Adv. Stat. Clim. Meteorol. Oceanogr., 11, 23–58, 2025
https://doi.org/10.5194/ascmo-11-23-2025

ASCMO
Open Access

Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation

Romain Pic[1], **Clément Dombry**[1], **Philippe Naveau**[2], **and Maxime Taillardat**[3]

[1]Université Marie et Louis Pasteur, CNRS, LmB (UMR 6623), 25000 Besançon, France
[2]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, EstimR, IPSL & U Paris-Saclay, Gif-sur-Yvette, France
[3]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Correspondence: Romain Pic (romain.pic@univ-fcomte.fr)

Received: 9 July 2024 – Revised: 23 December 2024 – Accepted: 14 January 2025 – Published: 13 March 2025

Abstract. Proper scoring rules are an essential tool to assess the predictive performance of probabilistic forecasts. However, propriety alone does not ensure an informative characterization of predictive performance, and it is recommended to compare forecasts using multiple scoring rules. With that in mind, interpretable scoring rules providing complementary information are necessary. We formalize a framework based on aggregation and transformation to build interpretable multivariate proper scoring rules. Aggregation-and-transformation-based scoring rules can target application-specific features of probabilistic forecasts, which improves the characterization of the predictive performance. This framework is illustrated through examples taken from the weather forecasting literature, and numerical experiments are used to showcase its benefits in a controlled setting. Additionally, the framework is tested on real-world data of postprocessed wind speed forecasts over central Europe. In particular, we show that it can help bridge the gap between proper scoring rules and spatial verification tools.

1 Introduction

Probabilistic forecasting allows for issuing forecasts carrying information about the prediction uncertainty. It has become an essential tool in numerous applied fields, such as weather and climate prediction (Vannitsem et al., 2021; Palmer, 2012), earthquake forecasting (Jordan et al., 2011; Schorlemmer et al., 2018), electricity price forecasting (Nowotarski and Weron, 2018), and renewable energies (Pinson, 2013; Gneiting et al., 2023). Moreover, it is slowly reaching fields further from historical applications of forecasting, such as epidemiology predictions (Bosse et al., 2023) or breast cancer recurrence prediction (Al Masry et al., 2023). In weather forecasting, probabilistic forecasts often take the form of ensemble forecasts in which the dispersion among members captures forecast uncertainty.

The development of probabilistic forecasts has induced the need for appropriate verification methods. Forecast verification fulfills two main purposes: quantifying how good a forecast is given available observations and allowing one to rank different forecasts according to their predictive performance. Scoring rules provide a single value to compare forecasts with observations. Propriety is a property of scoring rules that encourages forecasters to follow their true beliefs and that prevents hedging. Proper scoring rules allow for the assessment of calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996). Calibration is the statistical compatibility between forecasts and observations. Sharpness is the uncertainty of the forecast itself. Propriety is a necessary property of good scoring rules, but it does not guarantee that a scoring rule provides an informative characterization of predictive performance. In univariate and multivariate settings, numerous studies have proven that no scoring rule has it all, and thus, different scoring rules should be used to get a better understanding of the predictive performance of forecasts (see, e.g., Scheuerer and Hamill, 2015; Taillardat, 2021; Bjerregård et al., 2021). With that in mind, Scheuerer and Hamill (2015) "strongly recommend that several different scores be always considered before drawing conclusions". This amplifies the need for numerous complementary proper

scoring rules that are well understood to facilitate forecast verification. In that direction, Dorninger et al. (2018) states that "gaining an in-depth understanding of forecast performance depends on grasping the full meaning of the verification results". Interpretability of proper scoring rules can arise from being induced by a consistent scoring function for a functional (e.g., the squared error is induced by a scoring function consistent for the mean; Gneiting, 2011), knowing what aspects of the forecast the scoring rule is able to distinguish (e.g., the Dawid–Sebastiani score only discriminates forecasts based on their mean and variance; Dawid and Sebastiani, 1999) or knowing the limitations of a certain proper scoring rule (e.g., the variogram score is incapable of discriminating two forecasts that only differ by a constant bias; Scheuerer and Hamill, 2015). In this context, interpretable proper scoring rules become verification methods of choice as the ranking of forecasts they produce can be more informative than the ranking of a more complex but less interpretable scoring rule. Section 2 provides an in-depth explanation of this in the case of univariate scoring rules. It is worth noting that the interpretability of a scoring rule can also arise from its decomposition into meaningful terms (see, e.g., Bröcker, 2009). This type of interpretability can be used complementarily to the framework proposed in this article.

Scheuerer and Hamill (2015) proposed the variogram score to target the verification of the dependence structure. The variogram score of order p ($p > 0$) is defined as

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

where X_i and X_j are, respectively, the ith and jth components of the random vector $\boldsymbol{X} \in \mathbb{R}^d$ following F, w_{ij} is the set of non-negative weights, and $\boldsymbol{y} \in \mathbb{R}^d$ is an observation. The construction of the variogram score relies on two main principles. First, the variogram score is the weighted sum of scoring rules acting on the distribution of $\boldsymbol{X}_{i,j} = (X_i, X_j)$ and on paired components of the set of observations $\boldsymbol{y}_{i,j}$. This *aggregation* principle allows the combination of proper scoring rules and summarizes them into a proper scoring rule acting on the whole distribution F and observation $\boldsymbol{y}$. Second, the scoring rules composing the weighted sum can be seen as a standard proper scoring rule applied to transformations of both forecasts and observations. Let us denote $\gamma_{i,j} : \boldsymbol{x} \longmapsto |x_i - x_j|^p$ as the transformation related to the variogram of order p, and then the variogram score can be rewritten as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \mathrm{SE}(\gamma_{i,j}(F), \gamma_{i,j}(\boldsymbol{y})),$$

where $\mathrm{SE}(F, y) = (\mathbb{E}_F[X] - y)^2$ is the univariate squared error and $\gamma_{i,j}(F)$ is the distribution of $\gamma_{i,j}(\boldsymbol{X})$ for $\boldsymbol{X}$ following F. This second principle is the *transformation* principle, allowing us to build transformation-based proper scoring rules that can benefit from interpretability arising from a transformation (here, the variogram transformation $\gamma_{i,j}$) and the simplicity and interoperability of the proper scoring rule they rely on (here, the squared error).

We provide an overview of univariate and multivariate proper scoring rules through the lens of interpretability and by mentioning their known benefits and limitations. We formalize these two principles of aggregation and transformation to construct interpretable proper scoring rules for multivariate forecasts. To illustrate the use of these principles, we provide examples of transformation-and-aggregation-based scoring rules from the literature on probabilistic forecast verification and original propositions. The examples are backed with application-specific illustrations of their relevance. We conduct a simulation study to show in a controlled setting how transformation-and-aggregation-based scoring rules can be used. Moreover, the framework is confronted with real-world data in a case study of wind speed forecasts over Europe. Additionally, we show how the aggregation and transformation principles can help to bridge the gap between the proper scoring rule framework and the spatial verification tools (Gilleland et al., 2009; Dorninger et al., 2018).

The remainder of this article is organized as follows. Section 2 gives a general overview of verification methods for univariate and multivariate forecasts. Section 3 introduces the framework of proper scoring rules based on aggregation and transformation for multivariate forecasts. Section 4 provides examples of aggregation-and-transformation-based scoring rules. Then, Sect. 5 showcases through different simulation setups the interpretability of the aggregation-and-transformation-based framework. Section 6 confronts the proposed framework with real-world data. Finally, Sect. 7 provides a summary as well as a discussion on the verification of multivariate forecasts. Throughout the article, we focus on spatial forecasts for simplicity. However, the points made remain valid for any multivariate forecasts, including spatial forecasts, temporal forecasts, multivariable forecasts, or any combination of these categories (e.g., spatio-temporal forecasts of multiple variables).

The code associated with the numerical experiments of Sect. 5 and the case study of Sect. 6 is publicly available (https://github.com/pic-romain/aggregation-transformation, last access: 6 March 2025). The implementation is in `R` and relies mainly on the packages `scoringRules` (Jordan et al., 2019), `RandomFields` (Schlather et al., 2015), and `MultivCalibration` (Allen et al., 2024).

2 Overview of verification tools for univariate and multivariate forecasts

2.1 Calibration, sharpness, and propriety

Gneiting et al. (2007) proposed a paradigm for the evaluation of probabilistic forecasts: "maximizing the sharpness of the predictive distributions subject to calibration". *Calibra-*

tion is the statistical compatibility between the forecast and the observations. *Sharpness* is the concentration of the forecast and is a property of the forecast itself. In other words, the paradigm aims at minimizing the uncertainty of the forecast given that the forecast is statistically consistent with the observations. Tsyplakov (2011) states that the notion of calibration in the paradigm is too vague, but it holds if the definition of calibration is refined. This principle for the evaluation of probabilistic forecasts has reached a consensus in the field of probabilistic forecasting (see, e.g., Gneiting and Katzfuss, 2014; Thorarinsdottir and Schuhen, 2018). The paradigm proposed in Gneiting et al. (2007) is not the first mention of the link between sharpness and calibration: for example, Murphy and Winkler (1987) mentioned the relation between refinement (i.e., sharpness) and calibration.

For univariate forecasts, multiple definitions of calibration are available depending on the setting. The most used definition is *probabilistic calibration*, and, broadly speaking, it consists of computing the rank of observations among samples of the forecast and checking for uniformity with respect to observations. If the forecast is calibrated, observations should not be distinguishable from forecast samples, and thus, the distribution of their ranks should be uniform. Probabilistic calibration can be assessed by probability integral transform (PIT) histograms (Dawid, 1984) or rank histograms (Anderson, 1996; Talagrand et al., 1997) for ensemble forecasts when observations are stationary (i.e., their distribution is the same across time). PIT and rank histograms are popular diagnostic tools thanks to their interpretability. The shape of the PIT or rank histogram gives information about the type of (potential) miscalibration: a triangular-shaped histogram suggests that the probabilistic forecast has a systematic bias, a ∪-shaped histogram suggests that the probabilistic forecast is underdispersed, and a ∩-shaped histogram suggests that the probabilistic forecast is overdispersed. Moreover, probabilistic calibration implies that rank histograms should be uniform, but uniformity is not sufficient. For example, rank histograms should also be uniform conditionally on different forecast scenarios (e.g., conditionally on the value of the observations available when the forecast is issued). Additionally, under certain hypotheses, calibration tools have been developed to consider real-world limitations, such as serial dependence (Bröcker and Ben Bouallègue, 2020). Statistical tests have been developed to check the uniformity of rank histograms (Jolliffe and Primo, 2008). Readers interested in a more in-depth understanding of univariate forecast calibration are encouraged to consult Tsyplakov (2013, 2020).

For multivariate forecasts, a popular approach relies on a similar principle: first, multivariate forecast samples are transformed into univariate quantities using so-called pre-rank functions, and then the calibration is assessed by techniques used in the univariate case (see, e.g., Gneiting et al., 2008). Pre-rank functions may be interpretable and allow for targeting the calibration of specific aspects of the forecast, such as the dependence structure. Readers interested in the calibration of multivariate forecasts can refer to Allen et al. (2024) for a comprehensive review of multivariate calibration.

A scoring rule S assigns a real-valued quantity $S(F, \boldsymbol{y})$ to a forecast–observation pair $(F, \boldsymbol{y})$, where $F \in \mathcal{F}$ is a probabilistic forecast and $\boldsymbol{y} \in \mathbb{R}^d$ is an observation. In the negative-oriented convention, a scoring rule S is *proper relative to the class* $\mathcal{F}$ if

$$\mathbb{E}_G[S(G, \boldsymbol{Y})] \leq \mathbb{E}_G[S(F, \boldsymbol{Y})] \tag{1}$$

for all $F, G \in \mathcal{F}$, where $\mathbb{E}_G[\dots]$ is the expectation with respect to $\boldsymbol{Y} \sim G$. In simple terms, a scoring rule is proper relative to a class of distribution if its expected value is minimal when the true distribution is predicted for any distribution within the class. Forecasts minimizing the expected scoring rule are said to be *optimal*, and other forecasts are said to be *sub-optimal*. Moreover, the scoring rule S is *strictly proper relative to the class* $\mathcal{F}$ if the equality in Eq. (1) holds if and only if $F = G$. This ensures the characterization of the ideal forecast (i.e., there is a unique optimal forecast and it is the true distribution). Moreover, proper scoring rules are powerful tools as they allow for the assessment of calibration and sharpness simultaneously (Winkler, 1977; Winkler et al., 1996). Sharpness can be assessed individually using the entropy associated with proper scoring rules, defined by $e_S(F) = \mathbb{E}_F[S(F, \boldsymbol{Y})]$. The sharper the forecast, the smaller its entropy. Strictly proper scoring rules can also be used to infer the parameters of a parametric probabilistic forecast (see, e.g., Gneiting et al., 2005; Pacchiardi et al., 2024).

2.2 Univariate scoring rules

We recall a selection of univariate scoring rules as a means to explain key concepts involved in the multivariate scoring rules construction framework proposed in Sect. 3. For $d \geq 1$, let $\mathcal{P}(\mathbb{R}^d)$ denote the class of probabilities on $\mathbb{R}^d$ and let $\mathcal{P}_\alpha(\mathbb{R}^d)$ denote the class of probabilities with a finite moment of order α. In this section on univariate scoring rules, we consider the case $d = 1$ and $F \in \mathcal{P}(\mathbb{R})$ denotes a probabilistic forecast in the form of its cumulative distribution function (CDF) and $y \in \mathbb{R}$ denotes an observation.

The simplest scoring rules can be derived from scoring functions used to assess point forecasts. The squared error (SE) is the most popular one and is known through its averaged value (the mean squared error; MSE) or the square root of its average (the root mean squared error; RMSE) which has the advantage of being expressed in the same units as the observations. As a scoring rule, the SE is expressed as

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

where μ_F denotes the mean of the predicted distribution F. The SE solely discriminates the mean of the forecast (see Sect. B1); optimal forecasts for SE match the mean of the

true distribution. The SE is proper relative to $\mathcal{P}_2(\mathbb{R})$, the class of probabilities on $\mathbb{R}$ with a finite second moment (i.e., finite variance). Note that the SE cannot be strictly proper as the equality of mean does not imply the equality of distributions.

Another well-known scoring rule is the absolute error (AE) defined by

$$\mathrm{AE}(F, y) = |\mathrm{med}(F) - y|, \tag{3}$$

where $\mathrm{med}(F)$ is the median of the predicted distribution F. The mean absolute error (MAE), the average of the absolute error, is the most often seen form of the AE and it is also expressed in the same units as the observations. Optimal forecasts are forecasts that have a median equal to the median of the true distribution. The AE is proper relative to $\mathcal{P}_1(\mathbb{R})$ but not strictly proper. Similarly, the quantile score (QS), also known as the pinball loss, is a scoring rule focusing on quantiles of level α defined by

$$\mathrm{QS}_\alpha(F, y) = (1_{y \le F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y), \tag{4}$$

where $0 < \alpha < 1$ is a probability level and $F^{-1}(\alpha)$ is the predicted quantile of level α. The case $\alpha = 0.5$ corresponds to the AE up to a factor of 2. The QS of level α is proper relative to $\mathcal{P}_1(\mathbb{R})$ but not strictly proper since optimal forecasts are ones correctly predicting the quantile of level α (see, e.g., Friederichs and Hense, 2008).

Another summary statistic of interest is the exceedance of a threshold $t \in \mathbb{R}$. The Brier score (BS; Brier, 1950) was initially introduced for binary predictions but also allows for evaluating forecasts based on the exceedance of a threshold t. For probabilistic forecasts, the BS is defined as

$$\mathrm{BS}_t(F, y) = ((1 - F(t)) - 1_{y>t})^2 = (F(t) - 1_{y \le t})^2, \tag{5}$$

where $1 - F(t)$ is the predicted probability that the threshold t is exceeded. The BS is proper relative to $\mathcal{P}(\mathbb{R})$ but not strictly proper. Binary events (e.g., exceedance of thresholds) are relevant in weather forecasting as they are used, for example, in operational settings for decision-making.

All the scoring rules presented above are proper but not strictly proper since they only compare forecasts through specific summary statistics instead of the whole distribution. Nonetheless, they are still used as they allow forecasters to verify specific characteristics of the forecast: the mean, the median, the quantile of level α, or the exceedance of a threshold t. The simplicity and the specificity of these scoring rules make them interpretable, thus making them essential verification tools. They are used as diagnostic tools to check valuable characteristics of forecasts.

Some univariate scoring rules contain a summary statistic: for example, the formulas of the QS (Eq. 4) or the BS (Eq. 5) contain the exceedance of a threshold t and the quantile of level α, respectively. They can be seen as a scoring function applied to a summary statistic. This duality can be understood through the link between scoring functions and scoring rules through consistent functionals as presented in Gneiting (2011) or Sect. 2.2 in Lerch et al. (2017).

Other summary statistics can be of interest depending on applications. Nonetheless, it is worth noting that misspecifications of numerous summary statistics cannot be targeted because of their *non-elicitability*. Non-elicitability of a transformation implies that no proper scoring rule can be constructed such that optimal forecasts are forecasts where the transformation is equal to the one of the true distribution. For example, the variance is known to be non-elicitable; however, it is jointly elicitable with the mean (see, e.g., Brehmer, 2017). Readers interested in details regarding elicitable, non-elicitable, and jointly elicitable transformations may refer to Gneiting (2011), Brehmer and Strokorb (2019), and references therein.

A strictly proper scoring rule should compare the whole distribution and not only specific summary statistics. The continuous ranked probability score (CRPS; Matheson and Winkler, 1976) is the most popular univariate scoring rule in weather forecasting applications and can be expressed by the following:

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'| \tag{6}$$

$$= \int_{\mathbb{R}} \mathrm{BS}_z(F, y)\mathrm{d}z \tag{7}$$

$$= 2\int_0^1 \mathrm{QS}_\alpha(F, y)\mathrm{d}\alpha, \tag{8}$$

where $y \in \mathbb{R}$ and X and X' are independent random variables following F with a finite first moment. Equations (7) and (8) show that the CRPS is linked with the BS and the QS. Broadly speaking, as the QS discriminates a quantile associated with a specific level, integrating the QS across all levels discriminates the quantile function that fully characterizes univariate distributions. Similarly, integrating the BS across all thresholds discriminates the cumulative distribution function that also fully characterizes univariate distributions. The CRPS is a strictly proper scoring rule relative to $\mathcal{P}_1(\mathbb{R})$. In addition, Eq. (6) indicates that the CRPS values have the same units as observations. In the case of deterministic forecasts, the CRPS reduces to the absolute error in its scoring function form (Hersbach, 2000). The use of the CRPS for ensemble forecast is straightforward using expectations as in Eq. (6). Ferro et al. (2008), and Zamo and Naveau (2017) studied estimators of the CRPS for ensemble forecasts.

In addition to scoring rules based on scoring functions, some scoring rules use the moments of the probabilistic forecast F. The SE (Eq. 2) depends on the forecast only through its mean μ_F. The Dawid–Sebastiani score (DSS; Dawid and Sebastiani, 1999) is a scoring rule depending on the forecast F only through its first two central moments. The DSS is

expressed as

$$\mathrm{DSS}(F, y) = 2\log(\sigma_F) + \frac{(\mu_F - y)^2}{{\sigma_F}^2}, \tag{9}$$

where μ_F and ${\sigma_F}^2$ are the mean and the variance of the distribution F. The DSS is proper relative to $\mathcal{P}_2(\mathbb{R})$ but not strictly proper since optimal forecasts only need to correctly predict the first two central moments (see Sect. B1). Dawid and Sebastiani (1999) proposed a more general class of proper scoring rules but the DSS, as defined in Eq. (9), can be seen as a special case of the logarithmic score (up to an additive constant), introduced in Appendix A.

Another scoring rule relying on the central moments of the probabilistic forecast F up to order three is the error-spread score (ESS; Christensen et al., 2014). The ESS is defined as

$$\mathrm{ES}S(F, y) = ({\sigma_F}^2 - (\mu_F - y)^2 - (\mu_F - y)\sigma_F\gamma_F)^2, \tag{10}$$

where μ_F, σ_F^2, and γ_F are the mean, the variance, and the skewness of the probabilistic forecast F. The ESS is proper relative to $\mathcal{P}_4(\mathbb{R})$. As for the other scoring rules only based on moments of the forecast presented above, the expected ESS compares the probabilistic forecast F with the true distribution only via their four first moments (see Sect. B1). Scoring rules based on central moments of higher order could be built following the process described in Christensen et al. (2014). Such scoring rules benefit from the interpretability induced by their construction and the ease of application to ensemble forecasts. However, they would also inherit the limitation of being only proper.

Additional scoring rules relying on the existence of the probability density function (PDF) of the forecasts are presented in Appendix A. Readers may refer to the various reviews of scoring rules available (see, e.g., Bröcker and Smith, 2007; Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014; Thorarinsdottir and Schuhen, 2018; Alexander et al., 2022). Formulas of the expected scoring rules presented are available in Sect. B1.

Strictly proper scoring rules can be seen as more powerful than proper scoring rules. This is theoretically true when the interest is in identifying the ideal forecast (i.e., the true distribution). Regardless, in practice, scoring rules are also used to rank probabilistic forecasts and diagnostic tools, and with that in mind, a given ranking of forecasts in terms of the expectation of a strictly proper scoring rule (such as the CRPS) is harder to interpret than a ranking in terms of the expectation of a proper but more interpretable scoring rule (such as the SE). The SE is known to discriminate the mean, and thus, a better rank in terms of expected SE implies a better prediction of the mean of the true distribution. Conversely, a better ranking in terms of CRPS implies a better prediction of the whole prediction, but it might not be useful as is, and other verification tools are needed to know what caused this ranking. When forecasts are not calibrated, there seems to be a trade-off between interpretability and strict propriety. This becomes more prominent in a multivariate setting as forecasts are more complex to characterize. However, simpler interpretable scoring rules and strictly proper scoring rules can be used complementarily. The framework proposed in Sect. 3 aims at helping the construction of interpretable proper scoring rules.

2.3 Multivariate scoring rules

In a multivariate setting, forecasters cannot solely use univariate scoring rules as they are not able to distinguish forecasts beyond their 1-dimensional marginals. Univariate scoring rules cannot discriminate the dependence structure between the univariate margins. In the following, we consider $F \in \mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$ a multivariate probabilistic forecast and $\boldsymbol{y} \in \mathbb{R}^d$ an observation.

Even if there is no natural ordering in the multivariate case, the notions of median and quantile can be adapted using level sets, and then scoring rules using these quantities can be constructed (see, e.g., Meng et al., 2023). Nonetheless, as the mean is well defined, the squared error (SE) can be defined in the multivariate setting:

$$\mathrm{SE}(F, \boldsymbol{y}) = \|\boldsymbol{\mu}_F - \boldsymbol{y}\|_2^2, \tag{11}$$

where $\boldsymbol{\mu}_F$ is the mean vector of the distribution F. Similar to the univariate case, the SE is proper relative to $\mathcal{P}_2(\mathbb{R}^d)$. Moments are well defined in the multivariate case allowing the multivariate version of the Dawid–Sebastiani score to be defined. The Dawid–Sebastiani score (DSS) was proposed in Dawid and Sebastiani (1999) as

$$\mathrm{DSS}(F, \boldsymbol{y}) = \log(\det\boldsymbol{\Sigma}_F) + (\boldsymbol{\mu}_F - \boldsymbol{y})^T\boldsymbol{\Sigma}_F^{-1}(\boldsymbol{\mu}_F - \boldsymbol{y}),$$

where $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ are the mean vector and the covariance matrix of the distribution F. The DSS is proper relative to $\mathcal{P}_2(\mathbb{R}^d)$. The second term in the DSS is the squared Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}_F$.

