Forecast score distributions with imperfect observations

Julie Bessac, Philippe Naveau

Adv. Stat. Clim. Meteorol. Oceanogr., 7, 53–71, 2021
https://doi.org/10.5194/ascmo-7-53-2021

ASCMO
Open Access

Forecast score distributions with imperfect observations

Julie Bessac[1] **and Philippe Naveau**[2]

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA
[2]Laboratoire de Sciences du Climat et de l'Environnement, IPSL-CNRS, Gif-sur-Yvette, 91191, France

Correspondence: Julie Bessac (jbessac@anl.gov)

Received: 15 February 2021 – Revised: 12 July 2021 – Accepted: 28 July 2021 – Published: 23 September 2021

Abstract. The field of statistics has become one of the mathematical foundations in forecast evaluation studies, especially with regard to computing scoring rules. The classical paradigm of scoring rules is to discriminate between two different forecasts by comparing them with observations. The probability distribution of the observed record is assumed to be perfect as a verification benchmark. In practice, however, observations are almost always tainted by errors and uncertainties. These may be due to homogenization problems, instrumental deficiencies, the need for indirect reconstructions from other sources (e.g., radar data), model errors in gridded products like reanalysis, or any other data-recording issues. If the yardstick used to compare forecasts is imprecise, one can wonder whether such types of errors may or may not have a strong influence on decisions based on classical scoring rules. We propose a new scoring rule scheme in the context of models that incorporate errors of the verification data. We rely on existing scoring rules and incorporate uncertainty and error of the verification data through a hidden variable and the conditional expectation of scores when they are viewed as a random variable. The proposed scoring framework is applied to standard setups, mainly an additive Gaussian noise model and a multiplicative Gamma noise model. These classical examples provide known and tractable conditional distributions and, consequently, allow us to interpret explicit expressions of our score. By considering scores to be random variables, one can access the entire range of their distribution. In particular, we illustrate that the commonly used mean score can be a misleading representative of the distribution when the latter is highly skewed or has heavy tails. In a simulation study, through the power of a statistical test, we demonstrate the ability of the newly proposed score to better discriminate between forecasts when verification data are subject to uncertainty compared with the scores used in practice. We apply the benefit of accounting for the uncertainty of the verification data in the scoring procedure on a dataset of surface wind speed from measurements and numerical model outputs. Finally, we open some discussions on the use of this proposed scoring framework for non-explicit conditional distributions.

1 Introduction

Probabilistic forecast evaluation generally involves the comparison of a probabilistic forecast cumulative distribution function (cdf) F (in the following, we assume that F admits a probability density function f) with verification data y that could be of diverse origins (Jolliffe and Stephenson, 2004; Gneiting et al., 2007). In this context verification data are, most of the time, but not exclusively, observational data. Questions related to the quality and variability of observational data have been raised and tackled in different scientific contexts. For instance, data assimilation requires careful estimation of uncertainties related to both numerical models and observations (Daley, 1993; Waller et al., 2014; Janjić et al., 2017). Apart from a few studies (see, e.g., Hamill and Juras, 2006; Ferro, 2017), error and uncertainty associated with the verification data have rarely been addressed in forecast evaluation. Nonetheless, errors in verification data can

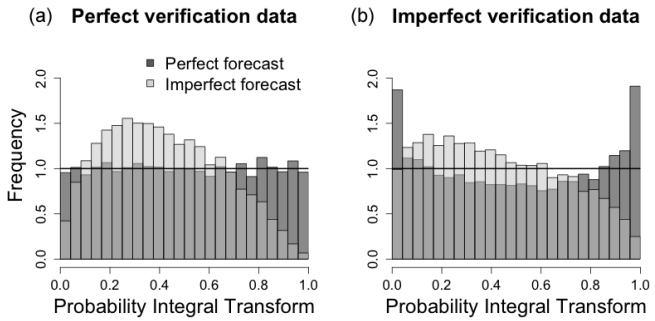

Figure 1. "Perfect" (dark grey) and "imperfect" (light grey) synthetic forecasts, with respective distributions $\mathcal{N}(0,2)$ and $\mathcal{N}(0.5,4)$, are compared with "perfect" verification data **(a)** with distribution $\mathcal{N}(0,2)$ and imperfect verification data Y and **(b)** with distribution $\mathcal{N}(0,3)$. Comparing a perfect forecast (forecast with the same distribution as the true data) to corrupted verification data leads to an apparent underdispersion of the perfect forecast.

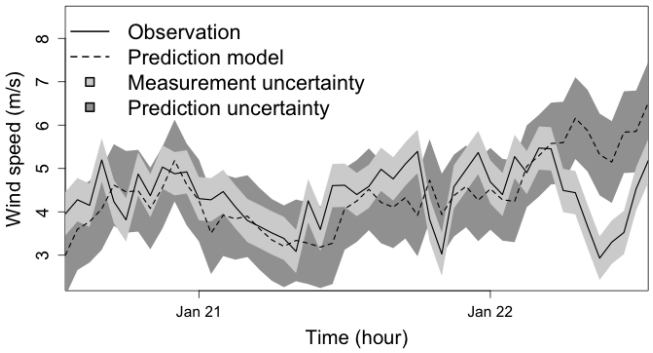

Figure 2. Time series of surface wind speed measurements of two days of January 2012 in Wisconsin, USA. The solid line represents ground measurements and the light grey shaded area represents observational uncertainty ($\pm\sigma$, with $\sigma = 0.5$ defined in Pinson and Hagedorn, 2012). The dashed line represents a numerical model output and the model uncertainty in dark grey shade. The model uncertainty refers to the standard deviation computed on the available forecast time series.

lead to severe mischaracterization of probabilistic forecasts, as illustrated in Fig. 1, where a perfect forecast (a forecast with the same distribution as the true data) appears underdispersed when the verification data have a larger variance than the true process. In the following, the underlying hidden process that is materialized by model simulations or measured through instruments will be referred to as the true process. Forecast evaluation can be performed qualitatively through visual inspection of statistics of the data such as in Fig. 1 (see, e.g., Bröcker and Ben Bouallègue, 2020) and quantitatively through the use of scalar functions called scores (Gneiting et al., 2007; Gneiting and Raftery, 2007). In this work, we focus on the latter, and we illustrate and propose a framework based on hidden variables to embed imperfect verification data in scoring functions via priors on the verification data and on the hidden state.

1.1 Motivating example

To illustrate the proposed work, we consider surface wind data from a previous work (Bessac et al., 2018). Time series of ground measurements from the NOAA Automated Surface Observing System (ASOS) network are available at ftp://ftp.ncdc.noaa.gov/pub/data/asos-onemin, last access: 8 September 2021 and are extracted at 1 min resolution. We focus on January 2012 in this study; the data are filtered via a moving-average procedure and considered at an hourly level, leading to 744 data points. These ground-station data are considered verification data in the following. Outputs from numerical weather prediction (NWP) forecasts are generated by using WRF v3.6 (Skamarock et al., 2008), a state-of-the-art numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs. The NWP forecasts are initialized by using the North American Regional Reanalysis fields dataset that covers the North American continent with a resolution of 10 min of a degree, 29 pressure levels (1000–100 hPa, excluding the surface), every 3 h from the year 1979 until the present. Simulations are started every day during January 2012 with a forecast lead time of 24 h and cover the continental United States on a grid of 25×25 km with a time resolution of 10 min. Only one trajectory is simulated in this study. As observed in Fig. 2, the uncertainty associated with observation data can affect the evaluation of the forecast. As an extension, if two forecasts were to fall within the uncertainty range of the observations, it would require a non-trivial choice from the forecaster to rank forecasts.

We will apply our proposed scoring framework to this dataset in Sect. 5.2 and compute scores embedding this uncertainty.

1.2 Imperfect verification data and underlying physical state

In verification and recalibration of ensemble forecasts, an essential verification step is to find data that precisely identify the forecast events of interest, the so-called verification dataset (see, e.g., Jolliffe and Stephenson, 2004). Jolliffe and Stephenson (2004), in Sect. 1.3 of their book, discussed the uncertainty associated with verification data, such as sampling uncertainty, direct measurement uncertainty, or changes in locations in the verification dataset. Verification data arise from various sources and consequently present various types of errors and uncertainties, such as measurement error (Dirkson et al., 2019), design, and quality: e.g., gridded data versus weather stations, homogenization problems, subsampling variability (Mittermaier and Stephenson, 2015), or mismatching resolutions. Resolution mismatching is becoming an issue due to weather or regional climate models running at ever finer resolution with increased fidelity for which

observational data are rarely available for evaluation, hence requiring us to account for up- or down-scaling error. In order to provide accurate forecast evaluations, it is crucial to distinguish and account for these types of errors. For instance, the effect of observational error can be stronger at a short-term horizon when forecast error is smaller. Additive models have been used for the observational error (Ciach and Krajewski, 1999; Saetra et al., 2004) or used to enhance coarser resolutions to allow comparison with fine-resolution data (Mittermaier and Stephenson, 2015). In the following, we will present an additive and a multiplicative framework based on a hidden variable to embed uncertainty in data sources.

A common way of modeling errors is by representing the truth as a hidden underlying process also called the state (Kalman, 1960; Kalman and Bucy, 1961). Subsequently each source of data is seen as a proxy of the hidden state and modeled as a function of it. This forms the basis of data assimilation models where the desired state of the atmosphere is estimated through the knowledge of physics-based models and observational data that are both seen as versions of the non-observed true physical state. In the following, we base our scoring framework on the decomposition of the verification data as a function of a hidden true state, referred to as X, and an error term.

