The Promise of Diversity: Distribution-based Hydrologic Model Evaluation and Diagnostics

Jasper A. Vrugt*†

March 26, 2023

*Department of Civil and Environmental Engineering, University of California, Irvine, California, USA.
†Corresponding author: jasper@uci.edu

Highlights

1. This paper advocates the use of Monte Carlo-based simulation distributions for hydrologic model evaluation and model diagnostics.

2. Distribution forecasts coalesce information about model behavior, robustness, sensitivity and uncertainty that is not available in single-valued model output.

3. *Strictly proper* scoring rules provide a rigorous information-theoretic underpinning to hydrologic model evaluation and diagnostics.

4. The Bregman divergence ensures *strict propriety* of scoring rules and their associated divergence functions.

5. Propriety and elicitability offer useful working paradigms for the development, use and standardization of hydrologic scoring functions and scoring rules.

6. We present divergence scores for flood frequency analysis and flow duration and recession curves.

Abstract

Distribution forecasts over future quantities or events are routinely made in hydrology but evade model evaluation by a need for ensemble simulation and lack of knowledge on how to properly evaluate simulation distributions against data. Predictive distributions P derived from (quasi)-Bayesian methods are usually traded for a (likelihood-weighted) mean or median prediction to accommodate error measures (scoring functions) such as the mean absolute error or the mean squared error that depend both on the point forecasts and the realizing observations. Point in case is the so-called KG efficiency or KGE of *Gupta et al.* (2009) and improvements thereof (*Lamontagne et al.*, 2020), which have rapidly gained popularity as scoring functions among hydrologists as alternative to the infamous *Nash and Sutcliffe* (1970) efficiency, but are equally exclusive in how they quantify model performance using only single-valued simulated output of the quantities of interest. This paper advocates the use of simulation distributions for hydrologic model evaluation and model diagnostics. Distribution evaluation is supported by information-theoretic arguments and puts into modeling practice the social justice narrative of diversity, equity and inclusion for different simulations. We discuss past developments that led to the current state-of-the-art of forecast verification in hydrology and bring to the fore scoring rules for model evaluation and diagnostics. *Strictly proper* scoring rules condense a distribution forecast to a single reward value for the materialized outcome(s) and have a strong underpinning in statistical, decision and information theory. We review scoring rules for dichotomous and categorical events, quantiles (intervals) and density forecasts, discuss the importance of scoring rule propriety and address diagnostic aspects of a distribution forecast such as sharpness, reliability and entropy. The usefulness and power of scoring rules is demonstrated on simple benchmark problems and discharge distributions simulated with conceptual watershed models using Generalized Likelihood Uncertainty Estimation and Bayesian model averaging. We also link scoring rules to model diagnostics and present *strictly proper* divergence scores for flood frequency analysis and flow duration and recession curves. Scoring rules offer a rigorous information-theoretic underpinning to model evaluation and diagnostics and provide statistically principled means for (Bayesian) model selection and the analysis of hydrograph functionals, flood frequencies and extreme events.

Keywords: Ensemble prediction, distribution forecast, information theory, scoring rules, divergence score, propriety, sharpness, reliability, uncertainty, entropy, integral transform, logarithmic score, continuous ranked probability score, recession analysis, flow duration curve, signatures, watershed models

1 Introduction and Scope

The topic of model evaluation has received considerable attention in the hydrologic and water resources literature over the past decades. Model evaluation is an integral part of the model development process and involves comparing simulated system behaviour with observations in pursuit of a qualitative and/or quantitative understanding of their similarities and differences and how well the model approximates reality according to some error measure. This process must acknowledge differences in extent, support and spacing of modeled and observed quantities (*Grayson and Blöschl*, 2001). We follow *Gneiting* (2011) and use the terminology of a scoring function on a general sample space Ω for an error measure.

Definition 1. *A scoring function is any real-valued function* $s : \Omega \times \Omega \to \mathbb{R}$ *where* $s(y, \omega)$ *represents the loss or penalty when the point forecast* $y \in \Omega$ *is issued and the observation* $\omega \in \Omega$ *materializes.*

Thus, scoring functions such as the ubiquitous squared error, $s_{\mathrm{SE}}(y, \omega) = (\omega - y)^2$, absolute error $s_{\mathrm{AE}}(y, \omega) = |\omega - y|$, and *Nash and Sutcliffe* (1970) efficiency, $s_{\mathrm{NS}}(y, \omega) = 1 - (\omega - y)^2/(\omega - \mu_\omega)^2$, measure the performance of a point forecast y, where μ_ω is the mean of the verifying data and $\epsilon = \omega - y$ is the so-called residual. The $s_{\mathrm{SE}}(y, \omega)$ and $s_{\mathrm{AE}}(y, \omega)$ scoring functions easily generalize to n-vectors of forecasts and observations. We use the error terminology for $s_{\mathrm{SE}}(y, \omega)$ and $s_{\mathrm{AE}}(y, \omega)$ though it would be more precise to use the wording of squared and absolute residual (*Vrugt and de Oliveira*, 2022). For the purpose of this discussion, we shall classify the research on model evaluation into two different groups including (i) residual-based approaches and (ii) non-residual-based methods. Methods in this second group may still use residuals to quantify model performance but these are usually not discharge residuals but residuals of some hydrograph functional.

The residual-based approach has its roots in linear regression theory and determines model performance using goodness-of-fit measures of simulated and measured watershed behavior. This includes the use of (a) formal objective functions, likelihood functions and summary metrics that result from rigorous application of statistical principles to the assumed probabilistic properties of the residuals within the context of weighted and/or generalized least squares (*Tasker*, 1980; *Stedinger and Tasker*, 1985; *Kavetski et al.*, 2006a,b), maximum likelihood and Bayesian estimation (*Sorooshian and Dracup*, 1980; *Kuczera*, 1983; *Bates and Campbell*, 2001; *Schoups and Vrugt*, 2010; *Scharnagl et al.*, 2015; *Ammann et al.*, 2019; *Vrugt et al.*, 2022) approximate Bayesian computation (*Nott et al.*, 2012; *Vrugt and Sadegh*, 2013a; *Sadegh*

and Vrugt, 2013) and information-theoretic approaches (*Neuman*, 2003; *Ye et al.*, 2008; *Weijs et al.*, 2010a; *Lu et al.*, 2011; *Schöniger et al.*, 2014; *Pachepsky et al.*, 2016; *Volpi et al.*, 2017), (b) informal metrics of the quality-of-fit and pseudo-likelihood functions within the context of model training (*Nash and Sutcliffe*, 1970; *Gupta et al.*, 2009; *Pool et al.*, 2018; *Knoben et al.*, 2019; *Lamontagne et al.*, 2020; *Schwemmle et al.*, 2021), informal Bayesian approaches (*Beven and Binley*, 1992; *Freer et al.*, 1996; *Beven and Freer*, 2001) and multi-criteria model calibration (*Gupta et al.*, 1998; *Boyle et al.*, 2000), and (c) tolerable ranges of the residuals within the context of limits of acceptability (*Beven*, 2006; *Vrugt and Beven*, 2018), regional sensitivity analysis (*Spear and Hornberger*, 1980; *Spear et al.*, 2020), dynamic identifiability analysis (*Wagener et al.*, 2003) and the parameter identification method based on the localization of information (*Vrugt et al.*, 2002). Let it be evident that residual-based methods are often used for parameter estimation purposes so as to train/calibrate models to mimic the observed data as closely and consistently as possible. While it is desirable that the model can reproduce historical observations, the aggregation of the residuals into a single performance index implies a significant loss of information about specific system behaviours. As a result, residual-based methods provide only limited guidance on model improvement. If the model is expected to match a certain functional of the hydrograph, it is critical that the scoring function be consistent for it, in the sense that the expected score is maximized (or minimized, if appropriate) when following the directive. Functionals that incentivize a truthful model description are called elicitable in decision-theory.

Non-residual-based methods are usually much more geared towards hypothesis testing and learning by exploiting hydrological context and theory, though they may be used for model training as well. These methods do not work with the residuals of simulated and measured watershed behavior but rather quantify model performance using multiple different hydrologic functionals of the streamflow hydrograph. Examples of such numeral descriptors of the catchment response to rainfall are the runoff ratio, baseflow index and flow duration curve. *Shamir et al.* (2005) laid the foundation of this approach in their work on hydrograph indices - and this approach advanced and matured further into what is now known as model diagnostics (*Gupta et al.*, 2008). As hydrologic functionals of the streamflow hydrograph are elicitable, discrepancies between measured and simulated functionals (signatures) are symptoms of model malfunctioning, and if able to relate each functional to a specific process, will guide model improvement in a more meaningful and systematic way. At its strongest, diagnostic evaluation will point clearly towards

the aspects of the model that need improvement, and give guidance towards the manner of improvement (*Gupta et al.*, 2008; *Yilmaz et al.*, 2008; *Westerberg et al.*, 2011).