To define a strictly proper scoring rule for multivariate forecast, Gneiting and Raftery (2007) proposed the energy score (ES) as a generalization of the CRPS to the multivariate case. The ES is defined by

$$\mathrm{ES}_\alpha(F, \boldsymbol{y}) = \mathbb{E}_F\|\boldsymbol{X} - \boldsymbol{y}\|_2^\alpha - \frac{1}{2}\mathbb{E}_F\|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha, \tag{12}$$

where $\alpha \in (0, 2)$ and $F \in \mathcal{P}_\alpha(\mathbb{R}^d)$, the class of probabilities on $\mathbb{R}^d$ such that the moment of order α is finite. The definition of the ES is related to the kernel form of the CRPS (Eq. 6), to which the ES reduces for $d = 1$ and $\alpha = 1$. As pointed out in Gneiting and Raftery (2007), in the limiting case $\alpha = 2$, the ES becomes the SE (Eq. 11). The ES is strictly proper relative to $\mathcal{P}_\alpha(\mathbb{R}^d)$ (Székely, 2003; Gneiting and Raftery, 2007) and is suited for ensemble forecasts (Gneiting et al., 2008). Moreover, the parameter α gives some flexibility: a small value of α can be chosen and still lead to a strictly proper scoring rule, for example,

when higher-order moments are ill-defined. The discrimination ability of the ES has been studied in numerous studies (see, e.g., Pinson and Girard, 2012; Pinson and Tastu, 2013; Scheuerer and Hamill, 2015). Pinson and Girard (2012) studied the ability of the ES to discriminate among rival sets of scenarios (i.e., forecasts) of wind power generation. In the case of bivariate Gaussian processes, Pinson and Tastu (2013) illustrated that the ES appears to be more sensitive to misspecifications of the mean rather than misspecifications of the variance or dependence structure. The lack of sensitivity to misspecifications of the dependence structure has been confirmed in Scheuerer and Hamill (2015) using multivariate Gaussian random vectors of higher dimension. Moreover, the discriminatory power of the ES deteriorates in higher dimensions (Pinson and Tastu, 2013).

To overcome the discriminatory limitation of the ES, Scheuerer and Hamill (2015) proposed the variogram score (VS), a score targeting the verification of the dependence structure. The VS of order p is defined as

$$\mathrm{VS}_p(F, \boldsymbol{y}) = \sum_{i,j=1}^{d} w_{ij} \left(\mathbb{E}_F \left[|X_i - X_j|^p \right] - |y_i - y_j|^p \right)^2, \quad (13)$$

where X_i and X_j are, respectively, the ith and jth components of the random vector $\boldsymbol{X}$ following F, w_{ij} are nonnegative weights and $p > 0$. The variogram score capitalizes on the variogram used in spatial statistics to access the dependence structure. The VS cannot detect an equal bias across all components. The VS of order p is proper relative to the class of probabilities on $\mathbb{R}^d$ such that the $2p$th moments of all univariate margins are finite. The weight values w_{ij} can be selected to emphasize or depreciate certain pair interactions. For example, in a spatial context, it can be expected the dependence between pairs decays with the distance: choosing the weights proportional to the inverse of the distance between locations can increase the signal-to-noise ratio and improve the discriminatory power of the VS (Scheuerer and Hamill, 2015).

Multivariate counterparts of univariate scoring rules relying on the existence of forecast PDFs are presented and discussed in Appendix A. Additionally, other multivariate scoring rules have been proposed among which the marginal-copula score (Ziel and Berk, 2019) or wavelet-based scoring rules (see, e.g., Buschow et al., 2019), which are briefly mentioned in Sect. 4 in light of the aggregation-and-transformation-based framework. However, fewer multivariate scoring rules have been proposed compared to the univariate setting. These scoring rules are briefly mentioned in Sect. 4 in light of the proper scoring rule construction framework proposed in this article. Section B2 provides formulas for the expected multivariate scoring rules presented above.

2.4 Spatial verification tools

Spatial forecasts are a very important group of multivariate forecasts as they are involved in various applications (e.g., weather or renewable energy forecasting). Spatial fields are often characterized by high dimensionality and potentially strong correlations between neighboring locations. These characteristics make the verification of spatial forecasts very demanding in terms of discriminating misspecified dependence structures, for example. In the case of spatial forecasts, it is known that traditional verification methods (e.g., grid point-by-grid point verification) may result in a double penalty. The *double-penalty effect* was pinned in Ebert (2008) and refers to the fact that if a forecast presents a spatial (or temporal) shift with respect to observations, the error made would be penalized twice: once where the event was observed and again where the forecast predicted it. In particular, high-resolution forecasts are more penalized than less realistic blurry forecasts. The double-penalty effect may also affect spatio-temporal forecasts in general.

In parallel with the development of scoring rules, various application-focused spatial verification methods have been developed to evaluate weather forecasts. The efforts toward improving spatial verification methods have been guided by two projects: the intercomparison project (ICP; Gilleland et al., 2009) and its second phase, called Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT; Dorninger et al., 2018). These projects resulted in the comparison of spatial verification methods with a particular focus on understanding their limitations and clarifying their interpretability. Only a few links exist between the approaches studied in these projects (and the work they induced) and the proper scoring rule framework. In particular, Casati et al. (2022) noted "a lack of representation of novel spatial verification methods for ensemble prediction systems". In general, there is a clear lack of methods focusing on the spatial verification of probabilistic forecasts. Moreover, to help bridge the gap between the two communities, we would like to recall the approach of spatial verification tools in the light of the scoring rule framework introduced above.

One of the goals of the ICP was to provide insights into how to develop methods robust to the double-penalty effect. In particular, Gilleland et al. (2009) proposed a classification of spatial verification tools updated later in Dorninger et al. (2018), resulting in a five-category classification. The classes differ in the computing principle they rely on. Not all spatial verification tools mentioned in these studies can be applied to probabilistic forecasts, some of them can solely be applied to deterministic forecasts. In the following description of the classes, we try to focus on methods suited to probabilistic forecasts or at least the special case of ensemble forecasts.

Neighborhood-based methods consist of applying a smoothing filter to the forecast and observation fields to prevent the double-penalty effect. The smoothing filter can take various forms (e.g., a minimum, a maximum, a mean,

or a Gaussian filter) and be applied over a given neighborhood. For example, Stein and Stoop (2022) proposed a neighborhood-based CRPS for ensemble forecasts gathering forecasts and observations made within the neighborhood of the location considered. The use of a neighborhood prevents the double-penalty effect from taking place at scales smaller than that of the neighborhood. In this general definition, neighborhood-based methods can lead to proper scoring rules; in particular, see the notion of patches in Sect. 4.

Scale-separation techniques denote methods for which the verification is obtained after comparing forecast and observation fields across different scales. The scale-separation process can be seen as several single-bandpass spatial filters (e.g., projection onto a base of wavelets as wavelet-based scoring rules; Buschow et al., 2019). However, to obtain proper scoring rules, the comparison of the scale-specific characteristics needs to be performed using a proper scoring rule. Section 4 provides a discussion on wavelet-based scoring rules and their propriety.

Object-based methods rely on the identification of objects of interest and the comparison of the objects obtained in the forecast and observation fields. Object identification is application-dependent and can take the form of objects that forecasters are familiar with (e.g., storm cells for precipitation forecasts). A well-known verification tool within this class is the structure–amplitude–location (SAL; Wernli et al., 2008) method which has been generalized to ensemble forecasts in Radanovics et al. (2018). The three components of the ensemble SAL do not lead to proper scoring rules. They rely on the mean of the forecast within scoring functions inconsistent with the mean. Thus, the ideal forecast does not minimize the expected value. Nonetheless, the three components of the SAL method could be adapted to use proper scoring rules sensitive to the misspecification of the same features.

Field-deformation techniques consist of deforming the forecasts field into the observation field (the similarity between the fields can be ensured by a metric of interest). The field of distortion associated with the morphing of the forecast field into the observation field becomes a measure of the predictive performance of the forecast (see, e.g., Han and Szunyogh, 2018).

Distance measures between binary images, such as exceedance of a threshold of interest, of the forecast and observation fields. These methods are inspired by development in image processing (e.g., Baddeley's delta measure, Gilleland, 2011).

These five categories partially overlap as it can be argued that some methods belong to multiple categories (e.g., some distance measures techniques can be seen as a mix of field deformation and object-based). They define different principles that can be used to build verification tools that are not subject to the double-penalty effect. The reader may refer to Dorninger et al. (2018) and references therein for details on the classification and the spatial verification methods not used thereafter. The frontier between the aforementioned spatial verification methods and the proper scoring rule framework is porous with, for example, wavelet-based scoring rules belonging to both. It appears that numerous spatial verification methods seek interpretability, and we believe that this is not incompatible with the use of proper scoring rules. We propose the following framework to facilitate the construction of interpretable proper scoring rules.

3 A framework for interpretable proper scoring rules

We define a framework to design proper scoring rules for multivariate forecasts. Its definition is motivated by remarks on the multivariate forecast literature and operational use. There seems to be a growing consensus around the fact that no single verification method has it all (see, e.g., Bjerregård et al., 2021). Most of the studies comparing forecast verification methods highlight that verification procedures should not be reduced to the use of a single method and that each procedure needs to be well suited to the context (see, e.g., Scheuerer and Hamill, 2015; Thorarinsdottir and Schuhen, 2018). Moreover, from a more theoretical point of view, (strict) propriety does not ensure discrimination ability, and different (strictly) proper scoring rules can lead to different rankings of sub-optimal forecasts. Proper scoring rules may have multiple optimal forecasts, and, in a general setting, no guarantee is given on their relevance. Moreover, strict propriety ensures that the optimal forecast is unique and that it is the ideal forecast (i.e., the true distribution); however, no guarantee is available for forecasts within the vicinity of the minimum in the general case. This is particularly problematic since, in practice, the unavailability of the ideal distribution makes it impossible to know if the minimum expected score is achieved. In the case of calibrated forecasts, the expected scoring rule is the entropy of the forecast, and the ranking of forecasts is thus linked to the information carried by the forecast (see Corollary 4, Holzmann and Eulert, 2014, for the complete result).

Standard verification procedures gradually increase the complexity of the quantities verified. Procedures often start by verifying simple quantities such as quantiles, mean, or binary events (e.g., prediction of dry/wet events for precipitation). If multiple forecasts have a satisfying performance for these quantities, marginal distributions of the multivariate forecast can be verified using univariate scoring rules. Finally, multivariate-related quantities, such as the dependence structure, can be verified through multivariate scoring rules. Forecasters rely on multiple verification methods to evaluate a forecast, and ideally, the verification method should be interpretable by targeting specific aspects of the distribution or thanks to the forecaster's experience. This type of verification, or diagnostic, procedure allows the forecaster to understand what characterizes the predictive performance of a forecast instead of directly looking at a strictly proper scor-

ing rule giving an encapsulated summary of the predictive performance.

As mentioned in Sect. 2.1, various multivariate forecast calibration methods rely on the calibration of univariate quantities obtained by dimension reduction techniques. As the general principle of multivariate calibration leans on studying the calibration of quantities obtained by pre-rank functions, Allen et al. (2024) argue that calibration procedures should not rely on a single pre-rank function and should instead use multiple simple pre-rank functions and leverage the interpretability of the associated PIT/rank histograms. A similar principle can be applied to increase the interpretability of verification methods based on scoring rules.

As general multivariate strictly proper scoring rules fail to distinguish forecasts for arbitrary misspecifications and they may lead to different ranking of sub-optimal forecasts, multivariate verification could benefit from using multiple proper scoring rules targeting specific aspects of the forecasts. Thereby, forecasters know which aspect of the observations are well predicted by the forecast and can update their forecast or select the best forecast among others in the light of this better understanding of the forecast. To facilitate the construction of interpretable proper scoring rules, we define a framework based on two principles: transformation and aggregation.

The transformation principle consists of transforming both the forecast and observation before applying a scoring rule. Heinrich-Mertsching et al. (2024) introduced this general principle in the context of point processes. In particular, they present scoring rules based on summary statistics targeting the clustering behavior or the intensity of the processes. In a more general context, the use of transformations was disseminated in the literature for several years (see Sect. 4). Proposition 1 shows how transformations can be used to construct proper scoring rules.

Proposition 1. *Let $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$, and let $F \in \mathcal{F}$ be a forecast and $\boldsymbol{y} \in \mathbb{R}^d$ an observation. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be a transformation, and let S be a scoring rule on $\mathbb{R}^k$ that is proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), \boldsymbol{X} \sim F \in \mathcal{F}\}$. Then, the scoring rule*

$$S_T(F, \boldsymbol{y}) = S(T(F), T(\boldsymbol{y}))$$

is proper relative to $\mathcal{F}$. If S is strictly proper relative to $T(\mathcal{F})$ and T is injective, then the resulting scoring rule S_T is strictly proper relative to $\mathcal{F}$.

To gain interpretability, it is natural to have dimension-reducing transformations (i.e., $k < d$), which generally leads to T not being injective and S_T not being strictly proper. Nonetheless, as expressed previously, interpretability is important, and it can mostly be leveraged if the transformation simplifies the multivariate quantities. Particularly, it is generally preferred to choose $k = 1$ to make the quantity easier to interpret and focus on specific information contained in the forecast or the observation. Straightforward transformations can be projections on a k-dimensional margin or a summary statistic relevant to the application, such as the total over a catchment area in the case of precipitation. Simple transformations may be preferred for their interpretability, and their potential lack of general discrimination ability can be made up for by multiple simpler transformations. Numerous examples of transformations are presented, discussed, and linked to the literature and applications in Sect. 4. The proof of Proposition 1 is provided in Sect. E1.

The second principle is the aggregation of scoring rules. Aggregation can be used on scoring rules to combine them and obtain a single scoring rule summarizing the evaluation. Note that Dawid and Musio (2014) introduced the notion of *composite score*, which is related to the aggregation principle but is closer to the combined application of both principles. Proposition 2 presents a general aggregation principle to build proper scoring rules. This principle has been known since proper scoring rules were introduced.

Proposition 2. *Let $\mathcal{S} = \{S_i\}_{1 \le i \le m}$ be a set of proper scoring rules relative to $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$. Let $\boldsymbol{w} = \{w_i\}_{1 \le i \le m}$ be non-negative weights. Then, the scoring rule*

$$S_{\mathcal{S},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i S_i(F, \boldsymbol{y})$$

is proper relative to $\mathcal{F}$. If at least one scoring rule S_i is strictly proper relative to $\mathcal{F}$ and $w_i > 0$, the aggregated scoring rule $S_{\mathcal{S},\boldsymbol{w}}$ is strictly proper relative to $\mathcal{F}$.

It is worth noting that Proposition 2 does not specify any strict condition for the scoring rules used. For example, the scoring rules aggregated do not need to be the same, do not need to be expressed in the same units, or even act on the same objects. Aggregated scoring rules can be used to summarize the evaluation of univariate probabilistic forecasts (e.g., aggregation of CRPS at different locations) or to summarize complementary scoring rules (e.g., aggregation of the Brier score and a threshold-weighted CRPS). Unless stated otherwise, for simplicity, we restrict ourselves to cases where the aggregated scoring rules are of the same type.

Bolin and Wallin (2023) showed that the aggregation of scoring rules can lead to unintuitive behaviors. For the aggregation of univariate scoring rules, they showed that scoring rules do not necessarily have the same dependence on the scale of the forecasted phenomenon: this leads to scoring rules putting more (or less) emphasis on the forecasts with larger scales. They define and propose local scale-invariant scoring rules to make scale-agnostic scoring rules. When performing aggregation, it is important to be aware of potential preferences or biases of the scoring rules.

We only consider aggregation of proper scoring rules through a weighted sum. To conserve (strict) propriety of scoring rules, aggregations can take, more generally, the form

of (strictly) isotonic transformations, such as a multiplicative structure when positive scoring rules are considered (Ziel and Berk, 2019).

The two principles of Proposition 1 and Proposition 2 can be used simultaneously to create proper scoring rules based on both aggregation and transformation as presented in Corollary 1.

Corollary 1. *Let $\mathcal{T} = \{T_i\}_{1 \le i \le m}$ be a set of transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$. Let $\mathcal{S}_\mathcal{T} = \{S_{T_i}\}_{1 \le i \le m}$ be a set of proper scoring rules where S is proper relative to $T_i(\mathcal{F})$ for all $1 \le i \le m$. Let $\boldsymbol{w} = \{w_i\}_{1 \le i \le m}$ be non-negative weights. Then, the scoring rule*

$$S_{\mathcal{S}_\mathcal{T}, \boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{m} w_i S_{T_i}(F, \boldsymbol{y})$$

is proper relative to $\mathcal{F}$.

Strict propriety relative to $\mathcal{F}$ of the resulting scoring rule is obtained as soon as there exists $1 \le i \le m$ such that S is strictly proper relative to $T_i(\mathcal{F})$, T_i is injective, and $w_i > 0$. The result of Corollary 1 can be extended to transformations with images in different dimensions and paired with different scoring rules (see Appendix C).

Any kernel score (which encapsulates the BS, the CRPS, the ES, and the VS) can be expressed as an aggregation of squared errors between transformations of the forecast–observation pair; see Appendix D. As we see in the examples developed in the following section, numerous scoring rules used in the literature are based on these two principles of aggregation and transformation.

4 Applications of the aggregation and transformation principles

4.1 Projections

Certainly, the most direct type of transformation is projections of forecasts and observations on their k-dimensional marginals. We denote T_i as the projection on the ith component such that $T_i(\boldsymbol{X}) = X_i$ for all $\boldsymbol{X} \in \mathbb{R}^d$. This allows the forecaster to assess the predictive performance of a forecast for a specific univariate marginal independently of the other variables. If S is a univariate scoring rule proper relative to $\mathcal{P}(\mathbb{R})$, then Proposition 1 leads to S_{T_i} being proper relative to $\mathcal{P}(\mathbb{R}^d)$. The resulting scoring rule S_{T_i} can be useful if a given marginal is of particular interest (e.g., location of high interest in a spatial forecast). However, it can be more interesting to aggregate such scoring rules across all 1-dimensional marginals. This leads to the following scoring rule:

$$S_{\mathcal{S}_\mathcal{T}, \boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^{d} w_i S_{T_i}(F, \boldsymbol{y}),$$

where $\mathcal{S}_\mathcal{T}$ is $\{S_{T_i}\}_{1 \le i \le d}$. This setting is popular for assessing the performance of multivariate forecasts, and we briefly present examples from the literature falling under this setting. Aggregation of CRPS (Eq. 6) across locations and/or lead times is common practice for plots or comparison tables with uniform weights (Gneiting et al., 2005; Taillardat et al., 2016; Rasp and Lerch, 2018; Schulz and Lerch, 2022; Lerch and Polsterer, 2022; Hu et al., 2023) or with more complex schemes such as weights proportional to the cosine of the latitude (Ben Bouallègue et al., 2024b). The SE (Eq. 2) and AE (Eq. 3) can be aggregated to obtain RMSE and MAE, respectively (Delle Monache et al., 2013; Gneiting et al., 2005; Lerch and Polsterer, 2022; Pathak et al., 2022). Bremnes (2019) aggregated QSs (Eq. 4) across stations and different quantile levels of interest with uniform weights. Note that the multivariate SE (Eq. 11) can be rewritten as the sum of univariate SE across 1-marginals: $\mathrm{SE}(F, \boldsymbol{y}) = \|\boldsymbol{\mu}_F - \boldsymbol{y}\|_2^2 = \sum_{i=1}^{d} \mathrm{SE}_{T_i}(F, \boldsymbol{y})$.

The second simplest choice is the 2-dimensional case, allowing for a focus on pair dependency. We denote $T_{(i,j)}$ as the projection on the ith and jth components (i.e., the (i, j) pair of components) such that $T_{(i,j)}(\boldsymbol{X}) = \boldsymbol{X}_{i,j} = (X_i, X_j)$. In this setting, S has to be a bivariate proper scoring rule to construct a proper scoring rule $S_{T_{(i,j)}}$. The aggregation of such scoring rules becomes

$$S_{\mathcal{S}_\mathcal{T}, \mathbf{w}}(F, \boldsymbol{y}) = \sum_{\substack{i,j=1 \\ i \ne j}}^{d} w_{i,j} S_{T_{(i,j)}}(F, \boldsymbol{y}).$$

As suggested in Scheuerer and Hamill (2015) for the VS (Eq. 13), the weight values $w_{i,j}$ can be chosen appropriately to optimize the signal-to-noise ratio. For example, in a spatial setting where the dependence between locations is believed to decrease with the distance separating them, the weight values $w_{i,j}$ can be chosen to be proportional to the inverse of the distance. This bivariate setting is less used in the literature; we present two articles using or mentioning scoring rules within this scope. In a general multivariate setting, Ziel and Berk (2019) suggest the use of a marginal-copula scoring rule where the copula score is the bivariate copula energy score (i.e., the aggregation of the energy scores across all the regularized pairs). To focus on the verification of the temporal dependence of spatio-temporal forecasts, Ben Bouallègue et al. (2024b) use the bivariate energy score over consecutive lead times.

In a more general setup, we consider projection on k-dimensional marginals. In order to reduce the number of transformation-based scores to aggregate, it is standard to focus on localized marginals (e.g., belonging to patches of a given spatial size). Denote $\mathcal{P} = \{P_i\}_{1 \le i \le m}$ as a set of valid patches (for some criterion or of a given size) and $\mathcal{S}_\mathcal{P}$ as the set of transformation-based scores associated with the projections on the patches $\mathcal{P}$. Given a multivariate scoring rule S proper relative to $\mathcal{P}(\mathbb{R}^k)$, we can construct the following

aggregated score:

$$S_{\mathcal{S}_\mathcal{P},\boldsymbol{w}}(F,\boldsymbol{y}) = \sum_{P\in\mathcal{P}} w_P S_P(F,\boldsymbol{y}).$$

This construction can be used to create a scoring rule only considering the dependence of localized components given that the patches are defined in that sense. The use of patches has similar benefits as the weighting of pairs given a belief on their correlations: obtain a better signal-to-noise ratio and improve the discrimination of the resulting scoring rule. For example, Pacchiardi et al. (2024) introduced patched energy scores as scoring rules to minimize in order to train a generative neural network. The patched energy scores are defined for $S = \mathrm{ES}$ and square patches spaced by a given stride. In a general setting, the patched ES, resulting from the aggregation of the ES (with $\alpha = 1$) over the set of patches $\mathcal{P}$, is defined as

$$\mathrm{ES}_{\mathcal{P},\boldsymbol{w}_\mathcal{P}}(F,\boldsymbol{y}) = \sum_{P\in\mathcal{P}} w_P \mathrm{ES}_1(F_P,\boldsymbol{y}_P), \tag{14}$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P \in \mathcal{P}$, and F_P is the marginal of F over the patch P. To make the scoring more interpretable, it is preferable to consider patches with a fixed size and uniform weights ($w_P = 1/|\mathcal{P}|$). The patched ES reduces to the aggregated CRPS and the ES when the patches span a single location and all the locations, respectively.

Patch-based scoring rules appear as a natural member of the neighborhood-based methods of the spatial verification classification mentioned in Sect. 2.4. Given that the patches are correctly chosen (e.g., of a size appropriate to the problem at hand), patch-based scoring rules are not subject to the double-penalty effect.

As noticeable by the low number of examples available in the literature, aggregation (and direct use) of scoring rules based on projection in dimension $k \geq 2$ is not standard practice, probably because such projections may lack interpretability. Instead, to assess the multivariate aspects of a forecast, scoring rules relying on summary statistics are often favored.