1.3 Related literature

Uncertainty and errors arise from various sources, such as the forecast and/or the verification data (predictability quality, uncertainty, errors, dependence and correlation, time-varying distribution), and the computational approximation of scores. A distribution-based verification approach was initially proposed by Murphy and Winkler (1987), where joint distributions for forecast and observation account for the information and interaction of both datasets. Wilks (2010) studied the effect of serial correlation of forecasts and observations on the sampling distributions of Brier score and, in particular, serially correlated forecasts inflate its variance. Bolin and Wallin (2019) discussed the misleading use of average scores, in particular for the continuous ranked probability score (CRPS) that is shown to be scale-dependent, when forecasts have varying predictability such as in non-stationary cases or exhibit outliers. Bröcker and Smith (2007), in their conclusion, highlighted the need to generalize scores when verification data are uncertain in the context of properness and locality. More specifically, robustness and scale-invariance corrections of the CRPS are proposed to account for outliers and varying predictability in scoring schemes. Concerning the impact of the forecast density approximation from finite ensembles, Zamo and Naveau (2018) compared four CRPS estimators and highlighted recommendations in terms of the type of ensemble, whether random or a set of quantiles. In addition, several works focus on embedding verification data errors and uncertainties in scoring setups. Some methods aimed at correcting the verification data and using regular scoring metrics, such as perturbed ensemble methods (Anderson, 1996; Hamill, 2001; Bowler, 2008; Gorgas and Dorninger, 2012). Other works modeled observations as random variables and expressed scoring metrics in that context (Candille and Talagrand, 2008; Pappenberger et al., 2009; Pinson and Hagedorn, 2012). Some approaches directly modified the expression of metrics (Hamill and Juras, 2006; Ferro, 2017), and others (see, e.g., Hamill and Juras, 2006) accounted for varying climatology by sub-dividing the sample into stationary ones. In Ciach and Krajewski (1999) and Saetra et al. (2004), additive errors were embedded in scores via convolution. Analogously to the Brier score decomposition of Murphy (1973), Weijs and Van De Giesen (2011) and Weijs et al. (2010) decomposed the Kullback–Leibler divergence score and the cross-entropy score into uncertainty into reliability and resolution components in order to account for uncertain verification data. Kleen (2019) discussed score-scale sensitivity to additive measurement errors and proposed a measure of discrepancy between scores computed on uncorrupted and corrupted verification data.

1.4 Proposed scoring framework

The following paper proposes an idealized framework to express commonly used scores with observational uncertainty and errors. The new framework relies on the decomposition of the verification data y into a "true" hidden state x and an error term and on the representation of scores as a random variable when the verification data are seen as a random variable. Information about the hidden state and the verification data is embedded via priors in the scoring framework, sharing analogies with classical Bayesian analysis (Gelman et al., 2013). More precisely, the proposed framework relies on the knowledge about the conditional distribution of the "true" hidden state given the verification data or about information which allows us to calculate that conditional distribution.

Section 2 introduces a scoring framework that accounts for errors in the verification data for scores used in practice. Sections 3 and 4, respectively, implement the additive and multiplicative error-embedding cases for the logarithmic score (log score) and CRPS. Section 5 illustrates the merits of scores in a simulation context and in a real application case. Finally, Sect. 6 provides a final discussion and insights into future works. In particular, we discuss the possible generalization of the proposed scoring framework when the decomposition of the verification data y into a hidden state x and an error does not fall into an additive or multiplicative setting as in Sects. 3 and 4.

2 Scoring rules under uncertainty

In the following, we propose a version of scores based on the conditional expectation of what is defined as an "ideal" score given the verification data tainted by errors, when scores are viewed as random variables. The idea of conditional expec-

tation was also used in Ferro (2017) but with a different conditioning and is discussed below as a comparison.

2.1 Scores as random variables

In order to build the proposed score version as well as to interpret scores further than through their mean and assess for their uncertainty, we will rely on scores represented as random variables. In practice, scores are averaged over all available verification data y, and the uncertainty associated with this averaged score is mostly neglected. However, this uncertainty is revealed to be significant, as pointed out in Dirkson et al. (2019), where confidence intervals were computed through bootstrapping. In order to assess the probability distribution of score s_0, we assume Y is a random variable representing the verification data y and write a score as a random variable $s_0(F, Y)$, where F is the forecast cdf to be evaluated. This representation gives access to the score distribution and enables us to explore the uncertainty of the latter. A few works in the literature have already considered scores to be random variables and performed the associated analysis. In Diebold and Mariano (2002) and Jolliffe (2007), score distributions were considered to build confidence intervals and hypothesis testing to assess differences in score values when comparing forecasts to the climatology. In Wilks (2010), the effect of serial correlation on the sampling distributions of Brier scores was investigated. Pinson and Hagedorn (2012) illustrated and discussed score distributions across different prediction horizons and across different levels of observational noise.

2.2 Hidden state and scoring

Scores used in practice are evaluated on verification data y, $s_0(F, y)$. We define the "ideal" score as $s_0(F, x)$, where x is the realization of the hidden underlying "true" state that gives rise to the verification data y. Ideally, one would use x to provide the best assessment of forecast F quality through $s_0(F, x)$; however, since x is not available, we consider the best approximation of $s_0(F, x)$ given the observation y in terms of the L^2 norm via the conditional expectation. For a given score s_0, we propose the following score version $s_\vee(., y)$:

$$s_\vee(F, y) = \mathbb{E}(s_0(F, X)|Y = y), \tag{1}$$

where X is the true hidden state, Y is the random variable representing the available observation used as verification data, and F is the forecast cdf to be evaluated. One can view this scoring framework incorporating information about the discrepancy between the true state and the verification data in terms of errors and uncertainty in a Bayesian setting. To compute Eq. (1), we assume that the distributional features of the law of X and the conditional law of $[Y|X]$, where $[.]$ and $[.|.]$ denote, respectively, marginal and conditional distributions, are directly available or can be obtained. This is the classical framework used in data assimilation when both observational and state equations are given, and the issue is to infer the realization x given observations y; see also our example in Sect. 5.2. Under this setup, the following properties hold for the score $s_\vee$:

$$\begin{aligned}&\mathbb{E}_X\left[s_0(F, X)\right] = \mathbb{E}_Y\left[s_\vee(F, Y)\right],\\ &\mathbb{V}_X\left[s_0(F, X)\right] \geq \mathbb{V}_Y\left[s_\vee(F, Y)\right], \text{ for any forecast cdf } F.\end{aligned} \tag{2}$$

Details of the computations are found in Appendix A. The first equality guarantees that any propriety attached to the mean value of s_0 is preserved with $s_\vee$. The second inequality arises from the law of total variance and implies a reduced dispersion of the corrected score compared to the ideal score. This can be explained by the prior knowledge on the verification data that reduces uncertainty in the score. These properties are illustrated with simulated examples and real data in Sect. 5.

As a comparison, Ferro (2017) proposed a mathematical scheme to correct a score when error is present in the verification data y. Ferro's modified score, denoted $s_F(f, y)$ in this work, is derived from a classical score, say $s_0(f, x)$, where x is the hidden true state. With these notations, the corrected score $s_F(f, y)$ is built such that it respects the following conditional expectation:

$$s_0(f, x) = \mathbb{E}(s_F(f, Y)|X = x).$$

In other words, the score $s_F(f, y)$ computed from the ys provides the same value on average as the proper score computed from the unobserved true state xs. The modified score s_F explicitly targets biases in the mean induced by imperfect verification data. In terms of assumptions, we note that the conditional law of Y given X needs to be known in order to compute $s_0(f, x)$ from $s_F(f, y)$; e.g., see Definitions 2 and 3 in Ferro (2017). In particular, in Ferro (2017) the correction framework is illustrated with the logarithmic score in the case of a Gaussian error model, i.e., $[Y|X = x] \sim \mathcal{N}(x, \omega^2)$.

Implementation and generalization

The proposed score reveals desirable mathematical properties of unbiasedness and variance reduction while accounting for the error in the verification data; however, it relies on the knowledge of $[X|Y]$ or equivalently of $[Y|X]$ and $[X]$. We assume that the decomposition of Y into a hidden state X and an error is given and is fixed with respect of the evaluation framework. Additionally, the new framework relies on the necessity to integrate $s_0(f, x)$ against conditional distributions, which might require some quadrature approximations in practice when closed formulations are not available. In this work, we assume we have access to the climatology distribution, and we rely on this assumption to compute the distribution of X or priors of its distribution. However, depending on the context and as exemplified in Sect. 5.2, alternative definitions and computations of X can be considered,

as for instance relying on known measurement error models. Nonetheless, as illustrated in Sects. 3 and 4, the simplifying key to the score derivation in Eq. (1) is to use Bayesian conjugacy when applicable. This is illustrated in the following sections with a Gaussian additive case and a Gamma multiplicative case. Although the Bayesian conjugacy is a simplifying assumption, as in most Bayesian settings it is not a necessary condition, and for cases with non-explicit conjugate priors, all Bayesian and/or assimilation tools (e.g., via sampling algorithms such as Markov chain Monte Carlo methods (Robert and Casella, 2013) or non-parametric approaches such as Dirichlet processes) could be used to draw samples from $[X|Y]$ and estimate the distribution $[s_0(F,X)|Y=y]$ for a given $s_0(F,.)$. Finally, in the following we assume that distributions have known parameters; however, this assumption can be relaxed by considering prior distributions on each involved parameter via hierarchical Bayesian modeling. The scope of this paper is to provide a conceptual tool, and the question of parameter estimation, besides the example in Sect. 5, is not treated in detail. In Sect. 5, we apply the proposed score derivation to the aforementioned surface wind data example described in the introduction. In Sect. 6, we discuss the challenges of computing scores as defined in Eq. (1) or in Ferro (2017) in more general cases of state-space models that are not limited to additive or multiplicative cases.

3 Gaussian additive case

As discussed earlier, a commonly used setup in applications is when errors are additive and their natural companion distribution is Gaussian. Hereafter, we derive the score from Eq. (1) for the commonly used log score and CRPS in the Gaussian additive case, where the hidden state X and the verification data Y are linked through the following system:

$$\text{Model (A)} \begin{cases} X \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), \\ Y = X + \mathcal{N}\left(0, \omega^2\right), \end{cases}$$

where Y is the observed verification data, X is the hidden true state, and all Gaussian variables are assumed to be independent. In the following, for simplicity we consider $\mathbb{E}(X) = \mathbb{E}(Y)$; however, one can update the model easily to mean-biased verification data Y. Parameters are supposed to be known from the applications; one could use priors on the parameters when estimates are not available. In the following, we express different versions of the log score and CRPS: the ideal version, the used-in-practice version, and the error-embedding version from Eq. (1). In this case, as well as in Sect. 4, since conditional distributions are expressed through Bayesian conjugacy, most computational efforts rely on integrating the scores against the conditional distributions.

3.1 Log-score versions

For a Gaussian predictive probability distribution function (pdf) f with mean μ and variance σ^2, the log score is defined by

$$s_0(f,x) = \log\sigma + \frac{1}{2\sigma^2}(x-\mu)^2 + \frac{1}{2}\log 2\pi \tag{3}$$

and has been widely used in the literature. Ideally, one would access the true state X and evaluate forecast against its realizations x; however, since X is not accessible, scores are computed against observations y:

$$s_0(f,y) = \log\sigma + \frac{1}{2\sigma^2}(y-\mu)^2 + \frac{1}{2}\log 2\pi. \tag{4}$$

Applying basic properties of Gaussian conditioning, our score defined by Eq. (1) can be written as

$$\begin{aligned} s_\vee(f,y) = &\log\sigma + \frac{1}{2\sigma^2}\left\{\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}\right. \\ &\left. + (\overline{y}-\mu)^2\right\} + \frac{1}{2}\log 2\pi, \\ &\text{where } \overline{y} = \frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}y. \end{aligned} \tag{5}$$

In particular, $\overline{y} = \mathbb{E}(X|Y=y)$ arises from the conditional expectation of $s_\vee(f,Y)$ in Eq. (3) and is a weighted sum that updates the prior information about $X \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with the observation $Y \sim \mathcal{N}(\mu_0, \sigma_0^2+\omega^2)$. In the following equations, $\overline{y}$ represents the same quantity. Details of the computations are found in Appendix B and rely on the integration of Eq. (3) against the conditional distribution of $[X|Y=y]$.