Our classification of model evaluation methods into residual and non-residual based approaches suffices for the purpose of this paper but is not complete. One may discern a group of residual-based model evaluation strategies with a diagnostic intent. Examples of this third group are model evaluation methods that couple ubiquitous scoring functions with wavelets (*Rathinasamy et al.*, 2014), self-organizing maps (*Reusser et al.*, 2009), interval deviation (*Chen et al.*, 2014) and information theory (*Gong et al.*, 2013).

While the diagnostics approach has helped establish a new philosophy and/or paradigm for hydrologic model evaluation, as a community, we continue to hold on to and rely too much on, deterministic, non-inclusive, measures of model performance. Point in case of these scoring functions is the so-called KG efficiency or KGE of *Gupta et al.* (2009) and improvements thereof (*Lamontagne et al.*, 2020), which have gained rapid popularity among hydrologists as alternative to the infamous *Nash and Sutcliffe* (1970) efficiency, but are equally exclusive in how they quantify model performance using only single-valued simulated output of the quantities of interest. In fact, the confidence intervals of the KG efficiency are not defined in a formal mathematical sense preventing a rigorous statistical description of model parameter and simulation uncertainty (*Vrugt and de Oliveira*, 2022). A little more than a decade ago, *Guttorp* (2011) formulated a vision of how climate models should be evaluated against data (P. 820)

> "*... Climate models are difficult to compare to data. Often climatologists compute some summary statistic, such as global annual mean temperature, and compare climate models using observed (or rather estimated) forcings to the observed (or rather estimated) temperatures. However, it seems more appropriate to compare the distribution (over time and space) of climate model output to the corresponding distribution of observed data, as opposed to point estimates with or without confidence intervals.*"

and this change from point to distributional evaluation finds support in hydrology through information-theoretic arguments (*Weijs et al.* 2010a, P. 2545)

> "*... models should preferably be explicitly probabilistic and calibrated to maximize the information they provide.*"

and parallels sociocultural developments related to equity, inclusiveness and diversity. A recent report of McKinsey & Company titled *Diversity wins: How inclusion matters* (*Hunt et al.*, 2020) concludes that companies with a more inclusive and greater workplace diversity outperform their competitors, increase employee engagement and achieve higher profits. These developments inspire a paradigm change in how we evaluate our models.

Distribution forecasts express diversity in the form of a probability distribution over future quantities or events (*Dawid*, 1984). Strictly speaking, the modeled outcomes in this paper are not forecasts as we use measured values of the exogenous variables. This, however, is inconsequential to the premise of this paper as methods discussed are equally applicable to simulation distributions. Such forecasts are routinely made in weather and climate prediction (*Palmer*, 2002; *Gneiting et al.*, 2005), computational finance (*Duffie and Pan*, 1997), macroeconomics (*Garratt et al.*, 2003; *Granger*, 2005) and hydrology (*Welles et al.*, 2007; *Thielen et al.*, 2008) but evade model evaluation by a need for ensemble simulation and lack of knowledge on how to properly evaluate a simulation distribution against measured data. Advances in computer hardware, software and theory have tremendously advanced our ability to quantify the uncertainty of hydrologic model output (*Beven and Binley*, 1992; *Kuczera and Parent*, 1998; *Vrugt et al.*, 2003; *Kavetski et al.*, 2006a; *Schoups and Vrugt*, 2010; *Vrugt*, 2016; *Vrugt et al.*, 2022) but we typically discard distribution forecasts P of (quasi)-Bayesian methods in hydrologic model evaluation and quantify model performance using some set-valued mapping $P \to T(P) \subseteq \Omega$ to the real line $\mathbb{R}$ with the mean, likelihood-weighted mean or median prediction of the sample space Ω as key examples. This point-valued mapping necessarily implies a huge loss of information about model performance. Exception to this are the works of *Volpi et al.* (2017) and *Brunetti et al.* (2017) who relied on the integrated or marginal likelihood for hydrologic model evaluation and selection.

Forecast verification is an active field of research in the climate, atmospheric and ocean sciences and is concerned with evaluating, verifying and determining the predictive power of prognostic model forecasts (*Murphy and Katz*, 1985; *Storch and Zwiers*, 1999; *Jolliffe and Stephenson*, 2011). Scoring rules have long been used to evaluate the accuracy of forecast probabilities after observing the occurrence, or nonoccurrence, of predicted events of dichotomous, categorical and continuous variables (*Gneiting and Raftery*, 2007).

Definition 2. *A scoring rule is any extended real-valued function* $S : \mathscr{P} \times \Omega \to \overline{\mathbb{R}} \equiv [-\infty, \infty]$ *such that*

$S(P, \omega)$ is $\mathscr{P}$-quasi-integrable for all $P \in \mathscr{P}$ and measures the reward (or loss) when the distribution forecast P is issued and observation $\omega \in \Omega$ materializes.

Thus, a scoring rule $S(P, \omega)$ measures the performance of a distribution forecast P in a single reward (or loss) value and reduces to a scoring function $s(y, \omega)$ for a point forecast. Most scoring rules are real-valued, thus, take on values in $\mathbb{R}$ with exceptions such as the ignorance score (*Roulston and Smith*, 2002) or logarithmic rule (*Good*, 1952), which can attain scores of infinity and minus infinity, respectively, and, thus, operate in $\overline{\mathbb{R}}$. The attractive statistical and information-theoretic properties of scoring rules benefits ranking of likelihood functions (*Vrugt et al.*, 2022), hypothesis testing with watershed models and, as we show in this paper, hydrologic model evaluation. All these are desirable qualities of scoring rules given the plethora of hydrologic models used by researchers and practitioners (*Clark et al.*, 2008; *Schoups et al.*, 2010; *Fenicia et al.*, 2011).

Weijs et al. (2010a) presents a convincing example so as to why scoring rules such as the *Brier* (1950) score and continuous ranked probability score (CRPS) of *Matheson and Winkler* (1976) should be used for hydrologic model calibration and evaluation. Otherwise, model training is overly susceptible to misinformation and/or incomplete (unfinished) learning. Despite their compelling plea, scoring rules such as the CRPS have only found sporadic application in hydrology, usually for evaluating ensemble forecast skill (*Vrugt et al.*, 2006; *Laio and Tamea*, 2007; *Girons Lopez et al.*, 2021). A simulation distribution coalesces model responses across the (prior/posterior) parameter and/or input space and contains information about model behavior, robustness, sensitivity and uncertainty that is not available in single-valued model output. Thus, scoring function-based model evaluation strategies imply an inherent loss of information about model functioning. This paper advocates the use of distribution-based model evaluation and diagnostics. This paper is concerned with the basic question of how we should evaluate distributions generated via simulation. We bring scoring rules to the attention of hydrologists and demonstrate their power, usefulness and applicability to hydrologic model evaluation and model diagnostics. We introduce *strictly proper* scoring rules for flow duration and recession curves and the analysis of flood frequencies and extreme events. To understand the different scoring rules for dichotomous, categorical and continuous variables, convey their relationship with information theory, explain the importance of elicitability and scoring rule propriety and sufficiency, we must review different concepts from probability and information theory. Hopefully, our work inspires others to delve deeper into the

topic of scoring rules and seek the advantages of distribution-based model evaluation and diagnostics over the current practice of deterministic model evaluation.

The remainder of this paper is organized as follows. In Section 2, we formalize our mathematical/statistical treatment of probability and discuss our use of symbols and notation. Section 3 reviews the use of information theory, specifically relative entropy, applicable to ideal situations with known distribution functions of each verifying observation. Sections 4 - 7 discuss the more common and realistic situation in which we do not have knowledge of the distribution $Q \in \mathscr{P}$ which materializes with the measurement $\omega \in \Omega$. Section 4 illustrates the insufficiency of common methods used in the hydrologic literature for evaluating distribution forecasts. This is followed by Section 5, which relates information theory to scoring rules for distribution forecasts of categorical (discrete) variables and their extension in Section 6 to density forecasts of continuous variables. Section 7 is considered with diagnostic analysis and revisits the decomposition of *strictly proper* scoring rules into an uncertainty, sharpness (resolution) and reliability term. The different sections are permeated with case studies using conceptual watershed models and measurements of the rainfall-discharge transformation. The penultimate Section 8 presents a future outlook about the use of scoring rules in diagnostic model evaluation, Bayesian model selection and the prediction of extreme events. We introduce divergence scores for the flow duration curve and *Brutsaert and Nieber* (1977) recession analysis and present analytic expressions for CRPS and logarithmic scoring rules for the Pearson type III distribution used in flood frequency analysis. Section 9 concludes this paper with a summary of our main findings.

