS. However, in contrast to EMOS models, DRNs are heavily over-parametrized, and (CRPS-based) neural network training presents a non-convex optimization problem. Hence, when training multiple DRNs, the resulting neural networks will generally differ in their parameters due to randomness in both the initial weight values and the stochastic gradient descent algorithm. To circumvent this, we fit the DRN 100 times and analyze the resulting scores and calibration metrics across these 100 models, allowing us to understand how the results vary across different learned parameter configurations. This is repeated when training the DRN using each regularization term.

Fig. 6 shows the calibration and tail calibration of the 100 DRN models trained by minimizing the CRPS without regularization, as well as the resulting CRPS in the test data. There is low variation in the CRPS of the 100 DRN models and in the distribution of the corresponding PIT values, with all models yielding well-calibrated forecasts. However, the resulting forecasts can behave very differently in terms of tail calibration, despite the fairly rigid parametric assumptions of the DRN. Some models yield forecasts that appear reasonably well calibrated in the tails, while others exhibit severe miscalibration. Both the CRPS and the overall calibration are insensitive to this variation in tail calibration. This variation stems solely from training randomness, highlighting the need for strategies to reduce and better control such variability.

We finetune the CRPS-trained DRN models using the different regularization terms discussed above (for details, see Appendix B.2). Fig. 7 shows the distribution of the CRPS, twCRPS, and measures of miscalibration for the 100 DRN models trained using the CRPS with the different regularization terms, for different values of γ. All models yield similarly accurate forecasts when assessed using the CRPS and twCRPS. Penalizing tail miscalibration

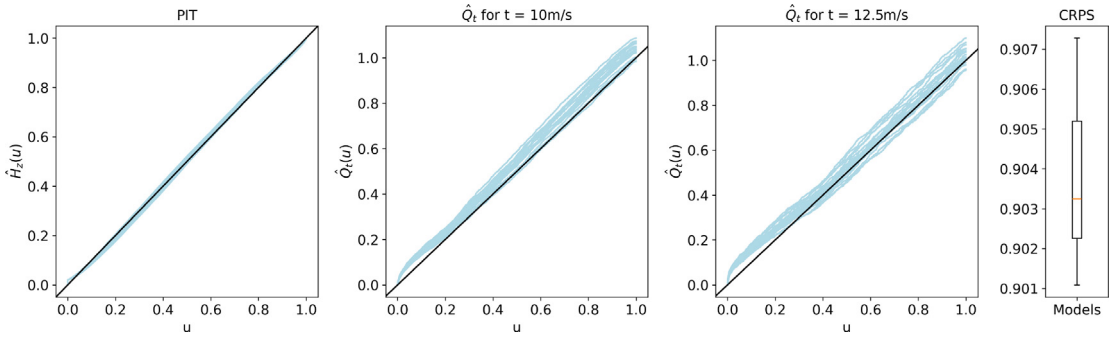

Fig. 6. Empirical distribution function of PIT values (first panel) and tail calibration diagnostic plots (second and third panels) for the 100 CRPS-trained DRN models, each shown by a light blue line. Tail calibration plots are shown at two thresholds: $t = 10$ m/s and $t = 12.5$ m/s. The diagonal black lines denote perfect calibration. The final panel shows the distribution of the CRPS for the 100 DRN models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

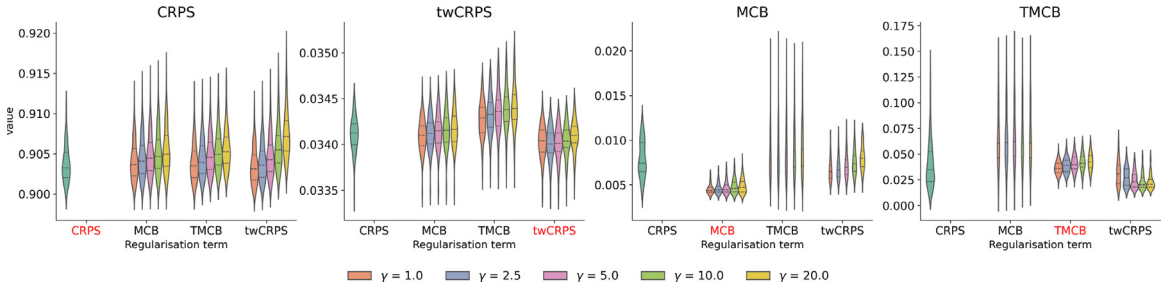

Fig. 7. Violin plots of metrics for the 100 DRN models trained using different regularization terms. Different panels correspond to different evaluation metrics, while the *x*-axis labels correspond to the different regularization terms. Red *x*-axis labels indicate when the evaluation metric matches the regularization metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

or twCRPS during training does not appear to substantially decrease forecast accuracy in this study, though it generally does not improve twCRPS. However, training with these penalties substantially reduces variation in tail miscalibration measures across the 100 DRN models.

Table 3 shows the relative improvement in each metric compared to the CRPS-trained baseline, for $\gamma = 5.0$. The improvement is based on the average evaluation metric across the 100 DRN models. The results are qualitatively similar to those obtained for the EMOS forecasts. Penalizing overall miscalibration improves the calibration of the resulting forecasts, at the expense of accuracy and tail calibration. Conversely, penalizing tail miscalibration during model training improves tail calibration at the expense of overall calibration. Adding the twCRPS to the loss function yields yet larger improvements in the tail calibration of the DRN models, as well as a slight improvement in overall calibration.

The improvement in tail calibration is not consistent across the 100 DRN models and depends on the degree of miscalibration of the CRPS-trained model. Fig. 8 displays the (absolute) improvement in tail calibration against the tail miscalibration of the CRPS-trained model. Results are shown for models trained with regularization, using the twCRPS and the tail miscalibration, for $\gamma = 5.0$. Intuitively, the larger the tail miscalibration of the CRPS-trained model, the larger the improvements obtained by penalizing tail miscalibration during model training. When the CRPS-trained model already issues reasonably well-calibrated forecasts in the tails, the additional regularization term can introduce unnecessary uncertainty into parameter estimation, slightly hindering the tail calibration of the resulting forecasts.

6.4. Conditional generative models

Conditional generative models (CGMs, Chen et al., 2024) provide a non-parametric forecasting framework that constructs predictive distributions by learning a mapping among the covariates, a latent random variable, and the target variable. The map is typically modeled using a neural network. CGMs do not rely on parametric assumptions, giving them more flexibility than EMOS and DRN models. However, while EMOS and DRNs output a continuous parametric distribution, the CGM predictive distribution is only available via a sample of possible outcomes. We refer to Chen et al. (2024), Pacchiardi et al. (2024), Schillinger et al. (2025) for more background on CGMs.