As a comparison, the corrected score from Ferro (2017) is expressed as $s_F(f,y) = \log\sigma + \frac{(y-\mu)^2-\omega^2}{2\sigma^2} + \frac{1}{2}\log 2\pi$, with the same notations and under the assumption $[Y|X=x] \sim \mathcal{N}(x,\omega^2)$. In this case, the distribution of the hidden state x is not involved; however, y is not scale-corrected, and only a location correction is applied compared to Eq. (5).

3.2 CRPS versions

Besides the logarithmic score, the CRPS is another classical scoring rule used in weather forecast centers. It is defined as

$$c_0(f,x) = \mathbb{E}|Z-x| - \frac{1}{2}\mathbb{E}|Z-Z'|, \tag{6}$$

where Z and Z' are independent and identically distributed copy random variables with a continuous pdf f. The CRPS can be rewritten as $c_0(f,x) = x + 2\mathbb{E}(Z-x)_+ - 2\mathbb{E}(Z\overline{F}(Z))$, where $(Z-x)_+$ represents the positive part of $Z-x$ and $\overline{F}(x) = 1-F(x)$ corresponds to the survival function associated with the cdf F. For example, the CRPS for a Gaussian forecast with parameters μ and σ is equal to

$$\begin{aligned} c_0(f,x) = x + 2\sigma &\left[\phi\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\overline{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right] \\ &- \left[\mu + \frac{\sigma}{\sqrt{\pi}}\right], \end{aligned} \tag{7}$$

where ϕ and Φ are the pdf and cdf of a standard normal distribution (Gneiting et al., 2005; Taillardat et al., 2016). Similarly to Eq. (4), in practice one evaluates Eq. (7) against observations y since the hidden state X is unobserved.

Under the Gaussian additive Model (A), the proposed CRPS defined by Eq. (1) is written as

$$c_\vee(f,y) = \overline{y} + 2\sigma_\omega\left[\phi\left(\frac{\overline{y}-\mu}{\sigma_\omega}\right) - \frac{\overline{y}-\mu}{\sigma_\omega}\overline{\Phi}\left(\frac{\overline{y}-\mu}{\sigma_\omega}\right)\right] - \left[\mu + \frac{\sigma}{\sqrt{\pi}}\right], \tag{8}$$

where $\sigma_\omega^2 = \sigma^2 + \frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}$ and $\overline{y} = \frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}$ as defined above. Details of the computations are found in Appendix C and rely on similar principles to the above paragraph.

3.3 Score distributions

Under the Gaussian additive Model (A), the random variables associated with the proposed log scores defined by Eq. (4) and (5) are written as

$$s_0(f,Y) \overset{d}{=} a_0 + b_0\chi_0^2 \quad \text{and} \quad s_\vee(f,Y) \overset{d}{=} a_\vee + b_\vee\chi_\vee^2, \tag{9}$$

where $\overset{d}{=}$ means equality in distribution and χ_0^2 and $\chi_\vee^2$ non-central chi-squared random variables with 1 DOF (degree of freedom) and respective non-centrality parameters λ_0 and $\lambda_\vee$. The explicit expressions of the constants λ_0, $\lambda_\vee$, a_0, $a_\vee$, b_0, and $b_\vee$ are found in Appendix D. The distribution of $s_0(f,X)$ can be derived similarly. Figure 3 illustrates the distributions in Eq. (9) for various values of the noise parameter ω. The distributions are very peaked due to the single degree of freedom of the chi-squared distribution; moreover, their bulks are far from the true mean of the ideal score of $s_0(.,X)$, challenging the use of the mean score to compare forecasts. The concept of propriety is based on averaging scores; however, the asymmetry and long right tails of the non-central chi-squared densities make the mean a non-reliable statistic to represent such distributions. Bolin and Wallin (2019) discussed the misleading use of averaged scores in the context of time-varying predictability where different scales of prediction errors arise, generating different orders of magnitude of evaluated scores. However, the newly proposed scores exhibit a bulk of their distribution closer to the mean and with a reduced variance as stated in Eq. (2), leading to more confidence in the mean when the latter is considered.

Similarly, under the additive Gaussian Model (A), the random variable associated with the proposed CRPS defined by Eq. (8) is written as

$$c_\vee(f,Y) = \overline{Y} + 2\sigma_\omega\left[\phi\left(\frac{\overline{Y}-\mu}{\sigma_\omega}\right) - \frac{\overline{Y}-\mu}{\sigma_\omega}\overline{\Phi}\left(\frac{\overline{Y}-\mu}{\sigma_\omega}\right)\right] - \left[\mu + \frac{\sigma}{\sqrt{\pi}}\right], \tag{10}$$

where $\sigma_\omega^2 = \sigma^2 + \frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}$ and the random variable $\overline{Y} = \frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}Y$ follow a Gaussian pdf with mean μ_0 and variance $\sigma_0^2 \times \frac{\sigma_0^2}{\sigma_0^2+\omega^2}$. The distribution of Eq. (10) does not belong to any known parametric families; however, it is still possible to characterize the score distribution through sampling when the distribution of Y is available. Finally, having access to distributions like Eq. (9) or Eq. (10) gives access to the whole range of uncertainty of the score distributions, helping to derive statistics that are more representative than the mean as pointed out above and to compute confidence intervals without bootstrapping approximations as in Wilks (2010) and Dirkson et al. (2019). Finding adequate representatives of a score distribution that shows reliable discriminative skills is beyond the scope of this work. Nevertheless, in Appendix G we take forward the concept of score distributions and apply it to computing distances between score distributions in order to assess their discriminative skills.

4 Multiplicative Gamma case

The Gaussian assumption is appropriate when dealing with averages, for example, mean temperatures; however, the normal hypothesis cannot be justified for positive and skewed variables such as precipitation intensities. For instance, multiplicative models are often used in hydrology to account for various errors and uncertainty (measurement errors, process uncertainty, and unknown quantities); see Kavetski et al. (2006a, b) and McMillan et al. (2011). An often-used alternative in such cases is to use a Gamma distribution, which works fairly well in practice to represent the bulk of rainfall intensities. Hence, we assume in this section that the true but unobserved X follows a Gamma distribution with parameters α_0 and β_0: $f_X(x) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}x^{\alpha_0-1}\exp(-\beta_0 x)$, for $x > 0$. For positive random variables such as precipitation, additive models cannot be used to introduce errors. Instead, we prefer to use a multiplicative model of the type

$$\text{Model (B)} \begin{cases} X \sim \Gamma(\alpha_0, \beta_0), \\ Y = X \times \epsilon, \end{cases} \tag{11}$$

where ϵ is a positive random variable independent of X. To make feasible computations, we model the error ϵ as an inverse Gamma pdf with parameters a and b: $f_\epsilon(u) = \frac{b^a}{\Gamma(a)}u^{-a-1}\exp\left(-\frac{b}{u}\right)$, for $u > 0$. The basic conjugate prior properties of such Gamma and inverse Gamma distributions allows us to easily derive the pdf $[X|Y=y]$. Analogously to Sect. 3, we express the log score and CRPS within this multiplicative Gamma model in the following paragraphs.

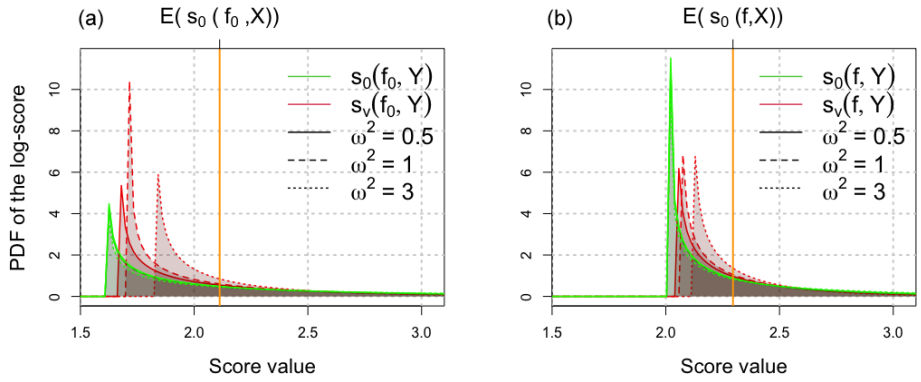

Figure 3. Probability distribution functions (pdfs) from Eq. (9) of the log score used in practice $s_0(.,Y)$ (green) and the proposed score $s_\vee(.,Y)$ (red) computed for the perfect forecast f_0 on the left ($\mu = \mu_0 = 0$ and $\sigma = \sigma_0 = 2$) and imperfect forecasts f on the right ($\mu = 1$ and $\sigma = 3$). The mean of the ideal score is depicted with an orange line: **(a)** $\mathbb{E}(s_0(f_0, X))$ and **(b)** $\mathbb{E}(s_0(f, X))$ on the right. The following distributions are used: $X \sim \mathcal{N}(0, 4)$ and $Y \sim \mathcal{N}(0, 4+\omega^2)$ with several levels of observational noise $\omega^2 = 0.5$, 1, and 3.

4.1 Log-score versions

Let us consider a Gamma distribution f with parameters $\alpha > 0$ and $\beta > 0$ for the prediction. With obvious notations, the log score for this forecast f becomes

$$s_0(f,x) = (1-\alpha)\log x + \beta x - \alpha\log\beta + \log\Gamma(\alpha). \tag{12}$$

Under the Gamma multiplicative model (B), the random variable associated with the corrected log scores defined by Eqs. (1) and (12) is expressed as

$$s_\vee(f,Y) = (1-\alpha)(\psi(\alpha_0+a) - \log(\beta_0 + b/Y)) + \beta\frac{\alpha_0+a}{\beta_0+b/Y} - \alpha\log\beta + \log\Gamma(\alpha), \tag{13}$$

where $\psi(x)$ represents the digamma function defined as the logarithmic derivative of the Gamma function, namely, $\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$. Details of the computations are found in Appendix E.