The topic of this paper warrants a strong statistical/information-theoretic treatment. But those discouraged by our statistical treatment of this topic are directed to the hydrologic case studies of this paper and the `ScoringRules` toolbox in MATLAB. Proofs and computational details have been deferred to Appendices. Specifically, Appendix A reviews Gibbs' inequality. Appendices B, C, D, F, G, I and J derive closed-form expressions for the relative entropy and CRPS of certain parametric *forecast* and *true* distributions. Appendix E presents precipitation data. Finally, Appendix H details our computational and numerical implementation of the HYMOD, Hmodel and SAC-SMA conceptual watershed models.

2 Preliminaries

One of the major purposes of hydrologic modeling is to predict watershed behavior under future conditions. We can shed much light on hydrologic theory, process knowledge, computational implementation, and aleatory and epistemic uncertainty by formalizing what is involved in making such forecasts and by assessing our methods on their empirical success at this task (e.g. *Dawid* 1984). Statistics provides methods for the characterization of the uncertainty associated with future events or quantities. If compelled by the interpretation of *Ramsey* (1926) and *de Finetti* (2017) that probability is a subjective degree of belief, then the laws and theorems of probability theory will suffice to revise these subjective probabilities (induction/learning) and express predictive uncertainty. Consequently, the probabilistic forecasts in this paper are probability distributions over future events. We wish to quantify the statistical consistency of the forecasts. This is a joint property of the forecasts and the events or values that materialize.

Before we proceed with a discussion of the underlying ideas, theory and principles, we first expose our treatment of probability and explain our notation. We consider a *probabilistic forecast*, P, to be a probability measure on the set of all possible outcomes of an experiment, the so-called sample space Ω. Let Σ be a nonempty collection of subsets of Ω closed under complement, countable unions, and countable intersections and let $\mathscr{P}$ be a convex class of probability measures on (Ω, Σ). A *probabilistic forecast* is a set function $P \in \mathscr{P}$ from Σ to the real number line $\mathbb{R} = (-\infty, \infty)$ which assigns probabilities $P \in [0, 1]$ to any subset $\Sigma \subseteq \Omega$, called an event, in a countably additive manner so that the entire sample space has probability of one. Similarly, the *true forecast* $Q \in \mathscr{P}$ assigns probabilities $Q \in [0, 1]$ to all events $\Sigma \subseteq \Omega$, with unit sum of all probabilities, $Q(\Omega) = 1$.

The measure theoretic treatment of probability allows us to simultaneously treat discrete and continuous probability distributions. In most of our examples, the sample space Ω will consist of real numbers and Ω is synonymous to a random variable of interest, say, next day's peak discharge. The forecaster's task is to quote a distribution $P \in \mathscr{P}$ which characterizes the uncertainty of Ω. Once the watershed has revealed $\omega \in \Omega$ the forecaster will obtain a reward $S(P, \omega)$ depending on both the quoted distribution P and the materialized value ω of the peak discharge. The utility or reward function S is a so-called *scoring rule* and plays a critical role in forecast evaluation. Note that a sign change reverses S to a loss, cost or penalty function. A scoring rule is *strictly proper* if, regardless of the *true* distribution Q of

the peak discharge, the forecaster's expected reward is maximized if and only if he/she quotes $P = Q$. Such scoring rules encourage the forecaster to be honest and report the true set of probabilities. As in *McCarthy* (1956) we assume that the forecaster cannot influence the event predicted beyond the usual experimentation, data collection and modeling involved.

In recent decades, probabilistic forecasting methods have systematically found their way into the hydrologic literature for quantifying predictive uncertainty. The topic of forecast verification has not received such methodical treatment (but with exception of *Laio and Tamea* (2007)) and, as a result, methods for evaluating distribution forecasts, if used at all, are often used haphazardly. To expose the requirements of a meaningful evaluation of distribution forecast and establish the relationship with information theory we have to be comprehensive in our statistical treatment. If deemed appropriate, we spare the main text from lengthy mathematical derivations and defer such technicalities and the description of data, models and computational procedures to appendices. A lowercase italic font (a) is used for scalars, a lowercase bold font ($\mathbf{a}$) for vectors and an uppercase bold font ($\mathbf{A}$) for matrices. Moreover, the standard mathematical convention of functions, say $f(x)$ and $g(x)$, will be used throughout. The symbol ω is used for verifying measurement; thus, we write $\omega_1, \ldots, \omega_n$ for the time series of materialized outcomes. Statistical distributions are designated common symbols. If x is an observable random quantity with a normal distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, we write $x \sim \mathcal{N}(\mu, \sigma^2)$; if the distribution of x is binomial with number of trials $n \in \mathbb{N}_+$ and probability of success $p \in [0, 1]$ we write, $x \sim \mathcal{B}(n, p)$ and use $x \sim \mathcal{U}(a, b)$ for the continuous uniform distribution on the closed-interval $[a, b]$, where $a, b \in \mathbb{R}$ and $a < b$. We designate a probability density function (PDF) with a lowercase f and a cumulative distribution function (CDF) with an uppercase F. Thus, $f_{\mathcal{N}}(x, \mu, \sigma^2)$ and $F_{\mathcal{N}}(x, \mu, \sigma^2)$ signify the PDF and CDF of the normal distribution, respectively. We use the vertical bar "|" for conditional expectation. Thus, $p(x|\boldsymbol{\omega})$ is the conditional PDF of x given the n-vector of verifying measurements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ with $p(x|\boldsymbol{\omega}) \geq 0$ and $\int_\Omega p(x|\boldsymbol{\omega})\, \mathrm{d}x = 1$.

3 Relative entropy

Let us assume that we have exact knowledge of the distribution Q of the measurement, ω, that will materialize at some future time. This situation, albeit uncommon, is a logical starting point for our

discussion. In mathematical statistics, the *Kullback and Leibler* (1951) divergence, $d_{\mathrm{KL}}(Q, P)$, also known as relative entropy or I-divergence (*Csiszar*, 1975), measures in a single value the distance between the *forecast distribution* P and a reference or *true distribution* Q (*Kullback and Leibler*, 1951; *Kullback*, 1959). Divergence is a physical measure of information gain in communication theory and machine learning. For the time being we adapt the common practice in information theory and precede the *probabilistic forecast* P with the *true distribution* Q. These two arguments are swapped later in the context of scoring rules.

3.1 Continuous random variables

For distributions Q and P of a continuous random variable in sample space Ω, the relative entropy from P to Q is defined as follows (*Jaynes*, 1963)

$$d_{\mathrm{KL}}(Q, P) = \mathbb{E}_Q\left[\log_{\mathfrak{b}}\left(\frac{Q(x)}{P(x)}\right)\right] = \int_{x\,\in\,\Omega} Q(x)\log_{\mathfrak{b}}\left(\frac{Q(x)}{P(x)}\right)\mathrm{d}x, \tag{1}$$

where $Q(x)$ and $P(x)$ are the probabilities of Q and P evaluated at the event $x \in \Omega$ and $Q(\Omega) = 1$ and $P(\Omega) = 1$. In applications, Q typically signifies the *true distribution* of data, observations, or possibly, some exactly defined theoretical distribution, whereas P is an approximation thereof obtained from paper-and-pencil calculation, computer modeling and/or other quantitative means. Note that our assignment of the symbols Q and P to the true and forecast distribution, respectively, is reversed to common practice in information theory but consistent with the statistical literature on forecast evaluation. The symbol $\mathfrak{b}$ is used for the base of the logarithm. Common values of $\mathfrak{b}$ are 2, $e = 2.7182818\ldots$ (Euler's number) and 10 and give units of the (relative) entropy in bits (or shannons), nats and hartleys (also referred to as dits or bans), respectively. In what follows we do not affix the base $\mathfrak{b}$ of the logarithm and assume units of bits in our colloquial references to entropy.

The relative entropy $d_{\mathrm{KL}}(Q, P)$ is defined only if the ratio $Q(x)/P(x)$ of the two probability measures, the so-called Radon-Nikodym derivative, $\mathrm{d}Q/\mathrm{d}P$, exists. This means that there does not exist an event $x \in \Omega$ for which $Q(x) > 0$ and $P(x) = 0$, otherwise we must divide by zero. As $\log_{\mathfrak{b}}(a/b) = \log_{\mathfrak{b}}(a) - \log_{\mathfrak{b}}(b)$, the familiar information-theoretic expression for the relative entropy is

$$\begin{aligned} d_{\mathrm{KL}}(Q, P) &= \int_{x\,\in\,\Omega}\Big(Q(x)\log_{\mathfrak{b}}\big(Q(x)\big) - Q(x)\log_{\mathfrak{b}}\big(P(x)\big)\Big)\mathrm{d}x \\ &= \underbrace{\int_{x\,\in\,\Omega} Q(x)\log_{\mathfrak{b}}\big(Q(x)\big)\mathrm{d}x}_{-\mathbb{H}(Q)} - \underbrace{\int_{x\,\in\,\Omega} Q(x)\log_{\mathfrak{b}}\big(P(x)\big)\mathrm{d}x}_{-\mathbb{H}(Q,\,P)} \end{aligned}$$

$$= \mathbb{H}(Q, P) - \mathbb{H}(Q). \tag{2}$$

where $\mathbb{H}(Q, P)$ is the so-called cross-entropy between the *true distribution* Q and the *probabilistic forecast* P and $\mathbb{H}(Q)$ is the Shannon entropy of the *true distribution* Q itself (*Shannon*, 1948a,b). The cross-entropy measures the number of bits required to represent or transmit an average event from distribution Q compared to distribution P. If $Q \neq P$ the cross-entropy $\mathbb{H}(Q, P)$ will always exceed the entropy $\mathbb{H}(Q)$ and $d_{\mathrm{KL}}(Q, P) > 0$. This is known as Gibbs' inequality, a common proof of which is reiterated in Appendix A. If $P = Q$ and our *distribution forecast* P matches exactly the *true distribution* Q then $\mathbb{H}(Q, P)$ will equal $\mathbb{H}(Q)$ and $d_{\mathrm{KL}}(Q, P)$ is zero. Thus, $d_{\mathrm{KL}}(Q, P) = 0$ if and only if $P = Q$. In conclusion, the closer the value of the relative entropy $d_{\mathrm{KL}}(Q, P)$ to zero, the more similar Q and P will be.