4.2 Summary statistics

Summary statistics are a central tool of statisticians' toolboxes as they provide interpretable and understandable quantities that can be linked to the behavior of the phenomenon studied. Moreover, their interpretability can be enhanced by the forecaster's experience, and this can be leveraged when constructing scoring rules based on them. Summary statistics are commonly present during the verification procedure and this can be extended by the use of new scoring rules derived from any summary statistic of interest. For example, numerous summary statistics can come in handy when studying precipitations over a region covered by gridded observation and forecasts. Firstly, it is common practice to focus on binary events, such as the exceedance of a threshold (e.g., the presence or absence of precipitation). This can be studied using the BS (Eq. 5) on all 1-dimensional marginals, as mentioned in the previous subsection, but also in a multivariate manner through the fraction of threshold exceedances (FTE) over patches as presented further. Regarding precipitation, it is standard to be interested in the prediction of total precipitation over a spatial region or a time period. This transformation of the field can be leveraged to construct a scoring rule. Finally, it is important to verify that the spatial structure of the forecast matches the spatial structure of observations. The spatial structure can be (partially) summarized by the variogram or by wavelet transformations. The predictive performance for the spatial structure can be assessed by their associated scoring rules: the VS of order p (Eq. 13) and the wavelet-based score (Buschow et al., 2019). Other summary statistics can be of interest to the phenomenon studied, Heinrich-Mertsching et al. (2024) present summary statistics specific to point processes focusing on clustering and intensity.

The best-known summary statistic is certainly the mean. In spatial statistics, it can be used to avoid double penalization when we are less interested in the exact location of the forecast but rather in a regional prediction. The transformation associated with the mean is

$$\mathrm{mean}_P(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P} X_i, \tag{15}$$

where P denotes a patch and $|P|$ its dimension. $\mathrm{mean}_P(\boldsymbol{X})$ is the average value of $\boldsymbol{X}$ over the spatial patch P. Proposition 1 ensures that this transformation can be used to construct proper scoring rules. The scoring rule involved in the construction has to be univariate; however, the choice depends on the general properties preferred. For example, the SE would focus on the mean of the transformed quantity, whereas the AE would target its median. We propose the aggregated CRPS of the spatial mean, which is defined as

$$\begin{aligned}\mathrm{CRPS}_{\mathrm{mean}_\mathcal{P},\boldsymbol{w}_\mathcal{P}}(F,\boldsymbol{y}) &= \sum_{P\in\mathcal{P}} w_P \ \mathrm{CRPS}_{\mathrm{mean}_P}(F,\boldsymbol{y}) \\ &= \sum_{P\in\mathcal{P}} w_P \ \mathrm{CRPS}(\mathrm{mean}_P(F), \mathrm{mean}_P(\boldsymbol{y})),\end{aligned} \tag{16}$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P \in \mathcal{P}$, and mean_P is the spatial mean over the patch P (Eq. 15). Practical details regarding the insensitivity to the double-penalty effect and the choice of patches are given in Sect. 5.4.

It is worth noting that the total can be derived by the mean transformation by removing the prefactor:

$$\mathrm{total}_P(\boldsymbol{X}) = \sum_{i\in P} X_i.$$

In the case of precipitation, the total is more used than the mean since the total precipitation over a river basin can be decisive in evaluating flood risk. For example, one could construct an adapted version of the amplitude component of the

SAL method (Wernli et al., 2008; Radanovics et al., 2018) using the SE if the mean total precipitation is of interest. Gneiting (2011) presents other possible links between the quantity of interest and the scoring rule associated. Similarly, the transformations associated with the minimum and the maximum over a patch P can be obtained:

$$\min_P(\boldsymbol{X}) = \min_{i\in P}(X_i),$$
$$\max_P(\boldsymbol{X}) = \max_{i\in P}(X_i).$$

The maximum or minimum can be useful when considering extreme events. It can help understand if the severity of an event is well captured. For example, as minimum and maximum temperatures affect crop yields (see, e.g., Agnolucci et al., 2020), it can be of particular interest that a weather forecast within an agricultural model correctly predicts the minimum and maximum temperatures. After studying the mean, it is natural to think of the moments of higher order. We can define the transformation associated with the variance over a patch P as

$$\mathrm{Var}_P(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P}(X_i - \mathrm{mean}_P(\boldsymbol{X}))^2.$$

The variance transformation can provide information on the fluctuations, or variability, of $\boldsymbol{X}$ over a patch and be used to assess the prediction of the local variability by the forecast. In a more general setup, it can be of interest to use a transformation related to the moment of order n, and the transformation associated follows naturally:

$$\mathrm{M}_{n,P}(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P}X_i^n.$$

More application-oriented transformations are the central or standardized moments (e.g., skewness or kurtosis). Their transformations can be obtained directly from estimators. As underlined in Heinrich-Mertsching et al. (2024), since Proposition 1 applies to any transformation, there is no condition on having an unbiased estimator to obtain proper scoring rules.

Threshold exceedance plays an important role in decision-making such as weather alerts. For example, MeteoSwiss' heat warning levels are based on the exceedance of daily mean temperature over three consecutive days (Allen et al., 2023a). They can be defined by the simultaneous exceedance of a certain threshold, and the fraction of threshold exceedance (FTE) is the summary statistic associated.

$$\mathrm{FTE}_{P,t}(\boldsymbol{X}) = \frac{1}{|P|}\sum_{i\in P}1_{\{X_i\geq t\}} \tag{17}$$

FTEs can be used as an extension of univariate threshold exceedances, and it prevents the double-penalty effect. FTEs may be used to target compound events (e.g., the simultaneous exceedances of a threshold at multiple locations of interest). Roberts and Lean (2008) used an FTE-based SE over different sizes of neighborhoods (patches) to verify at which scale forecasts become skillful. To assess extreme precipitation forecasts, Rivoire et al. (2023) introduces scores for extremes with temporal and spatial aggregation separately. Extreme events are defined as values higher than the seasonal 95 % quantile. In the subseasonal-to-seasonal range, the temporal patches are 7 d windows centered on the extreme event, and the spatial patches are square boxes of 150 km × 150 km centered on the extreme event. The final scores are transformed BSs (Eq. 5) with a threshold of one event predicted across the patch. We propose the aggregated SE of the FTE, which is defined as

$$\begin{aligned}\mathrm{SE}_{\mathrm{FTE}_{\mathcal{P},t},\boldsymbol{w}_{\mathcal{P}}}(F,\boldsymbol{y}) &= \sum_{P\in\mathcal{P}} w_P\ \mathrm{SE}_{\mathrm{FTE}_{P,t}}(F,\boldsymbol{y})\\ &= \sum_{P\in\mathcal{P}} w_P\ \mathrm{SE}\big(\mathrm{FTE}_{P,t}(F),\mathrm{FTE}_{P,t}(\boldsymbol{y})\big)\\ &= \sum_{P\in\mathcal{P}} w_P\ \big(\mathbb{E}_F[\mathrm{FTE}_{P,t}(\boldsymbol{X})] - \mathrm{FTE}_{P,t}(\boldsymbol{y})\big)^2, \end{aligned}\tag{18}$$

where $\mathcal{P}$ is an ensemble of spatial patches, w_P is the weight associated with a patch $P\in\mathcal{P}$, and $\mathrm{FTE}_{P,t}$ is the fraction of threshold exceedance over the patch P and for the threshold t (Eq. 17). This scoring rule is proper and focuses on the prediction of the exceedance of a threshold t via the fraction of locations over a patch P exceeding said threshold. The resemblance with the Brier score is clear and the aggregated SE of FTE becomes the aggregated BS when patches of a single location are considered.

Correctly predicting the structure dependence is crucial in multivariate forecasting. Variograms are summary statistics representing the dependence structure. The variogram of order p of the pair (i,j) corresponds to the following transformation:

$$\gamma_{ij}^p(\boldsymbol{X}) = |X_i - X_j|^p.$$

As mentioned in the Introduction, using both the transformation and aggregation principles, we can recover the VS of order p (Eq. 13) introduced in Scheuerer and Hamill (2015):

$$\begin{aligned}\mathrm{VS}_p(F,\boldsymbol{y}) &= \sum_{i,j=1}^{d} w_{ij}\mathrm{SE}_{\gamma_{ij}^p}(F,\boldsymbol{y})\\ &= \sum_{i,j=1}^{d} w_{ij}\big(\mathbb{E}_F[|X_i - X_j|^p] - |y_i - y_j|^p\big)^2.\end{aligned}$$

Along with the well-known VS of order p, Scheuerer and Hamill (2015) introduced alternatives where the scoring rule applied on the transformation is the CRPS (Eq. 6) or the AE (Eq. 3) instead of the SE (Eq. 2). As mentioned previously, under the *intrinsic hypothesis* of Matheron (1963) (i.e., pairwise differences only depend on the distance between locations), the weights can be selected to obtain an optimal signal-to-noise ratio. Moreover, the weights could be selected to investigate a specific scale by giving a non-zero weight to pairs separated by a given distance.

In the case of spatial forecasts over a grid of size $d \times d$, a spatial version of the variogram transformation is available:

$$\gamma^p_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{X}) = |X_{\boldsymbol{i}} - X_{\boldsymbol{j}}|^p,$$

where $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{D} = \{1, \ldots, d\}^2$ are the coordinates of grid points. Under the intrinsic hypothesis of Matheron (1963), the variogram between grid points separated by the vector $\boldsymbol{h}$ can be estimated by

$$\gamma_{\boldsymbol{X}}(\boldsymbol{h}) = \frac{1}{2|\mathcal{D}(\boldsymbol{h})|} \sum_{\boldsymbol{i} \in \mathcal{D}(\boldsymbol{h})} \gamma_{\boldsymbol{i},\boldsymbol{i}+\boldsymbol{h}}(\boldsymbol{X}),$$

where $\mathcal{D}(\boldsymbol{h}) = \{\boldsymbol{i} \in \mathcal{D} : \boldsymbol{i} + \boldsymbol{h} \in \mathcal{D}\}$. This directed variogram can be used to target the verification of the anisotropy of the dependence structure. The isotropy transformation associated with the vector $\boldsymbol{h}$ can be defined by

$$T_{\text{iso},\boldsymbol{h}}(\boldsymbol{h}) = -\frac{\left(\gamma_{\boldsymbol{X}}(\boldsymbol{h}) - \gamma_{\boldsymbol{X}}(\boldsymbol{h}^\perp)\right)^2}{\frac{2\gamma_{\boldsymbol{X}}(\boldsymbol{h})^2}{|\mathcal{D}(\boldsymbol{h})|} + \frac{2\gamma_{\boldsymbol{X}}(\boldsymbol{h}^\perp)^2}{|\mathcal{D}(\boldsymbol{h}^\perp)|}}, \tag{19}$$

where $\boldsymbol{h}^\perp = (-h_2, h_1)$ is orthogonal to $\boldsymbol{h} := (h_1, h_2)$. This transformation is the isotropy pre-rank function proposed in Allen et al. (2024) when $\boldsymbol{h} = (h, 0)$. The isotropy transformation considers the orthogonal directions formed by the abscissa and ordinate axes and evaluates how the variogram changes between these directions. The transformation leads to negative or zero quantities, with values close to zero characterizing isotropy and negative values corresponding to the anisotropy of the variograms in the directions and at the scale involved.

We propose two scoring rules that are used in Sects. 5 and 6: the anisotropic score and the power-variation score. We define the anisotropic score (AS), in its general form, as

$$\begin{aligned}\mathrm{AS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{h}} w_{\boldsymbol{h}} \; S_{T_{\text{iso},\boldsymbol{h}}}(F, \boldsymbol{y}) \\ &= \sum_{\boldsymbol{h}} w_{\boldsymbol{h}} \; S(T_{\text{iso},\boldsymbol{h}}(F), T_{\text{iso},\boldsymbol{h}}(\boldsymbol{y})),\end{aligned} \tag{20}$$

where $T_{\text{iso},\boldsymbol{h}}$ is a transformation summarizing the anisotropy of a field (Eq. 19). The anisotropic score is constructed based on the transformation principle to target misspecifications of anisotropy in the dependence structure between forecast and observations.

We propose the power-variation score of order p (PVS), which is based on the power-variation transformation of order p to focus on the discrimination of the regularity of the random fields:

$$T_{p,\boldsymbol{s}}(\boldsymbol{X}) = |\boldsymbol{X}_{\boldsymbol{s}+(1,1)} - \boldsymbol{X}_{\boldsymbol{s}+(1,0)} - \boldsymbol{X}_{\boldsymbol{s}+(0,1)} + \boldsymbol{X}_{\boldsymbol{s}}|^p,$$

$$\begin{aligned}p\mathrm{VS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} \mathrm{SE}_{T_{p,\boldsymbol{s}}}(F, \boldsymbol{y}) \\ &= \sum_{\boldsymbol{s} \in \mathcal{D}^*} w_{\boldsymbol{s}} (\mathbb{E}_F[T_{p,\boldsymbol{s}}(\boldsymbol{X})] - T_{p,\boldsymbol{s}}(\boldsymbol{y}))^2,\end{aligned} \tag{21}$$

where $\mathcal{D}^*$ is the domain $\mathcal{D}$ restricted to grid points such that $T_{p,\boldsymbol{s}}$ is defined (i.e., $\mathcal{D}^* = \{1, \ldots, 19\} \times \{1, \ldots, 19\}$). Note that in the literature on fractional random fields, the power-variation of order p is an important characteristic used to characterize the roughness of a random field and is commonly used for estimation purposes; see Benassi et al. (2004), Basse-O'Connor (2021) and the references therein.

4.3 Other transformations

Transformations other than projections or summary statistics can be used to target forecast characteristics. For example, a transformation in the form of a change in coordinates or a change in scale (e.g., a logarithmic scale) can be used to obtain proper scoring rules. We highlight two families of scoring rules that can be seen as transformation-based scoring rules: wavelet-based scoring rules and threshold-weighted scoring rules.

Generally speaking, wavelet-based scoring rules are built thanks to a projection of forecast and observation fields onto a wavelet basis. Based on the wavelet coefficients, dimension reduction might be performed to target specific characteristics such as the dependence structure or the location. The resulting coefficients of the forecast fields are compared to the coefficients of the observations fields using scoring rules (e.g., squared error, SE, or energy score, ES). Wavelet transformations are (complex) transformations, and thus, the scoring rules associated fall within the scope of Proposition 1. In particular, Buschow et al. (2019) used a dimension reduction procedure resulting in the obtention of a mean and a scale spectra and used scoring rules to compare forecasts and observation spectra. For example, the ES of the mean spectrum is used and shows good discrimination ability when the scale structure is misspecified.

Note that Buschow et al. (2019) proposed two other wavelet-based scoring rules: one based on the earth mover's distance (EMD) of the scale histograms and one based on the distance in the scale histograms' center of mass. The EMD-based scoring rules are not proper since the EMD is not a proper scoring rule (Thorarinsdottir et al., 2013), and the so-called distance between centers of mass is not a distance but rather a difference in position, leading to an improper scoring rule. However, the ES-based scoring rules are proper and could be derived from scale histograms.

Despite their apparent complexity, wavelet transformations allow for targeting interpretable characteristics such as the location (Buschow, 2022), the scale structure (Buschow et al., 2019; Buschow and Friederichs, 2020) or the anisotropy (Buschow and Friederichs, 2021). The transformations proposed for the deterministic forecasts setting in most of these articles could be used as foundations for future work willing to propose wavelet-based proper scoring rules targeting the location, the scale structure, or the anisotropy.

As showcased in Heinrich-Mertsching et al. (2024) for a specific example and hinted in Allen et al. (2024), trans-

formations can also be used to emphasize certain outputs. Threshold weighting is one of the three main types of weighting conserving the propriety of scoring rules. Its name comes from the fact that it corresponds to a weighting over different thresholds in the case of CRPS (Eq. 7; Gneiting, 2011). Recall that given a conditionally negative definite kernel ρ, the associated kernel scoring rule S_ρ is proper relative to $\mathcal{P}_\rho$. Many popular scoring rules are kernel scores such as the BS (Eq. 5), the CRPS (Eq. 6), the ES (Eq. 12), and the VS (Eq. 13). By definition (Allen et al., 2023b, Definition 4), threshold-weighted kernel scores are constructed as follows:

$$\begin{aligned}\mathrm{tw}S_\rho(F, \boldsymbol{y}; v) &= \mathbb{E}_F[\rho(v(\boldsymbol{X}), v(\boldsymbol{y}))] \\ &\quad - \frac{1}{2}\mathbb{E}_F[\rho(v(\boldsymbol{X}), v(\boldsymbol{X}'))] - \frac{1}{2}\rho(v(\boldsymbol{y}), v(\boldsymbol{y})) \\ &= S_\rho(v(F), v(\boldsymbol{y})),\end{aligned}$$

where v is the chaining function capturing how the emphasis is put on certain outputs. With this explicit definition, it is obvious that threshold-weighted kernel scores are covered by the framework of Proposition 1. It can be noted that Proposition 4 in Allen et al. (2023b) states that strict propriety of the kernel score is preserved by the chaining function v if and only if v is injective. Weighted scoring rules allow for emphasizing particular outcomes: when studying extreme events, it is often of particular interest to focus on values larger than a given threshold t, and this can be achieved using the chaining rule $v(x) = 1_{x \geq t}$. Threshold-weighted scoring rules have been used in verification procedures in the literature; we illustrate its use through three different studies. Lerch and Thorarinsdottir (2013) aggregated across station threshold-weighted CRPS to compare the upper tail performance of different daily maximum wind speed forecasts. Chapman et al. (2022) aggregated the threshold-weighted CRPS across locations to study the improvement of statistical postprocessing techniques, the importance of predictors, and the influence of the size of the training set on the performance. Allen et al. (2023a) used threshold-weighted versions of the CRPS, the ES, and the VS to compare the predictive performance of forecasts regarding heat wave severity; the scoring rules were aggregated across stations. Readers may refer to Allen et al. (2023a) and Allen et al. (2023b) for insightful reviews of weighted scoring rules in both univariate and multivariate settings.

5 Simulation study

This section provides simulated examples to showcase the different uses of the framework introduced in Sect. 3 to construct interpretable proper scoring rules for multivariate forecasts. Four examples are developed. Firstly, a setup where the emphasis is put on 1-marginal verification is proposed. This setup serves as a means of recalling and showing the limitations of strictly proper scoring rules and the benefits of interpretable scoring rules in a concrete setting. Secondly, a standard multivariate setup is studied where popular multivariate scoring rules (i.e., VS and ES) are compared to a multivariate scoring rule aggregated over patches and an aggregation-and-transformation-based scoring rule in their discrimination ability regarding the dependence structure. Thirdly, a setup introducing anisotropy in both observations and forecasts is introduced. Fourthly, we propose a setup to test the sensitivity of scoring rules to the double-penalty effect, and we introduce scoring rules that can be built to be resilient to some manifestation of the double-penalty effect.

In these four numerical experiments, the spatial field is observed and predicted on a regular 20×20 grid $\mathcal{D} = \{1, \ldots, 20\} \times \{1, \ldots, 20\}$. Observations are realizations of a Gaussian random field $(G(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{D}}$ with zero mean and a power-exponential covariance defined as

$$\mathrm{cov}(G(\boldsymbol{s}), G(\boldsymbol{s}')) = {\sigma_0}^2 \exp\left(-\left(\frac{\|\boldsymbol{s} - \boldsymbol{s}'\|}{\lambda_0}\right)^{\beta_0}\right), \quad \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}, \tag{22}$$

where ${\sigma_0}^2$ is the variance, λ_0 is the range parameter, and β_0 is the smoothness (or roughness) parameter. The parameters are taken to be equal to $\sigma_0 = 1$, $\lambda_0 = 3$ and $\beta_0 = 1$.

In each numerical experiment, we compare a few predictive distributions, including the distribution generating observations and other ones deviating from the generative distributions in a specific way. These different predictive distributions are evaluated with different scoring rules, and the aim is to illustrate the discriminatory ability of the different scoring rules.

The simulation study uses 500 observations of the random field $(G(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{D}}$. The scoring rules are computed using exact formulas when possible (see Appendix F), and, when exact formulas are not available, they are computed based on ensemble forecasts of 100 members. Estimated expectations over the 500 observations are computed, and the experiment is repeated 10 times. The corresponding results are represented by box plots. The units of the scoring rules are rescaled by the average expected score of the true distribution (i.e., the ideal forecast). The statistical significance of the ranking between forecasts is tested using a Diebold–Mariano test (Diebold and Mariano, 1995). When deemed necessary, statistical significance is mentioned for a confidence level of 95 %.

5.1 Marginals

This first numerical experiment focuses on the prediction of the 1-dimensional marginal distributions and the aggregation of univariate scoring rules. For simplicity, we consider only stationary random fields so that the 1-marginal distribution is the same at all grid points. Although similar conclusions could be drawn from a univariate framework (i.e., with independent 1-dimensional rather than spatial observations), this example aims to clarify the notion of interpretability and

presents notions that is reused in the following examples. The verification of marginals, along with other simple quantities, is usually one of the first steps of any multivariate forecast verification process.

Observations follow the model of Eq. (22), and multiple competing forecasts are considered:

- the ideal forecast is the Gaussian distribution generating observations and is used as a reference;
- the biased forecast is a Gaussian predictive distribution with the same covariance structure as the observation but a different mean, $\mathbb{E}[F_{\text{bias}}(\boldsymbol{s})] = c = 0.255$;
- the overdispersed forecast and the underdispersed forecast are Gaussian predictive distributions from the same model as the observations, except for an overestimation ($\sigma = 1.4$) and an underestimation ($\sigma = 2/3$) of the variance, respectively;
- the location-scale Student forecast is used where the marginals follow location-scale Student t distributions with parameters $\mu = 0$ and $\mathrm{df} = 5$, and τ is such that the standard deviation is 0.745 and the covariance structure the same as in Eq. (22).

In order to compare the predictive performance of forecasts, we use scoring rules constructed by aggregating univariate scoring rules. Here, the aggregation is done with uniform weights since there is no prior knowledge on the locations. The univariate scoring rules considered are the continuous ranked probability score (CRPS), the Brier score (BS), the quantile score (QS), the squared error (SE), and the Dawid–Sebastiani score (DSS). Figure 1a compares five different forecasts based on their expected CRPS. It can be seen that all forecasts except for the ideal one have similar expected values and no sub-optimal forecast is significantly better than the others. In order to gain more insight into the predictive performance of the forecast, it is necessary to use other scoring rules. In practice, the distribution is unknown; thus, it is impossible to know if a forecast is optimal. It is only possible to provide a ranking linked to the *closeness* of the forecast with respect to the observations. The definition of closeness depends on the scoring rule used: for example, the CRPS defines closeness in terms of the integrated quadratic distance between the two cumulative distribution functions (see, e.g., Thorarinsdottir and Schuhen, 2018).

If the quantity of interest is the value of a quantile of a certain level α, the aggregated QS is an appropriate scoring rule. Figure 1b shows the expected aggregated QS for three different levels of α: $\alpha = 0.5$, $\alpha = 0.75$, and $\alpha = 0.95$. $\alpha = 0.5$ is associated with the prediction of the median, and, since all the forecasts are symmetric and only the biased forecast is not centered on zero, the other forecasts are equally the best and optimal forecasts. If the third quartile is of interest ($\alpha = 0.75$), the location-scale Student forecast appears as significantly the best (among the non-ideal). For the higher level of $\alpha = 0.95$, the biased forecast is significantly the best since its bias error seems to be compensated by its correct prediction of the variance. Depending on the level of interest, the best forecast varies; the only forecast that would appear to be the best regardless of the level α is the ideal forecast, as implied by Eq. (8).

If a quantity of interest is the exceedance of a threshold t at each location, then the aggregated BS is an interesting scoring rule. Figure 1c shows the expectation of aggregated BS for the different forecasts and for two different thresholds ($t = 0.5$ and $t = 1$). Among the non-ideal forecasts, there seems to be a clearer ranking than for the CRPS. The overdispersed forecast is significantly the best regarding the prediction of the exceedance of the threshold $t = 0.5$, and the biased forecast is significantly the best regarding the exceedance of $t = 1$. As for the aggregated quantile score, the best forecast depends on the threshold t considered and the only forecast that is the best regardless of the threshold t is the ideal one (see Eq. 7).