4.2 CRPS versions

For a Gamma forecast with parameters α and β, the corresponding CRPS (see, e.g., Taillardat et al., 2016; Scheuerer and Möller, 2015) is equal to

$$c_0(f,x) = \left[\frac{\alpha}{\beta} - \frac{1}{\beta B(.5,\alpha)}\right] - x + 2\left[\frac{x}{\beta}f(x) + \left(\frac{\alpha}{\beta} - x\right)\overline{F}(x)\right]. \tag{14}$$

Under the multiplicative Gamma model (B), the random variable associated with the CRPS Eq. (14) corrected by Eq. (1) is expressed as

$$c_\vee(f,Y) = \left[\frac{\alpha}{\beta} - \frac{1}{\beta B(5,\alpha)}\right] - \frac{\alpha_0+a}{\beta_0+\frac{b}{Y}} + 2\frac{\beta^{\alpha-1}(\beta_0+b/Y)^{\alpha_0+a}}{B(\alpha,\alpha_0+a)(\beta+\beta_0+b/Y)^{\alpha+\alpha_0+a}} + \frac{2(\beta_0+b/Y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}\int_0^{+\infty}\left(\frac{\alpha}{\beta}-x\right)\Gamma(\alpha,\beta x)x^{\alpha_0+a-1}\exp(-(\beta_0+b/Y)x)\,\mathrm{d}x. \tag{15}$$

Details of the computations are found in Appendix F. Similarly to Sect. 3.3, one can access the range of uncertainty of the proposed scores, Eqs. (13) and (15), when sampling from the distribution of Y is available. As an illustration, Fig. 4 shows the distributions of the three CRPSs presented in this section. Similarly to the previous section, one can see that the benefits of embedding the uncertainty of the verification data are noticeable in the variance reduction of distributions shaded in red and the smaller distance between the bulk of distributions in red and the mean value of the ideal score.

5 Applications and illustrations

The following section applies and illustrates through simulation studies the benefit of accounting for uncertainty in scores as presented in Sect. 2 through the power analysis of a hypothesis test and the numerical illustration of the motivating example with wind data from Sect. 1. In Appendix G, we illustrate further the consideration of score distributions via the Wasserstein distance.

5.1 Power analysis of hypothesis testing

In this section, a power analysis of the following hypothesis test is performed on simulated data following Diebold

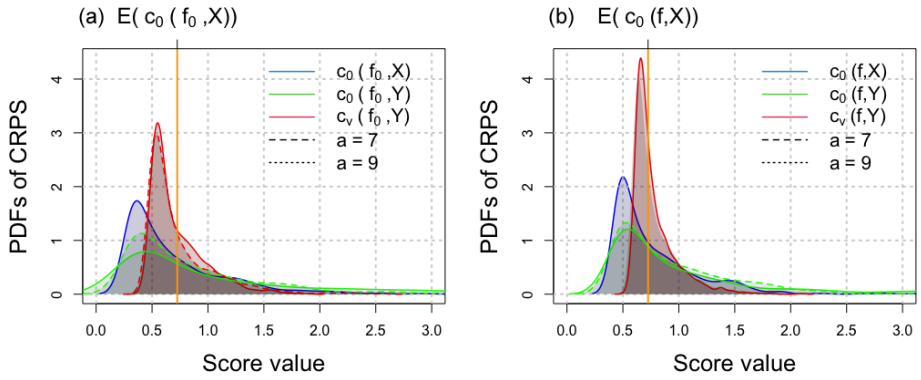

Figure 4. Estimated probability distribution of the CRPS under the multiplicative Gamma model; shown in blue is the distribution of $c_0(., X)$, shown in green is the distribution of $s_0(., Y)$, and shown in red is the distribution of $c_\vee(., Y)$, respectively, from Eqs. (14) and (15). The mean of the ideal score $\mathbb{E}(c_0(., X))$ is depicted with an orange line. **(a)** Score distributions for the perfect forecast f_0 with the same distribution as X ($\alpha = \alpha_0 = 7$ and $\beta = \beta_0 = 2$) validated against X and corrupted verification data Y with different levels of error: $a = 7$ and 9 and $b = 8$. **(b)** Score distributions for the imperfect forecast f with distribution parameters $\alpha = 4$ and $\beta = 1$ validated against X and corrupted verification data Y. The following parameters $\alpha_0 = 7$ and $\beta_0 = 2$ are used for the hidden state X.

and Mariano (2002) in order to investigate the discriminative skills of the proposed score from Eq. (1). In Diebold and Mariano (2002), hypothesis tests focused on discriminating forecasts from the climatology; in the current study, we test the ability of the proposed score to discriminate any forecast f from the perfect forecast f_0 (forecast having the "true" hidden state X distribution). We consider the reference score value of the perfect forecast f_0 to be the mean score evaluated against the true process $\mathbb{E}(s_0(f_0, X))$. Hypothesis tests are tuned to primarily minimize the error of type I (wrongly rejecting the null hypothesis); consequently, it is common practice to assess the power of a test. The power p is the probability of rejecting a false null hypothesis; the power is expressed as $1 - \beta$, where β is the error of type II (failing to reject a false null hypothesis) and is expressed as $p = P(\text{Rejecting } H_0 | H_1 \text{ true})$. The closer to 1 the power is, the better the test is at detecting a false null hypothesis. For a given forecast f, the considered hypothesis is expressed as

$$\begin{cases} H_0 : \text{forecast } f \text{ is perfect } (f = f_0) \\ \quad \text{through the score } s, \\ \quad \text{leading to } \mathbb{E}(s(f, .)) \text{ close to } \mathbb{E}(s_0(f_0, X)), \\ H_1 : \text{forecast } f \text{ is imperfect, leading to } \mathbb{E}(s(f, .)) \\ \text{far from } \mathbb{E}(s_0(f_0, X)), \end{cases} \tag{16}$$

where f_0 and f are, respectively, the perfect forecast and an imperfect forecast to be compared with the perfect forecast f_0. In the following, the score $s(f, .)$ will represent, respectively, $s_0(f, X)$, $s_0(f, Y)$, and $s_\vee(f, Y)$ in order to compare the ability of the ideal score, the score used in practice, and the proposed score. The parameters associated with the hypothesis are the parameters, μ and σ in the following additive Gaussian model, of the imperfect forecast f, and they are varied to compute the different powers. The statistical test corresponding to Eq. (16) is expressed as

$$\begin{cases} H_0 \text{ is accepted if } t \leq c, \\ H_0 \text{ is rejected if } t > c, \end{cases} \tag{17}$$

where $t = |\mathbb{E}(s(f, .)) - \mathbb{E}(s_0(f_0, X))|$ is the test statistics and c is defined via the error of type I $P(t > c | H_0 \text{ is true}) = \alpha$ with the level $\alpha = 0.05$ in the present study.

To illustrate this test with numerical simulations, the additive Gaussian Model (A) is considered for the log score, where the forecast, X, and Y are assumed to be normally distributed with an additive error. The expectation in Eq. (17) is approximated and computed over $N = 1000$ verification data points, the true state being X and the corrupted data Y. The approximated test statistic is denoted with $\hat{t}$. The power p is expressed as $P(\hat{t} > c | f \text{ is imperfect})$ and is computed over 10 000 samples of length N when parameters μ and σ of the forecast f are varied.

In Fig. 5, the above power p is shown for varying mean μ and standard deviation σ of the forecast f in order to demonstrate the ability of the proposed score $s_\vee(., Y)$ to better discriminate between forecasts than the commonly used score $s_0(., Y)$. One expects the power to be low around, respectively, μ_0 and σ_0, and as high as possible outside these values. We note that the ideal score s_0 and the proposed score $s_\vee$ have similar power for the test Eq. (17), suggesting similar discriminating skills for both scores. However, the $s_0(., Y)$ score commonly used in practice results in an ill-behaved power as the observational noise increases (from left to right), indicating the necessity to account for the uncertainty associated with the verification data. The ill-behaved power illustrates the propensity of the test based on $s_0(., Y)$ to reject H_0 too often and in particular wrongfully when the forecast f is perfect and equals f_0. In addition, the power associated with the score $s_0(., Y)$ fails to reach the nominal test level α due to the difference in means between $s_0(., X)$

and $s_0(., Y)$ ($\mathbb{E}(s_0(f, X)) \neq \mathbb{E}(s_0(f, Y))$ for any forecast f) caused by the errors in the verification data Y. This highlights the unreliability of scores evaluated against corrupted verification data. Both varying mean and standard deviation reveal similar conclusions regarding the ability of the proposed score to improve its discriminative skills over a score evaluated on corrupted evaluation data.

5.2 Application to wind speed prediction

As discussed in the motivating example of Sect. 1, we consider surface wind speed data from the previous work (Bessac et al., 2018) and associated probabilistic distributions. In Bessac et al. (2018), a joint distribution for observations, denoted here as X_{ref}, and NWP model outputs is proposed and based on Gaussian distributions in a space–time fashion. This joint model aimed at predicting surface wind speed based on the conditional distribution of X_{ref} given NWP model outputs. The data are Box–Cox-transformed in this study to approximate normal distributions. The model was fitted by maximum likelihood for an area covering parts of Wisconsin, Illinois, Indiana, and Michigan in the United States. In this study, we focus on one station in Wisconsin and recover the parameters of its marginal joint distribution of observations X_{ref} and NWP outputs from the joint spatiotemporal model. In the following, we evaluate with scores the probability distribution of the NWP model outputs depicted in Fig. 2. In Bessac et al. (2018), the target distribution was the fitted distribution of the observations X_{ref}; however, in the validation step of the predictive probabilistic model, the observations shown in Fig. 2 were used without accounting for their potential uncertainty and error. This leads to a discrepancy between the target variable X_{ref} and the verification data that we denote as Y_{obs}. From Pinson and Hagedorn (2012), a reasonable model for unbiased measurement error in wind speed is $\epsilon_{\text{obs}} \sim \mathcal{N}(0, 0.25)$. Subsequently to Sect. 3, we proposed the following additive framework to account for the observational noise in the scoring framework:

$$\begin{cases} X_{\text{ref}} \sim \mathcal{N}(\mu_0, \sigma_0^2), \mu_0 \text{ and } \sigma_0 \text{ retrieved from the joint model,} \\ Y_{\text{obs}} = X_{\text{ref}} + \epsilon_{\text{obs}}, \text{ with } \epsilon_{\text{obs}} \sim \mathcal{N}(0, 0.25), \end{cases}$$

where $\mu_0 = 2.55$ and $\sigma_0 = 1.23$ from the fitted distribution in Bessac et al. (2018). In Table 1, log scores and CRPS are computed as averages over the studied time series in the additive Gaussian case with a previously given formula in Sect. 3. The variance associated with each average score is provided in parentheses. Table 1 shows a significant decrease in the variance when the proposed score is used compared to the score commonly used in practice that does not account for measurement error. One can notice that the variance of the scores used in practice are considerably high, limiting the reliability of these computed scores for decision-making purposes. Additionally, the new mean scores are closer to the ideal mean log score and CRPS, showing the benefit of accounting for observational errors in the scoring framework.