If Q and P are strictly continuous on $\mathscr{P}$ and follow a known statistical distribution then it may be possible to derive an analytic expression for their distance, $d_{\mathrm{KL}}(Q, P)$ (see e.g. *Bouhlel and Dziri* 2019). Suppose, for example, that the *true* probability distribution Q of some quantity of interest Ω equals the uniform distribution, $Q = \mathcal{U}(a, b)$ on the closed interval $\Omega = [a, b]$ with PDF, $f_{\mathcal{U}}(\omega, a, b) = 1/(b - a)$ (see Figure 1). The distribution *forecast* P of Ω equals a symmetric triangular distribution, $P = \mathcal{T}(a, b)$, with midpoint $c = (a + b)/2$ and PDF

$$f_{\mathcal{T}}(x, a, b) = \frac{2(b - a) - 2|a + b - 2x|}{(b - a)^2}, \tag{3}$$

where $|\cdot|$ denotes the absolute value. We can enter the analytic expressions of the PDFs of Q and P into the integral of Equation (1) to yield $d_{\mathrm{KL}}(Q, P) = 1 - \log_{\mathfrak{b}}(2)$ in Appendix B. Note that $d_{\mathrm{KL}}(Q, P)$ does not depend on the width of the interval of Ω. Interestingly, for $\mathfrak{b} = 2$ the relative entropy $d_{\mathrm{KL}}(Q, P) = 0$ even though $Q \neq P$. This artifact is easily resolved with a temporary change to the base of the logarithm to yield, $d_{\mathrm{KL}}(Q, P) = \big(1 - \log_e(\mathfrak{b})\big)/\log_e(\mathfrak{b})$. If we then admit $\mathfrak{b} = 2$ we yield $d_{\mathrm{KL}}(Q, P) = 0.4427$ bits. If we swap the arguments Q and P in our derivation, we yield the relative entropy from Q to P instead. In Appendix B, we derive that $d_{\mathrm{KL}}(P, Q) = \log_{\mathfrak{b}}(2) - \frac{1}{2}$ or $d_{\mathrm{KL}}(P, Q) = 0.2787$ bits.

The analytic exercise demonstrates that the relative entropy does not satisfy the symmetry axiom of a metric $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ in a metric space $\mathcal{M}$. Indeed, the relative entropy $d_{\mathrm{KL}}(Q, P)$ from P to Q does not equal its counterpart $d_{\mathrm{KL}}(P, Q)$ from Q to P. But this is not all. The relative entropy also does not satisfy the triangle inequality, $d_{\mathrm{KL}}(Q, P) \leq d_{\mathrm{KL}}(Q, R) + d_{\mathrm{KL}}(R, P)$, which we will confirm later with an analytic example. This fourth and last axiom of a metric d in $\mathcal{M}$ states that the distance from Q to

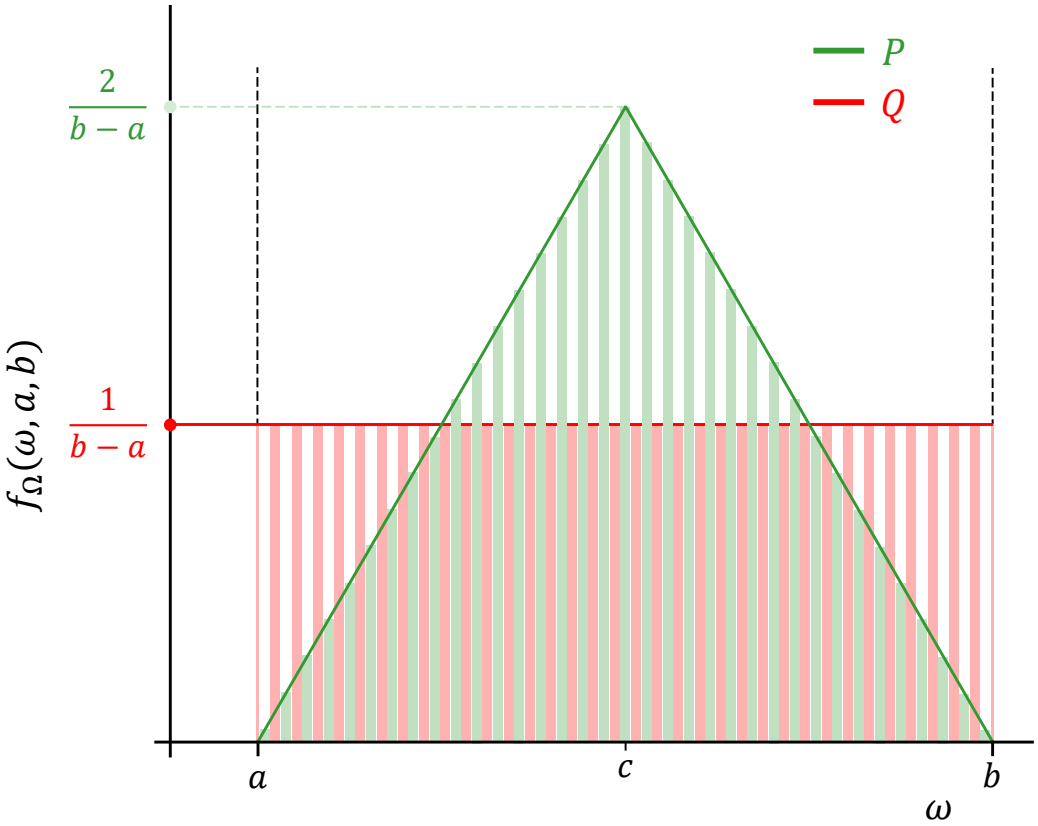

Figure 1: Probability density function of the uniform true distribution $Q = \mathcal{U}(a, b)$ and symmetric triangular forecast distribution $P = \mathcal{T}(a, b)$ of quantity $\Omega = [a, b]$.

P with a detour through R cannot be less than the geodesic distance (shortest path) between Q and P. This implies that relative entropy is not a metric or distance function in an Euclidean space with its usual physical or metaphorical notion of length but rather should be thought of as a divergence of two distributions. This statistical distance is commonly referred to in the statistical literature as the *Kullback and Leibler* (1951) divergence, hence our use of the subscript KL in $d_{\text{KL}}(Q, P)$. Furthermore, the fact that $d_{\text{KL}}(Q, P)$ is strictly nonnegative and zero only when $P = Q$ suggests its interpretation as a so-called *Bregman* (1967) *distance* or *divergence* of Q and P. We point forward to Figure 5 for a graphical exposition of the Bregman divergence but refrain from further explanation until we get to the topic of scoring rules in Section 5.

Bregman divergences are of paramount importance in scientific forecast evaluation as they ensure *strict propriety* of scoring rules and their associated divergence functions. Thus, $d_{\text{KL}}(Q, P) > 0$ if $P \neq Q$ and $d_{\text{KL}}(Q, P) = 0$ if and only if $P = Q$. Propriety, "*. . . the state or quality of conforming to conventionally accepted standards of behavior or morals* and/or *. . . conformity with what is required by a rule, principle, etc.*" follows directly from Jensen's inequality (the convex transformation of a mean is less than or equal

to the mean applied after convex transformation) and incentives a forecaster to issue $P = Q$ rather than any $P \neq Q$. In other words, only *strictly proper* scoring rules will lead us to the *true forecast distribution*, $Q \in \mathscr{P}$. The minimum divergence, $d(P, Q) = 0$ is achieved when $P = Q$, and this minimum is unique. This encourages a forecaster to be honest and to volunteer his or her true beliefs. *Proper* scoring rules also yield a minimum at $P = Q$ but this minimum may not be unique. A symmetrized variant of the *Kullback and Leibler* (1951) divergence

$$d_{\mathrm{J}}(P, Q) = d_{\mathrm{KL}}(Q, P) + d_{\mathrm{KL}}(P, Q) = \int_{x \in \Omega} \big(Q(x) - P(x)\big) \left(\frac{Q(x)}{P(x)}\right) \mathrm{d}x \tag{4}$$

is also known as J-divergence in honor of Sir Harold Jeffreys (*Jeffreys*, 1946) and used widely in pattern recognition and computer vision applications (*Chang et al.*, 2009; *Zheng and You*, 2013). Divergences are sometimes called divergence functions or discrepancy functions, or validation metrics (*Liu et al.*, 2011), even though our example has shown that they may not necessarily satisfy the requirements of a metric in mathematical sense (*Thorarinsdottir et al.*, 2013).