If the moments are of interest, the aggregated SE discriminates the first moment (i.e., the mean), and the aggregated DSS discriminates the first two moments (i.e., the mean and the variance). Figure 1d presents the expected values of these scoring rules for the different forecasts considered in this example. The aggregated SEs of all forecasts, except the biased forecast, are equal since they have the same (correct) marginal means. The aggregated DSS presents the biased forecast as significantly the best one (among non-ideal). This is caused by the combined discrimination of the first two moments of the Dawid–Sebastiani score (see Eq. 9 and Sect. B1).

5.2 Multivariate scores over patches

This second numerical experiment focuses on the prediction of the dependence structure. Observations are sampled from the model of Eq. (22), and we compare forecasts that differ only in their dependence structure through misspecification of the range parameter λ and the smoothness parameter β:

- the ideal forecast is the Gaussian distribution generating the observations;
- the small-range forecast and the large-range forecast are Gaussian predictive distributions from the same model (Eq. 22) as the observations, except for an underestimation ($\lambda = 1$) and an overestimation ($\lambda = 5$), respectively, of the range;
- the under-smoothed forecast and the over-smoothed forecast are Gaussian predictive distributions from the same model as the observations except for an underestimation ($\beta = 0.5$) and an overestimation ($\beta = 2$), respectively, of the smoothness.

Since the forecasts differ only in their dependence structure, scoring rules acting on the 1-dimensional marginals

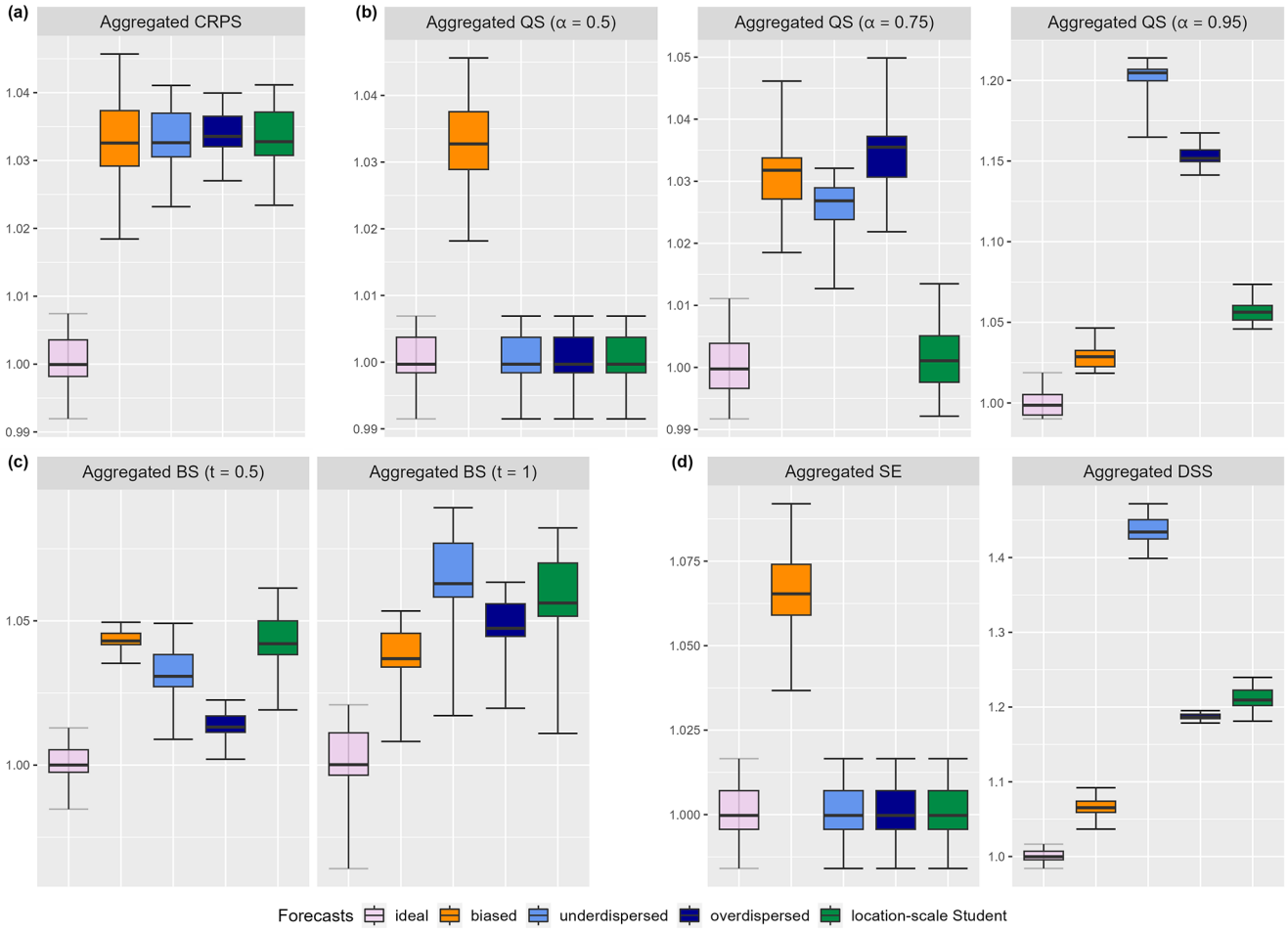

Figure 1. Expectation of aggregated univariate scoring rules: **(a)** the CRPS, **(b)** the quantile score, **(c)** the Brier score, and **(d)** the squared error and the Dawid–Sebastiani score for the ideal forecast (light violet), a biased forecast (orange), an underdispersed forecast (lighter blue), an overdispersed forecast (darker blue) and a local-scale Student forecast (green). More details are available in the main text.

would not be able to distinguish the ideal forecast from the others. We use the variogram score (VS) as a reference since it is known to be able to differentiate misspecifications of the dependence structure. We also use the patched ES (Eq. 14) with square patches of a given size s and uniform weights. Moreover, we consider the aggregated CRPS and the ES since they are limiting cases of the patched ES for 1×1 patches and a single patch over the whole domain $\mathcal{D}$, respectively. Additionally, we consider the power-variation score (PVS) of order p (Eq. 21). The PVS is meant to target misspecifications of the dependence structure at short scales and of roughness in forecasts.

In Fig. 2, the ES and the patched ES were computed using samples from the forecasts since closed expressions could not be derived. However, closed formulas for the VS and the PVS were derived and are available in Appendix F. As already shown in Scheuerer and Hamill (2015), the VS is able to significantly discriminate misspecification of the dependence structure induced by the range parameter λ (see Fig. 2a). Smaller orders of p (such as $p = 0.5$) appear as more informative than higher ones. Moreover, it is able to discriminate misspecifications induced by the smoothness parameter β (significantly for all orders p considered) even if it is less marked than for the misspecification of the range λ.

Figure 2b compares the forecasts using the p-variation score with $p \in \{0.5, 1, 2\}$. Note that the forecasts are provided in the same order as in the other panels. The PVS is able to (significantly) discriminate all four sub-optimal forecasts from the ideal forecast at all the orders of p. In the cases considered, the PVS has a stronger discriminating ability than the VS, in particular for the misspecification of the smoothness parameter β. The overall improvement in the discrimination ability of the PVS compared to the VS is because it only considers local pair interactions between grid points, which in the experimental setup considered greatly improves the signal-to-noise ratio compared to the VS. For example, it would be incapable of differentiating between two forecasts

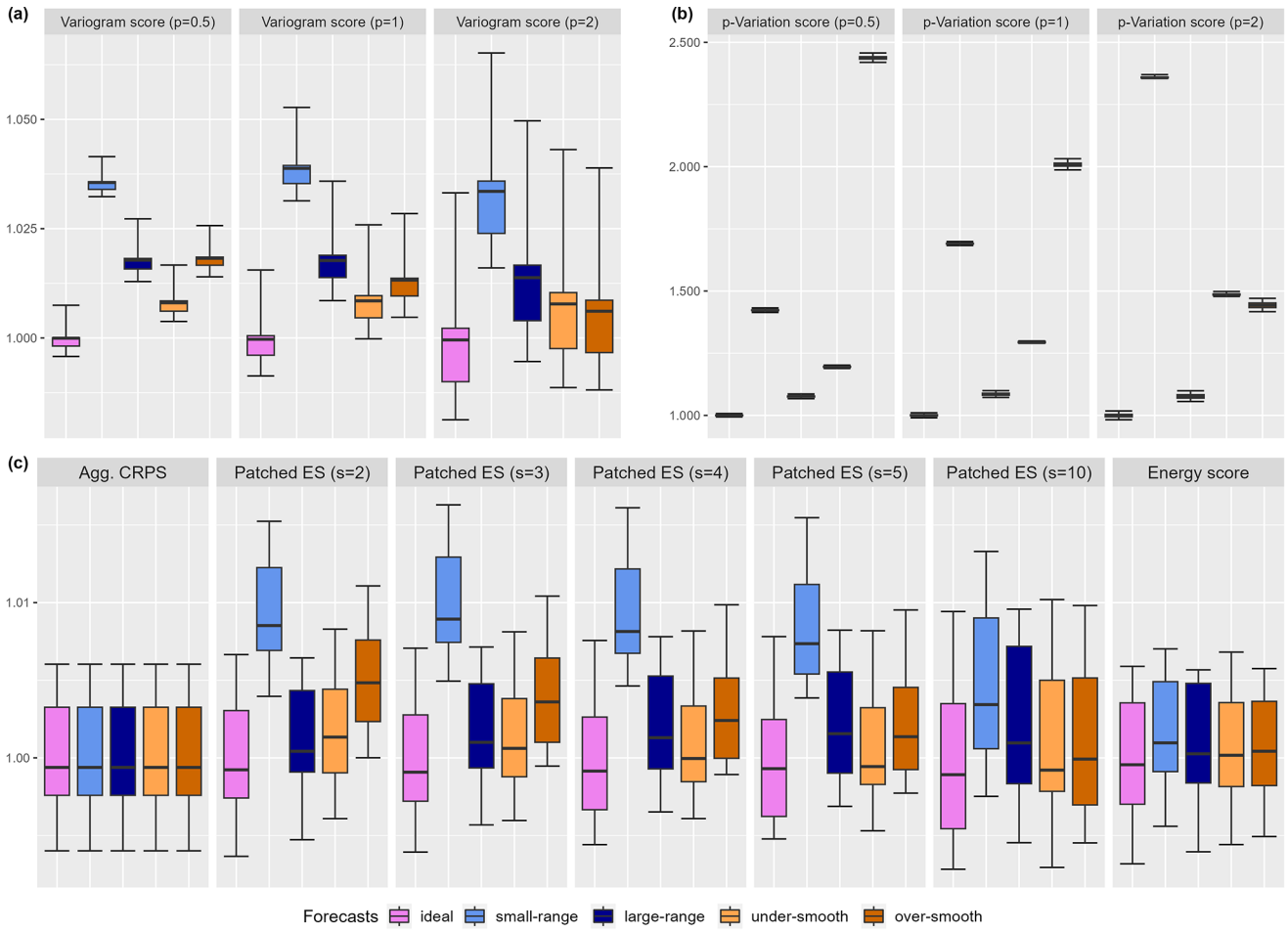

Figure 2. Expectation of scoring rules focused the dependence structure: **(a)** the variogram score, **(b)** the p-variation score and **(c)** the patched energy score (and its limiting cases: the aggregated CRPS and the energy score) for the ideal forecast (violet), the small-range forecast (lighter blue), the large-range forecast (darker blue), the under-smoothed forecast (lighter orange), and the over-smoothed forecast (darker orange). More details are available in the main text.

that only differ in their longer-range dependence structure, whereas the VS could.

Figure 2c shows that the patched ESs have a better discrimination ability than the ES. As expected by the clear analogy between the variogram score weights and the selection of valid patches, focusing on smaller patches improves the signal-to-noise ratio. For all patch size s values considered, the patched ES significantly differentiates the ideal forecast from the others. Whereas the ES does not significantly discriminate the misspecification of smoothness of the under-smoothed and over-smoothed forecasts. Nonetheless, the patched ES remains less sensitive than the VS to misspecifications in the dependence structure through the range parameter λ or the smoothness parameter β.

The VS relies on the aggregation and transformation principles and is able to discriminate misspecifications of the dependence structure. Similarly, the PVS is able to discriminate misspecifications of the dependence structure. Being based on more local transformations (i.e., p-variation transformation instead of variogram transformation), it has a greater discrimination ability than the VS in this experimental setup. In addition to this known application of the aggregation and transformation principles, it has been shown that multivariate transformations can be used to obtain patched scores that, in the case of the ES, lead to an improvement in the signal-to-noise ratio with respect to the original scoring rule.

5.3 Anisotropy

In this example, we focus on the anisotropy of the dependence structure. We introduce geometric anisotropy in observations and forecasts via the covariance function in the following way:

$$\mathrm{cov}(G(\boldsymbol{s}), G(\boldsymbol{s}')) = \exp\left(-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|_A}{\lambda_0}\right)\right),$$

with $\|\boldsymbol{s}-\boldsymbol{s}'\|_A = (\boldsymbol{s}-\boldsymbol{s}')^T A(\boldsymbol{s}-\boldsymbol{s}')$. The matrix $\mathbf{A}$ has the following form:

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \rho\sin\theta & \rho\cos\theta \end{bmatrix},$$

with $\theta \in [-\pi/2, \pi/2]$ being the direction of the anisotropy and ρ the ratio between the axes.

The observations follow the anisotropic version of the model in Eq. (22), where the covariance function presents the geometric anisotropy introduced above with $\lambda_0 = 3$ (as previously) and $\rho_0 = 2$ and $\theta_0 = \pi/4$. Multiple forecasts are considered that only differ in their prediction of the anisotropy in the model:

- the ideal forecast has the same distribution as the observations and is used as a reference;
- the small-angle forecast and the large-angle forecast have a correct ratio ρ but an under- and over-estimation of the angle, respectively (i.e., $\theta_{\text{small}} = 0$ and $\theta_{\text{large}} = \pi/2$);
- the isotropic forecast and the over-anisotropic forecast have a ratio of $\rho = 1$ and $\rho = 3$, respectively, but a correct angle θ.

Since these forecasts differ only in the anisotropy of their dependence structure, scoring rules not suited to discriminate the dependence structure would not be able to differentiate them. We compare two proper scoring rules: the variogram score and the anisotropic scoring rule. The variogram score is studied in two different settings: one where the weights are proportional to the inverse of the distance and one where the weights are proportional to the inverse of the anisotropic distance $\|\cdot\|_A$, which is supposed to be more informed since it is the quantity present in the covariance function. We use the anisotropic score (Eq. 20) in a special case of this where we do not aggregate across vectors $\boldsymbol{h}$ and where S is the squared error:

$$\begin{aligned} S_{T_{\text{iso},\boldsymbol{h}}}(F, \boldsymbol{y}) &= \text{SE}(T_{\text{iso},\boldsymbol{h}}(F), T_{\text{iso},\boldsymbol{h}}(\boldsymbol{y})) \\ &= \left(\mathbb{E}_{T_{\text{iso},\boldsymbol{h}}(F)}[\boldsymbol{X}] - T_{\text{iso},\boldsymbol{h}}(\boldsymbol{y})\right)^2. \end{aligned} \tag{23}$$

We consider vectors on the first bisector (i.e., of the form $\boldsymbol{h} = (h, h)$). The choice of this transformation instead of the transformation based on the anisotropy along the abscissa and ordinate is motivated by the fact that it leads to a clearer differentiation of the forecasts (not shown).

Figure 3a presents the variogram score of order $p = 0.5$ in its two variants. Both the standard VS and the informed VS can significantly distinguish the ideal forecast from the other forecasts but they have a weak sensitivity to misspecification of the geometric anisotropy. Even though the informed VS is supposed to increase the signal-to-noise ratio compared to the standard VS, it is not more sensitive to misspecifications in the experimental setup considered. Other orders of variograms were tested and worsened the discrimination ability of both standard and informed VS (not shown).

Figure 3b shows the AS (Eq. 23) with scales $1 \leq h \leq 5$ for the different forecasts and the aggregated AS (Eq. 20), where the scales are aggregated with weight $w_h = 1/h$. The anisotropic scores were computed using samples drawn from the forecasts; this explains why the ideal forecast may appear sub-optimal for some values of h (e.g., $h = 4$). As aimed by its construction, the AS is able to significantly distinguish the correct anisotropy behavior in the dependence structure for values of h up to $h = 3$ included. For $h = 4$, the AS does not significantly discriminate the isotropic forecast and the over-anisotropic forecast from the ideal one. The fact that $h = 1$ is the most sensitive to misspecifications is probably caused by the fact that the strength of the dependence structure decays with the distance (i.e., with h). This shows that the hyperparameter h plays an important role in having an informative AS (as do the weights and the order p for the variogram score). For $h = 2$ in particular, it can be seen that the AS is not sensitive to misspecifications of the ratio ρ and the angle θ in the same manner. This depends on the degree of misspecification but also on the hyperparameters of the AS. The aggregated AS allows us to avoid the selection of a scale h while conserving the discrimination ability of the lower values of h.

The anisotropic score is an interpretable scoring rule targeting the anisotropy of the dependence structure. However, it has the limitation of introducing hyperparameters in the form of the scale h and the axes along which the anisotropy is measured. Aggregation across scales and axes can circumvent the selection of these hyperparameters; however, a clever choice of weights is required to maintain the signal-to-noise ratio.

5.4 Double-penalty effect

In this example, we illustrate in a simple setting how scoring rules over patches can be robust to the double-penalty effect (see Sect. 2.4). The double-penalty effect is introduced in the form of forecasts that deviate from the ideal forecast by an additive or multiplicative noise term (i.e., nugget effect). The noises are centered uniforms such that the forecasts are correct on average but incorrect over each grid point.

Observations follow the model of Eq. (22) with the parameters $\sigma_0 = 1$, $\lambda_0 = 3$, and $\beta_0 = 1$. As per usual the ideal forecast, having the same distribution as the observations, is used as a reference. *Additive-noised forecasts* are the first type of forecast introduced to test the sensitivity of scoring rules to the form of the double-penalty effect (presented above). They differ from the ideal forecast through their marginals in the following way:

$$F_{\text{add}}(\boldsymbol{s}) = \mathcal{N}(\epsilon_{\boldsymbol{s}}, \sigma_0^2),$$

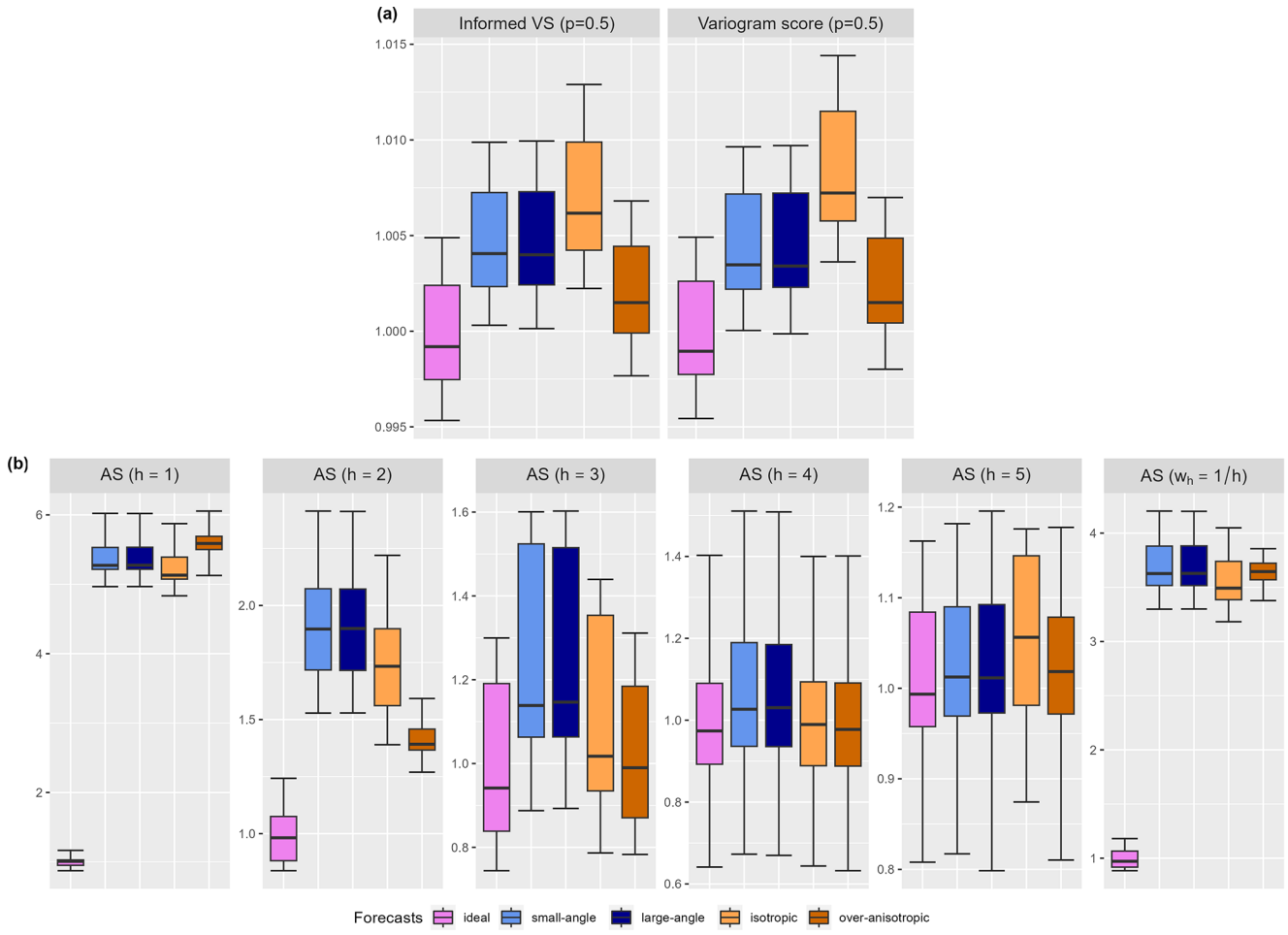

Figure 3. Expectation of interpretable proper scoring rules focused the dependence structure: **(a)** the variogram score and **(b)** the anisotropic score for the ideal forecast (violet), the small-angle forecast (lighter blue), the large-angle forecast (darker blue), the isotropic forecast (lighter orange) and the over-anisotropic forecast (darker orange). More details are available in the main text.

where $\epsilon_{\boldsymbol{s}} \sim \text{Unif}([-r, r])$ is a spatial white noise independent at each location $\boldsymbol{s} \in \mathcal{D}$. This has an effect on the mean of the marginals at each grid point. Three different noise range values are tested, $r \in \{0.1, 0.25, 0.5\}$. Similarly, *multiplicative-noised forecasts* that differ from the ideal forecast through their marginals are introduced:

$$F_{\text{mul}}(\boldsymbol{s}) = \mathcal{N}(0, \sigma^2(1+\eta_{\boldsymbol{s}})^2),$$

where $\eta_{\boldsymbol{s}} \sim \text{Unif}([-r, r])$ and $\boldsymbol{s} \in \mathcal{D}$. This has an effect on the variance of the marginals at each grid point and, thus, on the covariance. The same noise range values are tested, $r \in \{0.1, 0.25, 0.5\}$.

The aggregated CRPS is a naive scoring rule that is sensitive to the double-penalty effect. We use the aggregated CRPS of the spatial mean (Eq. 16). It is proper and has an interpretation similar to the aggregated CRPS, but the forecasts are only evaluated on the performance of their spatial mean. In order to make the scoring more interpretable, only square patches of a given size s are considered and the weight values w_P are uniform. The size of the patches s can be determined by multiple factors such as the physics of the problem, the constraints of the verification in the case of models on different scales, or the hypotheses on a different behavior below and above the scale of the patch (e.g., independent and identically distributed; Taillardat and Mestre, 2020). Note that the aggregated CRPS of the spatial mean is equal to the aggregated CRPS when patches of size $s = 1$ are considered.