Table 1. Average scores (log score and CRPS) computed for the predictive distribution of the NWP model; the associated standard deviation is given in parentheses. The statistics are computed over the entire studied time series. The mean ideal score $\mathbb{E}(s_0(f, X))$, the averaged score computed in practice against the measurements $\mathbb{E}(s_0(f, Y))$, and the proposed score $\mathbb{E}(s_\vee(f, X))$ embedding the error in the verification data are computed.

	Mean score	NWP
Log score	Ideal score $\mathbb{E}(s_0(f, X))$	1.76
	$\mathbb{E}(s_0(f, Y))$	1.97 (1.52)
	$\mathbb{E}(s_\vee(f, Y))$	1.81 (1.15)
CRPS	Ideal score $\mathbb{E}(c_0(f, X))$	0.73
	$\mathbb{E}(c_0(f, Y))$	0.82 (0.67)
	$\mathbb{E}(c_\vee(f, Y))$	0.73 (0.48)

Figure 6 shows the pdf of the scores considered in Table 1; the skewness and the large dispersion in the upper tail illustrate with wind speed data cases where the mean is potentially not an informative summary statistic of the whole distribution. The non-corrected version of the score has a large variance, raising the concern of reliability on scores when computed on error-prone data.

6 Conclusions

We have quantified, in terms of variance and even distribution, the need to account for the error associated with the verification data when evaluating probabilistic forecasts with scores. An additive framework and a multiplicative framework have been studied in detail to account for the error associated with verification data. Both setups involve a probabilistic model for the accounted errors and a probabilistic description of the underlying non-observed physical process. Although we look only at idealized cases where the involved distributions and their parameters are known, this approach enables us to understand the importance of accounting for the error associated with the verification data.

6.1 Scores with hidden states

The proposed scoring framework was applied to standard additive and multiplicative models for which, via Bayesian conjugacy, the expression of the new scores could be made explicit. However, if the prior on X is not conjugate to the distribution of Y, one can derive approximately the conditional distribution $[X|Y = y]$ through sampling algorithms such as Markov chain Monte Carlo methods. Additionally, one can relax the assumption of known parameters for the distributions of $[X]$ and $[Y|X]$ by considering priors on the parameters (e.g., priors on μ_0 and σ_0). In practice, we rely on the idea that the climatology and/or information on measurement errors can help express the distribution of X or its priors.

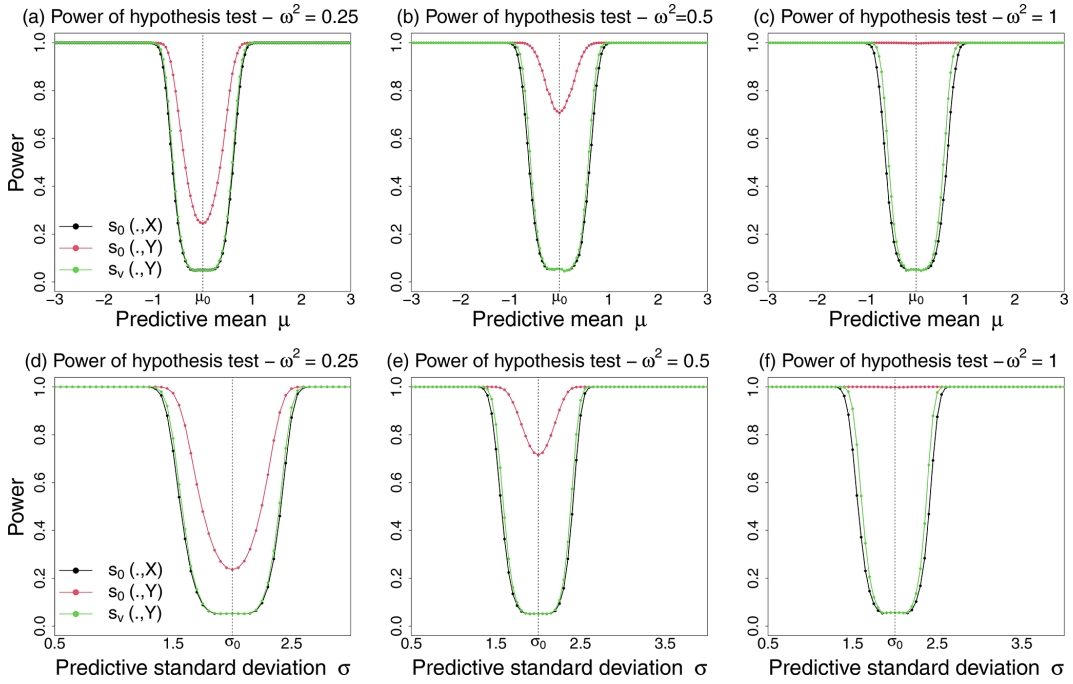

Figure 5. Power of the test Eq. (17) against varying predictive mean μ **(a, b, c)** and against varying predictive standard deviation σ **(d, e, f)** of the forecast f for different observational noise levels $\omega^2 = 0.25$ **(a, d)**, $\omega^2 = 0.5$ **(b, e)**, and $\omega^2 = 1$ **(c, f)**. In the simulations, the true state X is distributed as $\mathcal{N}(\mu_0 = 0, \sigma_0^2 = 4)$, and Y is distributed as $\mathcal{N}(0, \sigma_0^2 + \omega^2)$. The power is expected to be low around μ_0 (or σ_0) and as large as possible elsewhere.

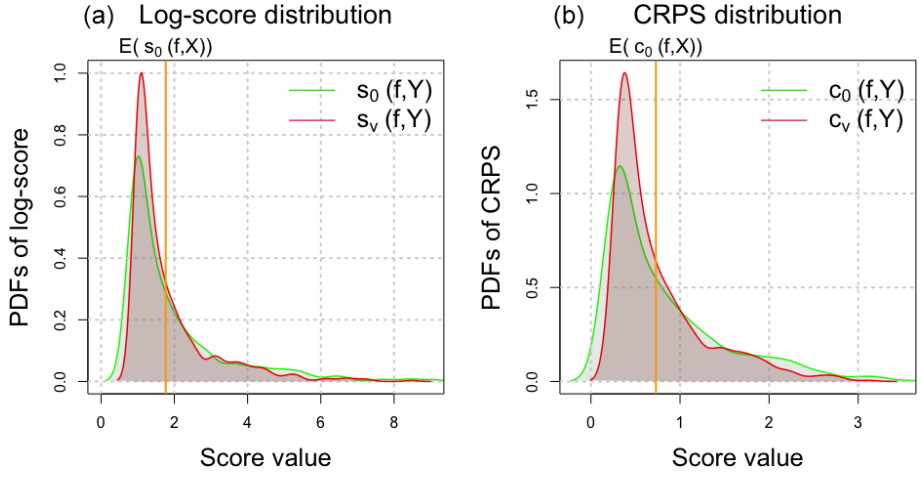

Figure 6. Empirical distribution of uncorrected (green) and corrected scores (red) for the log score **(a)** and CRPS **(b)** for the probabilistic distribution NWP model 1 evaluated against observations tainted with uncertainty. The ideal mean score value is illustrated by the orange vertical line.

In the current literature on scoring with uncertain verification data, most proposed works rely on additive errors as in Ciach and Krajewski (1999), Saetra et al. (2004), and Mittermaier and Stephenson (2015). Multiplicative error models for various error representations have been proposed in modeling studies, such as for precipitation (McMillan et al., 2011), but have not been considered in scoring correction frameworks. Furthermore, additive and multiplicative state-space models can be generalized to nonlinear and non-Gaussian state-space model specifications; see chap. 7 of Cressie and Wikle (2015) for a discussion and examples. More generally, one could consider the following state-space model specification:

$$\begin{cases} X = f(\eta), \\ Y = g(X, \epsilon), \end{cases}$$

where f and g are nonlinear functions describing the hidden state and the state-observation mapping, and η and ϵ are stochastic terms representing, respectively, the stochasticity of the hidden state and the observational error. This generalized specification of the state-observation model could be integrated in future work in the proposed scoring framework via Bayesian specifications of the new score in order to account for prior information on the verification data Y and its uncertainty and priors on the true state X.

6.2 Scores as random variables

Finally, the study raises the important point of investigating the distribution of scores when the verification data are considered to be a random variable. Indeed, investigating the means of scores may not provide sufficient information to compare between score discriminative capabilities. This has been pointed and investigated in Taillardat et al. (2019) in the context of evaluating extremes. This topic has also been extensively discussed by Bolin and Wallin (2019) in the context of varying predictability that generates non-homogeneity in the score values that is poorly represented by an average. One could choose to take into account the uncertainty associated with the inference of distribution parameters and compute different statistics of the distribution rather than the mean. In this case, a Bayesian setup similar to the current work could elegantly integrate the different hierarchies of knowledge, uncertainties, and a priori information to generalize the notion of scores.

Appendix A: Proof of Eq. (2)

For any random variable, say U, its mean can be written conditionally to the random variable y in the following way:

$$\mathbb{E}[U] = \mathbb{E}\big[\mathbb{E}\big[U|Y=y\big]\big].$$

In our case, the variables $U = s_o(f,X)$ and $s_\vee(f,y) = \mathbb{E}\big[U|Y=y\big]$. This gives $\mathbb{E}\big[s_\vee(f,Y)\big] = \mathbb{E}\big[s_o(f,X)\big]$. To show the inequality Eq. (2), we use the classical variance decomposition

$$\mathbb{V}[U] = \mathbb{V}\big[\mathbb{E}\big[U|Y=y\big]\big] + \mathbb{E}\big[\mathbb{V}\big[U|Y=y\big]\big].$$

With our notations, we have

$$\begin{aligned}\mathbb{V}\big[s_o(f,X)\big] &= \mathbb{V}\big[\mathbb{E}\big[s_o(f,X)|Y=y\big]\big] \\ &\quad + \mathbb{E}\big[\mathbb{V}\big[s_o(f,X)|Y=y\big]\big] \\ &= \mathbb{V}\big[s_\vee(f,Y)\big] + \text{ a non-negative term.}\end{aligned}$$

This leads to

$$\mathbb{V}\big[s_o(f,X)\big] \geq \mathbb{V}\big[s_\vee(f,Y)\big].$$

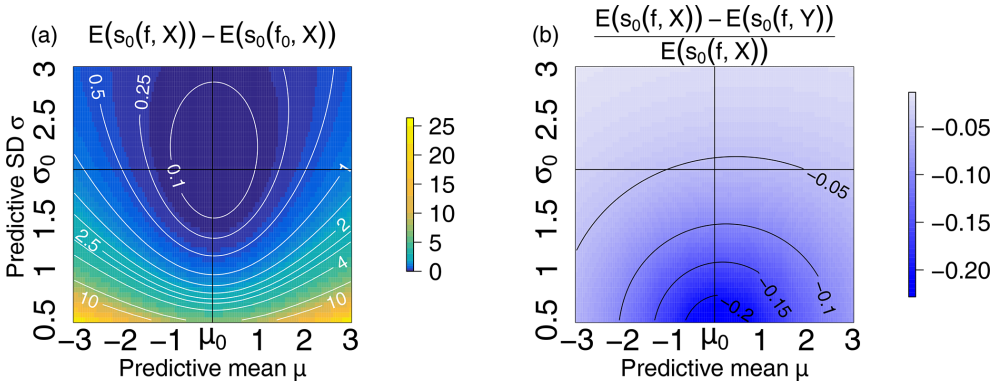

Figure A1. **(a)** Mean log score for an imperfect forecast f minus the mean log score of the perfect forecast f_0 when evaluated against perfect data X; the imperfect forecast $f \sim \mathcal{N}(\mu,\sigma)$ has a varying mean μ (x axis) and varying standard deviation σ (y axis). **(b)** Relative difference between $\mathbb{E}(s_0(f,X))$ and $\mathbb{E}(s_0(f,Y))$.