If the *true distribution* Q and *probabilistic forecast* P are multivariate normal in $\mathbb{R}^b$ with means, $\boldsymbol{\mu}_Q, \boldsymbol{\mu}_P \in \mathbb{R}^{b \times 1}$, and non-singular $b \times b$ covariance matrices $\boldsymbol{\Sigma}_Q$ and $\boldsymbol{\Sigma}_P$, respectively, then the relative entropy in units of nats becomes (see Appendix C)

$$d_{\mathrm{KL}}\big(\mathcal{N}_b(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q), \mathcal{N}_b(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)\big) = \tfrac{1}{2}\big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - b + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + (\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)\big] \tag{5}$$

where $|\cdot|$ is the determinant operator, the symbol $\top$ denotes transpose and the trace function, $\mathrm{tr}(\mathbf{A})$, returns the sum of the elements on the main diagonal of the $b \times b$ matrix, $\mathbf{A} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0$. Equation (5) is also known as the *Dawid and Sebastiani* (1999) divergence, $d_{\mathrm{DDS}}(P, Q)$, more of which later in the context of multivariate scoring rules. Note that the arguments in $d_{\mathrm{DDS}}(P, Q)$ have reversed to satisfy convention used in the statistical literature. This analytic expression of the KL-divergence is *strictly proper* only for normal probability measures, which are uniquely characterized by their respective means and covariance matrices. If Q and P are univariate normal then Equation (5) reduces to

$$d_{\mathrm{KL}}\big(\mathcal{N}_1(\mu_Q, \sigma_Q^2), \mathcal{N}_1(\mu_P, \sigma_P^2)\big) = \tfrac{1}{2}\log_e\left(\frac{\sigma_P^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + (\mu_Q - \mu_P)^2 - \sigma_P^2}{2\sigma_P^2}. \tag{6}$$

The relative entropy $d_{\mathrm{KL}}(Q, P)$ is not defined when $P(x) = 0$ and $Q(x) > 0$. To illustrate this, suppose $Q = \mathcal{N}(\mu_Q, \sigma_Q^2)$ and $P = \mathcal{U}(a_P, b_P)$ on a finite sample space $\Omega = [a, b]$. In Appendix D we derive an

analytic expression for the KL-divergence of Q from P

$$d_{\mathrm{KL}}\big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{U}(a_P,b_P)\big) = \mathbb{H}\big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{U}(a_P,b_P)\big) - \mathbb{H}\big(\mathcal{N}(\mu_Q,\sigma_Q^2)\big)$$
$$= \log_e(b_P - a_P) - \tfrac{1}{2}\log_e(2e\pi\sigma_Q^2), \tag{7}$$

and may go to infinity with support of P on the extended real line, $\overline{\mathbb{R}}$. Note that if we link Q and P using $\sigma = (b_P - a_P)/\nu$ with $\nu \in \mathbb{R}_+$ then the relative entropy simplifies to $d_{\mathrm{KL}}\big(Q,P\big) = \log_e(\nu) - \frac{1}{2}\log_e(2e\pi)$.

Now suppose that we have three distributions, $Q = \mathcal{N}(\mu_Q,\sigma_Q^2)$, $P = \mathcal{U}(a_P,b_P)$ and $R = \mathcal{N}(\mu_R,\sigma_R^2)$. Equations (6) and (7) will help demonstrate that the relative entropy does not honor the triangle inequality, $d_{\mathrm{KL}}(Q,P) \leq d_{\mathrm{KL}}(Q,R) + d_{\mathrm{KL}}(R,P)$. Indeed, we yield

$$d_{\mathrm{KL}}\big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{U}(a_P,d_P)\big) \leq d_{\mathrm{KL}}\big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{N}(\mu_R,\sigma_R^2)\big) + d_{\mathrm{KL}}\big(\mathcal{N}(\mu_R,\sigma_R^2),\mathcal{U}(a_P,d_P)\big)$$
$$\log_e(b_P - a_P) - \tfrac{1}{2}\log_e(2e\pi\sigma_Q^2) \leq \tfrac{1}{2}\log_e\left(\frac{\sigma_R^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + (\mu_Q - \mu_R)^2 - \sigma_R^2}{2\sigma_R^2} + \log_e(b_P - a_P) - \tfrac{1}{2}\log_e(2e\pi\sigma_R^2)$$
$$-\tfrac{1}{2}\log_e(\sigma_Q^2) \leq \tfrac{1}{2}\log_e\left(\frac{\sigma_R^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + (\mu_Q - \mu_R)^2 - \sigma_R^2}{2\sigma_R^2} - \tfrac{1}{2}\log_e(\sigma_R^2)$$
$$\implies \sigma_Q^2 + (\mu_Q - \mu_R)^2 - \sigma_R^2 \geq 0. \tag{8}$$

The trivial example, $\sigma_Q^2 < \sigma_R^2$ and $\mu_Q = \mu_R$, violates the triangle inequality. To convey the fundamental asymmetry in the relation between Q and P it is common to refer to $d_{\mathrm{KL}}(Q,P)$ as the relative entropy of Q with respect to P or the information gain from Q over P.

3.2 Discrete random variables

For discrete probability distributions Q and P the sample space, $\Omega = \{\omega_1,\ldots,\omega_m\}$ consists of a finite number m of mutually exclusive and collectively exhaustive events, ω, and a probabilistic forecast is a probability vector $\mathbf{p} = (p_1,\ldots,p_m)^\top$ defined on the convex class $\mathscr{P} = \mathscr{P}_m$

$$\mathscr{P}_m = \big\{\mathbf{p} = (p_1,\ldots,p_m)^\top : p_1 + \ldots + p_m = 1 \text{ and } p_k \geq 0 \text{ for all } k\big\}. \tag{9}$$

where $\top$ denotes transpose. It is further assumed that the vector $\mathbf{q} = (q_1,\ldots,q_m)^\top$ reports the *true probabilities* of the m events, $\big\{\mathbf{q} \in \mathbb{R}_+^{m\times 1} : \mathbf{1}^\top\mathbf{q} = 1\big\}$, where $\mathbf{1}_m$ is a $m \times 1$ vector of ones.

For discrete probability distributions P and Q on sample space Ω, Equation (2) reduces to

$$d_{\mathrm{KL}}(\mathbf{q},\mathbf{p}) = \mathbb{H}(\mathbf{q},\mathbf{p}) - \mathbb{H}(\mathbf{q}), \tag{10}$$

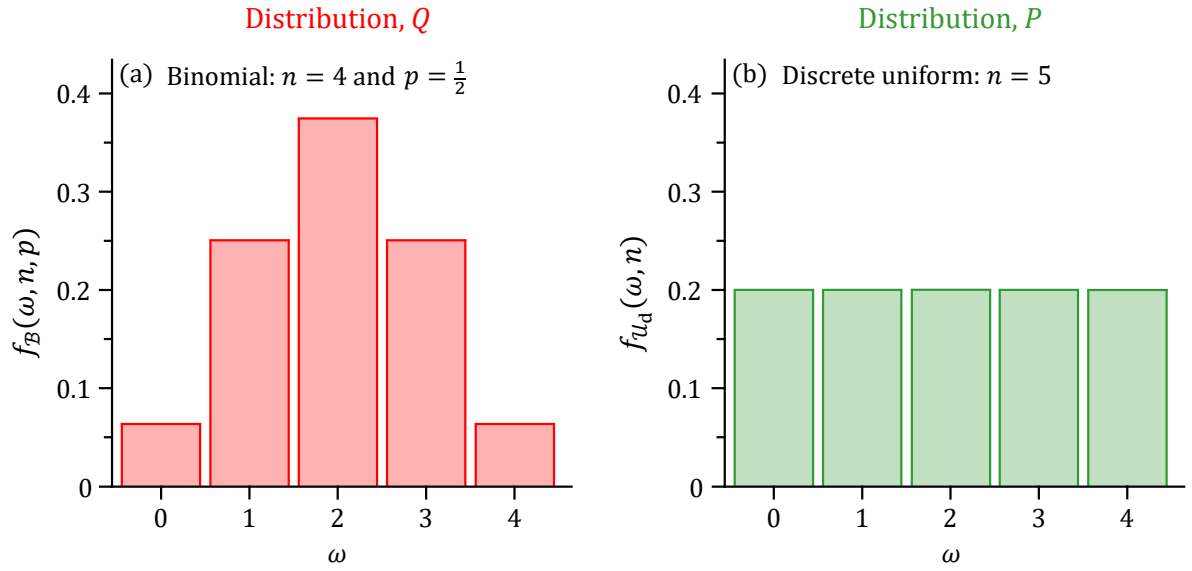

Figure 2: Illustration of the computation of the relative entropy $d_{\mathrm{KL}}(Q, P)$ on a sample space $\Omega = \{0, 1, 2, 3, 4\}$: (a) *true distribution*, $Q = \mathcal{B}(n, p)$, and (b) *forecast distribution*, $P = \mathcal{U}_{\mathrm{d}}(n)$. According to data, Ω follows a binomial distribution with $n = 4$, $p = \frac{1}{2}$ and PMF, $f_{\mathcal{B}}(\omega, n, p) = c(n, \omega)p^{\omega}(1-p)^{n-\omega}$, where $c(a, b) = a!/b!(a-b)!$ denotes the binomial coefficient and ! is the factorial function. Theory predicts a discrete uniform distribution for ω with equal density, $f_{\mathcal{U}_{\mathrm{d}}}(\omega, n) = 1/n$, for all $n = 5$ values.