If a quantity of interest is the exceedance of a threshold t, the scoring rule associated with that is the Brier score (Eq. 5). We compare the aggregated BS with its multivariate counterpart over patches: the aggregated SE of the fraction of threshold exceedance (Eq. 18).

In Fig. 4, the values of the aggregated SE of FTE have been obtained by sampling the forecasts' distribution. Figure 4a compares the aggregated CRPS and the aggregated CRPS of the spatial mean for different values of the patch size s. For all the scoring rules, we observe an increase in the expected value with the increase in the range of the noise r.

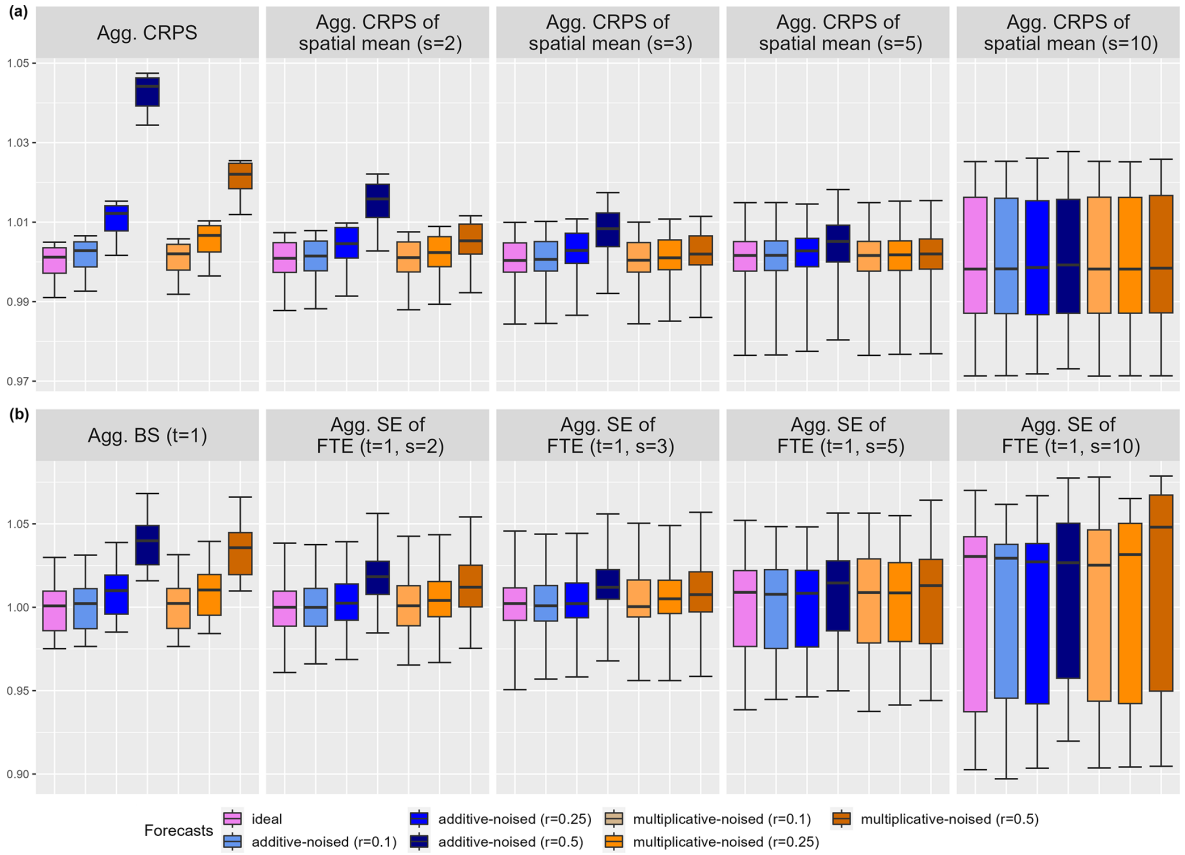

Figure 4. Expectation of scoring rules tested on their sensitivity to double-penalty effect: **(a)** the aggregated CRPS and the aggregated CRPS of the spatial mean, and **(b)** the aggregated Brier score and the aggregated squared error of fraction of threshold exceedances for the ideal forecast (violet), the additive-noised forecasts (shades of blue), and the multiplicative-noised forecasts (shades of orange). For the noised forecasts, darker colors correspond to larger values of the range $r \in \{0.1,\ 0.25,\ 0.5\}$. More details are available in the main text.

As expected, the aggregated CRPS is very sensitive to noise in the mean or the variance and, thus, is prone to the double-penalty effect. The aggregated CRPS of the spatial mean is less sensitive to noise of the mean or the variance. Moreover, different patch sizes allow us to select the spatial scale below which we want to avoid a double penalty. Given that the distribution of the noise is fixed in this simulation (i.e., uniform), patch size is related to the level of random fluctuations (i.e., the range r) tolerated by the scoring rule before significant discrimination with respect to the ideal forecast. It is worth noting that the range r of the noise leads to a stronger increase in the values of these CRPS-related scoring rules when the noise is on the mean rather than on the variance.

Figure 4b compares the aggregated BS and the aggregated squared error of the fraction of threshold exceedances. For simplicity, we fix the threshold at $t = 1$. The aggregated BS is, as expected, sensitive to noise in the mean or the variance, and an increase in the range of the noise leads to an increase in the expected value of the score. The aggregated SE of FTE acts as a natural extension of the aggregated BS to patches and provides scoring rules that are less sensitive to noise on the mean or the variance. The sensitivity evolves differently with the increase in the patch size s compared to the aggregated CRPS of the spatial mean since the aggregated SE of FTE measures the effect on the average exceedance over a patch. The range r of the noise apparently leads to a comparable increase in the values of the aggregated SE of FTE when the noise is additive or multiplicative.

The use of transformations over patches is similar to neighborhood-based methods in the spatial verification tools framework. Even though avoiding the double-penalty effect is not restricted to tools that do not penalize forecasts below a certain scale, this simulation setup presents a type of test relevant to probability forecasts. The patched-based scoring rules proposed here are not by themselves a sufficient verification tool since they are insensitive to some unrealistic forecast (e.g., if the mean value over the patch is correct, but deviations may be as large as possible and lead to unchanged values of the scoring rule). As for any other scoring rule, they should be used with other scoring rules.

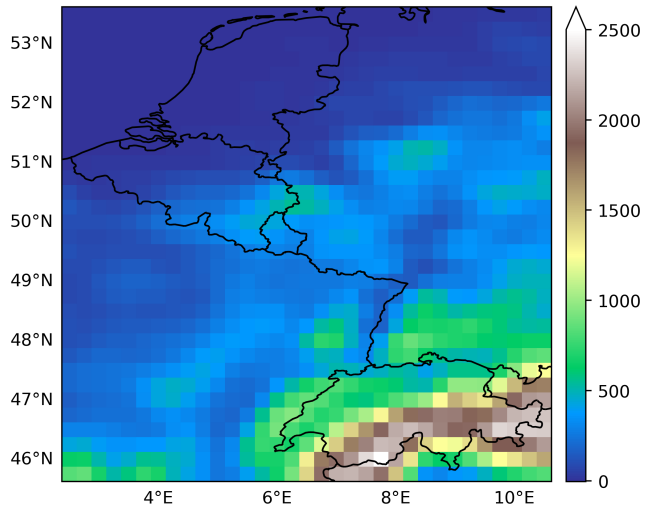

Figure 5. Model orography obtained by dividing the model surface geopotential by $g = 9.80665\,\mathrm{m\,s^{-2}}$ (as in Fig. 2a of Demaeyer et al., 2023).

6 Case study: postprocessed wind speed forecasts from EUPPBench

6.1 Data

We consider 10 m wind speed forecasts and reanalysis from the `MultivCalibration` package (Allen et al., 2024) relying on the European Meteorological Network's (EUMETNET) postprocessing benchmark dataset (EUPPBench; Demaeyer et al., 2023). EUPPBench has been developed to provide common ground to compare postprocessing techniques, but it also provides a common ground to illustrate forecast evaluation methods.

The dataset considered uses 20 years of reforecasts (simply forecasts thereafter) issued by the European Center for Medium-range Weather Forecasts' (ECMWF) Integrated Forecasting System (IFS). The forecasts are 11-member ensemble forecasts with a lead time of 5 d and are compared to ERA5 reanalyses (Hersbach et al., 2020) for their evaluation. Both forecasts and reanalyses are on a regular longitude–latitude grid with a resolution of 0.25° (approximately 25 km) over a part of central Europe (45.75–53.5° N, 2.5–10.5° E). Figure 5 shows the region covered by the dataset.

The dataset provides raw ensemble forecasts issued by IFS as well as two postprocessed forecasts and we use multivariate scoring rules to evaluate and compare these three different forecasts. Both postprocessed forecasts are obtained with a two-step procedure: (i) a standard ensemble model output statistics (EMOS; Gneiting et al., 2005) approach is applied at each grid point to postprocess the marginal distributions; (ii) then, the multivariate dependencies are retrieved either using ensemble copula coupling (ECC; Schefzik et al., 2013) or Schaake shuffle (ScS; Clark et al., 2004). The EMOS approach assumes that the predicted wind speeds follow a truncated logistic distribution (Scheuerer and Möller, 2015) with location and scale parameters linearly dependent on the ensemble mean forecast and the ensemble standard deviation, respectively. The parameters are estimated using the first 15 years of forecast-reanalysis pairs. The multivariate dependence is lost in the univariate postprocessing. It is retrieved by reordering evenly spaced quantiles from the postprocessed distribution at each grid point according to an appropriate dependence template. ECC uses the raw ensemble forecast as a template. While ScS uses a random sample of past observations (in our case reanalyses) to construct the dependence template.

The three forecasts (IFS, ECC, and ScS) are then compared on the remaining 5 years of unseen data (corresponding to 1045 forecast-reanalysis pairs). For more details on the dataset, readers may refer to Allen et al. (2024), Demaeyer et al. (2023), and references therein.

In Sect. 5, the multiple settings considered are controlled and the ideal forecast is known. However, in a real-world setting (such as the one considered in this case study), some limitations appear. First, the ideal forecast corresponding to the true distribution of the variable of interest is unknown. This implies that, when comparing competing forecasts, it is likely that no single forecast is better overall and that all forecasts considered present misspecifications. Scoring rules can help to describe which aspects are best captured by the different forecasts. Second, the quality of the estimation of expected scoring rules (and of their comparison) depends on the quantity and the quality of the data available for verification. Ideally, the verification data should be consistent and be composed of enough realizations.

Given the heterogeneity of the wind speed over the domain, some regions might have a stronger influence on aggregated scoring rules. However, if any prior knowledge is available, it can be used in the weights of the aggregated scoring rules, or the individual contribution can be investigated separately.

We compare the three forecasts using a standard verification procedure and show how interpretable scoring rules constructed using the aggregation-and-transformation-based framework can fit within the procedure and enable the characterization of differences between forecasts.

6.2 Results

Forecasts are compared using multiple scoring rules, and, as in Sect. 5, Diebold–Mariano tests (Diebold and Mariano, 1995) are used to test the statistical significance of the ranking between forecasts. Since the scoring rules considered are proper, the comparison of the expected scoring rules of two forecasts can be summarized by the *skill score*. For a proper scoring rule S, the skill score of a forecast F with respect to

a reference forecast F_{ref} is defined as

$$\text{SS}(F, F_{\text{ref}}) = \frac{\mathbb{E}_G[S(F_{\text{ref}}, \boldsymbol{Y})] - \mathbb{E}_G[S(F, \boldsymbol{Y})]}{\mathbb{E}_G[S(F_{\text{ref}}, \boldsymbol{Y})]}, \tag{24}$$

where G is the distribution of the observations and $\mathbb{E}_G[\dots]$ is the expectation with respect to $Y \sim G$. The skill score is positive if the forecast F improves the expected score with respect to the reference forecast F_{ref} and negative otherwise. The skill score can be expressed in percentage. Given that we consider the postprocessing of wind speed forecasts, the reference of choice is the raw ensemble (IFS) that the postprocessed forecasts (ECC and ScS) aim to improve upon.

We start by evaluating forecasts using aggregated univariate scoring rules to compare the performance of the forecasts on 1-dimensional marginals. Table 1 relates the average values of the scoring rules considered for IFS, ECC, and ScS. Values are reported in bold if they correspond to the statistically significant lowest value for a confidence level of 95 %. Both postprocessed forecasts have the same values since they provide the same 1-dimensional predictive distributions and differ only in their dependence structure. Thus, the difference observed between the raw forecast and the postprocessed forecasts is only the effect of the univariate EMOS postprocessing. The postprocessing of marginals using EMOS significantly improves the predictive performance overall (as shown by the aggregated CRPS) but also in terms of specific characteristics of the wind speed forecasts: the first two moments (aggregated SE and aggregated DSS), exceedance of selected thresholds (aggregated BS with thresholds $t \in \{2.5, 5, 7.5, 10\}\,\text{m s}^{-1}$), and quantiles of selected levels (aggregated QS of levels $\alpha \in \{0.7, 0.8, 0.9\}$).

We next proceed with scoring rules assessing the multivariate/spatial structure of the forecasts. We compute the ES and aggregation-and-transformation-based scoring rules that are introduced above and used in the numerical experiments (see Sect. 5). Table 2 relates the average values for the three competing forecasts. The scoring rules based on patches only consider a patch size of 3×3 grid points. This corresponds to a patch size close to an average administrative region; forecasts over such administrative regions are considered when issuing warnings (see, e.g., EUMETNET, 2024). The ES (Eq. 12), which is strictly proper, evaluates the overall multivariate predictive performance and ranks ECC as the best forecast. On the opposite, the patched ES (Eq. 14) considered shows ScS as the significantly best forecast when the spatial dependence is limited on a patch of 3×3 grid points. This seems to indicate that ScS produces a better forecast of the dependence structure at a short range while ECC is better at larger ranges. For both scoring rules, the postprocessed forecasts ECC and ScS significantly improve upon the raw forecast IFS.

The VS (Eq. 13) with different values of the parameter p shows that ECC seems to have a better predictive performance of the dependence structure when aggregating across all spatial scales. Where the PVS (Eq. 21) shows that ScS appears to have a better prediction of the dependence structure at smaller scales and in terms of roughness. These results corroborate the previous finding that ECC and ScS better account for the short-range and large-range spatial dependence, respectively. Quite surprisingly, for the short-range dependence and roughness as scored with PVS, IFS performs significantly better than ECC. Anisotropy is another characteristic of the dependence structure that is targeted by AS (Eq. 23). Given the large-scale circulations over the region of interest, the eastward direction can be considered to be well suited to investigate the anisotropy of wind speeds. Hence, we consider AS with $\boldsymbol{h} = (h, 0)$. For $h = 1$ (i.e., a lag of one grid point), IFS appears to have the best prediction of the anisotropy dependence structure. Similarly, Allen et al. (2024) have found that ECC and ScS are not able to improve the calibration of the dependence structure anisotropy at a lag of $h = 1$ compared to IFS. For a different lag, $h = 2$, ScS and IFS have comparable performances and, at $h = 3$, ScS becomes significantly the best in terms of AS.

Finally, we compare forecasts using scoring rules that are robust to the double-penalty effect: the aggregated CRPS of the spatial mean (Eq. 16) and the aggregated SE of FTE (Eq. 18). The CRPS of the spatial mean shows ScS as the significantly best where ECC and ScS have the same performance in terms of aggregated CRPS. This indicates that, if location errors at the scale of 3×3 patches are tolerated, ScS is the best forecast in terms of the mean wind speed over patches. Note that when investigating peak wind speed (or wind gusts), the preferred transformation is the maximum over patches. The aggregated SE of FTE shows that ECC and ScS have the best performance regarding the exceedance of various values of the threshold t over patches. As for the aggregated univariate scoring rules, the aggregated SE of FTE is invariant to the reordering of forecasts at each grid point and thus leads to strictly the same values for ECC and ScS.

The aggregation-and-transformation-based scoring rules can explain and mitigate the ranking of competing forecasts in terms of their overall performance associated with a strictly proper scoring rule (such as the ES). In particular, they allow us to investigate specific characteristics when the overall performance does not grasp the complexity of the differences between competing forecasts.

7 Conclusions

Verification of probabilistic forecasts is an essential but complex step of all forecasting procedures. Scoring rules may appear as the perfect tool to compare forecast performance since, when proper, they can simultaneously assess calibration and sharpness. However, propriety, even if strict, does not ensure that a scoring rule is relevant to the problem at hand. With that in mind, we agree with the recommendation of Scheuerer and Hamill (2015) that "several different scores be always considered before drawing conclusions". This is

Table 1. Skill scores of aggregated scoring rules acting on the 1-marginals with respect to IFS. The values of the threshold t of the aggregated Brier scores are expressed in $\mathrm{m\,s^{-1}}$. Values reported in bold identify the significantly best forecast (according to a Diebold–Mariano test with a confidence level of 95 %).

	Agg. SE	Agg. DSS	Agg. BS				Agg. QS			Agg. CRPS
			$t = 2.5$	$t = 5$	$t = 7.5$	$t = 10$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	
ECC	**7.03 %**	**33.64 %**	**4.87 %**	**6.55 %**	**7.55 %**	**6.96 %**	**5.82 %**	**6.65 %**	**8.62 %**	**6.64 %**
ScS	**7.03 %**	**33.64 %**	**4.87 %**	**6.55 %**	**7.55 %**	**6.96 %**	**5.82 %**	**6.65 %**	**8.62 %**	**6.64 %**

Table 2. Skill scores of multivariate scoring rules, including aggregation-and-transformation-based scoring rules, with respect to IFS. The values of the threshold t of the aggregated Brier scores are expressed in $\mathrm{m\,s^{-1}}$. Values reported in bold identify the significantly best forecast (according to a Diebold–Mariano test with a confidence level of 95 %). More details on the construction of the scoring rules are provided in the text.

	ES	Patched ES	VS			PVS		
			$p = 0.5$	$p = 1$	$p = 2$	$p = 0.5$	$p = 1$	$p = 2$
ECC	**3.93 %**	5.77 %	**5.20 %**	**6.44 %**	**6.38 %**	−18.62 %	−22.02 %	−13.36 %
ScS	3.61 %	**6.01 %**	3.44 %	5.19 %	5.70 %	**9.90 %**	**13.63 %**	**20.80 %**

	AS			Agg. CRPS of	Agg. SE of FTE			
	$h = 1$	$h = 2$	$h = 3$	spatial mean	$t = 2.5$	$t = 5$	$t = 7.5$	$t = 10$
IFS	**0.00 %**	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
ECC	−11.31 %	−7.28 %	−5.00 %	4.86 %	**2.77 %**	**5.88 %**	**7.47 %**	**7.73 %**
ScS	−15.30 %	0.04 %	**2.56 %**	**5.25 %**	**2.77 %**	**5.88 %**	**7.47 %**	**7.73 %**

even more important in a multivariate setting where forecasts are characterized by more complex objects.

We proposed a framework to construct proper scoring rules in a multivariate setting using aggregation and transformation principles. Aggregation-and-transformation-based scoring rules can improve the conclusions drawn since they enable the verification of specific aspects of the forecast (e.g., anisotropy of the dependence structure). This has been illustrated using examples from the literature and original ones. Moreover, more practical usages have been showcased in multiple numerical experiment settings and a case study of wind speed forecasts over central Europe. In particular, we have observed in the numerical experiments that the variogram score is more sensitive to large-scale misspecifications of the dependency, and the power-variation score is more sensitive to small-scale ones. This has been confirmed in the case study with the different ranking of the postprocessed forecasts based on the Schaake shuffle and ensemble copula coupling. Overall, we have shown that the aggregation and transformation principles can be used to construct scoring rules that are proper, interpretable, and not affected by the double-penalty effect. This could be a starting point to help bridge the gap between the proper scoring rule community and the spatial verification tools community.

As the interest in machine learning-based weather forecast is increasing (see, e.g., Ben Bouallègue et al., 2024a), multiple approaches have performance comparable to ECMWF deterministic high-resolution forecasts (Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2022; Chen et al., 2023). The natural extension to probabilistic forecast is already developing and enabled by publicly available benchmark datasets such as WeatherBench 2 (Rasp et al., 2024). Aggregation-and-transformation-based methods can help ensure that parameter inference does not hedge certain important aspects of the multivariate probabilistic forecasts.

There seems to be a trade-off between discrimination ability and strict propriety. Discrimination ability comes from the ability of scoring rules to differentiate misspecification of certain characteristics. By definition, the expectation of strictly proper scoring rules is minimized when the probabilistic forecast is the true distribution. Nonetheless, it does not guarantee that this global minimum is steep in any misspecification direction. However, interpretable scoring rules can discriminate the misspecification of their target characteristic. We believe that both interpretable and strictly proper scoring rules should coexist in any verification procedure and provide complementary information. Forecasters should be aware of the limitations and benefits of the verification tools they use and how they relate to the application of interest in order to take advantage of the available information. Regarding the theoretical existence of a trade-off between discrimination ability and strict propriety, the question is open, and its exploration in future research could help untangle this link.

Appendix A: Additional scoring rules

A1 Univariate scoring rules

When the probabilistic forecast F has a probability density function (PDF) f, scoring rules of a different type can be defined. Let $\mathcal{L}_\alpha(\mathbb{R})$ denote the class of probabilities on $\mathbb{R}$ that are absolutely continuous with respect to μ (usually taken as the Lebesgue measure) and have μ density f such that

$$\|f\|_\alpha = \left(\int_{\mathbb{R}} f(x)^\alpha \mu(\mathrm{d}x) \right)^{1/\alpha} < \infty.$$

The most popular scoring rule based on the PDF is the logarithmic score (also known as ignorance score; Good, 1952; Roulston and Smith, 2002). The logarithmic score is defined as

$$\mathrm{logS}(F, y) = -\log(f(y)), \tag{A1}$$

for y such that $f(y) > 0$. In its formulation, the logarithmic score is different from the scoring rules seen previously. Good (1952) proposed the logarithmic score knowing its link with the theory of information: its entropy is the Shannon entropy (Shannon, 1948), and its expectation is related to the Kullback–Leibler divergence (Kullback and Leibler, 1951) (see Sect. B1). The logarithmic score is strictly proper relative to the class $\mathcal{L}_1(\mathbb{R})$. Moreover, inference via minimization of the expected logarithmic score is equivalent to maximum likelihood estimation (see, e.g., Dawid et al., 2015). The logarithmic score belongs to the family of *local scoring rules*, which are scoring rules only depending on y, $f(y)$, and its derivatives up to a finite order. Another local scoring rule is the Hyvärinen score (HS; also known as the gradient scoring rule; Hyvärinen, 2005), and it is defined as

$$\mathrm{HS}(F, y) = 2\frac{f''(y)}{f(y)} - \frac{f'(y)^2}{f(y)^2}$$

for y such that $f(y) > 0$. The Hyvärinen score is proper relative to the subclass of $\mathcal{P}(\mathbb{R})$ such that the density f exists, is twice continuously differentiable, and satisfies $f'(x)/f(x) \to 0$ as $|x| \to \infty$. It is worth noticing that the Hyvärinen score can be computed even if f is only known up to a scale factor (e.g., up to a normalizing constant). This property allows for circumventing the use of Monte Carlo methods or approximations of the normalizing constant when it is unavailable or hard to compute. This is a property of local proper scoring rules, except for the logarithmic score (Parry et al., 2012). Readers eager to learn more about local proper scoring rules may refer to Parry et al. (2012) and Ehm and Gneiting (2012).