Appendix B: Proof of Eq. (5)

To express the score proposed in Eq. (1), one needs to derive the conditional distribution $[X|Y=y]$ from Model (A). More precisely, the Gaussian conditional distribution of X given $Y=y$ is equal to

$$[X|Y=y] \sim \mathcal{N}\left(\overline{y}, \frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}\right),$$

where $\overline{y}$ is a weighted sum that updates the prior information about $X \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with the observation $Y \sim \mathcal{N}(\mu_0, \sigma_0^2 + \omega^2)$:

$$\overline{Y} = \frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}Y \sim \mathcal{N}\left(\mu_0, \sigma_0^2 \times \frac{\sigma_0^2}{\sigma_0^2+\omega^2}\right).$$

Combining this information with Eqs. (1) and (3) leads to

$$\begin{aligned}
s_\vee(f,y) &= \log\sigma + \frac{1}{2\sigma^2}\left\{\mathbb{E}\left[(X-\mu)^2|Y=y\right]\right\} + \frac{1}{2}\log 2\pi \\
&= \log\sigma + \frac{1}{2\sigma^2}\left\{\mathbb{V}[X|Y=y] + (\mathbb{E}[X|Y=y]-\mu)^2\right\} \\
&\quad + \frac{1}{2}\log 2\pi \\
&= \log\sigma + \frac{1}{2\sigma^2}\left\{\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}\right. \\
&\quad \left. + \left(\frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}y - \mu\right)^2\right\} + \frac{1}{2}\log 2\pi.
\end{aligned}$$

By construction, we have

$$\mathbb{E}_Y\left(s_\vee(f,Y)\right) = \mathbb{E}_{\overline{Y}}\left(s_\vee(f,\overline{Y})\right) = \mathbb{E}_X\left(s_0(f,X)\right).$$

This means that, to obtain the right score value, we can first compute $\overline{Y}$ as the best estimator of the unobserved X and then use it in the corrected score $s_\vee(f,\overline{Y})$.

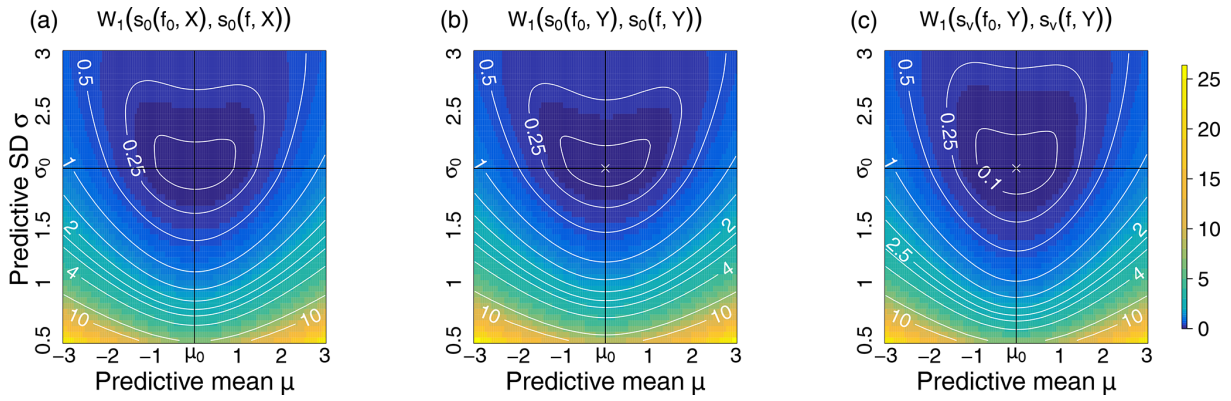

Figure B1. Bottom row: Wasserstein distance $W_1(s_0(f_0, .), s_0(f, .))$ between log-score s_0 distributions evaluated at the perfect forecast f_0 and the imperfect forecast f with varying predictive mean μ (x axis) and varying predictive standard deviation σ (y axis). From **(a)** to **(c)**: log scores are evaluated against the hidden true state X via $s_0(f, X)$, against Y tainted by observational noise of level $\omega^2 = 1$ via $s_0(f, Y)$ and through the corrected log-score version $s_\vee(f, Y)$. The verification data X and perfect forecast f_0 are distributed according to $f_0 \sim \mathcal{N}(0, 4)$. On the central and right surfaces, the white cross "X" indicates the numerical minimum of each surface.

Appendix C: Proof of Eq. (8)

To compute the corrected CRPS, one needs to calculate the conditional expectation of $c_0(f,X)$ under the distribution of $[X|Y=y]$. We first compute the expectation $E(c_0(f,X))$ and then substitute X by $\overline{Y}$ and its distribution with mean $a=\overline{y}$ and standard deviation $b=\sqrt{\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}}$. From Eq. (7) we obtain

$$\mathbb{E}(c_0(f,X)) = \mathbb{E}(X) + 2\sigma\left[\mathbb{E}\left(\phi\left(\frac{X-\mu}{\sigma}\right)\right) - \mathbb{E}\left(\frac{X-\mu}{\sigma}\overline{\Phi}_0\left(\frac{X-\mu}{\sigma}\right)\right)\right] - \left[\mu+\frac{\sigma}{\sqrt{\pi}}\right].$$

If X follows a normal distribution with mean a and variance b^2, that is, $X=a+bZ$ with Z a standard random variable, then we can define the continuous function $h(z)=\overline{\Phi}\left(\frac{a+bz-\mu}{\sigma}\right)$, with $h'(z)=-\frac{b}{\sigma}\phi\left(\frac{a+bz-\mu}{\sigma}\right)$. Then, we apply Stein's lemma (Stein, 1981), which states that $\mathbb{E}\left[h'(Z)\right]=\mathbb{E}[Zh(Z)]$, because Z is a standard random variable. It follows with the notations $t=\frac{b^2}{2\sigma^2}$ and $\lambda=\frac{a-\mu}{\sigma}$ that

$$\begin{aligned}\mathbb{E}\left[\frac{X-\mu}{\sigma}\overline{\Phi}\left(\frac{X-\mu}{\sigma}\right)\right] &= \lambda\mathbb{E}\left[\overline{\Phi}\left(\lambda+\frac{b}{\sigma}Z\right)\right] + \frac{b}{\sigma}\mathbb{E}[Zh(Z)],\\ &= \lambda\mathbb{E}\left(\mathbb{P}\left[Z'>\left(\lambda+\frac{b}{\sigma}Z\right)\right]\right) + \frac{b}{\sigma}\mathbb{E}\left[h'(Z)\right],\end{aligned}$$

where Z' has a standard normal distribution

$$= \lambda\mathbb{E}\left(\mathbb{P}\left[Z'-\frac{b}{\sigma}Z>\lambda\right]\right) - \frac{b^2}{\sigma^2}\mathbb{E}\left[\phi\left(\frac{a+bZ-\mu}{\sigma}\right)\right],$$

with

$$\begin{aligned}\lambda\mathbb{P}\left[Z'-\frac{b}{\sigma}Z>\lambda\right] &= \lambda\overline{\Phi}\left[\frac{\lambda}{\sqrt{1+b^2/\sigma^2}}\right]\\ &= \lambda\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right]\\ &= \frac{a-\mu}{\sigma}\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right].\end{aligned}$$

Then

$$\mathbb{E}\left[\frac{X-\mu}{\sigma}\overline{\Phi}\left(\frac{X-\mu}{\sigma}\right)\right] = \frac{a-\mu}{\sigma}\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right] - \frac{b^2}{\sigma^2}\mathbb{E}\left[\phi\left(\frac{a+bZ-\mu}{\sigma}\right)\right]$$

and

$$\begin{aligned}\mathbb{E}\left[\phi\left(\frac{a+bZ-\mu}{\sigma}\right)\right] &= \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\exp\left(-\frac{1}{2}\left(\frac{a+bZ-\mu}{\sigma}\right)^2\right)\right)\\ &= \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\exp\left(-\frac{b^2}{2\sigma^2}\left(Z+\frac{a-\mu}{b}\right)^2\right)\right).\end{aligned}$$

$\left(Z+\frac{(a-\mu)}{b}\right)^2$ is a non-centered chi-squared distribution with 1 degree of freedom and a non-central parameter $(\frac{a-\mu}{b})^2$ with a known moment-generating function

$$G\left(t;k=1,\lambda=\left(\frac{a-\mu}{b}\right)^2\right) = \frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{k/2}}.$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\phi\left(\frac{a+bZ-\mu}{\sigma}\right)\right] &= \frac{1}{\sqrt{2\pi}}\\ &\quad G\left(t=\frac{-b^2}{2\sigma^2};k=1,\lambda=\left(\frac{a-\mu}{b}\right)^2\right)\\ &= \frac{1}{\sqrt{2\pi}}\frac{\exp\left(\frac{\frac{-(a-\mu)^2}{b^2}\frac{b^2}{2\sigma^2}}{1+\frac{b^2}{\sigma^2}}\right)}{\sqrt{\left(1+\frac{b^2}{\sigma^2}\right)}}\\ &= \frac{\sigma}{\sqrt{2\pi}\sqrt{\sigma^2+b^2}}\\ &\quad \exp\left(\frac{-(a-\mu)^2}{2\left(\sigma^2+b^2\right)}\right).\end{aligned}$$