As for the DRN, we fit one CGM to data from all stations, using an embedding layer to encode spatial information. This is repeated 100 times to account for uncertainty in the model training. The CGM model architecture is similar to Chen et al. (2024), except that while they designed CGMs to obtain multivariate predictive distributions, we implement them here in the univariate setting; that is, the map that we learn using the CGM is a function from the covariates and latent variable to the univariate

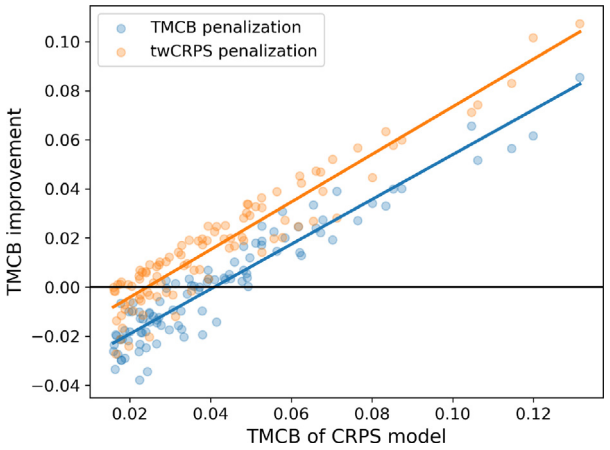

Fig. 8. Improvement in TMCB obtained from the regularized model ($\gamma = 5.0$) relative to the baseline CRPS-trained model, plotted against the TMCB of the baseline model. Results are shown for twCRPS (orange) and tail miscalibration (blue) regularization. Each point corresponds to one of the 100 DRN models, with the solid lines representing linear regression fits to these 100 points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

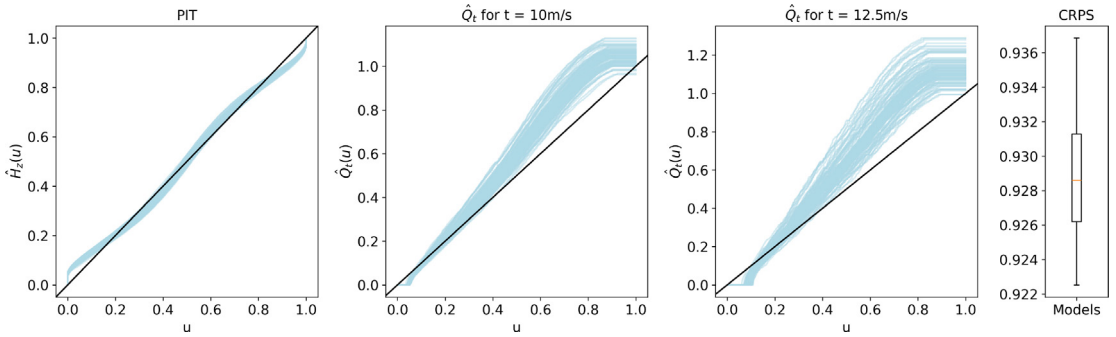

Fig. 9. Empirical distribution function of PIT values (first panel) and tail calibration diagnostic plots (second and third panels) for the 100 CRPS-trained CGMs, each shown by a light blue line. Tail calibration plots are shown at two thresholds: $t = 10$ m/s and $t = 12.5$ m/s. The diagonal black lines denote perfect calibration. The final panel shows the distribution of the CRPS for the 100 CGMs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Relative improvement of average evaluation metrics for the 100 regularized DRN models compared to the 100 baseline CRPS-trained DRN models. Results are shown for $\gamma = 5.0$. The largest improvement for each metric is displayed in bold.

		Regularization term		
		MCB	TMCB	twCRPS
Proper scores	CRPS skill (%)	−0.12	−0.12	−0.11
	twCRPS skill (%)	−0.11	−0.73	**0.28**
Calibration	MCB skill (%)	**40.97**	−15.83	10.17
	TMCB skill (%)	−59.08	2.33	**48.90**

wind speed at a given time and location. We sample 250 values from the generative model, resulting in predictive distributions that are reasonably smooth. Parameter estimation is performed using sample-based estimates of the CRPS and twCRPS, and an additional smoothing of the PIT and CPIT values is also applied during model training to ensure that the gradients of the miscalibration penalties exist. This is discussed in Appendix D.

Fig. 9 shows calibration plots for predictive distributions obtained from the 100 CGM models, trained using the CRPS. The models show very little variability in overall calibration and CRPS. Minor deviations of the PIT values from the diagonal suggest that the CGM forecasts are not perfectly calibrated. The CGM forecasts are generally more accurate than the EMOS forecasts but less accurate than the DRN forecasts, as assessed using the average CRPS. There is again a large variability in the tail calibration of the CRPS-trained CGM forecasts. However, all 100 models yield miscalibrated forecasts for extreme outcomes.

We again fine-tune the CRPS-trained CGM models using the tail-calibration penalties. Fig. 10 shows the distribution of the evaluation metrics for the 100 CGM models trained using the CRPS with the different regularization terms, for different values of γ. Table 4 presents the

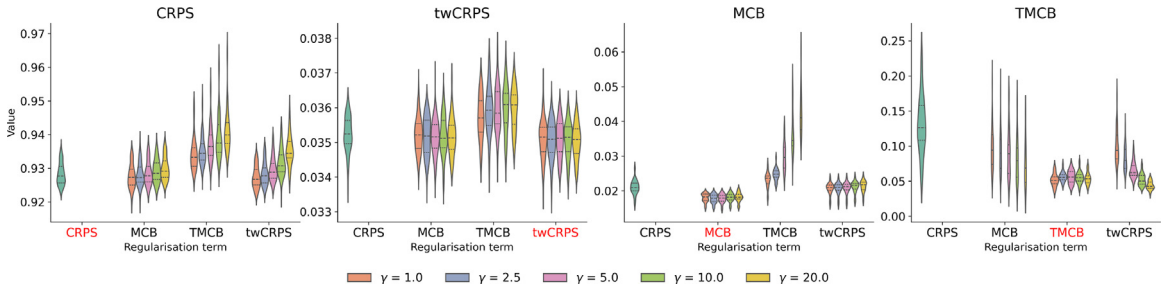

Fig. 10. Violin plots of metrics for the 100 CGMs trained using different regularization terms. Different panels correspond to different evaluation metrics, while the *x*-axis labels correspond to the different regularization terms. Red *x*-axis labels correspond to when the evaluation metric is the same as the metric used for regularization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Relative improvement of average evaluation metrics for the 100 regularized CGMs compared to the 100 baseline CRPS-trained GCMs. Results are shown for $\gamma = 5.0$. The largest improvement for each metric is displayed in bold.

		Regularization term		
		MCB	TMCB	twCRPS
Proper scores	CRPS skill (%)	−0.02	−0.98	−0.15
	twCRPS skill (%)	0.26	−2.03	**0.43**
Calibration	MCB skill (%)	**15.45**	−42.35	1.15
	TMCB skill (%)	32.80	**56.88**	49.28

relative improvement in average metrics over the baseline CRPS-trained CGMs for $\gamma = 5.0$. The results are qualitatively similar to those obtained for EMOS and the DRNs: penalizing twCRPS or tail miscalibration during model training results in slightly worse CRPS and overall calibration, but can significantly improve the tail calibration of the resulting forecasts; the variation in tail miscalibration across the 100 models is also significantly decreased via these regularization terms; twCRPS regularization has less impact on the CRPS and overall calibration, but generally improves tail calibration, although less than the tail calibration regularization. Penalizing overall miscalibration during model training helps address the slight miscalibration of the original CGM forecasts, while also slightly improving twCRPS and tail calibration, likely because the increased spread of the resulting predictive distributions means fewer observations fall outside the range of the 250 samples from the model. The improvement in tail calibration again depends on the original miscalibration of the CRPS-trained CGM (Fig. 11).