The logarithmic score and the Hyvärinen score do not allow f to be zero. To overcome this limitation, scoring rules expressed in terms of the L_α norm have been proposed. The quadratic score is defined as

$$\mathrm{QuadS}(F, y) = \|f\|_2^2 - 2f(y),$$

where $\|f\|_2^2 = \int_{\mathbb{R}} f(y)^2 \mathrm{d}y$. The quadratic score is strictly proper relative to the class $\mathcal{L}_2(\mathbb{R})$.

The pseudospherical score is defined as

$$\mathrm{PseudoS}(F, y) = -f(y)^{\alpha-1} / \|f\|_\alpha^{\alpha-1},$$

with $\alpha > 1$. For $\alpha = 2$, it reduces to the spherical score (see, e.g., Jose, 2007). The pseudospherical score is strictly proper relative to the class $\mathcal{L}_\alpha(\mathbb{R})$. The four scoring rules presented above have been criticized as they do not encourage a high probability in the vicinity of the observation y (Gneiting and Raftery, 2007). In particular, as the logarithmic score is more sensitive to outliers, probabilistic forecasts inferred by its minimization may be overdispersive (Gneiting et al., 2005). Moreover, the PDF is not always available, for example, in the case of ensemble forecasts.

A2 Multivariate scoring rules

When the PDF f of the probabilistic forecast F is available, multivariate versions of the univariate scoring rules based on the PDF are available. The multivariate versions of the scoring rules have the same properties and limitations as their univariate counterpart. The logarithmic score (Eq. A1) has a natural multivariate version:

$$\mathrm{logS}(F, \boldsymbol{y}) = -\log(f(\boldsymbol{y}))$$

for $\boldsymbol{y}$ such that $f(\boldsymbol{y}) > 0$. The logarithmic score is strictly proper relative to the class $\mathcal{L}_1(\mathbb{R}^d)$.

The Hyvärinen score (HS; Hyvärinen, 2005) was initially proposed in its multivariate form:

$$\mathrm{HS}(F, \boldsymbol{y}) = 2\Delta \log(f(\boldsymbol{y})) + |\nabla \log(f(\boldsymbol{y}))|^2$$

for $\boldsymbol{y}$ such that $f(\boldsymbol{y}) > 0$, where Δ is the Laplace operator (i.e., the sum of the second-order partial derivatives) and ∇ is the gradient operator (i.e., vector of the first-order partial derivatives). In the multivariate setting, the HS can also be computed if the predicted PDF is known up to a normalizing constant. The HS is proper relative to the subclass of $\mathcal{P}(\mathbb{R}^d)$ such that the density f exists, is twice continuously differentiable, and satisfies $\|\nabla \log(f(\boldsymbol{x}))\| \to 0$ as $\|\boldsymbol{x}\| \to \infty$.

The quadratic score and pseudospherical score are directly suited to the multivariate setting:

$$\mathrm{QuadS}(F, \boldsymbol{y}) = \|f\|_2^2 - 2f(\boldsymbol{y}),$$
$$\mathrm{PseudoS}(F, \boldsymbol{y}) = -f(\boldsymbol{y})^{\alpha-1} / \|f\|_\alpha^{\alpha-1},$$

where $\|f\|_\alpha = (\int_{\mathbb{R}^d} f(\boldsymbol{x})^\alpha \mathrm{d}\boldsymbol{x})^{1/\alpha}$. The quadratic score is strictly proper relative to the class $\mathcal{L}_2(\mathbb{R}^d)$. The pseudospherical score is strictly proper relative to the class $\mathcal{L}_\alpha(\mathbb{R}^d)$.

Appendix B: Expected scoring rules

B1 Univariate scoring rules

B1.1 Squared error

For any $F, G \in \mathcal{P}_2(\mathbb{R})$, the expectation of the squared error (Eq. 2) is

$$\mathbb{E}_G[\mathrm{SE}(F,Y)] = (\mu_F - \mu_G)^2 + {\sigma_G}^2,$$

where μ_F is the mean of the distribution F and μ_G and ${\sigma_G}^2$ are the mean and the variance of the distribution G.

Proof.

$$\begin{aligned}\mathbb{E}_G[\mathrm{SE}(F,Y)] &= \mathbb{E}_G[(\mu_F - Y)^2] \\ &= \mu_F^2 - 2\ \mu_F \mathbb{E}_G[Y] + \mathbb{E}_G[Y^2]\end{aligned}$$

Using the fact that $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$, the following applies:

$$\begin{aligned}\mathbb{E}_G[\mathrm{SE}(F,Y)] &= \mu_F^2 - 2\ \mu_F \mu_G + \sigma_G^2 + \mu_G^2 \\ &= (\mu_F - \mu_G)^2 + \sigma_G^2.\end{aligned}$$

□

B1.2 Quantile score

For any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the quantile score of level α (Eq. 4) is

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] = &\int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &- \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= \mathbb{E}_G[\mathrm{QS}_\alpha(G,Y)] \\ &+ \Big\{ (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &- \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (y - G^{-1}(\alpha)) G(\mathrm{d}y) \Big\}.\end{aligned}$$

Proof. Inspired by the proof of the propriety of the quantile score in Friederichs and Hense (2008).

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] &= \int_{\mathbb{R}} (1_{y \le F^{-1}(\alpha)} - \alpha)(F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= \int_{-\infty}^{F^{-1}(\alpha)} (1-\alpha)(F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &+ \int_{F^{-1}(\alpha)}^{+\infty} (-\alpha)(F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= \int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &- \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - y) G(\mathrm{d}y)\end{aligned}$$

Then, using $F^{-1}(\alpha) - y = (F^{-1}(\alpha) - G^{-1}(\alpha)) + (G^{-1}(\alpha) - y)$,

$$\begin{aligned}\mathbb{E}_G[\mathrm{QS}_\alpha(F,Y)] = &\int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - G^{-1}(\alpha)) G(\mathrm{d}y) \\ &- \alpha \int_{\mathbb{R}} (F^{-1}(\alpha) - G^{-1}(\alpha)) G(\mathrm{d}y) \\ &+ \int_{-\infty}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &- \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &+ \int_{-\infty}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &- \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &+ \int_{-\infty}^{G^{-1}(\alpha)} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &+ \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (G^{-1}(\alpha) - y) G(\mathrm{d}y) - \alpha \int_{\mathbb{R}} (G^{-1}(\alpha) - y) G(\mathrm{d}y) \\ &= (G(F^{-1}(\alpha)) - \alpha)(F^{-1}(\alpha) - G^{-1}(\alpha)) \\ &+ \mathbb{E}_G[\mathrm{QS}_\alpha(G,Y)] \\ &- \int_{G^{-1}(\alpha)}^{F^{-1}(\alpha)} (y - G^{-1}(\alpha)) G(\mathrm{d}y).\end{aligned}$$

□

B1.3 Absolute error

First of all, for $F \in \mathcal{P}_1(\mathbb{R})$ and $y \in \mathbb{R}$, the absolute error (Eq. 3) is equal to twice the quantile score of level $\alpha = 0.5$:

$$\mathrm{AE}(F, y) = |\mathrm{med}(F) - y| = 2\ \mathrm{QS}_{0.5}(F, y),$$

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

It can be deduced that, for any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the absolute error is

$$\begin{aligned}
\mathbb{E}_G[\mathrm{AE}(F, Y)] &= \mathbb{E}_G[|\mathrm{med}(F) - Y|] \\
&= 2 \int_{-\infty}^{\mathrm{med}(F)} (\mathrm{med}(F) - y) G(\mathrm{d}y) \\
&\quad - 2\alpha \int_{\mathbb{R}} (\mathrm{med}(F) - y) G(\mathrm{d}y) \\
&= \mathbb{E}_G[\mathrm{AE}(G, Y)] \\
&\quad + 2\left\{ (G(\mathrm{med}(F)) - \alpha)(\mathrm{med}(F) - \mathrm{med}(G)) \right. \\
&\quad \left. - \int_{\mathrm{med}(G)}^{\mathrm{med}(F)} (y - \mathrm{med}(G)) G(\mathrm{d}y) \right\}.
\end{aligned}$$

B1.4 Brier score

For any $F, G \in \mathcal{P}(\mathbb{R})$, the expectation of the Brier score (Eq. 5) is

$$\mathbb{E}_G[\mathrm{BS}_t(F, Y)] = (F(t) - G(t))^2 + G(t)(1 - G(t)).$$

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{BS}_t(F, Y)] &= \mathbb{E}_G[(F(t) - 1_{Y \le t})^2] \\
&= F(t)^2 - 2F(t)\mathbb{E}_G[1_{Y \le t}] + \mathbb{E}_G[{1_{Y \le t}}^2] \\
&= F(t)^2 - 2F(t)G(t) + G(t) \\
&= F(t)^2 - 2F(t)G(t) + G(t)^2 - G(t)^2 + G(t) \\
&= (F(t) - G(t))^2 + G(t)(1 - G(t))
\end{aligned}$$

□

B1.5 Continuous ranked probability score

For any $F, G \in \mathcal{P}_1(\mathbb{R})$, the expectation of the CRPS (Eq. 7) is

$$\begin{aligned}
\mathbb{E}_G[\mathrm{CRPS}(F, Y)] &= \mathbb{E}_{F,G}|X - Y| - \frac{1}{2}\mathbb{E}_F|X - X'| \\
&= \int_{\mathbb{R}} (F(z) - G(z))^2 \mathrm{d}z \\
&\quad + \int_{\mathbb{R}} G(z)(1 - G(z)) \mathrm{d}z,
\end{aligned}$$

where the second term of the last line is the entropy of the CRPS.

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{CRPS}(F, Y)] &= \mathbb{E}_G\left[\int_{\mathbb{R}} (F(z) - 1_{y \le z})^2 \mathrm{d}z \right] \\
&= \int_{\mathbb{R}} \mathbb{E}_G\left[(F(z) - 1_{y \le z})^2 \right] \mathrm{d}z \\
&= \int_{\mathbb{R}} \mathbb{E}_G\left[F(z)^2 - 2F(z)1_{y \le z} + 1^2_{y \le z} \right] \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{ F(z)^2 - 2F(z)\mathbb{E}_G\left[1_{y \le z}\right] \right. \\
&\quad \left. + \mathbb{E}_G\left[1_{y \le z}\right] \right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{ F(z)^2 - 2F(z)G(z) + G(z) \right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} \left\{ F(z)^2 - 2F(z)G(z) + G(z)^2 \right. \\
&\quad \left. - G(z)^2 + G(z) \right\} \mathrm{d}z \\
&= \int_{\mathbb{R}} (F(z) - G(z))^2 \mathrm{d}z + \int_{\mathbb{R}} G(z)(1 - G(z)) \mathrm{d}z
\end{aligned}$$

□

B1.6 Dawid–Sebastiani score

For any $F, G \in \mathcal{P}_2(\mathbb{R})$, the expectation of the Dawid–Sebastiani score (Eq. 9) is

$$\mathbb{E}_G[\mathrm{DSS}(F, Y)] = \frac{(\mu_F - \mu_G)^2}{{\sigma_F}^2} + \frac{{\sigma_G}^2}{{\sigma_F}^2} + 2\log\sigma_F.$$

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{DSS}(F, Y)] &= \mathbb{E}_G\left[\frac{(Y - \mu_F)^2}{{\sigma_F}^2} + 2\log\sigma_F \right] \\
&= \frac{\mathbb{E}_G\left[(Y - \mu_F)^2\right]}{{\sigma_F}^2} + 2\log\sigma_F
\end{aligned}$$

Noticing that $\mathbb{E}_G\left[(Y - \mu_F)^2\right] = \mathbb{E}_G\left[\mathrm{SE}(F, Y)\right]$, the following applies:

$$\mathbb{E}_G[\mathrm{DSS}(F, Y)] = \frac{(\mu_F - \mu_G)^2 + {\sigma_G}^2}{{\sigma_F}^2} + 2\log\sigma_F.$$

□

B1.7 Error-spread score

For any $F, G \in \mathcal{P}_4(\mathbb{R})$, the expectation of the error-spread score (Eq. 10) is

$$\begin{aligned}\mathbb{E}_G[\mathrm{ESS}(F,Y)] = &\left[({\sigma_G}^2 - {\sigma_F}^2) + (\mu_G - \mu_F)^2 - \sigma_F\gamma_F(\mu_G - \mu_F)\right]^2 \\ &+ {\sigma_G}^2\left[2(\mu_G - \mu_F) + (\sigma_G\gamma_G - \sigma_F\gamma_F)\right]^2 \\ &+ {\sigma_G}^4({\beta_G} - {\gamma_G}^2 - 1),\end{aligned}$$

where μ_F, σ_F^2, and γ_F are the mean, the variance, and the skewness of the probabilistic forecast F. Similarly, μ_G, σ_G^2, γ_G, and β_G are the first four centered moments of the distribution G. The proof is available in Appendix B of Christensen et al. (2014).

B1.8 Logarithmic score

For any $F, G \in \mathcal{P}(\mathbb{R})$ such that F and G have probability density functions in the class $\mathcal{L}_1(\mathbb{R})$, the expectation of the logarithmic score (Eq. A1) is

$$\mathbb{E}_G[\mathrm{LogS}(F,Y)] = D_{\mathrm{KL}}(G||F) + \mathrm{H}(F),$$

where $D_{\mathrm{KL}}(G||F)$ is the Kullback–Leibler divergence from F to G and $\mathrm{H}(F)$ is the Shannon entropy of F. The proof is straightforward given that the Kullback–Leibler divergence and Shannon entropy are defined as

$$D_{\mathrm{KL}}(G||F) = \int_{\mathbb{R}} g(y)\log\left(\frac{g(y)}{f(y)}\right)\mathrm{d}y,$$

$$\mathrm{H}(F) = \int_{\mathbb{R}} f(y)\log(f(y))\mathrm{d}y.$$

B1.9 Hyvärinen score

For F, G such that their densities, f and g, exist, are twice continuously differentiable, and satisfy $f'(x)/f(x) \to 0$ as $|x| \to \infty$ and $g'(x)/g(x) \to 0$ as $|x| \to \infty$, the expectation of the Hyvärinen score is

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= \int_{\mathbb{R}} \left(\frac{f'(y)^2}{f(y)^2} - 2\frac{f'(y)g'(y)}{f(y)g(y)}\right) g(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} \left(\frac{f'(y)}{f(y)} - \frac{g'(y)}{g(y)}\right)^2 g(y)\mathrm{d}y - \int_{\mathbb{R}} \frac{g'(y)^2}{g(y)^2} g(y)\mathrm{d}y,\end{aligned}$$

where the last formula shows the entropy of the Hyvärinen score (second term on the right-hand side).

Proof.

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= \mathbb{E}\left[2\frac{f''(Y)}{f(Y)} - \frac{f'(Y)^2}{f(y)^2}\right] \\ &= 2\int_{\mathbb{R}} \frac{f''(y)}{f(y)} g(y)\mathrm{d}y - \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y\end{aligned}$$

Integrating by part the integral of the first term on the right-hand side leads to

$$\begin{aligned}\int_{\mathbb{R}} \frac{f''(y)}{f(y)} g(y)\mathrm{d}y &= \left[\frac{f'(y)}{f(y)} g(y)\right]_{-\infty}^{+\infty} \\ &\quad - \int_{\mathbb{R}} f'(y)\frac{g'(y)f(y) - g(y)f'(y)}{f(y)^2}\mathrm{d}y \\ &= -\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y \\ &\quad + \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y.\end{aligned}$$

The boundary term is null since $f'(x)/f(x) \to 0$ as $|x| \to \infty$ and g is a probability density function.

Thus, the following applies:

$$\begin{aligned}\mathbb{E}_G[\mathrm{HS}(F,Y)] &= -2\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y \\ &\quad + 2\int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y - \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y \\ &= -2\int_{\mathbb{R}} \frac{f'(y)g'(y)}{f(y)g(y)} g(y)\mathrm{d}y \\ &\quad + \int_{\mathbb{R}} \frac{f'(y)^2}{f(y)^2} g(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} \left(\frac{f'(y)^2}{f(y)^2} - 2\frac{f'(y)g'(y)}{f(y)g(y)}\right) g(y)\mathrm{d}y.\end{aligned}$$

□

B1.10 Quadratic score

For any $F, G \in \mathcal{L}_2(\mathbb{R})$, the expectation of the quadratic score is

$$\mathbb{E}_G[\mathrm{QuadS}(F,Y)] = \|f\|_2^2 - 2\langle f, g\rangle,$$

where $\langle f, g\rangle = \int_{\mathbb{R}} f(y)g(y)\mathrm{d}y$.

B1.11 Pseudospherical score

For any $F, G \in \mathcal{L}_\alpha(\mathbb{R})$, the expectation of the quadratic score is

$$\mathbb{E}_G[\mathrm{PseudoS}(F,Y)] = -\frac{\langle f^{\alpha-1}, g\rangle}{\|f\|_\alpha^{\alpha-1}},$$

where $\langle f^{\alpha-1}, g\rangle = \int_{\mathbb{R}} f(y)^{\alpha-1} g(y)\mathrm{d}y$.

B2 Multivariate scoring rules

B2.1 Squared error

For any $F, G \in \mathcal{P}_2(\mathbb{R}^d)$, the expectation of the squared error (Eq. 11) is

$$\mathbb{E}_G[\mathrm{SE}(F, \boldsymbol{Y})] = \|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_G),$$

where $\boldsymbol{\mu}_F$ is the mean vector of the distribution F and $\boldsymbol{\mu}_G$ and $\boldsymbol{\Sigma}_G$ are the mean vector and the covariance matrix of the distribution G.

Proof. Let T_i denote the projection on the ith margin.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{SE}(F, \boldsymbol{Y})] &= \mathbb{E}_G[\|\boldsymbol{\mu}_F - \boldsymbol{Y}\|_2^2] \\
&= \mathbb{E}_G\left[\sum_{i=1}^{d} (\boldsymbol{\mu}_{T_i(F)} - T_i(\boldsymbol{Y}))^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{T_i(G)}[\mathrm{SE}(T_i(F), Y)] \\
&= \sum_{i=1}^{d} \left((\mu_{T_i(F)} - \mu_{T_i(G)})^2 + \sigma^2_{T_i(G)}\right) \\
&= \|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_G)
\end{aligned}$$

□

B2.2 Dawid–Sebastiani score

For any $F, G \in \mathcal{P}_2(\mathbb{R}^d)$, the expectation of the Dawid–Sebastiani score is

$$\begin{aligned}
\mathbb{E}_G[\mathrm{DSS}(F, \boldsymbol{Y})] = \log(\det \boldsymbol{\Sigma}_F) + (\boldsymbol{\mu}_F - \boldsymbol{\mu}_G)^T \boldsymbol{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - \boldsymbol{\mu}_G) \\
+ \mathrm{tr}(\boldsymbol{\Sigma}_G \boldsymbol{\Sigma}_F^{-1}).
\end{aligned}$$

The proof is available in the original article (Dawid and Sebastiani, 1999).

B2.3 Energy score

In a general setting, the expected energy score does not simplify. For any $F, G \in \mathcal{P}_\alpha(\mathbb{R}^d)$, the expected energy score (Eq. 12) is

$$\mathbb{E}_G[\mathrm{ES}_\alpha(F, \boldsymbol{Y})] = \mathbb{E}_{F,G}\|\boldsymbol{X} - \boldsymbol{Y}\|_2^\alpha - \frac{1}{2}\mathbb{E}_F\|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha.$$

B2.4 Variogram score

For any $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that the $2p$th moments of all their univariate margins are finite, the expected variogram score of order p (Eq. 13) is

$$\begin{aligned}
\mathbb{E}_G[\mathrm{VS}_p(F, \boldsymbol{Y})] = \sum_{i,j=1}^{d} w_{ij} \big(&\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 \\
&- 2\mathbb{E}_F\left[|X_i - X_j|^p\right]\mathbb{E}_G[|Y_i - Y_j|^p] \\
&+ \mathbb{E}_G[|Y_i - Y_j|^{2p}]\big).
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}_G[\mathrm{VS}_p(F, \boldsymbol{Y})] &= \mathbb{E}_G\left[\sum_{i,j=1}^{d} w_{ij}\left(\mathbb{E}_F\left[|X_i - X_j|^p\right] - |Y_i - Y_j|^p\right)^2\right] \\
&= \mathbb{E}_G\Bigg[\sum_{i,j=1}^{d} w_{ij}\big(\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 \\
&\quad - 2\mathbb{E}_F\left[|X_i - X_j|^p\right]|Y_i - Y_j|^p + |Y_i - Y_j|^{2p}\big)\Bigg] \\
&= \sum_{i,j=1}^{d} w_{ij}\big(\mathbb{E}_F\left[|X_i - X_j|^p\right]^2 \\
&\quad - 2\mathbb{E}_F\left[|X_i - X_j|^p\right]\mathbb{E}_G[|Y_i - Y_j|^p] \\
&\quad + \mathbb{E}_G[|Y_i - Y_j|^{2p}]\big)
\end{aligned}$$

□

B2.5 Logarithmic score

For any $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that F and G have probability density functions that belong to $\mathcal{L}_1(\mathbb{R}^d)$, the expectation of the logarithmic score is analogous to its univariate version:

$$\mathbb{E}_G[\mathrm{LogS}(F, \boldsymbol{Y})] = D_{\mathrm{KL}}(G||F) + \mathrm{H}(F),$$

where $D_{\mathrm{KL}}(G||F)$ is the Kullback–Leibler divergence from F to G and $\mathrm{H}(F)$ is the Shannon entropy of F.

$$D_{\mathrm{KL}}(G||F) = \int_{\mathbb{R}^d} g(\boldsymbol{x}) \log\left(\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

$$\mathrm{H}(F) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \log(f(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}$$

B2.6 Hyvärinen score

For $F, G \in \mathcal{P}(\mathbb{R}^d)$ such that their probability density functions f and g such that they are twice continuously differentiable and satisfying $\nabla f(\boldsymbol{x}) \to 0$ and $\nabla g(\boldsymbol{x}) \to 0$ as $\|\boldsymbol{x}\| \to \infty$, the expectation of the Hyvärinen score is

$$\begin{aligned}
\mathbb{E}_G[\mathrm{HS}(F, \boldsymbol{Y})] = \int_{\mathbb{R}^d} g(\boldsymbol{x}) \langle \nabla \log(f(\boldsymbol{x})) - 2\nabla \log(g(\boldsymbol{x})), \\
\nabla \log(f(\boldsymbol{x}))\rangle g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},
\end{aligned}$$

where ∇ is the gradient operator and $\langle \cdot, \cdot \rangle$ is the scalar product. The proof is similar to the proof for the univariate case using integration by parts and Stoke's theorem (Parry et al., 2012).

B2.7 Quadratic score

For any $F, G \in \mathcal{L}_2(\mathbb{R}^d)$, the expectation of the quadratic score is analogous to its univariate version

$$\mathbb{E}_G[\mathrm{QuadS}(F, \boldsymbol{Y})] = \|f\|_2^2 - 2\langle f, g \rangle,$$

where $\langle f, g \rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$.

B2.8 Pseudospherical score

For any $F, G \in \mathcal{L}_\alpha(\mathbb{R}^d)$, the expectation of the quadratic score is analogous to its univariate version

$$\mathbb{E}_G[\mathrm{PseudoS}(F, \boldsymbol{Y})] = -\frac{\langle f^{\alpha-1}, g\rangle}{\|f\|_\alpha^{\alpha-1}},$$

where $\langle f^{\alpha-1}, g\rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x})^{\alpha-1} g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Appendix C: General form of Corollary 1

Corollary 2. *Let $\mathcal{T} = \{T_i\}_{1\le i\le m}$ be a set of transformations from $\mathbb{R}^d$ to $\mathbb{R}^k$. Let $\mathcal{S} = \{S_i\}_{1\le i\le m}$ be a set of proper scoring rules such that S_i is proper relative to $T_i(\mathcal{F})$ for all $1 \le i \le m$. Let $\boldsymbol{w} = \{w_i\}_{1\le i\le m}$ be a set of non-negative weights. Then the scoring rule*

$$S_{\mathcal{S}_\mathcal{T},\boldsymbol{w}}(F, \boldsymbol{y}) = \sum_{i=1}^m w_i S_{i\,T_i}(F, \boldsymbol{y}) = \sum_{i=1}^m w_i S_i(T_i(F), T_i(\boldsymbol{y}))$$

is proper relative to $\mathcal{F}$.