We obtain

$$\begin{aligned}E(c_0(f,X)) &= E(X) + 2\sigma\left[\left(1+\frac{b^2}{\sigma^2}\right)E\left(\phi\left(\frac{X-\mu}{\sigma}\right)\right) - \frac{a-\mu}{\sigma}\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right]\right] - \left(\mu+\frac{\sigma}{\sqrt{\pi}}\right)\\ &= E(X) + 2\sigma\left[\left(1+\frac{b^2}{\sigma^2}\right)\frac{\sigma}{\sqrt{2\pi(\sigma^2+b^2)}}\right.\\ &\quad \left.\exp\left(\frac{-(a-\mu)^2}{2(\sigma^2+b^2)}\right) - \frac{a-\mu}{\sigma}\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right]\right]\\ &\quad - \left(\mu+\frac{\sigma}{\sqrt{\pi}}\right)\\ &= E(X) + 2\left[\frac{\sqrt{\sigma^2+b^2}}{\sqrt{2\pi}}\exp\left(\frac{-(a-\mu)^2}{2(\sigma^2+b^2)}\right)\right.\\ &\quad \left. - (a-\mu)\overline{\Phi}\left[\frac{a-\mu}{\sqrt{\sigma^2+b^2}}\right]\right]\\ &\quad - \left(\mu+\frac{\sigma}{\sqrt{\pi}}\right).\end{aligned}$$

The expression of Eq. (8) is obtained by substituting X with $\overline{Y}$ and its Gaussian distribution with mean $a=\overline{y}$ and standard deviation $b=\sqrt{\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}}$ in the expression Eq. (6).

This gives

$$
\begin{aligned}
c_\vee(f,\overline{y}) &= E(c_0(f,X)|Y=y)\\
&= \overline{y} - \left(\mu + \frac{\sigma}{\sqrt{\pi}}\right)\\
&+ 2\left(\frac{\sqrt{\sigma^2+\frac{\sigma_0^2\omega^2}{\sigma_0^2+\omega^2}}}{\sqrt{2\pi}}\exp\left(-\frac{(\overline{y}-\mu)^2}{2(\sigma^2+\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2})}\right)\right.\\
&\left.-(\overline{y}-\mu)\,\Phi\left(\frac{\overline{y}-\mu}{\sqrt{\sigma^2+\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2}}}\right)\right).
\end{aligned}
$$

Appendix D: Proof of Eq. (9)

For Model (A), both random variables Y and $\overline{Y} = \frac{\omega^2}{\sigma_0^2+\omega^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\omega^2}Y$ are normally distributed with the same mean μ_0 but different variances, $\sigma_0^2+\omega^2$ and $\left(\frac{\sigma_0^2}{\sigma_0^2+\omega^2}\right)^2(\sigma_0^2+\omega^2)$, respectively. Since a chi-squared distribution can be defined as the square of a Gaussian random variable, it follows from Eq. (5) that

$$s_0(f,Y) \overset{d}{=} a_0 + b_0\chi_0^2 \text{ and } s_\vee(f,Y) \overset{d}{=} a_\vee + b_\vee\chi_\vee^2,$$

where $\overset{d}{=}$ means equality in distribution and

$$
\begin{aligned}
a_0 &= \log\sigma + \frac{1}{2}\log 2\pi \quad \text{and} \quad b_0 = \frac{\sigma_0^2+\omega^2}{2\sigma^2},\\
a_\vee &= \log\sigma + \frac{1}{2\sigma^2}\frac{\omega^2\sigma_0^2}{\sigma_0^2+\omega^2} + \frac{1}{2}\log 2\pi \quad \text{and}\\
b_\vee &= \frac{\sigma_0^2+\omega^2}{2\sigma^2}\left(\frac{\sigma_0^2}{\sigma_0^2+\omega^2}\right)^2,
\end{aligned}
$$

with χ_0^2 and $\chi_\vee^2$ representing a non-central chi-squared random variable with 1 degree of freedom and a non-centrality parameter:

$$\lambda_0 = \frac{(\mu_0-\mu)^2}{\sigma_0^2+\omega^2} \text{ and } \lambda_\vee = \frac{(\mu_0-\mu)^2}{\sigma_0^2+\omega^2}\left(\frac{\sigma_0^2+\omega^2}{\sigma_0^2}\right)^2.$$

Appendix E: Proof of Eq. (13)

In Model (B), the basic conjugate prior properties of such Gamma and inverse Gamma distributions allow us to say that $[X|Y=y]$ now follows a Gamma distribution with parameters α_0+a and β_0+b/y:

$$
\begin{aligned}
f_{X|Y=y}(x|y) &= \frac{(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha_0+a)}x^{\alpha_0+a-1}\\
&\exp(-x(\beta_0+b/y)), \text{ for } x>0.
\end{aligned}
$$

It follows that the proposed corrected score is

$$
\begin{aligned}
s_\vee(f,y) &= \mathbb{E}\left[s_o(f,X)|Y=y\right]\\
&= (1-\alpha)\mathbb{E}\left[\log(X)|Y=y\right]\\
&+ \beta\mathbb{E}\left[X|Y=y\right] - \alpha\log\beta + \log\Gamma(\alpha),\\
&= (1-\alpha)\left(\psi(\alpha_0+a) - \log\left(\beta_0+\frac{b}{y}\right)\right)\\
&+ \beta\frac{\alpha_0+a}{\beta_0+\frac{b}{y}} - \alpha\log\beta + \log\Gamma(\alpha).
\end{aligned}
$$

Indeed, $\mathbb{E}\left[X|Y=y\right] = \frac{\alpha_0+a}{\beta_0+\frac{b}{y}}$ and $\mathbb{E}\left[\log(X)|Y=y\right] = \psi(\alpha_0+a) - \log\left(\beta_0+\frac{b}{y}\right)$, where $\psi(x)$ represents the digamma function defined as the logarithmic derivative of the Gamma function, namely, $\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$.

Appendix F: Proof of Eq. (15)

From Eq. (14) we obtain

$$
\begin{aligned}
c_\vee(f,y) &= \mathbb{E}\left[X|Y=y\right] - \left[\frac{\alpha}{\beta} + \frac{1}{\beta B(.5,\alpha)}\right]\\
&+ 2\left[\mathbb{E}\left[\frac{X}{\beta}f(X)|Y=y\right]\right.\\
&\left.+\mathbb{E}\left[\left(\frac{\alpha}{\beta}-X\right)\overline{F}(X)|Y=y\right]\right].
\end{aligned}
$$

Since the conditional distribution of $[X|Y=y]$ is known,

$$
\begin{aligned}
\mathbb{E}\left[\frac{X}{\beta}f(X)|Y=y\right] &= \frac{1}{\beta}\int_0^{+\infty} xf(x)\frac{(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha_0+a)}x^{\alpha_0+a-1}\\
&\exp(-x(\beta_0+b/y))\,\mathrm{d}x\\
&= \frac{\beta^{\alpha-1}(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}\int_0^{+\infty} x^{\alpha+\alpha_0+a-1}\\
&\exp(-(\beta+\beta_0+b/y)x)\,\mathrm{d}x\\
&= \frac{\beta^{\alpha-1}(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}\\
&\frac{1}{(\beta+\beta_0+b/y)^{\alpha+\alpha_0+a}}\int_0^{+\infty} u^{\alpha+\alpha_0+a-1}\\
&\exp(-u)\mathrm{d}u\\
&= \frac{\beta^{\alpha-1}(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}\\
&\frac{\Gamma(\alpha+\alpha_0+a)}{(\beta+\beta_0+b/y)^{\alpha+\alpha_0+a}}\\
&= \frac{\beta^{\alpha-1}(\beta_0+b/y)^{\alpha_0+a}}{B(\alpha,\alpha_0+a)(\beta+\beta_0+b/y)^{\alpha+\alpha_0+a}},
\end{aligned}
$$

and the last term is

$$\mathbb{E}\left(\left(\frac{\alpha}{\beta}-X\right)\overline{F}(X)|Y=y\right)=\frac{(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha_0+a)}\int_0^{+\infty}\left(\frac{\alpha}{\beta}-x\right)$$
$$\left(\int_x^{+\infty}\frac{\beta^\alpha}{\Gamma(\alpha)}u^{\alpha-1}\exp(-\beta u)\mathrm{d}u\right)$$
$$\times x^{\alpha_0+a-1}\exp(-(\beta_0+b/y)x)\mathrm{d}x$$

with

$$\int_x^{+\infty}\frac{\beta^\alpha}{\Gamma(\alpha)}u^{\alpha-1}\exp(-\beta u)\mathrm{d}u=\frac{1}{\Gamma(\alpha)}\int_{\beta x}^{+\infty}v^{\alpha-1}\exp(-v)\mathrm{d}v$$
$$=\frac{(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}\int_0^{+\infty}\left(\frac{\alpha}{\beta}-x\right)\Gamma(\alpha,\beta x)x^{\alpha_0+a-1}$$
$$\exp(-(\beta_0+b/y)x)\mathrm{d}x,$$

where $\Gamma(\alpha,\beta x)=\int_{\beta x}^{+\infty}v^{\alpha-1}\exp(-v)\mathrm{d}v$ is the upper incomplete Gamma function.

The entire expression of the corrected CRPS is expressed as

$$c_\vee(f,y)=\left[\frac{\alpha}{\beta}-\frac{1}{\beta B(.5,\alpha)}\right]-\frac{\alpha_0+a}{\beta_0+\frac{b}{y}}$$
$$+2\frac{\beta^{\alpha-1}(\beta_0+b/y)^{\alpha_0+a}}{B(\alpha,\alpha_0+a)(\beta+\beta_0+b/y)^{\alpha+\alpha_0+a}}$$
$$+2\frac{(\beta_0+b/y)^{\alpha_0+a}}{\Gamma(\alpha)\Gamma(\alpha_0+a)}$$
$$\int_0^{+\infty}\left(\frac{\alpha}{\beta}-x\right)\Gamma(\alpha,\beta x)x^{\alpha_0+a-1}$$
$$\exp(-(\beta_0+b/y)x)\mathrm{d}x.$$

Appendix G: Additional results: distance between score distributions

In order to further study the impact of imperfect verification data and to take full advantages of the score distributions, we compute the Wasserstein distance (Muskulus and Verduyn-Lunel, 2011; Santambrogio, 2015; Robin et al., 2017) between several score distributions and compare it to the commonly used score average. In particular, through their full distributions we investigate the discriminative skills of scores compared to the use of their mean only. The p-Wasserstein distance between two probability measures P and Q on $\mathbb{R}$ with finite p moments is given by $W_p(P,Q):=\left(\inf_{\gamma\in\Gamma(P,Q)}\int_{M\times M}\mathrm{d}(x,y)^p\,\mathrm{d}\gamma(x,y)\right)^{1/p}$, where $\Gamma(P,Q)$ is the set of all joint probability measures on $\mathbb{R}\times\mathbb{R}$ whose marginals are P and Q. In the one-dimensional case, as here, the Wasserstein distance can be computed as $W_p(F,G)=\left(\int_0^1|F^{-1}(u)-G^{-1}(u)|^p\mathrm{d}u\right)^{1/p}$, with F and G the cdfs of P and Q to be compared, F^{-1} and G^{-1} their generalized inverse (or quantile function), and, in our case $p=1$. The R-package transport (Schuhmacher et al., 2020) is used to compute Wasserstein distances. Figure A1 shows the mean of the log score minus its minimum $\mathbb{E}(s_0(f_0,X))$ and the relative difference between the ideal mean log score and the mean log score evaluated against imperfect verification data. One can first observe the flatness of the mean log score around its minimum, indicating a lack of discriminative skills of the score mean when comparing several forecasts. Secondly, the discrepancy between the score evaluated against perfect and imperfect verification data indicates the effects of error-prone verification data as discussed earlier.