7. Conclusions

This work studies how the loss functions used to train probabilistic forecast models can be adapted to target extreme events. While parameters of probabilistic models are typically estimated by optimizing a proper scoring rule over a training dataset, if the model class is not well-specified or only a small training dataset is available, the choice of scoring rule will influence the behavior of the resulting forecasts. In particular, there is no guarantee that the model's predictions will be calibrated or tail-calibrated.

In an application to wind speed forecasting, we demonstrate that three state-of-the-art statistical and machine learning models fail to issue calibrated forecasts for extreme outcomes. This includes forecasts issued by simple parametric approaches, distributional regression networks, and conditional generative models. We therefore consider how the training of these methods could be adapted to improve the reliability of the resulting forecasts.

We consider two approaches. The first uses a weighted scoring rule to target extreme outcomes during model training, as studied by Wessel et al. (2025). The motivation is that if forecasts for extreme events are to be evaluated using weighted scoring rules, it makes sense to deploy them when fitting the forecast model; the resulting forecasts should be more accurate when assessed using the same scoring rules with which they are trained. Alternatively, the (tail-) miscalibration could be penalized more directly by adding a measure of (tail-) miscalibration as a regularization term. This measure of miscalibration does not capture the information content of the forecast and can therefore lead to greater deterioration in forecast accuracy, though it generally produces more reliable forecasts. We find that both training with a tail-calibration penalty and using a weighted score improve out-of-sample tail calibration. This, however, can come at the expense of overall probabilistic calibration and forecast skill.

We studied extreme wind-speed events defined as exceedances of a moderately high threshold, chosen based on relevant quantiles of historical observations. However, such threshold exceedances do not necessarily correspond to impactful events in reality. While it would be desirable to consider higher thresholds, such an analysis is prohibited by data availability, which remains a crucial challenge for characterizing and forecasting extreme events. The motivation behind this work was to demonstrate

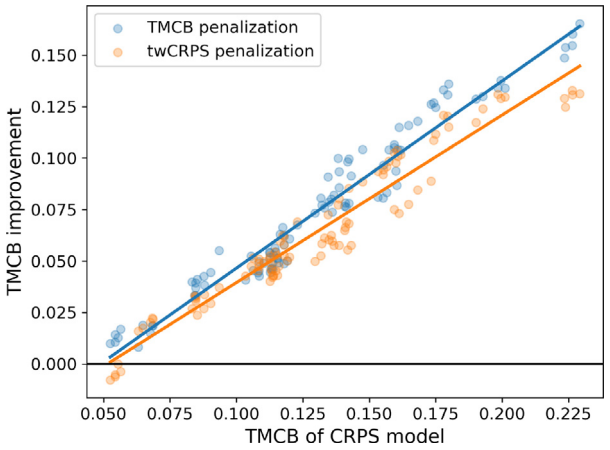

Fig. 11. Improvement in TMCB obtained from the regularized model ($\gamma = 5.0$) relative to the baseline CRPS-trained model, plotted against the TMCB of the baseline model. Results are shown for twCRPS (orange) and tail miscalibration (blue) regularization. Each point corresponds to one of the 100 CGM models, with the solid lines representing linear regression fits to these 100 points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

how the loss function can be adapted to obtain better-calibrated forecasts for particular outcomes. Future work should assess how the proposed regularization strategies perform as the amount of training data changes, which would, for example, help to understand the scalability and robustness of the results in data-rich settings.

Nonetheless, while we focus on extreme events, the notion of tail calibration readily extends to other regions of the outcome space (Allen, Bhend et al., 2023; Mitchell & Weale, 2023). Local miscalibration within these regions could similarly be penalized during model training to obtain more reliable forecasts for these particular outcomes. This could be used to obtain more reliable forecasts for outcomes within a particular range, below a threshold of interest, or in disjoint regions of interest. The framework could additionally be extended to multivariate probabilistic forecasts. However, the calibration of multivariate forecasts has received much less attention in the literature, making it difficult to define notions of multivariate tail calibration. On the other hand, Allen, Ginsbourger et al. (2023) introduce weighted multivariate scoring rules, and it would be interesting to assess whether they can be employed during model training to yield more accurate and reliable forecasts for multivariate extremes.

Other models, beyond the three considered in Section 6, can similarly be regularized to penalize (tail) miscalibration. Forecasting methods that do not rely on optimizing an objective function, such as k-nearest neighbors, cannot, since regularization relies on having a parameter estimate that can be pushed towards a better-calibrated solution. This regularization approach is also not necessary for forecast methods that are guaranteed to be calibrated in the training data by construction, such as isotonic distributional regression (IDR; Henzi et al., 2021) and binning or analog methods (Allen, Gavrilopoulos et al., 2025, Proposition 2.4). However, the regularization approach is also applicable to methods relying on objective functions not constructed from proper scoring rules.

The (tail) miscalibration regularization term penalizes probabilistic (tail) miscalibration, which is the notion of forecast calibration most commonly assessed in practice. Probabilistic calibration is a fairly weak notion of calibration (Gneiting & Resin, 2023), though it can therefore be interpreted as a minimum requirement that probabilistic forecasts should satisfy. While it would be preferable to penalize a stronger notion of calibration, such as auto-calibration (Gneiting & Ranjan, 2013; Tsyplakov, 2011), such notions are generally difficult to compute efficiently (or at all), limiting their applicability in loss functions.

The methods used herein weakly enforce tail calibration by adding an additional regularization term to the loss function. Future work could consider methods to enforce strong tail calibration. For example, conformal prediction has recently received considerable attention in the machine learning literature as a means of issuing predictions with theoretical, out-of-sample calibration guarantees (see Angelopoulos et al., 2024; Vovk et al., 2022, for textbook introductions). Conformal predictive systems are generally not constructed using optimal score estimators, but it may be possible to introduce conformal prediction strategies tailored to predicting extremes.

Similarly, the idea of regularizing the loss function relies on the notion that there are two objective functions to be minimized. In this paper, we combine these loss functions and study how forecast behavior changes as the mix of these losses varies. However, one could also treat this as a multi-objective optimization problem, in which multiple loss functions are optimized simultaneously. Multi-objective optimization is already commonly used in fields such as hydrology and climate science; Zerenner et al. (2021), for example, simultaneously optimizes the CRPS and the root-mean-squared error of corresponding point predictions. Instead, a proper scoring rule, a measure of forecast miscalibration, and a measure of tail miscalibration could be included as three objective functions that are to be minimized simultaneously. In the neural network setting, multiple gradient descent methods (e.g. Désidéri, 2012; Fliege & Svaiter, 2000; Mercier et al., 2018) could be used for this.

CRediT authorship contribution statement

Jakob Benjamin Wessel: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Maybritt Schillinger:** Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Frank Kwasniok:** Writing – review & editing, Supervision, Funding acquisition. **Sam Allen:** Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Jakob Wessel was supported during this work by an Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership (DTP) under grant award number 2696930. Sam Allen gratefully acknowledges funding from the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting". The authors thank Johanna Ziegel and two anonymous reviewers for their insightful comments on previous versions of the paper.