Appendix D: Decomposition of kernel scores

We briefly discuss the link between the transformation and aggregation principles for scoring rules and the specific class of kernel scores. A kernel on $\mathbb{R}^d$ is a measurable function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying the following two properties:

i. (symmetry) $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\boldsymbol{x}_2, \boldsymbol{x}_1)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$;

ii. (non-negativity) $\sum_{1\le i\le j\le n} a_i a_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \ge 0$ for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$ for all $n \in \mathbb{N}$.

The kernel score S_k associated with the kernel k is defined on the space of predictive distributions

$$\mathcal{P}_k = \left\{ F \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \sqrt{k(x,x)}F(\mathrm{d}x) < +\infty \right\}$$

by

$$\begin{aligned} S_k(F, \boldsymbol{y}) &= \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} k(x_1, x_2)(F-\delta_{\boldsymbol{y}})(\mathrm{d}x_1)(F-\delta_{\boldsymbol{y}})(\mathrm{d}x_2), \\ &= \frac{1}{2}\mathbb{E}_F[k(\boldsymbol{X}, \boldsymbol{X}')] + \frac{1}{2}k(\boldsymbol{y}, \boldsymbol{y}) - \mathbb{E}_F[k(\boldsymbol{X}, \boldsymbol{y})], \end{aligned} \tag{D1}$$

where $\boldsymbol{y} \in \mathbb{R}^d$, $\delta_{\boldsymbol{y}}$ denotes the Dirac mass at $\boldsymbol{y}$, and $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random variables following F. Importantly, S_k is proper on $\mathcal{P}_k$ and, for an ensemble forecast $F = \frac{1}{M}\sum_{m=1}^M \delta_{\boldsymbol{x}_m}$ with M members, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$, it takes the simple form

$$\begin{aligned} S_k(F, \boldsymbol{y}) = \frac{1}{2M^2}\sum_{1\le m_1, m_2\le M} k(\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}) + \frac{1}{2}k(\boldsymbol{y}, \boldsymbol{y}) \\ - \frac{1}{M}\sum_{m=1}^M k(\boldsymbol{x}_m, \boldsymbol{y}), \end{aligned} \tag{D2}$$

making scoring rules particularly useful for ensemble forecasts.

The CRPS is surely the most widely used kernel score. Equation (6) shows that it is associated with the kernel $k(x_1, x_2) = |x_1| + |x_2| - |x_1 - x_2|$ (the function $|x_1 - x_2|$ is conditionally semi-definite negative so that k is non-negative). For more details on kernel scores, the reader should refer to Gneiting et al. (2005) or Steinwart and Ziegel (2021).

The following proposition reveals that a kernel score can always be expressed as an aggregation of squared errors (SEs) between transformations of the forecast–observation pair.

Proposition 3. *Let S_k be the kernel score associated with the kernel k. Then there exists a sequence of transformations $T_l : \mathbb{R}^d \to \mathbb{R}$, $l \ge 1$, such that*

$$S_k(F, \boldsymbol{y}) = \frac{1}{2}\sum_{l\ge1} \mathrm{SE}(T_l(F), T_l(\boldsymbol{y}))$$

for any predictive distribution $F \in \mathcal{P}_k$ and observation $\boldsymbol{y} \in \mathbb{R}^d$.

In particular, the series on the right-hand side is always finite. The proof is provided in Sect. E2 and relies on the reproducing kernel Hilbert space (RKHS) representation of kernel scores. In particular, we see that the sequence $(T_l)_{l\ge1}$ can be chosen as an orthonormal basis of the RKHS associated with the kernel k.

This representation of kernel scores can be useful to understand more deeply the comparison of the predictive forecast F and observation $\boldsymbol{y}$. While the definition (Eq. D1) is quite abstract, the series representation can be rewritten as

$$S_k(F, \boldsymbol{y}) = \sum_{l\ge1}\big(\mathbb{E}_F[T_l(\boldsymbol{X})] - T_l(\boldsymbol{y})\big)^2,$$

with $\boldsymbol{X}$ being a random variable following F. In other words, for $l \ge 1$, the observed value $T_l(\boldsymbol{y})$ is compared to the predicted value $T_l(\boldsymbol{X})$ under the predictive distribution F using the SE; then all these contributions are aggregated in a series forming the kernel score.

To give more intuition, we study two important cases in dimension $d = 1$. The details of the computations are provided in Sect. E3. For the Gaussian kernel score associated with the kernel

$$k(x_1, x_2) = \exp(-(x_1 - x_2)^2/2)$$

some computations yield the series representation

$$S_k(F, y) = \frac{1}{2}\sum_{l\ge0}\frac{1}{l!}\Big(\mathbb{E}_F[X^l e^{-X^2/2}] - y^l e^{-y^2/2}\Big)^2$$

so that this score compares the probabilistic forecast F and the observation y through the transforms

$$T_l(x) = \frac{1}{\sqrt{l!}}x^l e^{-x^2/2}, \quad l \ge 0.$$

For the CRPS, a possible series representation is obtained thanks to the following wavelet basis of functions: let $T^0(x) = x1_{[0,1)}(x) + 1_{[1,+\infty)}(x)$ (plateau function) and $T^1(x) = \big(1/2 - |x - 1/2|\big)1_{[0,1]}(x)$ (triangle function) and consider the collection of functions

$$T_l^0(x) = T^0(x-l), \quad T_{l,m}^1(x) = 2^{-m/2}T^1(2^m x - l),$$
$$l \in \mathbb{Z}, m \geq 0,$$

where $l \in \mathbb{Z}$ is a position parameter and $m \geq 0$ a scale parameter. Then, the CRPS can be written as

$$\begin{aligned}\mathrm{CRPS}(F,y) &= \sum_{l\in\mathbb{Z}} \mathrm{SE}(T_l^0(F), T_l^0(y)) \\ &\quad + \sum_{l\in\mathbb{Z}}\sum_{m\geq 0} \mathrm{SE}(T_{l,m}^1(F), T_{l,m}^1(y)) \\ &= \sum_{l\in\mathbb{Z}} \Big(\mathbb{E}_F[T^0(X-l)] - T^0(y-l)\Big)^2 \\ &\quad + \sum_{l\in\mathbb{Z}}\sum_{m\geq 0} 2^{-m}\Big(\mathbb{E}_F[T^1(2^m X - l)] - T(2^m y - l)\Big)^2.\end{aligned}$$

We can see that the CRPS compares forecast and observation through the SE after applying the plateau and triangle transformations for multiple positions and scales and then aggregates all the contributions.

Appendix E: Proofs

E1 Proposition 1

Proof of Proposition 1. Let $\mathcal{F} \subset \mathcal{P}(\mathbb{R}^d)$ be a class of probabilities on $\mathbb{R}^d$, and let $F \in \mathcal{F}$ be a forecast and $\boldsymbol{y} \in \mathbb{R}^d$ an observation. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be a transformation, and let S be a scoring rule on $\mathbb{R}^k$ that is proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), \boldsymbol{X} \sim F \in \mathcal{F}\}$.

$$\begin{aligned}\mathbb{E}_G[S_T(F,\boldsymbol{Y})] &= \mathbb{E}_G[S(T(F),T(\boldsymbol{Y}))] \\ &= \mathbb{E}_{T(G)}[S(T(F),\boldsymbol{Y})]\end{aligned}$$

Given that $T(F), T(G) \in T(\mathcal{F})$ and S is proper relative to $T(\mathcal{F})$, the following applies:

$$\begin{aligned}&\mathbb{E}_{T(G)}[S(T(G),\boldsymbol{Y})] \leq \mathbb{E}_{T(G)}[S(T(F),\boldsymbol{Y})] \\ \Leftrightarrow &\mathbb{E}_G[S_T(G,\boldsymbol{Y})] \leq \mathbb{E}_G[S_T(F,\boldsymbol{Y})]. \qquad \text{(E1)}\end{aligned}$$

□

Proof of the strict propriety case in Proposition 1. The notations are the same as in the proof above, except the following. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be an injective transformation, and let S be a scoring rule on $\mathbb{R}^k$ that is strictly proper relative to $T(\mathcal{F}) = \{\mathcal{L}(T(\boldsymbol{X})), \boldsymbol{X} \sim F \in \mathcal{F}\}$.

The equality in Eq. (E1) leads to

$$\begin{aligned}&\mathbb{E}_G[S_T(G,\boldsymbol{Y})] = \mathbb{E}_G[S_T(F,\boldsymbol{Y})] \\ \Leftrightarrow &\mathbb{E}_G[S(T(G),T(\boldsymbol{Y}))] = \mathbb{E}_G[S(T(F),T(\boldsymbol{Y}))] \\ \Leftrightarrow &\mathbb{E}_{T(G)}[S(T(G),\boldsymbol{Y})] = \mathbb{E}_{T(G)}[S(T(F),\boldsymbol{Y})].\end{aligned}$$

The fact that S is strictly proper relative to $T(\mathcal{F})$ leads to $T(F) = T(G)$, and finally, since T is injective, we have $F = G$.

□

E2 Proposition 3

Proof of Proposition 3. The proof relies on the reproducing kernel Hilbert space (RKHS) representation of the kernel score S_k. For a background on kernel score, maximum mean discrepancies, and RKHS, we refer to Smola et al. (2007) or Steinwart and Christmann (2008, Sect. 4).

Let $\mathcal{H}_k$ denote the RKHS associated with k. We recall that $\mathcal{H}_k$ contains all the functions $k(\boldsymbol{x}, \cdot)$ and that the inner product on $\mathcal{H}_k$ satisfies the property

$$\langle k(\boldsymbol{x}_1, \cdot), k(\boldsymbol{x}_2, \cdot)\rangle_{\mathcal{H}_k} = k(\boldsymbol{x}_1, \boldsymbol{x}_2).$$

The *kernel mean embedding* is a linear application $\Psi_k : \mathcal{P}_k \to \mathcal{H}_k$ mapping an admissible distribution $F \in \mathcal{P}_k$ to a function $\Psi_k(F)$ in the RKHS and such that the image of the point measure $\delta_{\boldsymbol{x}}$ is $k(\boldsymbol{x}, \cdot)$. Equation (D2) giving the kernel score for an ensemble prediction $F = \frac{1}{M}\sum_{m=1}^M \delta_{\boldsymbol{x}_m}$ can be written as

$$\begin{aligned}S_k(F,\boldsymbol{y}) &= \frac{1}{2}\langle \Psi_k(F) - \Psi_k(\delta_{\boldsymbol{y}}), \Psi_k(F) - \Psi_k(\delta_{\boldsymbol{y}})\rangle_{\mathcal{H}_k} \\ &= \frac{1}{2}\|\Psi_k(F - \delta_{\boldsymbol{y}})\|_{\mathcal{H}_k}^2.\end{aligned}$$

The properties of the kernel mean embedding ensure that this relation still holds for any $F \in \mathcal{P}_k$. As a consequence, if $(T_l)_{l\geq 1}$ is an Hilbertian basis of $\mathcal{H}_k$, we have

$$\begin{aligned}S_k(F,y) &= \frac{1}{2}\|\Psi_k(F - \delta_{\boldsymbol{y}})\|_{\mathcal{H}_k}^2 \\ &= \frac{1}{2}\sum_{l\geq 1}\langle \Psi_k(F - \delta_{\boldsymbol{y}}), T_l\rangle_{\mathcal{H}_k}^2.\end{aligned}$$

Finally, the properties of the kernel mean embedding ensure that, for all $T \in \mathcal{H}_k$, the following applies:

$$\langle \Psi_k(F - \delta_{\boldsymbol{y}}), T\rangle_{\mathcal{H}_k} = \int_{\mathbb{R}^d} T(\boldsymbol{x})(F - \delta_{\boldsymbol{y}})(\mathrm{d},\boldsymbol{x}) = \mathbb{E}_F[T(\boldsymbol{X})] - T(\boldsymbol{y})$$

whence the result follows.

□

E3 Proof of examples illustrating Proposition 3

Next, we illustrate the Proposition 3 and provide some computations in two cases: the Gaussian kernel score and the continuous rank probability score (CRPS).

E3.1 Gaussian kernel score

This is the scoring rule related to the Gaussian kernel:

$$k(x_1, x_2) = \exp(-(x_1 - x_2)^2/2), \quad x_1, x_2 \in \mathbb{R}.$$

Using a series expansion of the exponential function, we have

$$k(x_1,x_2)=\mathrm{e}^{-x_1^2/2}\mathrm{e}^{-x_2^2/2}\sum_{l\geq 0}\frac{(x_1x_2)^l}{l!}=\sum_{l\geq 0}T_l(x_1)T_l(x_2),$$

with T_l the transformation defined, for $l\geq 0$, by

$$T_l(x)=\frac{1}{\sqrt{l!}}\mathrm{e}^{-x^2/2}x^l.$$

As a consequence, the Gaussian kernel score writes for all $F\in\mathcal{P}(\mathbb{R})$ and $y\in\mathbb{R}$; that is, the following applies:

$$\begin{aligned}S_k(F,y)&=\frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}}k(x_1,x_2)(F-\delta_y)(\mathrm{d}x_1)(F-\delta_y)(\mathrm{d}x_2)\\&=\frac{1}{2}\int_{\mathbb{R}\times\mathbb{R}}\Big(\sum_{l\geq 0}T_l(x_1)T_l(x_2)\Big)(F-\delta_y)(\mathrm{d}x_1)(F-\delta_y)(\mathrm{d}x_2)\\&=\frac{1}{2}\sum_{l\geq 0}\Big(\int_{\mathbb{R}}T_l(x)(F-\delta_y)(\mathrm{d}x)\Big)^2\\&=\frac{1}{2}\sum_{l\geq 0}\Big(\mathbb{E}_F[T_l(X)]-T_l(y)\Big)^2.\end{aligned}$$

E3.2 Continuous ranked probability score

The CRPS is the scoring rule with kernel $k(x_1,x_2)=|x_1|+|x_2|-|x_1-x_2|$. This kernel is the covariance of the Brownian motion on $\mathbb{R}$ and its RKHS is known to be the Sobolev space $\mathrm{H}^1=\mathrm{H}^1(\mathbb{R})$; see Berlinet and Thomas-Agnan (2004). We recall the definition of the Sobolev space:

$$\mathrm{H}^1=\Big\{f\in\mathcal{C}(\mathbb{R},\mathbb{R}):f(0)=0,\dot{f}\in L^2(\mathbb{R})\Big\},$$

where $\dot{f}$ denotes the derivative of f assumed to be defined almost everywhere and square-integrable. The inner product on H^1 is defined by

$$\langle f_1,f_2\rangle_{H^1}=\int_{\mathbb{R}}\dot{f}_1(x)\dot{f}_2(x)\mathrm{d}x,$$

and one can easily check the fundamental relation

$$\langle k(x_1,\cdot),k(x_2,\cdot)\rangle_{H^1}=\int_{\mathbb{R}}\dot{k}(x_1,x)\dot{k}(x_2,x)\mathrm{d}x=k(x_1,x_2).$$

Here the derivative $\dot{k}(x_1,x)=1_{[0,x_1]}(x)$ is taken with respect to the second variable x. Then, we consider the Haar system defined as the following collection of functions:

$$H_l^0(x)=H^0(x-l)\quad\text{and}\quad H_{l,m}^1(x)=2^{m/2}H^1(2^mx-l),$$
$$l\in\mathbb{Z},\ m\geq 0,$$

with $H^0(x)=1_{[0,1)}(x)$ and $H^1(x)=1_{[0,1/2)}(x)-1_{[1/2,1)}(x)$. Since the Haar system is an orthonormal basis of the space $L^2(\mathbb{R})$, and the map $f\in H^1\longmapsto\dot{f}\in L^2$ is an isomorphism between Hilbert spaces, we obtain an orthonormal basis of $\mathrm{H}^1(\mathbb{R})$ by considering the primitives vanishing at 0 of the Haar basis functions. Setting $T^0(x)=x1_{[0,1)}(x)+1_{[1,+\infty)}(x)$ and $T^1(x)=\big(1/2-|x-1/2|\big)1_{[0,1]}(x)$ as the primitive functions of H^0 and H^1, respectively, we obtain the following system:

$$T_l^0(x)=T^0(x-l),\quad T_{l,m}^1(x)=2^{-m/2}T^1(2^mx-l),$$
$$l\in\mathbb{Z},\ m\geq 0.$$

The series representation of the CRPS is then deduced from Proposition 3 and its proof since the collection $\{T_{l,m}:l\in\mathbb{Z},m\geq 0\}$, is an orthonormal basis of the RKHS associated with the kernel k of the CRPS.

Appendix F: Scoring rules of the simulation study

The following formulas are deduced for a probabilistic forecast F taking the form of the Gaussian random field model of Eq. (22). The formulas of the aggregated univariate scoring rules can be obtained from the formulas in Gneiting and Raftery (2007) and Jordan et al. (2019) and, thus, are not presented here. We focus on the expression of the variogram score and the CRPS of the spatial mean.

F1 Variogram score

$$\mathrm{VS}_p(F,\boldsymbol{y})=\sum_{s,s'\in\mathcal{D}}w_{ss'}\big(\mathbb{E}_F[|X_s-X_{s'}|^p]-|y_s-y_{s'}|^p\big)^2$$

For $X\sim\mathcal{N}(\mu,\sigma^2)$, the absolute moment is as follows (Winkelbauer, 2014):

$$\mathbb{E}[|X|^\nu]=\sigma^\nu 2^{\nu/2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}}{}_1F_1\left(-\nu/2,1/2;-\frac{\mu^2}{2\sigma^2}\right),\quad\text{(F1)}$$

where ${}_1F_1$ is the confluent hypergeometric function of the first kind. For $X\sim F$, the following applies:

$$\begin{aligned}X_s-X_{s'}&\sim\mathcal{N}(\mu_s-\mu_{s'},{\sigma_s}^2+{\sigma_{s'}}^2-2\mathrm{cov}(F_s,F_{s'}))\\&\sim\mathcal{N}\left(0,2\sigma^2\left(1-e^{-\left(\frac{\|s-s'\|}{\lambda}\right)^\beta}\right)\right).\end{aligned}$$

This leads to

$$\begin{aligned}\mathbb{E}_G[|X_{\boldsymbol{s}} - X_{\boldsymbol{s}'}|^p] &= \left(2\sigma^2\left(1 - e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta}\right)\right)^{p/2} \\ &\quad 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}{}_1F_1\Bigg(-p/2, 1/2; \\ &\quad -\frac{(\mu_{\boldsymbol{s}} - \mu_{\boldsymbol{s}'})^2}{4\sigma^2(1 - e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta})}\Bigg) \\ &= 2^p\sigma^p\left(1 - e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta}\right)^{p/2} \\ &\quad \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}{}_1F_1(-p/2, 1/2; 0) \\ &= 2^p\sigma^p\left(1 - e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta}\right)^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.\end{aligned}$$

Finally, the following applies:

$$\begin{aligned}\mathrm{VS}_p(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s}\boldsymbol{s}'\in\mathcal{D}} w_{\boldsymbol{s}\boldsymbol{s}'}\left(\mathbb{E}_G[|X_{\boldsymbol{s}} - X_{\boldsymbol{s}'}|^p] - |y_{\boldsymbol{s}} - y_{\boldsymbol{s}'}|^p\right)^2 \\ &= \sum_{\boldsymbol{s}\boldsymbol{s}'\in\mathcal{D}} w_{\boldsymbol{s}\boldsymbol{s}'}\Bigg(\left(2\sigma^2(1 - e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta})\right)^{p/2} \\ &\quad 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} - |y_{\boldsymbol{s}} - y_{\boldsymbol{s}'}|^p\Bigg)^2.\end{aligned}$$

F2 Power-variation score

$$\begin{aligned}\mathrm{PVS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}}\mathrm{SE}_{T_{p,\boldsymbol{s}}}(F, \boldsymbol{y}) \\ &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}}(\mathbb{E}_F[T_{p,\boldsymbol{s}}(\boldsymbol{X})] \\ &\quad - T_{p,\boldsymbol{s}}(\boldsymbol{y}))^2\end{aligned}$$

Denote $Z = \boldsymbol{X}_{\boldsymbol{s}+(1,1)} - \boldsymbol{X}_{\boldsymbol{s}+(1,0)} - \boldsymbol{X}_{\boldsymbol{s}+(0,1)} + \boldsymbol{X}_{\boldsymbol{s}}$. For $X \sim F$, we have $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$, with

$$\mu_Z = \mu_{\boldsymbol{s}+(1,1)} - \mu_{\boldsymbol{s}+(1,0)} - \mu_{\boldsymbol{s}+(0,1)} + \mu_{\boldsymbol{s}} = 0$$

and

$$\begin{aligned}\sigma_Z^2 &= \sigma^2_{\boldsymbol{s}+(1,1)} + \sigma^2_{\boldsymbol{s}+(1,0)} + \sigma^2_{\boldsymbol{s}+(0,1)} + \sigma^2_{\boldsymbol{s}} \\ &\quad - 2\mathrm{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s}+(1,0))) \\ &\quad - 2\mathrm{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s}+(0,1)) \\ &\quad + 2\mathrm{cov}(F(\boldsymbol{s}+(1,1)), F(\boldsymbol{s})) \\ &\quad + 2\mathrm{cov}(F(\boldsymbol{s}+(1,0)), F(\boldsymbol{s}+(0,1))) \\ &\quad - 2\mathrm{cov}(F(\boldsymbol{s}+(1,0)), F(\boldsymbol{s})) \\ &\quad - 2\mathrm{cov}(F(\boldsymbol{s}+(0,1)), F(\boldsymbol{s})) \\ &= 4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta}).\end{aligned}$$

Using Eq. (F1), this leads to

$$\begin{aligned}\mathbb{E}_F[T_{p,\boldsymbol{s}}(\boldsymbol{X})] &= \left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} \\ &\quad 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}{}_1F_1(-p/2, 1/2; 0) \\ &= \left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} \\ &\quad 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{PVS}(F, \boldsymbol{y}) &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}}\mathrm{SE}_{T_{p,\boldsymbol{s}}}(F, \boldsymbol{y}) \\ &= \sum_{\boldsymbol{s}\in\mathcal{D}^*} w_{\boldsymbol{s}}\Bigg(\left(4\sigma^2(1 + e^{-(\sqrt{2}/\lambda)^\beta} - 2e^{-(1/\lambda)^\beta})\right)^{p/2} \\ &\quad 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} - |y_{\boldsymbol{s}+(1,1)} - y_{\boldsymbol{s}+(1,0)} - y_{\boldsymbol{s}+(0,1)} + y_{\boldsymbol{s}}|^p\Bigg)^2.\end{aligned}$$

F3 CRPS of the spatial mean

The CRPS of the spatial mean is defined as

$$\begin{aligned}\mathrm{CRPS}_{\mathrm{mean}_{\mathcal{P}}, \boldsymbol{w}_{\mathcal{P}}}(F, \boldsymbol{y}) &= \sum_{P\in\mathcal{P}} w_P\mathrm{CRPS}_{\mathrm{mean}_P}(F, \boldsymbol{y}) \\ &= \sum_{P\in\mathcal{P}} w_P\mathrm{CRPS}(\mathrm{mean}_P(F), \mathrm{mean}_P(\boldsymbol{y})),\end{aligned}$$

where $\mathcal{P}$ is an ensemble of spatial patches and w_P is the weight associated with a patch $P \in \mathcal{P}$. The mean of Gaussian marginals follows a Gaussian distribution:

$$\mathrm{mean}_P(F) \sim \mathcal{N}\left(\sum_{\boldsymbol{s}\in P}\mu_{\boldsymbol{s}}, \frac{\sigma^2}{|P|^2}\sum_{\boldsymbol{s},\boldsymbol{s}'\in P} e^{-\left(\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{\lambda}\right)^\beta}\right) = \mathcal{N}(\mu_P, \sigma_P^2),$$

where $|P|$ is the cardinal of the patch P (i.e., the number of grid points belonging to P).