Figure B1 shows the Wasserstein distance between the distributions of three scores evaluated in different contexts (a perfect forecast and an imperfect forecast, true and error-prone verification data). Three different log scores are considered the ideal log score $s_0(.,X)$, the log score used in practice $s_0(.,Y)$, and the proposed corrected score $s_\vee(.,Y)$. One can notice that Wasserstein distances exhibit stronger gradients (contour lines are narrower) than the mean log score in Fig. A1. In particular, the surfaces delimited by a given contour level are smaller for the proposed score than for the other scores; for instance, the area inside the contour of level $z=0.1$ is larger for the mean log score in Fig. A1 than for the Wasserstein distance between the score distance. This indicates that considering the entire score distribution has the potential to improve the discriminative skills of scoring procedures. In particular, imperfect forecasts departing from the perfect forecast will be more sharply discriminated with the Wasserstein distance computed on score distributions.

Finally, when considering Wasserstein distances associated with the score evaluated on imperfect verification data, the minimum of the distances (indicated by white crosses "X" in the central and right panels) is close to the "true" minimum (intersection of $x=\mu_0$ and $y=\sigma_0$). This indicates some robustness of the Wasserstein distance between the score distributions when errors are present in the verification data. Similar results are obtained for the CRPS and are not reported here. As stated earlier, developing metrics to express the discriminative skills of a score is beyond the scope of this work.

Code and data availability. Codes for this study can be found at https://github.com/jbessac/uncertainty_scoring (last access: 8 September 2021, Bessac, 2021). Data are ground measurements from the NOAA Automated Surface Observing System (ASOS) network and are available at ftp://ftp.ncdc.noaa.gov/pub/data/asos-onemin (last access: 8 September 2021, National Centers for Environmental Information, 2021).

Author contributions. JB and PN contributed to developing methodological aspects of the paper. PN provided expertise on scoring. JB conducted numerical analyses. All the authors edited and wrote portions of the paper.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We thank Aurélien Ribes for his helpful comments and discussions.

Financial support. This collaboration is supported by the initiative Make Our Planet Great Again, through the Office of Science and Technology of the Embassy of France in the United States. The effort of Julie Bessac is based in part on work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR; grant no. DE-AC02-06CH11347). Part of Philippe Naveau's work has been supported by the European DAMOCLES-COST-ACTION on compound events and also by French national programs (FRAISE-LEFE/INSU, MELODY-ANR, ANR-11-IDEX-0004-17-EURE-0006, and ANR T-REX AAP CE40).

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

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Clim., 9, 1518–1530, 1996.

Bessac, J., Constantinescu, E., and Anitescu, M.: Stochastic simulation of predictive space–time scenarios of wind speed using observations and physical model outputs, Ann. Appl. Stat., 12, 432–458, 2018.

Bessac, J.: Codes for scoring under uncertain verification data, available at: https://github.com/jbessac/uncertainty_scoring, GitHub [code], last access: 8 September 2021.

Bolin, D. and Wallin, J.: Scale invariant proper scoring rules Scale dependence: Why the average CRPS often is inappropriate for ranking probabilistic forecasts, arXiv preprint arXiv:1912.05642, available at: https://arxiv.org/abs/1912.05642 (last access: 8 September 2021), 2019.

Bowler, N. E.: Accounting for the effect of observation errors on verification of MOGREPS, Meteorol. Appl., 15, 199–205, 2008.

Bröcker, J. and Ben Bouallègue, Z.: Stratified rank histograms for ensemble forecast verification under serial dependence, Q. J. Roy. Meteorol. 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, 2007.

Candille, G. and Talagrand, O.: Retracted and replaced: Impact of observational error on the validation of ensemble prediction systems, Q. J. Roy. Meteorol. Soc., 134, 509–521, 2008.

Ciach, G. J. and Krajewski, W. F.: On the estimation of radar rainfall error variance, Adv. Water Resour., 22, 585–595, 1999.

Cressie, N. and Wikle, C. K.: Statistics for spatio-temporal data, John Wiley & Sons, Hoboken, N.J., 2015.

Daley, R.: Estimating observation error statistics for atmospheric data assimilation, Ann. Geophys., 11, 634–647, 1993.

Diebold, F. X. and Mariano, R. S.: Comparing predictive accuracy, J. Bus. Econ. Stat., 20, 134–144, 2002.

Dirkson, A., Merryfield, W. J., and Monahan, A. H.: Calibrated probabilistic forecasts of Arctic sea ice concentration, J. Clim., 32, 1251–1271, 2019.

Ferro, C. A. T.: Measuring forecast performance in the presence of observation error, Q. J. Roy. Meteorol. Soc., 143, 2665–2676, https://doi.org/10.1002/qj.3115, 2017.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: Bayesian data analysis, CRC press, 2013.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.

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

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. Roy. Stat.l Soc. Ser. B, 69, 243–268, 2007.

Gorgas, T. and Dorninger, M.: Quantifying verification uncertainty by reference data variation, Meteorol. Z., 21, 259–277, 2012.

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

Hamill, T. M. and Juras, J.: Measuring forecast skill: Is it real skill or is it the varying climatology?, Q. J. Roy. Meteorol. Soc., 132, 2905–2923, 2006.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Q. J. Roy. Meteorol. Soc., 144, 1257–1278, 2017.

Jolliffe, I. T.: Uncertainty and inference for verification measures, Weather Forecast., 22, 637–650, 2007.

Jolliffe, T. and Stephenson, D. B.: Forecast verification: A practitioner's guide in atmospheric science, edited by: Wiley, I., Chichester, Weather, 59, 132–132, https://doi.org/10.1256/wea.123.03, 2004.

Kalman, R. E.: A new approach to linear prediction and filtering problems, Transactions of the ASME, J. Basic Eng., 82, 35–45, 1960.

Kalman, R. E. and Bucy, R. S.: New results in linear filtering and prediction theory, J. Basic Eng., 83, 95–108, 1961.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, 3, https://doi.org/10.1029/2005WR004368, 2006a.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, Water Resour. Res., 42, 3, https://doi.org/10.1029/2005WR004376, 2006b.

Kleen, O.: Measurement Error Sensitivity of Loss Functions for Distribution Forecasts, SSRN 3476461, https://doi.org/10.2139/ssrn.3476461, 2019.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, J. Hydrol., 400, 83–94, 2011.

Mittermaier, M. P. and Stephenson, D. B.: Inherent bounds on forecast accuracy due to observation uncertainty caused by temporal sampling, Mon. Weather Rev., 143, 4236–4243, 2015.

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Weather Rev., 115, 1330–1338, 1987.

Muskulus, M. and Verduyn-Lunel, S.: Wasserstein distances in the analysis of time series and dynamical systems, Physica D, 240, 45–58, 2011.

National Centers for Environmental Information, National Oceanic Atmospheric Administration, U.S. Department of Commerce: Automated Surface Observing Systems (ASOS) program, [code], available at: ftp://ftp.ncdc.noaa.gov/pub/data/asos-onemin, last access: 8 September 2021.

Pappenberger, F., Ghelli, A., Buizza, R., and Bodis, K.: The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications, J. Hydrometeorol., 10, 807–819, 2009.

Pinson, P. and Hagedorn, R.: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations, Meteorol. Appl., 19, 484–500, 2012.

Robert, C. and Casella, G.: Monte Carlo statistical methods, Springer Science & Business Media, 2013.

Robin, Y., Yiou, P., and Naveau, P.: Detecting changes in forced climate attractors with Wasserstein distance, Nonl. Process. Geophys., 24, 393–405, 2017.

Saetra, O., Hersbach, H., Bidlot, J.-R., and Richardson, D. S.: Effects of observation errors on the statistics for ensemble spread and reliability, Mon. Weather Rev., 132, 1487–1501, 2004.

Santambrogio, F.: Optimal transport for applied mathematicians, Vol. 87, Birkhäuser Basel, 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, 2015.

Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., Schmitzer, B., Schrieber, J., and Wilm, T.: transport: Computation of Optimal Transport Plans and Wasserstein Distances, R package version 0.12-2, https://cran.r-project.org/package=transport (last access: 8 September 2021), 2020.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X.-Y., Wang, W., and Powers, J.: A description of the Advanced Research WRF Version 3, Tech. Rep., https://doi.org/10.5065/D68S4MVH, 2008.

Stein, C. M.: Estimation of the mean of a multivariate normal distribution, Ann. Stat., 9, 1135–1151, https://doi.org/10.1214/aos/1176345632, 1981.

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.

Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R.: Extreme events evaluation using CRPS distributions, arXiv preprint arXiv:1905.04022, available at: https://arxiv.org/abs/1905.04022 (last access: 8 September 2021), 2019.

Waller, J. A., Dance, S. L., Lawless, A. S., and Nichols, N. K.: Estimating correlated observation error statistics using an ensemble transform Kalman filter, Tellus A, 66, 23294, https://doi.org/10.3402/tellusa.v66.23294, 2014.

Weijs, S. V. and Van De Giesen, N.: Accounting for observational uncertainty in forecast verification: an information-theoretical view on forecasts, observations, and truth, Mon. Weather Rev., 139, 2156–2162, 2011.

Weijs, S. V., Van Nooijen, R., and Van De Giesen, N.: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition, Mon. Weather Rev., 138, 3387–3399, 2010.

Wilks, D. S.: Sampling distributions of the Brier score and Brier skill score under serial dependence, Q. J. Roy. Meteorol. Soc., 136, 2109–2118, 2010.

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, 2018.