Appendix A. Tail miscalibration interpretation

Let $O_t = \mathbb{P}(Y > t)/\mathbb{E}[1 - F(t)]$, for $t \in \mathbb{R}$, and let H_{Z_t} be the conditional distribution function of $F_t(Y)$ given $Y > t$, with inverse $H_{Z_t}^{-1}$. From Definition 3, a forecast F is probabilistically tail calibrated at threshold t if $O_t = 1$ and $H_{Z_t}(u) = H_{Z_t}^{-1}(u) = u$, for all $u \in [0, 1]$. The forecast can be interpreted as having a tail that is too light if $O_t > 1$ and $H_{Z_t}(u) < u$ (or $H_{Z_t}^{-1}(u) > u$) for all $u \in (0, 1)$, and too heavy if $O_t < 1$ and $H_{Z_t}(u) > u$ (or $H_{Z_t}^{-1}(u) < u$) for all $u \in (0, 1)$. If $O_t > 1$ but $H_{Z_t}(u) > u$, then the forecast underpredicts the average exceedance probability at threshold t, but generally overpredicts the probability of higher threshold exceedances (the opposite is true if $O_t < 1$ but $H_{Z_t}(u) < u$). In this case, it is more difficult to make conclusions regarding whether the forecast exhibits a tail that is too light or too heavy.

Allen, Koh et al. (2025) note that a forecast is probabilistically tail calibrated if and only if $R_t(u) = O_t H_{Z_t}(u) = u$ for all $u \in [0, 1]$. This allows tail miscalibration to be assessed in practice via one plot of $u \mapsto R_t(u)$, rather than two separate plots of $t \mapsto O_t$ and $u \mapsto H_{Z_t}(u)$. However, as the authors remark, the resulting plot is more difficult to interpret, since errors in O_t and H_{Z_t} can cancel each other out in such a way that $R_t(u)$ is close (albeit not exactly equal) to u, despite biases in both components. For example, for a forecast that is too light-tailed ($O_t > 1$ and $H_{Z_t}(u) < u$ for all $u \in (0, 1)$), we cannot say whether $R_t(u)$ should be less than or greater than u. In fact, the plot of $u \mapsto R_t(u)$ will often cross the diagonal line, and the forecast may appear better tail calibrated compared to a forecast with the same $H_{Z_t}(u)$ but with $O_t = 1$. It is therefore difficult to conclude from the resulting plot that the forecast has a too-light tail. Analogous arguments hold for a forecast that is too heavy-tailed.

This also limits the applicability of this approach to weakly enforcing tail calibration. A measure of tail miscalibration could be derived by measuring the distance between $R_t(u)$ and u, which could be used as a regularization term to penalize tail miscalibration during model training, as proposed in Section 4.2. However, the canceling out of the biases in R_t can yield a loss function that is more difficult to optimize; this was therefore found to perform worse in practice than the approach based on $Q_t(u) = O_t H_{Z_t}^{-1}(u)$ outlined in Section 3.

The quantity Q_t partly mitigates the issues arising from the cancellation of biases in O_t and H_{Z_t}. For example, if the forecast is too light-tailed, then $O_t > 1$ and $H_{Z_t}^{-1}(u) > u$, and hence $Q_t(u) > u$. Similarly, if the forecast is too heavy-tailed, then $O_t < 1$, $H_{Z_t}^{-1}(u) < u$, and $Q_t(u) < u$. In this case, a larger bias in O_t or H_{Z_t} leads to a larger bias in Q_t. The resulting plot of $u \mapsto Q_t(u)$ is therefore much more interpretable than a plot of $u \mapsto R_t(u)$, since the biases in the individual components scale in the same direction. Of course, if $O_t > 1$ but $H_{Z_t}^{-1}(u) < u$, then $Q_t(u)$ may still be close to u despite the biases in the individual components (which would not be the case for $R_t(u)$), but this combination is less common in practice. The interpretation in this case is unclear anyway.

Similar arguments hold when O_t, H_{Z_t}, and Q_t are replaced with the finite-sample estimates $\hat{O}_t$, $\hat{H}_{z_t}$, and $\hat{Q}_t$ in Eq. (9), and R_t is replaced by

$$\hat{R}_t(u) = \frac{\sum_{i \in \mathcal{I}_t} \mathbb{1}\{z_{i,t} \leq u\}}{\sum_{i=1}^{n} [1 - F_i(t)]}, \quad u \in [0, 1].$$

To illustrate this, consider the following simple simulation experiment. We simulate 100,000 standard normal observations $Y \sim N(0, 1)$ and consider three forecasts:

- Forecast 1 – Overdispersed: $F_1 \sim N(0, 1.5)$
- Forecast 2 – Underdispersed: $F_2 \sim N(0, 0.85)$
- Forecast 3 – Heavy tailed: $F_3 \sim \text{Student}-t(3)$

We investigate tail calibration at the threshold $t = 1.0$. Fig. 12 displays the distribution function of the conditional PIT values, as well as $\hat{R}_t(u)$ and $\hat{Q}_t(u)$ as a function of u. One can see that for the forecasts F_1 and F_3, the distribution functions of CPIT values (left panel) lie above the diagonal line, indicating that the excess distributions are too heavy-tailed. At the same time, the corresponding distribution function for F_2 indicates a tail that is too light. However, this information is not available from the plot of $\hat{R}_t(u)$ (middle panel), where the curves for F_1 and F_2 lie on the opposite side of the diagonal line to before, and the curve for the heavy tailed forecast F_3 crosses the diagonal, appearing reasonably well tail calibrated despite both components ($\hat{H}_{z_t}$ and $\hat{O}_t$) indicating that the forecast is substantially miscalibrated in the tails. On the other hand, the diagnostic plot based on $\hat{Q}_t(u)$ (right panel) avoids this issue, with the errors in $\hat{H}_{z_t}$ amplified by the errors in $\hat{O}_t$.