Finally, the following applies:

$$\mathrm{CRPS}_{\mathrm{mean}_{\mathcal{P}}, \boldsymbol{w}_{\mathcal{P}}}(F, \boldsymbol{y}) = \sum_{P\in\mathcal{P}} w_P\mathrm{CRPS}(\mathcal{N}(\mu_P, \sigma_P^2), \mathrm{mean}_P(\boldsymbol{y})).$$

Code and data availability. The code used for the different numerical experiments of Sect. 5 and the case study of Sect. 6 is publicly available at https://github.com/pic-romain/aggregation-transformation (last access: 7 March 2025) and https://doi.org/10.5281/zenodo.14982271 (Pic, 2024). The data on which the case study of Sect. 6 relies are taken from the `MultivCalibration` package (https://doi.org/10.5281/zenodo.10201289, Allen, 2023) and EUPPBench (https://doi.org/10.5281/zenodo.7429236, Demaeyer, 2022).

Author contributions. RP prepared the original draft and wrote the code of the numerical experiments and the case study. RP, CD,

PN, and MT conceptualized, developed the methodology, and reviewed and edited the article.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. Sam Allen is thanked for fruitful discussions during the preparation of this article. The authors would like to thank Thordis Thorarinsdottir, Stéphane Vannitsem, and Eric Gilleland for providing feedback on a preliminary version of the article.

Financial support. The authors received support from the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and the Energy-oriented Centre of Excellence II (EoCoE-II; grant agreement 824158, funded within the EU Horizon2020 framework of the European Union). Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU, and this study has received funding from Agence Nationale de la Recherche – France 2030 as part of the PEPR TRACCS program under grant no. ANR-22-EXTR-0005 and the ANR EXSTA project (grant no. ANR-23-CE40-0009).

Review statement. This paper was edited by Mark Risser and reviewed by two anonymous referees.

References

Agnolucci, P., Rapti, C., Alexander, P., De Lipsis, V., Holland, R. A., Eigenbrod, F., and Ekins, P.: Impacts of rising temperatures and farm management practices on global yields of 18 crops, Nature Food, 1, 562–571, https://doi.org/10.1038/s43016-020-00148-x, 2020.

Al Masry, Z., Pic, R., Dombry, C., and Devalland, C.: A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles, Breast Cancer Res. Tr., https://doi.org/10.1007/s10549-023-07141-5, 2023.

Alexander, C., Coulon, M., Han, Y., and Meng, X.: Evaluating the discrimination ability of proper multi-variate scoring rules, Ann. Oper. Res., 334, 857–883, https://doi.org/10.1007/s10479-022-04611-9, 2022.

Allen, S.: sallen12/MultivCalibration: MultivCalibration v.1.0 (v.1.0), Zenodo [code, data set], https://doi.org/10.5281/zenodo.10201289, 2023.

Allen, S., Bhend, J., Martius, O., and Ziegel, J.: Weighted Verification Tools to Evaluate Univariate and Multivariate Probabilistic Forecasts for High-Impact Weather Events, Weather Forecast., 38, 499–516, https://doi.org/10.1175/waf-d-22-0161.1, 2023a.

Allen, S., Ginsbourger, D., and Ziegel, J.: Evaluating Forecasts for High-Impact Events Using Transformed Kernel Scores, SIAM/ASA Journal on Uncertainty Quantification, 11, 906–940, https://doi.org/10.1137/22m1532184, 2023b.

Allen, S., Ziegel, J., and Ginsbourger, D.: Assessing the calibration of multivariate probabilistic forecasts, Q. J. Roy. Meteor. Soc., 150, 1315–1335, https://doi.org/10.1002/qj.4647, 2024.

Anderson, J. L.: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations, J. Climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:amfpae>2.0.co;2, 1996.

Basse-O'Connor, A., Pilipauskaitė, V., and Podolskij, M.: Power variations for fractional type infinitely divisible random fields, Electron. J. Probab., 26, 1–35, https://doi.org/10.1214/21-EJP617, 2021.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, B. Am. Meteorol. Soc., 105, E864–E883, https://doi.org/10.1175/bams-d-23-0162.1, 2024a.

Ben Bouallègue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., and Chantry, M.: Improving Medium-Range Ensemble Weather Forecasts with Hierarchical Ensemble Transformers, Artificial Intelligence for the Earth Systems, 3, e230027, https://doi.org/10.1175/aies-d-23-0027.1, 2024b.

Benassi, A., Cohen, S., and Istas, J.: On roughness indices for fractional fields, Bernoulli, 10, 357–373, https://doi.org/10.3150/bj/1082380223, 2004.

Berlinet, A. and Thomas-Agnan, C.: Reproducing kernel Hilbert spaces in probability and statistics, with a preface by Persi Diaconis, Kluwer Academic Publishers, Boston, MA, ISBN 1-4020-7679-7, https://doi.org/10.1007/978-1-4419-9096-9, 2004.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bjerregård, M. B., Møller, J. K., and Madsen, H.: An introduction to multivariate probabilistic forecast evaluation, Energy and AI, 4, 100058, https://doi.org/10.1016/j.egyai.2021.100058, 2021.

Bolin, D. and Wallin, J.: Local scale invariance and robustness of proper scoring rules, Stat. Science, 38, 140–159, https://doi.org/10.1214/22-sts864, 2023.

Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J., and Funk, S.: Scoring epidemiological forecasts on transformed scales, PLOS Comput. Biol., 19, e1011393, https://doi.org/10.1371/journal.pcbi.1011393, 2023.

Brehmer, J.: Elicitability and its Application in Risk Management, arXiv [thesis], https://doi.org/10.48550/ARXIV.1707.09604, 2017.

Brehmer, J. R. and Strokorb, K.: Why scoring functions cannot assess tail properties, Electronic Journal of Statistics, 13, https://doi.org/10.1214/19-ejs1622, 2019.

Bremnes, J. B.: Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bern-

stein Polynomials, Mon. Weather Rev., 148, 403–414, https://doi.org/10.1175/mwr-d-19-0227.1, 2019.

Brier, G. W.: Verification of Forecasts Expressed in Terms of Probability, Mon. Weather Rev., 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2, 1950.

Bröcker, J.: Reliability, sufficiency, and the decomposition of proper scores, Q. J. Roy. Meteor. Soc., 135, 1512–1519, https://doi.org/10.1002/qj.456, 2009.

Bröcker, J. and Ben Bouallègue, Z.: Stratified rank histograms for ensemble forecast verification under serial dependence, Q. J. Roy. Meteor. Soc., 146, 1976–1990, https://doi.org/10.1002/qj.3778, 2020.

Bröcker, J. and Smith, L. A.: Scoring Probabilistic Forecasts: The Importance of Being Proper, Weather Forecast., 22, 382–388, https://doi.org/10.1175/waf966.1, 2007.

Buschow, S.: Measuring Displacement Errors with Complex Wavelets, Weather Forecast., 37, 953–970, https://doi.org/10.1175/waf-d-21-0180.1, 2022.

Buschow, S. and Friederichs, P.: Using wavelets to verify the scale structure of precipitation forecasts, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 13–30, https://doi.org/10.5194/ascmo-6-13-2020, 2020.

Buschow, S. and Friederichs, P.: SAD: Verifying the scale, anisotropy and direction of precipitation forecasts, Q. J. Roy. Meteor. Soc., 147, 1150–1169, https://doi.org/10.1002/qj.3964, 2021.

Buschow, S., Pidstrigach, J., and Friederichs, P.: Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0), Geosci. Model Dev., 12, 3401–3418, https://doi.org/10.5194/gmd-12-3401-2019, 2019.

Casati, B., Dorninger, M., Coelho, C. A. S., Ebert, E. E., Marsigli, C., Mittermaier, M. P., and Gilleland, E.: The 2020 International Verification Methods Workshop Online: Major Outcomes and Way Forward, B. Am. Meteorol. Soc., 103, E899–E910, https://doi.org/10.1175/bams-d-21-0126.1, 2022.

Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.-P., Lerch, S., and Hayatbini, N.: Probabilistic Predictions from Deterministic Atmospheric River Forecasts with Deep Learning, Mon. Weather Rev., 150, 215–234, https://doi.org/10.1175/mwr-d-21-0106.1, 2022.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W.: FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead, arXiv [preprint], https://doi.org/10.48550/ARXIV.2304.02948, 2023.

Christensen, H. M., Moroz, I. M., and Palmer, T. N.: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts, Q. J. Roy. Meteor. Soc., 141, 538–549, https://doi.org/10.1002/qj.2375, 2014.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields, J. Hydrometeorol., 5, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:tssamf>2.0.co;2, 2004.

Dawid, A. P.: Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach, J. R. Stat. Soc. Ser. A-G., 147, 278, https://doi.org/10.2307/2981683, 1984.

Dawid, A. P. and Musio, M.: Theory and applications of proper scoring rules, METRON, 72, 169–183, https://doi.org/10.1007/s40300-014-0039-y, 2014.

Dawid, A. P. and Sebastiani, P.: Coherent dispersion criteria for optimal experimental design, Ann. Stat., 27, 65–81, https://doi.org/10.1214/aos/1018031101, 1999.

Dawid, A. P., Musio, M., and Ventura, L.: Minimum Scoring Rule Inference, Scand. J. Stat., 43, 123–138, https://doi.org/10.1111/sjos.12168, 2015.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K.: Probabilistic Weather Prediction with an Analog Ensemble, Mon. Weather Rev., 141, 3498–3516, https://doi.org/10.1175/mwr-d-12-00281.1, 2013.

Demaeyer, J.: EUPPBench postprocessing benchmark dataset – gridded data – Part I (v1.0), Zenodo [data set], https://doi.org/10.5281/zenodo.7429236, 2022.

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A., Ben Bouallègue, Z., Chen, J., Dabernig, M., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Merše, J., Mlakar, P., Möller, A., Mestre, O., Taillardat, M., and Vannitsem, S.: The EUPPBench postprocessing benchmark dataset v1.0, Earth Syst. Sci. Data, 15, 2635–2653, https://doi.org/10.5194/essd-15-2635-2023, 2023.

Diebold, F. X. and Mariano, R. S.: Comparing Predictive Accuracy, J. Bus. Econ. Stat., 13, 253–263, https://doi.org/10.1080/07350015.1995.10524599, 1995.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The Setup of the MesoVICT Project, B. Am. Meteorol. Soc., 99, 1887–1906, https://doi.org/10.1175/bams-d-17-0164.1, 2018.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Ehm, W. and Gneiting, T.: Local proper scoring rules of order two, Ann. Stat., 40, 609–637, https://doi.org/10.1214/12-aos973, 2012.

EUMETNET: MeteoAlarm, https://www.meteoalarm.org/en/live/, last access: 16 October 2024.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, https://doi.org/10.1002/met.45, 2008.

Friederichs, P. and Hense, A.: A Probabilistic Forecast Approach for Daily Precipitation Totals, Weather Forecast., 23, 659–673, https://doi.org/10.1175/2007waf2007051.1, 2008.

Gilleland, E.: Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases, Weather Forecast., 26, 409–415, https://doi.org/10.1175/waf-d-10-05061.1, 2011.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Weather Forecast., 24, 1416–1430, https://doi.org/10.1175/2009waf2222269.1, 2009.

Gneiting, T.: Making and Evaluating Point Forecasts, J. Am. Stat. Assoc., 106, 746–762, https://doi.org/10.1198/jasa.2011.r10138, 2011.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annu. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, https://doi.org/10.1175/mwr2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness, J. R. Stat. Soc. B, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A.: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, TEST, 17, 211–235, https://doi.org/10.1007/s11749-008-0114-x, 2008.

Gneiting, T., Lerch, S., and Schulz, B.: Probabilistic solar forecasting: Benchmarks, post-processing, verification, Sol. Energy, 252, 72–80, https://doi.org/10.1016/j.solener.2022.12.054, 2023.

Good, I. J.: Rational Decisions, J. Roy. Stat. Soc. B Met., 14, 107–114, https://doi.org/10.1111/j.2517-6161.1952.tb00104.x, 1952.

Han, F. and Szunyogh, I.: A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability, Mon. Weather Rev., 146, 1303–1318, https://doi.org/10.1175/mwr-d-17-0040.1, 2018.

Heinrich-Mertsching, C., Thorarinsdottir, T. L., Guttorp, P., and Schneider, M.: Validation of point process predictions with proper scoring rules, Scand. J. Stat., 51, 1533–1566, https://doi.org/10.1111/sjos.12736, 2024.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Holzmann, H. and Eulert, M.: The role of the information set for forecasting – with applications to risk management, Ann. Appl. Stat., 8, 595–621, https://doi.org/10.1214/13-aoas709, 2014.

Hu, W., Ghazvinian, M., Chapman, W. E., Sengupta, A., Ralph, F. M., and Delle Monache, L.: Deep Learning Forecast Uncertainty for Precipitation over the Western United States, Mon. Weather Rev., 151, 1367–1385, https://doi.org/10.1175/mwr-d-22-0268.1, 2023.

Hyvärinen, A.: Estimation of Non-Normalized Statistical Models by Score Matching, J. Mach. Learn. Res., 6, 695–709, 2005.

Jolliffe, I. T. and Primo, C.: Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic, Mon. Weather Rev., 136, 2133–2139, https://doi.org/10.1175/2007mwr2219.1, 2008.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, J. Stat. Softw., 90, 1–37, https://doi.org/10.18637/jss.v090.i12, 2019.

Jordan, T. H., Chen, Y.-T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., and Zschau, J.: OPERATIONAL EARTHQUAKE FORECASTING. State of Knowledge and Guidelines for Utilization, Ann. Geophys.-Italy, 54, 316–391, https://doi.org/10.4401/ag-5350, 2011.

Jose, V. R.: A Characterization for the Spherical Scoring Rule, Theor. Decis., 66, 263–281, https://doi.org/10.1007/s11238-007-9067-x, 2007.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, arXiv [preprint], https://doi.org/10.48550/ARXIV.2202.07575, 2022.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, Ann. Math. Stat., 22, 79–86, https://doi.org/10.1214/aoms/1177729694, 1951.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: GraphCast: Learning skillful medium-range global weather forecasting, arXiv [preprint], https://doi.org/10.48550/ARXIV.2212.12794, 2022.

Lerch, S. and Polsterer, K. L.: Convolutional autoencoders for spatially-informed ensemble post-processing, in: ICLR 2022 – AI for Earth and Space Science Workshop, https://ai4earthscience.github.io/iclr-2022-workshop/camera_ready/iclr_2022_ai4ess_04.pdf (last access: 6 March 2025), 2022.

Lerch, S. and Thorarinsdottir, T. L.: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting, Tellus A, 65, 21206, https://doi.org/10.3402/tellusa.v65i0.21206, 2013.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster's Dilemma: Extreme Events and Forecast Evaluation, Stat. Sci., 32, 106–127, https://doi.org/10.1214/16-sts588, 2017.

Matheron, G.: Principles of geostatistics, Econ. Geol., 58, 1246–1266, https://doi.org/10.2113/gsecongeo.58.8.1246, 1963.

Matheson, J. E. and Winkler, R. L.: Scoring Rules for Continuous Probability Distributions, Manage. Sci., 22, 1087–1096, 1976.

Meng, X., Taylor, J. W., Ben Taieb, S., and Li, S.: Scores for Multivariate Distributions and Level Sets, Oper. Res., 344–362, https://doi.org/10.1287/opre.2020.0365, 2023.

Murphy, A. H. and Winkler, R. L.: A General Framework for Forecast Verification, Mon. Weather Rev., 115, 1330–1338, https://doi.org/10.1175/1520-0493(1987)115<1330:agfffv>2.0.co;2, 1987.

Nowotarski, J. and Weron, R.: Recent advances in electricity price forecasting: A review of probabilistic forecasting, Renew. Sust. Energ. Rev., 81, 1548–1568, https://doi.org/10.1016/j.rser.2017.05.234, 2018.

Pacchiardi, L., Adewoyin, R., Dueben, P., and Dutta, R.: Probabilistic Forecasting with Generative Networks via Scoring Rule Minimization, J. Mach. Learn. Res., 25, 1–64, 2024.

Palmer, T. N.: Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction, Q. J. Roy. Meteor. Soc., 138, 841–861, https://doi.org/10.1002/qj.1923, 2012.

Parry, M., Dawid, A. P., and Lauritzen, S.: Proper local scoring rules, Ann. Stat., 40, 561–592, https://doi.org/10.1214/12-aos971, 2012.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, arXiv [preprint], https://doi.org/10.48550/arXiv.2202.11214, 2022.

Pic, R.: aggregation-transformation, Zenodo [code], https://doi.org/10.5281/zenodo.14982271, 2024.

Pinson, P.: Wind Energy: Forecasting Challenges for Its Operational Management, Stat. Sci., 28, 564–585, https://doi.org/10.1214/13-sts445, 2013.

Pinson, P. and Girard, R.: Evaluating the quality of scenarios of short-term wind power generation, Appl. Energ., 96, 12–20, https://doi.org/10.1016/j.apenergy.2011.11.004, 2012.

Pinson, P. and Tastu, J.: Discrimination ability of the Energy score, DTU Compute – Technical Report, Technical University of Denmark, https://orbit.dtu.dk/files/56966842/tr13_15_Pinson_Tastu.pdf (last access: 6 March 2025), 2013.

Radanovics, S., Vidal, J.-P., and Sauquet, E.: Spatial Verification of Ensemble Precipitation: An Ensemble Version of SAL, Weather Forecast., 33, 1001–1020, https://doi.org/10.1175/waf-d-17-0162.1, 2018.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Mon. Weather Rev., 146, 3885–3900, https://doi.org/10.1175/mwr-d-18-0187.1, 2018.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallègue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, J. Adv. Model. Earth Sy., 16, e2023MS004019, https://doi.org/10.1029/2023MS004019, 2024.

Rivoire, P., Martius, O., Naveau, P., and Tuel, A.: Assessment of subseasonal-to-seasonal (S2S) ensemble extreme precipitation forecast skill over Europe, Nat. Hazards Earth Syst. Sci., 23, 2857–2871, https://doi.org/10.5194/nhess-23-2857-2023, 2023.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Mon. Weather Rev., 136, 78–97, https://doi.org/10.1175/2007mwr2123.1, 2008.

Roulston, M. S. and Smith, L. A.: Evaluating Probabilistic Forecasts Using Information Theory, Mon. Weather Rev., 130, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:epfuit>2.0.co;2, 2002.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling, Stat. Sci., 28, 616–640, https://doi.org/10.1214/13-sts443, 2013.

Scheuerer, M. and Hamill, T. M.: Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities∗, Mon. Weather Rev., 143, 1321–1334, https://doi.org/10.1175/mwr-d-14-00269.1, 2015.

Scheuerer, M. and Möller, D.: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics, Ann. Appl. Stat., 9, 1328–1349, https://doi.org/10.1214/15-aoas843, 2015.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K.: Analysis, Simulation and Prediction of Multivariate Random Fields with PackageRandomFields, J. Stat. Softw., 63, 1–25, https://doi.org/10.18637/jss.v063.i08, 2015.

Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P. J., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., and Zhuang, J.: The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities, Seismol. Res. Lett., 89, 1305–1313, https://doi.org/10.1785/0220180053, 2018.

Schulz, B. and Lerch, S.: Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison, Mon. Weather Rev., 150, 235–257, https://doi.org/10.1175/mwr-d-21-0150.1, 2022.

Shannon, C. E.: A Mathematical Theory of Communication, Bell Syst. Tech. J., 27, 623–656, https://doi.org/10.1002/j.1538-7305.1948.tb00917.x, 1948.

Smola, A., Gretton, A., Song, L., and Schölkopf, B.: A Hilbert Space Embedding for Distributions, in: Algorithmic Learning Theory, edited by: Hutter, M., Servedio, R. A., and Takimoto, E., 13–31, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-75225-7, 2007.

Stein, J. and Stoop, F.: Neighborhood-Based Ensemble Evaluation Using the CRPS, Mon. Weather Rev., 150, 1901–1914, https://doi.org/10.1175/mwr-d-21-0224.1, 2022.

Steinwart, I. and Christmann, A.: Support Vector Machines, Information Science and Statistics, Springer, New York, ISBN 978-0-387-77241-7, 2008.

Steinwart, I. and Ziegel, J. F.: Strictly proper kernel scores and characteristic kernels on compact spaces, Appl. Comput. Harmon. A., 51, 510–542, https://doi.org/10.1016/j.acha.2019.11.005, 2021.

Székely, G.: E-statistics: The Energy of Statistical Samples, techreport, Bowling Green State University, https://doi.org/10.13140/RG.2.1.5063.9761, 2003.

Taillardat, M.: Skewed and Mixture of Gaussian Distributions for Ensemble Postprocessing, Atmosphere, 12, 966, https://doi.org/10.3390/atmos12080966, 2021.

Taillardat, M. and Mestre, O.: From research to applications – examples of operational ensemble post-processing in France using machine learning, Nonlin. Processes Geophys., 27, 329–347, https://doi.org/10.5194/npg-27-329-2020, 2020.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, Mon. Weather Rev., 144, 2375–2393, https://doi.org/10.1175/mwr-d-15-0260.1, 2016.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Workshop on Predictability, 20–22 October 1997, 1–26, ECMWF, Shinfield Park, Reading, 1997.

Thorarinsdottir, T. L. and Schuhen, N.: Verification: Assessment of Calibration and Accuracy, 155–186, Elsevier, https://doi.org/10.1016/b978-0-12-812372-0.00006-6, 2018.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N.: Using Proper Divergence Functions to Evaluate Climate Models, SIAM/ASA Journal on Uncertainty Quantification, 1, 522–534, https://doi.org/10.1137/130907550, 2013.

Tsyplakov, A.: Evaluating Density Forecasts: A Comment, SSRN Electronic Journal, 1907799, https://doi.org/10.2139/ssrn.1907799, 2011.

Tsyplakov, A.: Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments, SSRN Electronic Journal, 2236605, https://doi.org/10.2139/ssrn.2236605, 2013.

Tsyplakov, A.: Evaluation of probabilistic forecasts: Conditional auto-calibration, https://www.sas.upenn.edu/~fdiebold/papers2/Tsyplakov_Auto_calibration_sent_eswc2020.pdf (last access: 6 March 2025), 2020.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., and Ylhaisi, J.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, B. Am. Meteorol. Soc., 102, E681–E699, https://doi.org/10.1175/bams-d-19-0308.1, 2021.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Weather Rev., 136, 4470–4487, https://doi.org/10.1175/2008mwr2415.1, 2008.

Winkelbauer, A.: Moments and Absolute Moments of the Normal Distribution, arXiv [preprint], https://doi.org/10.48550/ARXIV.1209.4340, 2014.

Winkler, R. L.: Rewarding Expertise in Probability Assessment, 127–140, Springer Netherlands, ISBN 9789401012768, https://doi.org/10.1007/978-94-010-1276-8_10, 1977.

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D.: Scoring rules and the evaluation of probabilities, Test, 5, 1–60, https://doi.org/10.1007/bf02562681, 1996.

Zamo, M. and Naveau, P.: Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts, Math. Geosci., 50, 209–234, https://doi.org/10.1007/s11004-017-9709-7, 2017.

Ziel, F. and Berk, K.: Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules, arXiv [preprint], https://doi.org/10.48550/arXiv.1910.07325, 2019.