and the integral of the relative entropy from P to Q becomes a nonnegative sum

$$d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p}) = \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{q_k}{p_k}\right), \tag{11}$$

which is equivalent to

$$d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p}) = -\sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{p_k}{q_k}\right). \tag{12}$$

To illustrate the computation and interpretation of the KL-divergence, please consider the probability mass functions (PMFs) of Q and P depicted in Figure 2. The *true distribution* Q of quantity Ω is a binomial distribution with $n = 4$ and $p = 0.5$ and the *distribution forecast* P is discrete uniform with equal density for $\omega = (0, 1, \ldots, 4)$. The relative entropy, $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p})$, may now be computed

$$\begin{aligned} d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p}) &= \sum_{\omega=0}^{4} f_{\mathcal{B}}(\omega, 4, \tfrac{1}{2}) \log_{\mathfrak{b}}\left(\frac{f_{\mathcal{B}}(\omega, 4, \frac{1}{2})}{f_{\mathcal{U}_{\mathrm{d}}}(\omega, 5)}\right) \\ &= \tfrac{1}{16}\log_{\mathfrak{b}}\left(\frac{\frac{1}{16}}{\frac{1}{5}}\right) + \tfrac{1}{4}\log_{\mathfrak{b}}\left(\frac{\frac{1}{4}}{\frac{1}{5}}\right) + \tfrac{6}{16}\log_{\mathfrak{b}}\left(\frac{\frac{6}{16}}{\frac{1}{5}}\right) + \tfrac{1}{4}\log_{\mathfrak{b}}\left(\frac{\frac{1}{4}}{\frac{1}{5}}\right) + \tfrac{1}{16}\log_{\mathfrak{b}}\left(\frac{\frac{1}{16}}{\frac{1}{5}}\right) \\ &= \tfrac{1}{8}\log_{\mathfrak{b}}\left(\tfrac{5}{16}\right) + \tfrac{1}{2}\log_{\mathfrak{b}}\left(\tfrac{5}{4}\right) + \tfrac{3}{8}\log_{\mathfrak{b}}\left(\tfrac{30}{16}\right), \end{aligned} \tag{13}$$

which with base of the logarithmic function equal to two leads to $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p}) \approx 0.2913$ bits. If we divide the so-obtained value of $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p})$ by $\log_2(e)$ then we yield the KL-divergence in units of nats. Note

that if $q_k = 0$ for some $\omega_k \in \Omega$, the summand, $q_k \log_{\mathfrak{b}}\big(q_k\big)$, is set to zero in accordance with the limit, $\lim_{q \downarrow 0} q \log_{\mathfrak{b}}(q) = 0$. If we swap the discrete distributions of Q and P we yield the so-called reverse KL-divergence

$$
\begin{aligned}
d_{\mathrm{KL}}(\mathbf{p}, \mathbf{q}) &= \sum_{\omega=0}^{4} f_{\mathcal{U}_{\mathrm{d}}}(\omega, 5) \log_{\mathfrak{b}}\left(\frac{f_{\mathcal{U}_{\mathrm{d}}}(\omega, 5)}{f_{\mathcal{B}}(\omega, 4, \frac{1}{2})}\right) \\
&= \tfrac{1}{5} \log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{16}}\right) + \tfrac{1}{5} \log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{4}}\right) + \tfrac{1}{5} \log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{6}{16}}\right) + \tfrac{1}{5} \log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{4}}\right) + \tfrac{1}{5} \log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{16}}\right) \\
&= \tfrac{2}{5} \log_{\mathfrak{b}}\Big(\tfrac{16}{5}\Big) + \tfrac{2}{5} \log_{\mathfrak{b}}\Big(\tfrac{4}{5}\Big) + \tfrac{1}{5} \log_{\mathfrak{b}}\Big(\tfrac{16}{30}\Big), \qquad (14)
\end{aligned}
$$

which is equal to $d_{\mathrm{KL}}(\mathbf{p}, \mathbf{q}) \approx 0.3611$ bits. This again confirms that $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p})$ is not a metric but rather a (Bregman) divergence, more of which later.

Expressed in the language of Bayesian inference, $d_{\mathrm{KL}}(Q, P)$ is a measure of the information gained by revising one's beliefs from the prior probability distribution P to the posterior probability distribution Q. In other words, it is the amount of information lost when P is used to approximate Q (*Burnham and Anderson*, 2002). We used the KL-divergence in our earlier work to quantify the gain of information when moving from the prior to the posterior parameter distribution (*Scharnagl et al.*, 2010). For example, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ equals a d-vector of unknown coefficients of some mathematical model or hypothesis, $\mathcal{H}$, with prior distribution, $p(\boldsymbol{\theta}|\mathcal{H})$ on the parameter (sample) space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$. When new data, $\mathcal{D}$, become available, we can update $p(\boldsymbol{\theta}|\mathcal{H})$ to a posterior distribution, $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$, using *Bayes* (1763) theorem

$$
p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})}. \qquad (15)
$$

where $p(\mathcal{D}|\mathcal{H})$ is a normalization constant and $L(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) \equiv p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H})$ denotes the likelihood. The entropy of the posterior distribution

$$
\mathbb{H}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})\big) = -\sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) \log_{\mathfrak{b}}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})\big), \qquad (16)
$$

may be less than, equal to, or greater than the entropy $\mathbb{H}\big(p(\boldsymbol{\theta}|\mathcal{H})\big)$ of the prior distribution. The relative entropy

$$
d_{\mathrm{KL}}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}), p(\boldsymbol{\theta}|\mathcal{H})\big) = \mathbb{H}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}), p(\boldsymbol{\theta}|\mathcal{H})\big) - \mathbb{H}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})\big), \qquad (17)
$$

measures the added message length (bits) that an original code with the prior distribution $p(\boldsymbol{\theta}|\mathcal{H})$ would require relative to a new code based on the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$. Or in the words of *Cover and Thomas* (2006) (Chapter 2, Page 19) "*... if we knew the true distribution Q of the random*

variable, we could construct a code with average description length $\mathbb{H}(Q)$*. If, instead, we used the code for a distribution* P*, we would need* $\mathbb{H}(Q) + d_{\mathrm{KL}}(Q, P)$ *bits on the average to describe the random variable*", where Q and P correspond to the posterior and prior distributions, respectively. Thus, $d_{\mathrm{KL}}\big(p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}), p(\boldsymbol{\theta}|\mathcal{H})\big)$, is synonymous with the information gain about $\boldsymbol{\theta}$ that has been learned by the new data $\mathcal{D}$. Note, that $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H})$ (and $p(\boldsymbol{\theta}|\mathcal{H})$ for that matter) must equal a probability measure on $\boldsymbol{\Theta}$ and, thus, $\sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) = 1$. Posterior distributions derived from MCMC methods lend itself well to computation of the Shannon entropy via Equation (16). If regularity conditions are satisfied then the Markov chain will visit each event $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}$ with frequency n_k proportional to its so-called Boltzmann weight, $Q(\boldsymbol{\theta}_k) \sim \exp(-U(\boldsymbol{\theta}_k)/k_{\mathrm{B}}T)$, where $U(\boldsymbol{\theta}_k)$ is the potential energy at event $\boldsymbol{\theta}_k$, T is the absolute temperature and $k_{\mathrm{B}} = 1.380649 \times 10^{-23}\ \mathrm{J\,K^{-1}}$ signifies the Boltzmann constant. The relative frequency of each of m posterior samples (events or micro-states) visited by the Markov chain is given by the Boltzmann distribution

$$Q(\boldsymbol{\theta}_k) = \frac{n_k}{m} = \frac{\exp\big(-U(\boldsymbol{\theta}_k)/k_{\mathrm{B}}T\big)}{\sum_{i=1}^{m} \exp\big(-U(\boldsymbol{\theta}_i)/k_{\mathrm{B}}T\big)}, \tag{18}$$

and defines a probability measure $Q \in \mathscr{P}$ with $Q(\boldsymbol{\Theta}) = 1$. For arbitrary distributions Q and P in $\mathbb{R}^d$ we must evaluate Equation (11) at a collection of points (events), $\mathbf{x} \in \Omega$. This is equivalent to numerical integration and different approaches exist to do so efficiently including dimension-adaptive quadrature rules (*Smolyak*, 1963; *Gerstner and Griebel*, 1998, 2003; *Jakeman and Narayan*, 2018), quasi-random sequences (*Halton*, 1960; *Niederreiter*, 1992; *Sobol' and Shukhman*, 1995) or pseudo-random sampling (*Hammersley and Handscomb*, 1960; *Volpi et al.*, 2017).