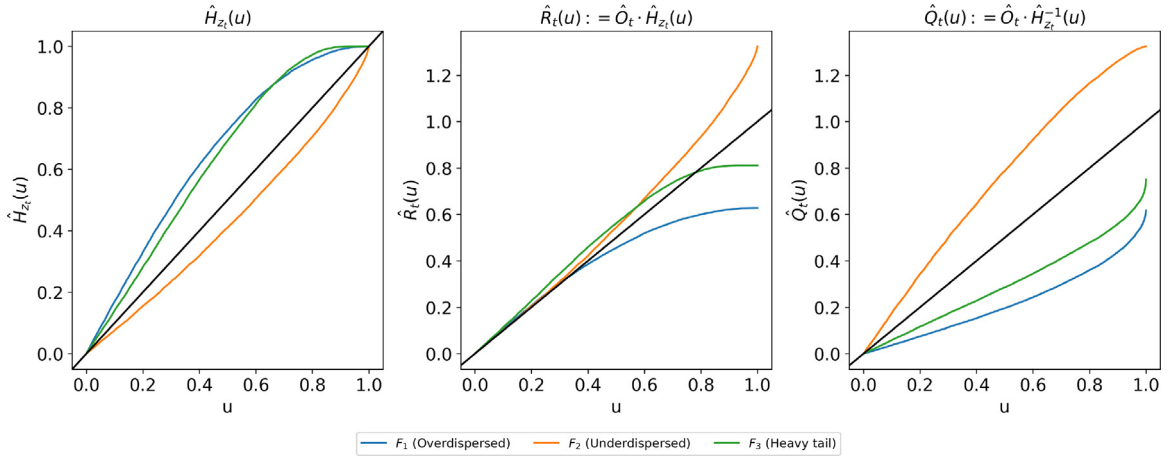

Fig. 12. Empirical distribution function $\hat{H}_{z_t}$ of conditional PIT values (left), $\hat{R}_t(u)$ against u (middle), and $\hat{Q}_t(u)$ against u (right), for the three miscalibrated forecasts F_1, F_2, and F_3, with $t = 1.0$. The estimates of O_t corresponding to the three forecasts are $\hat{O}_t = 0.63, \hat{O}_t = 1.34, \hat{O}_t = 0.82$ for F_1, F_2, F_3, respectively.

Appendix B. Model details

We provide some more details on the models used in Section 6. The code to implement all methods can be found at https://github.com/jakobwes/Enforcing-tail-calibration.

B.1. Ensemble model output statistics

Ensemble model output statistics (EMOS; Gneiting et al., 2005; Jewson et al., 2004) is a popular post-processing technique to recalibrate ensemble forecasts issued by a numerical weather prediction model. EMOS models the target variable, here wind speed W, as a parametric distribution whose location and scale parameters depend on the ensemble forecast mean and standard deviation at the corresponding time and location. Following Thorarinsdottir and Gneiting (2010), we use a truncated normal distribution $N_0(\mu, \sigma)$ for wind speed. The location parameter μ is modeled as a linear function of the ensemble mean m, while the (log) scale parameter $\log \sigma$ is modeled linearly in terms of the ensemble standard deviation s; for simplicity, we suppress the dependence of m and s on the time and location in the notation. We also include the day of year (doy) as a covariate to account for seasonal variations in forecast skill and biases:

$$W \mid m, s, doy \sim N_0(\mu, \sigma)$$
$$\mu = \alpha + \beta m + \lambda_{\mu,s} \sin\left(\tfrac{2\pi\ doy}{365.25}\right) + \lambda_{\mu,c} \cos\left(\tfrac{2\pi\ doy}{365.25}\right)$$
$$\log \sigma = \eta + \delta s + \lambda_{\sigma,s} \sin\left(\tfrac{2\pi\ doy}{365.25}\right) + \lambda_{\sigma,c} \cos\left(\tfrac{2\pi\ doy}{365.25}\right)$$

The parameters to estimate are then $\theta = (\alpha, \beta, \eta, \delta, \lambda_{\mu,s}, \lambda_{\mu,c}, \lambda_{\sigma,s}, \lambda_{\sigma,c})$. While the parameters could be estimated at each weather station individually, we use a semi-local post-processing approach (e.g. Lerch & Baran, 2016) to increase the amount of data available for model training, which is particularly relevant when forecasting extreme events, as discussed by Wessel et al. (2025). In this approach, individual EMOS models are fit to clusters of similar stations. This corresponds to location-wise fitting if one cluster per station is used, while smaller numbers of clusters allow data to be shared efficiently across stations.

We follow Wessel et al. (2025) and cluster based on the observed climatology. For this, we define K features as the $1/(K+1), \ldots, K/(K+1)$th quantiles of the observational distribution. Subsequently, k-means clustering is used to find clusters of similar stations. We use the "elbow" technique to select the number of clusters and identify four distinct clusters corresponding broadly to coastal, more-exposed coastal, inland, and more-exposed inland stations (see Fig. 3). EMOS models are subsequently fit to data from all stations in each cluster.

For parameter optimization, we use closed-form CRPS and twCRPS expressions for a truncated normal distribution (Jordan et al., 2019; Wessel et al., 2024). We follow the practice of the `crch` R-package (Messner et al., 2016). We use a Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton optimizer (Nocedal & Wright, 2006) with empirical gradient estimates, falling back on a Nelder–Mead optimizer (Nelder & Mead, 1965) should the BFGS optimizer fail. As optimization starting values, the results of a linear regression are used for the location parameters, whilst the scale intercept is initialized to the log-transformed standard error, and all other parameters are set to zero. As the optimizers rely on empirical gradients, the (tail) miscalibration penalties can easily be incorporated into the objective functions.

B.2. Distributional regression network

Distributional regression networks (Rasp & Lerch, 2018) are a nonlinear extension of EMOS. Wind speed is again assumed to follow a truncated normal distribution $N_0(\mu, \sigma)$, whose parameters are modeled not as linear functions, but in terms of feedforward neural networks. DRNs thus allow nonlinear dependencies between the covariate and the outcome to be captured, whilst maintaining the robustness that a parametric assumption offers.

The neural networks have two hidden layers with 16 units each and rectified linear unit (ReLU) activation functions. As the scale parameter σ is constrained to be positive, it is exponentially transformed after the final layer. We again use the ensemble mean, ensemble standard deviation, as well as the sine and cosine transformed normalized day of year ($\frac{2\pi \ doy}{365.25}$) as covariates. A single DRN is fitted to all stations; however, to make the post-processing spatially aware, the station index is used in a learned embedding layer to encode location information in two dimensions, which are used as covariates.

We use the closed-form expression of the CRPS for a truncated normal distribution for CRPS-based training. We use the Adam optimizer (Kingma & Ba, 2017), a stochastic gradient descent method, for parameter optimization, and use minibatching with a large batch size of 2048. We train for a fixed number of 200 epochs and find good convergence in all cases.

After CRPS-based training, we fine-tune the CRPS-pre-trained models using the (tail) calibration and twCRPS penalties with a threshold of $t = 12.5$ m/s and values of $\gamma = 1.0, 2.5, 5.0, 10.0, 20.0$. We do not use minibatching to compute the CRPS and calibration scores during finetuning, since the calibration penalties rely on the entire set of forecasts and observations from the training dataset to compute the empirical distribution of PIT or $\hat{Q}$ values. While this can be estimated within batches, these estimates can be unstable for small batch sizes and high thresholds, where only a small amount of data is available. Instead, we pass through the entire training dataset to compute the CRPS and calibration losses, as in the EMOS models, for 50 steps. Whilst full training with the full batch and calibration-penalized losses might also be considered, we opt to fine-tune CRPS-trained models, as this speeds up convergence. We again use an Adam optimizer for this fine-tuning step.

B.3. Conditional generative model

Conditional generative models (CGMs, Chen et al., 2024) provide a non-parametric alternative to DRN and EMOS models. They construct a predictive distribution by transforming samples from a latent distribution, here standard Normal, using neural networks. The predictive distribution is then available via samples. CGMs do not rely on parametric assumptions and can represent a wide range of predictive distributions.

We fit a single CGM to all stations, again using an embedding layer to encode spatial information in two dimensions. The CGM model architecture follows the structure proposed by Chen et al. (2024). As covariates, we again use the ensemble mean and standard deviation, sine- and cosine-transformed, normalized day of year, as well as location embeddings. The CGM model broadly consists of two parts: one that post-processes the ensemble mean and one that accounts for uncertainty by post-processing the latent noise and the ensemble standard deviation. These are added together and transformed with a softplus activation to ensure positivity of the predicted wind speed values. We use two hidden layers with 16 units each to adjust the ensemble mean, together with the external covariates. Similarly, the latent Gaussian noise is scaled up by the ensemble standard deviation and transformed, together with the exogenous covariates, again through two layers with 16 units.

Similar to the DRN case, we pre-train using the CRPS and finetune with the twCRPS and (tail) calibration penalties. Since the CGM outputs a predictive distribution via samples, we rely on the sample expression of the CRPS for a discrete predictive distribution in (20) for CRPS-based training. We use the Adam optimizer for parameter optimization and, as in the DRN case, minibatch training with a batch size of 2048. We train the models for a fixed 50 epochs and observe good convergence. At each step, we construct a predictive distribution from an ensemble of 250 samples and use it to compute the sample CRPS.

After CRPS-based training, we fine-tune using the (tail) calibration penalties and twCRPS with a threshold of $t = 12.5$ m/s and $\gamma = 1.0, 2.5, 5.0, 10.0, 20.0$. As for the CRPS, the PIT values, and the $\hat{Q}$ values for the miscalibration penalties, are computed using the empirical distribution functions based on the CGM sample. With a predictive distribution given through samples $\hat{y}_1, \dots, \hat{y}_n$ and an observation y, an estimate of the PIT value can be obtained through the normalized rank of y among $\hat{y}_1, \dots, \hat{y}_n$. However, ranking as a mathematical operation is not differentiable, posing a challenge for stochastic gradient descent methods used to train neural networks, which rely on gradients of the loss function with respect to the neural network parameters. Thus, we approximate the ranking via smooth approximations to step functions. Specifically, we approximate the PIT using

$$\begin{aligned} &\frac{1}{n}\left[\text{rank}(y, \{\hat{y}_1, \dots, \hat{y}_n\}) - 1\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{y}_i \leq y\} \approx \frac{1}{n}\sum_{i=1}^{n} \sigma\left(\frac{y - \hat{y}_i}{\nu}\right), \end{aligned} \tag{17}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and ν is a hyperparameter that ensures the quality of the approximation. For smaller values of ν, the approximation becomes more exact, whilst the objective remains smoother for larger values of ν. We choose $\nu = 1/100$.

As for the DRN model, batch-wise training is not possible when penalizing miscalibration, since the miscalibration terms depend on the forecasts and observations across the entire training dataset. However, fitting the CGM on the whole training dataset is computationally prohibitive, since it becomes too expensive to calculate the sample-based CRPS; this scales quadratically with the ensemble size M and is of complexity $\mathcal{O}(NM + NM^2)$ for N training samples, thus becoming infeasible for large batch sizes and ensemble sizes. To address this, we employ the fair CRPS for training (see (22) in Appendix C), as proposed in Lang et al. (2024). The fair CRPS is essentially an unbiased estimator of the population CRPS and is therefore less sensitive to sample size; it remains computationally tractable for an ensemble size of $M = 50$. We thus finetune with losses of the form $fCRPS + \gamma TMCB$, using the entire training dataset for each gradient estimate. We also use a fair twCRPS. Lang et al. (2024) find that the fair CRPS is degenerate when all but one ensemble members have

the same value, which can make training of the CGM unstable. However, this did not cause any problems here due to the higher floating-point precision used. Alternatively, one could use the almost-fair CRPS as introduced in Lang et al. (2024). The (tail) calibration penalties are computed as before using $M = 250$ samples.

Appendix C. Scoring rules for forecast samples

When F has a finite first moment, the CRPS defined in (2) can alternatively be written as

$$\mathrm{CRPS}(F, y) = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'| \tag{18}$$

where $X, X' \sim F$ are independent (Gneiting & Raftery, 2007). The twCRPS can similarly be expressed as

$$\mathrm{twCRPS}(F, y) = \mathbb{E}|v(X) - v(y)| - \frac{1}{2}\mathbb{E}|v(X) - v(X')|, \tag{19}$$

where v is an anti-derivative of the weight function w, that is, $v(x) - v(x') = \int_{x'}^{x} w(z)\,\mathrm{d}z$ for all $x, x' \in \mathbb{R}$ (Allen, Ginsbourger et al., 2023; Taillardat et al., 2023). When $w(x) = \mathbb{1}\{x \geq t\}$, for some threshold $t \in \mathbb{R}$, one choice for v is $v(x) = \max\{x, t\}$.

When the forecast F is a discrete predictive distribution, the CRPS and twCRPS can be obtained by replacing the expectations in the above expressions with sample means. For example, suppose F is the empirical distribution function corresponding to a sample of points $\hat{y}_1, \ldots, \hat{y}_M \in \mathbb{R}$, that is, $F(x) = \sum_{i=1}^{M} \mathbb{1}\{\hat{y}_i \leq x\}/M$. In this case, the CRPS becomes

$$\mathrm{CRPS}(F, y) = \frac{1}{M}\sum_{i=1}^{M}|\hat{y}_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|\hat{y}_i - \hat{y}_j|, \tag{20}$$

and the twCRPS becomes (Allen, Bhend et al., 2023)

$$\mathrm{twCRPS}(F, y) = \frac{1}{M}\sum_{i=1}^{M}|v(\hat{y}_i) - v(y)| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|v(\hat{y}_i) - v(\hat{y}_j)|. \tag{21}$$

These sample versions of the CRPS and twCRPS are used to evaluate the CGM forecasts in Section 6.

If the sample is treated as random draws from an underlying distribution rather than as a predictive distribution in its own right, then the sample CRPS can be adjusted to account for biases arising from sample size. Ferro et al. (2008) propose the fair CRPS (fCRPS) as

$$\mathrm{fCRPS}(F, y) = \frac{1}{M}\sum_{i=1}^{M}|\hat{y}_i - y| - \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M}|\hat{y}_i - \hat{y}_j|, \tag{22}$$

which essentially corresponds to an unbiased estimator of the population CRPS for the distribution underlying the samples (Ferro, 2014). A fair threshold-weighted CRPS can be obtained analogously. We use the fair CRPS to fine-tune the conditional generative models in Section 6, where larger training datasets require us to reduce the number of generated samples M from the model.

Appendix D. Regularizing with conditional PIT values

Tail miscalibration has traditionally also been assessed using histograms of conditional PIT (CPIT) values (Allen, Ginsbourger et al., 2023; Mitchell & Weale, 2023). These are values

$$z_{i,t} = \frac{F_i(y_i) - F_i(t)}{1 - F_i(t)}, \quad \text{for} \quad i \in \mathcal{I}_t, \tag{23}$$

where $\mathcal{I}_t = \{i \in \{1, \ldots, n\} : y_i > t\}$ is the set of indices for which the outcome exceeds the threshold. If the exceedance distribution is probabilistically calibrated, the empirical distribution function of conditional PIT values $\hat{H}_{z_t}$ should be (close to) uniform. Unlike the combined ratio in the $\hat{Q}$ values, however, the CPIT values do not account for (mis-)calibration of the occurrence of extreme events, but only for the probabilistic calibration of the intensity distribution (see Eq. (9)).