4 Insufficient scoring rules for density forecasts

In most practical situations we do not have knowledge of the underlying distribution $Q \in \mathscr{P}$ which materializes with the measurement $\omega \in \Omega$. Then, we must find ways to evaluate the distribution forecast, $P \in \mathscr{P}$, using only a single verifying observation, ω_{st}, at given *s*pace and *t*ime coordinates. The data may arrive *en bloc* in simulation mode or in natural (sequential) order in a forecasting problem. Then, it is the forecaster's task, at any time, to produce a distribution forecast for the next observation. The success at this task can be judged by using methods from probability forecasting. Before we address the intriguing topic of scoring rules, we must first review past developments that led to current perspectives.

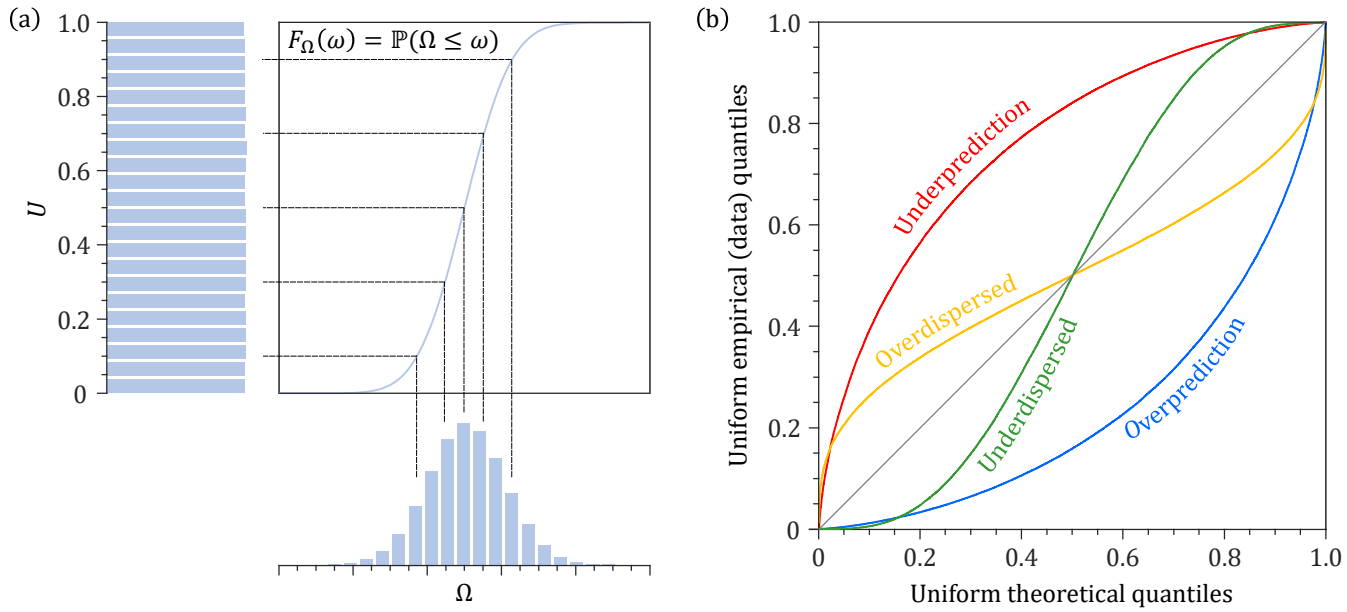

Figure 3: (a) Illustration of the probability integral transform and (b) interpretation of the so-called quantile-quantile plot (adapted from *Laio and Tamea* (2007)).

This is a necessary step into understanding the strengths and weaknesses of current metrics used in hydrology and the need for scoring rules.

The probability integral transform of *Dawid* (1984) is one of the earliest methods for evaluating the statistical coherency and association between a time series of forecast distributions $P_1, \ldots, P_n$ and observed outcomes $\omega_1, \ldots, \omega_n$. Integral transforms have a long history dating back to at least the *Rosenblatt* (1952) transformation and turn a vector of dependent random variables into a vector of independent uniform distributed values (see Figure 3a). Let Ω be a real-valued continuous random variable with CDF, $F_\Omega(\omega) = \mathbb{P}(\Omega \leq \omega)$. Then, random variable $U = F_\Omega(\omega)$ has a standard uniform distribution. Thus, if $\omega_1, \ldots, \omega_n$ are samples of Ω (dependent or not) then the values $u_i = F_\Omega(\omega_i)$; $i = (1, \ldots, n)$ of U will be uniformly distributed on the unit interval. Hence, the probability integral transform reduces the assessment of F_Ω to the question whether the sequence of u's behaves as a random sample of $\mathcal{U}[0, 1]$. Fig. 3b illustrates the consequences of using an incorrect distribution for Ω on the relationship between the theoretical quantiles of the standard uniform distribution and the quantiles of the empirical distribution function of the sampled data, $\omega_1, \ldots, \omega_n$. This so-called predictive quantile-quantile (Q-Q) plot (*Dawid*, 1984; *Casella and Berger*, 2002) is a common verification tool for probabilistic forecasts of meteorological (*Gneiting and Raftery*, 2007) and hydrologic (*Laio and Tamea*, 2007; *Thyer et al.*, 2009; *Renard et al.*, 2011) variables. This graph helps diagnose errors in ensemble mean (bias) and spread (dispersion) as

causes for the deviation from the theoretical 1:1 line for perfect distribution forecasts.

The uniformity of the u's can be tested formally using the Kolmogorov-Smirnov statistic (*Kolmogorov*, 1933; *Smirnov*, 1948) and we can inspect the u's for any sign of non-independence or a trend using the uniform condition test (*Cox and Lewis*, 1966). To simplify pairwise comparison of Q-Q plots, we can concatenate the deviations of the u's from the 1:1 line into a single numerical value or index. For example, *Renard et al.* (2010) introduced the so-called reliability R_{l} as affine transformation of the taxicab distance between the empirical quantiles, $u_t = F_{P_t}(\omega_t);\ t = (1, \ldots, n)$ and the corresponding quantiles of the standard uniform distribution

$$R_{\mathrm{l}} = \frac{2}{n} \sum_{t=1}^{n} \left| u'_t - \frac{t}{n} \right|, \tag{19}$$

where $u'_1, \ldots, u'_n$ denote the ordered values of $u_1, \ldots, u_n$, $|\cdot|$ is the absolute value operator, and the true quantiles jump up by $1/n$ at each of the n observations. The multiplier of two scales the index to the closed interval between 0 (most reliable) and 1 (least reliable). Note that Equation (19) is at odds with the formal definition of reliability derived from reliability diagrams of probability forecasts for dichotomous events (*Dimitriadis et al.*, 2021). This formal definition is presented in Section 7.

The predictive Q-Q plot provides a simple and assumption-free graphical summary of the reliability of distribution forecasts. This graph has become commonplace in the hydrologic literature for evaluating predictive distributions of precipitation (*Renard et al.*, 2011) and discharge (*Thyer et al.*, 2009; *Renard et al.*, 2011; *Evin et al.*, 2013), compare, contrast and rank different formulations of the likelihood function (*Evin et al.*, 2014; *McInerney et al.*, 2017, 2019) and characterize model input, output and structural errors (*Renard et al.*, 2010). But the Q-Q plot and reliability index R_{l} should not be used as sole determinants of the quality of distribution forecasts (*Gneiting et al.*, 2007; *Renard et al.*, 2011). In a thought-provoking example, *Hamill* (2001) demonstrated that the probability integral transform may yield a uniform histogram on the unit interval, even if every single forecast is biased. Thus, uniformity of the PIT values is a necessary but not a sufficient condition for determining ensemble reliability.

To address these limitations, *Gneiting et al.* (2007) proposed a more diagnostic approach to the evaluation of predictive performance that is based on maximizing the sharpness of the distribution forecasts subject to calibration. Within this context, sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. Calibration refers to the statistical consistency between the forecast distributions and the observations and is a joint property of the predictions and the outcomes

that materialize. The sharpness principle has a theoretical underpinning under the assumption of autocalibration (*Tsyplakov*, 2011) and has become a useful working paradigm for probabilistic forecasting and forecast evaluation.