Similarly to the MCB and TMCB, one can also obtain a measure of miscalibration in the CPIT values, by penalizing the distance of the empirical distribution of conditional PIT values $\hat{H}_{z_t}$ to uniformity:

$$\text{CPIT-MCB} = \int_0^1 |\hat{H}_{z_t}(u) - u|\,\mathrm{d}u \tag{24}$$

This can be penalized during training, similar to the MCB and TMCB.

Figs. 13 and 14 show the scores for the DRN and CGM models trained with all tail calibration penalties, including the CPIT-MCB. This includes the results in Figs. 7 and 10, as well as additional results for CPIT-MCB regularization and evaluation. Results for the EMOS have been omitted for brevity, but are largely consistent. CPIT-MCB penalization improves the CPIT-MCB for both model types. This, however, comes at a high cost in terms of CRPS, MCB, and TMCB, where the scores deteriorate significantly. The CPIT-MCB penalization improves the calibration of the exceedance distribution. However, it does so by shifting a large amount of probability mass beyond the threshold and strongly overpredicting threshold exceedances. The TMCB and CRPS both improve the CPIT-MCB and also lead to improvements in TMCB.

This behavior indicates the need to assess both the calibration of the threshold-exceedance (occurrence) forecast and the calibration of the exceedance distribution (intensity) when evaluating the calibration of extreme-event forecasts. This can be understood as a calibration version of the Forecaster's dilemma discussed by Lerch et al. (2017), who argue for the need to jointly evaluate both the intensity and occurrence of extreme-event forecasts. The TMCB addresses this by adjusting the CPIT values with the occurrence ratio.

Data and code availability

Code to reproduce the results herein is available at https://github.com/jakobwes/Enforcing-tail-calibration. Unfortunately, the NWP forecast and observation data used in the case study in Section 6 are proprietary to the UK Met Office and cannot be shared publicly by the authors.

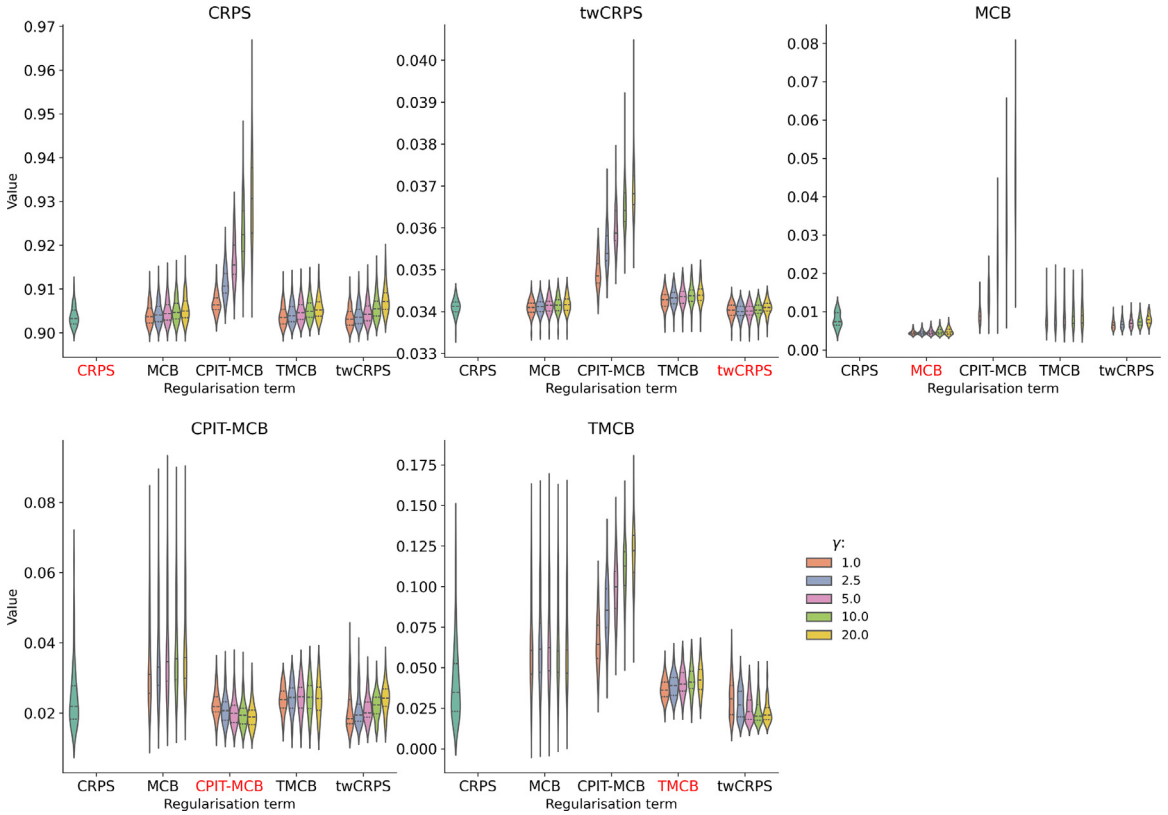

Fig. 13. Violin plot of metrics for the 100 DRN models trained using different regularization terms. Different panels correspond to different evaluation metrics, while the *x*-axis labels correspond to the different regularization terms. Red *x*-axis labels indicate when the evaluation metric matches the regularization metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

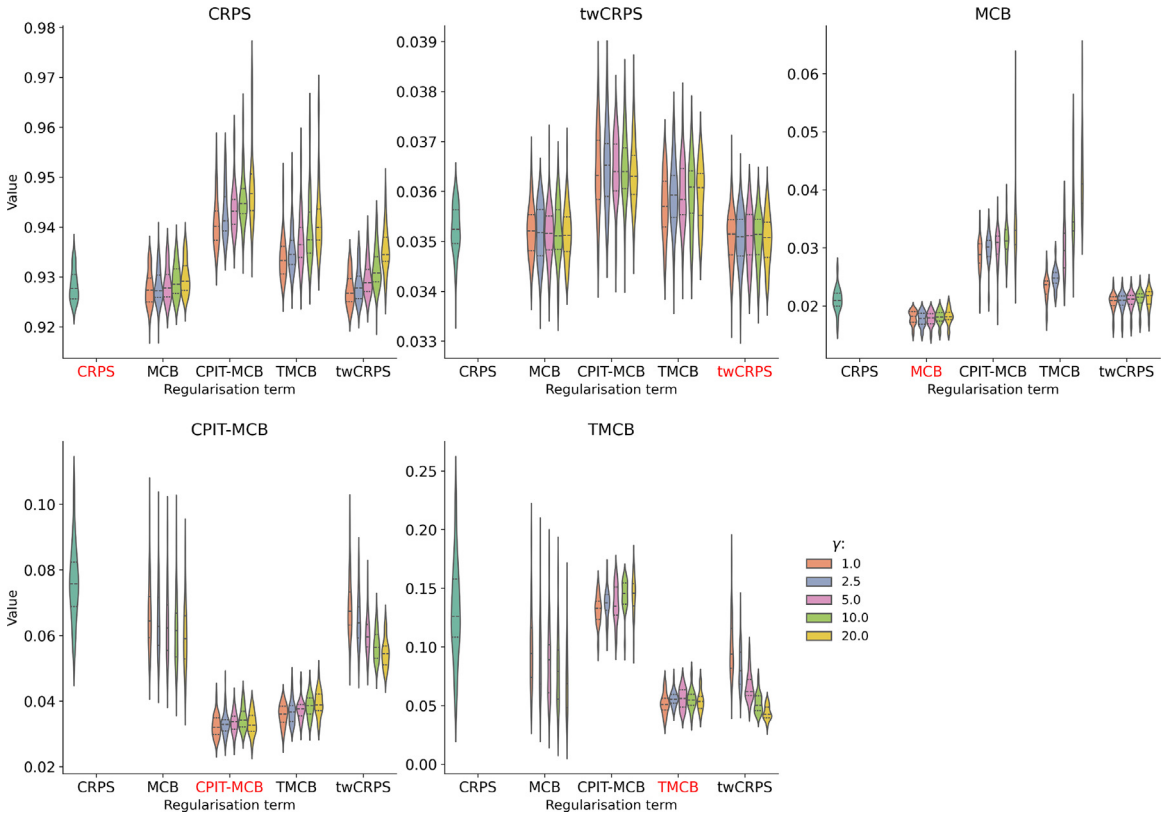

Fig. 14. Violin plot of metrics for the 100 CGM models trained using different regularization terms. Different panels correspond to different evaluation metrics, while the *x*-axis labels correspond to the different regularization terms. Red *x*-axis labels indicate when the evaluation metric matches the regularization metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Readers may request this data directly from the UK Met Office.

References

Allen, S., Bhend, J., Martius, O., & Ziegel, J. (2023). Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events. *Weather and Forecasting*, *38*, 499–516.

Allen, S., Gavrilopoulos, G., Henzi, A., Kleger, G.-R., & Ziegel, J. (2025). In-sample calibration yields conformal calibration guarantees. ArXiv preprint arXiv:2503.03841.

Allen, S., Ginsbourger, D., & Ziegel, J. (2023). Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, *11*, 906–940.

Allen, S., Koh, J., Segers, J., & Ziegel, J. (2025). Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, (ja), 1–20.

Angelopoulos, A. N., Barber, R. F., & Bates, S. (2024). *Theoretical foundations of conformal prediction*. Cambridge University Press (in press).

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, *135*, 1512–1519.

Buchweitz, E., Romano, J. V., & Tibshirani, R. J. (2025). Asymmetric penalties underlie proper loss functions in probabilistic forecasting. ArXiv preprint arXiv:2505.00937.

Chen, J., Janke, T., Steinke, F., & Lerch, S. (2024). Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, *18*, 159–183.

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

Dawid, A. P., Musio, M., & Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, *43*, 123–138.

de Punder, R., Diks, C. G., Laeven, R. J., & van Dijk, D. (2023). *Localizing strictly proper scoring rules*: *Technical report*, Tinbergen Institute Discussion Paper.

Désidéri, J.-A. (2012). Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, *350*, 313–318.

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

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

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1917–1923.

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

Fliege, J., & Svaiter, B. F. (2000). Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, *51*, 479–494.

Gebetsberger, M., Messner, J. W., Mayr, G. J., & Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, *146*(12), 4323–4338.

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

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

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

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

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

Gneiting, T., & Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, *17*, 3226–3286.

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

Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *83*, 963–993.

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

Jewson, S., Brix, A., & Ziehmann, C. (2004). A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts. *Atmospheric Science Letters*, *5*, 96–102.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, *90*, 1–37.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980.

Lang, S., Alexe, M., Clare, M. C. A., Roberts, C., Adewoyin, R., Boullègue, Z. B., Chantry, M., Dramsch, J., Dueben, P. D., Hahner, S., Maciel, P., Prieto-Nemesio, A., O'Brien, C., Pinault, F., Polster, J., Raoult, B., Tietsche, S., & Leutbecher, M. (2024). AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score. ArXiv preprint arXiv:2412.15832.

Lerch, S., & Baran, S. (2016). Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *66*, 29–51.

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

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

Mercier, Q., Poirion, F., & Désidéri, J.-A. (2018). A stochastic multiple gradient descent algorithm. *European Journal of Operational Research*, *271*, 808–817.

Messner, J. W., Mayr, G. J., & Zeileis, A. (2016). Heteroscedastic censored and truncated regression with crch.

Mitchell, J., & Weale, M. (2023). Censored density forecasts: Production and evaluation. *Journal of Applied Econometrics*, *38*, 714–734.

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

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*, 308–313.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization*. Springer New York.

Pacchiardi, L., Adewoyin, R. A., Dueben, P., & Dutta, R. (2024). Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, *25*, 1–64.

Porson, A. N., Carr, J. M., Hagelin, S., Darvell, R., North, R., Walters, D., Mylne, K. R., Mittermaier, M. P., Willington, S., & Macpherson, B. (2020). Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Quarterly Journal of the Royal Meteorological Society*, *146*, 3245–3265.

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

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

Schillinger, M., Samarin, M., Shen, X., Knutti, R., & Meinshausen, N. (2025). EnScale: Temporally-consistent multivariate generative downscaling via proper scoring rules. ArXiv preprint arXiv:2509.26258.

Taillardat, M., Fougères, A.-L., Naveau, P., & de Fondeville, R. (2023). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*, *39*, 1448–1459.

Thorarinsdottir, T. L., & Gneiting, T. (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*, 371–388.

Thorarinsdottir, T. L., & Schuhen, N. (2018). Chapter 6 - Verification: Assessment of calibration and accuracy. In *Statistical postprocessing of ensemble forecasts* (pp. 155–186). Elsevier.

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. Available at SSRN 1907799, http://dx.doi.org/10.2139/ssrn.1907799.

Vannitsem, S., Wilks, D. S., & Messner, J. (2018). *Statistical postprocessing of ensemble forecasts*. Elsevier.

Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic learning in a random world* (2nd ed.). Springer.

Waghmare, K., & Ziegel, J. (2025). Proper scoring rules for estimation and forecast evaluation. ArXiv preprint arXiv:2504.01781.

Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., ... Xavier, P. (2017). The met office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geoscientific Model Development*, *10*, 1487–1520.

Wessel, J. B., Ferro, C. A. T., Evans, G. R., & Kwasniok, F. (2025). Improving probabilistic forecasts of extreme wind speeds by training statistical post-processing models with weighted scoring rules. *Monthly Weather Review*, *153*, 1489–1511.

Wessel, J. B., Ferro, C. A. T., & Kwasniok, F. (2024). Lead-time-continuous statistical postprocessing of ensemble weather forecasts. *Quarterly Journal of the Royal Meteorological Society*, *150*, 2147–2167.

Wilks, D. S. (2018). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, *144*, 76–84.

Zerenner, T., Venema, V., Friederichs, P., & Simmer, C. (2021). Multi-objective downscaling of precipitation time series by genetic programming. *International Journal of Climatology*, *41*(14), 6162–6182.