Table 1 presents three other measures that have found application and use in hydrology for evaluating the accuracy of probabilistic forecasts. The coefficient of variation C_v is a dimensionless measure of the

Table 1: Time-averaged performance measures, $\overline{M}(\mathbf{P}, \boldsymbol{\omega})$, of distribution forecasts $\mathbf{P} = \{P_1, \ldots, P_n\}$ and verifying observations $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$.

Performance measure	Symbol	$\overline{M}(\mathbf{P}, \boldsymbol{\omega})$	Reference
Reliability[a]	R_l	$\frac{2}{n}\sum_{t=1}^{n}\left\lvert u'_t - \frac{t}{n}\right\rvert$	*Renard et al.* (2010)
Coefficient of variation[b]	C_v	$\frac{1}{n}\sum_{t=1}^{n}\frac{\sigma_{P_t}}{\mu_{P_t}}$	*Evin et al.* (2013)
Coverage[c,d]	C	$\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\{l_t \le \omega_t \le u_t\}$	*Dunsmore* (1968)
Width[d]	W	$\frac{1}{n}\sum_{t=1}^{n}(u_t - l_t)$	*Raftery et al.* (2005)

[a] $u_t = F_{P_t}(\omega_t)$; $t = (1, \ldots, n)$ and u's are sorted in ascending order to yield $u'_1, \ldots, u'_n$; $F_P(x)$ is the cumulative distribution function of P

[b] Replace with sample mean, m_P, and sample standard deviation, s_P, in case of an ensemble forecast

[c] The indicator function $\mathbb{1}\{a\}$ returns 1 if a is true and zero otherwise

[d] Lower, $l_t = F_{P_t}^{-1}(\alpha/2)$, and upper, $u_t = F_{P_t}^{-1}(1-\alpha/2)$, endpoints of $100(1-\alpha)\%$ prediction interval at $\alpha \in (0,1)$ significance level; $F_P^{-1}(x)$ is the inverse cumulative distribution (quantile) function of P

extent of variability (dispersion) in relation to the mean of the distribution. This measure should only be computed for data measured on so-called ratio scales which have a meaningful zero point. This measure is related to the conjectured sharpness principle of *Raftery et al.* (2005). Smaller values of the C_v are preferred subject to the intervals having the right coverage. The coverage, C, equals the fraction of observations inside the $\gamma = 100(1-\alpha)\%$ prediction intervals. To be statistically meaningful and robust, C should equal $1-\alpha$ at a significance level $\alpha \in (0,1)$. The width, W, measures the average size of the $\gamma\%$ prediction intervals. Despite their intuitive appeal and ease of interpretation, none of the performance metrics of Table 1 provides a complete evaluation of the forecast density, and may even be invariant to the *true distribution* Q (see Figure 4). Specifically, the width W and coefficient of variation C_v are properties of the predictive distribution only and, thus, do not guarantee honest forecasts. This invariance to the materialized outcome, ω, is easily illustrated using Fig. 4a and b. A shift of the distribution forecasts to

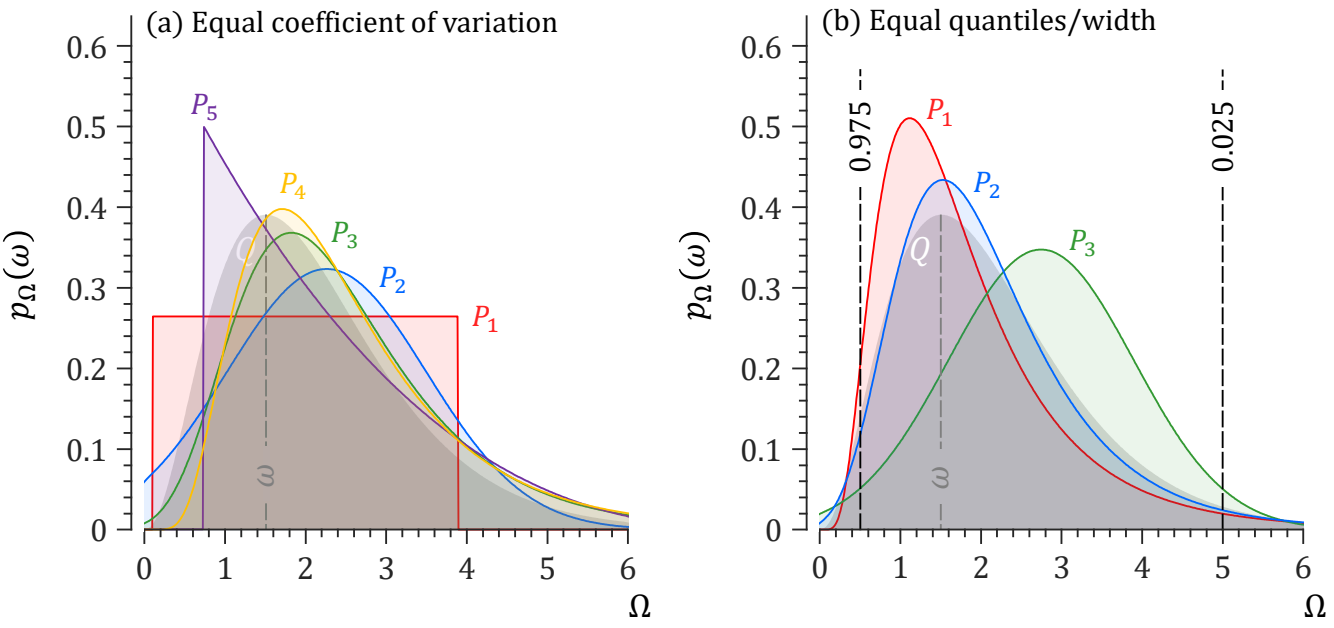

Figure 4: Hypothetical *true distribution* $Q = \mathcal{G}(3.36, 0.64)$ (gray) and different *probabilistic forecasts* P (in color) with an equal (a) coefficient of variation, $C_\text{v} = 0.546$, and (b) width, $W = 4.5$, at $\alpha = 0.05$ using using lower and upper quantiles of 0.5 and 5.0, respectively. The peak of the *true distribution* is equal to the verifying observation, ω. The distribution forecasts of the left graph are used in later studies: $P_1 = \mathcal{U}(0.11, 3.89)$, $P_2 = \mathcal{N}(2.26, 1.52)$, $P_3 = \mathcal{GEV}(0.04, 1, 1.86)$; $P_4 = \mathcal{LN}(0.80, 0.26)$ and $P_5 = \mathcal{GP}(-0.28, 2, 0.73)$.

the right will substantially reduce their overlap with the *true* distribution, but not affect in any way their widths and coefficients of variation, which will remain fixed at $W = 4.5$ and $C_\text{v} = 0.546$. The reliability and coverage measure only two aspects of the statistical consistency between the distributional forecasts and the observations. Even if the coverage C is adequate at a given significance level α, this does not guarantee accurate prediction intervals for other confidence levels (*Christoffersen*, 1998). This necessitates the simultaneous conditional calibration of many different quantile forecasts, which is a daunting task.

As should be evident from our discussion, the R_l, C_v, C and W performance metrics of Table 1 measure different and complementary aspects of the *distribution forecast* P. This is diagnostically appealing (we address this topic in Section 7), but frustrates forecast evaluation as we cannot aggregate the R_l, C_v, W and C criteria into a single performance index without assigning arbitrary weights. One can adopt "Paretian" theory of general equilibrium and use non-dominated sorting of the performance metrics to rank the distribution forecasts (*McInerney et al.*, 2017, 2019). This is a pragmatic solution but synonymous to an arbitrary and incomplete evaluation of distribution forecasts.

5 Scoring Rules

Scoring rules are indispensable in our search for the *true* forecast distribution Q, nevertheless, have not yet entered mainstream use in the hydrologic community. In terms of elicitation, the role of scoring rules is to encourage the assessor to make careful assessments and to be honest (*Garthwaite et al.*, 2005). In terms of evaluation, scoring rules measure the quality of the probabilistic forecasts, reward probability assessors for forecasting jobs, and rank competing forecast procedures. Meteorologists refer to this broad task as *forecast verification*. We briefly review the most important statistical concepts and explicate our notation.

5.1 Definition

A *probabilistic forecast* is any probability measure $P \in \mathscr{P}$ on Ω, the sample space, $P(\Omega) = 1$. A *point forecast* is a functional, i.e. a set-valued mapping $P \to T(P) \subseteq \Omega$, from a class of probability distributions P to the real line $\mathbb{R}$ with the mean or expectation, quantiles and expectiles being key examples. Per definition 1 on Page 4, a *scoring function* $s(x, \omega)$ is any real-valued function $s : \Omega \times \Omega \to \mathbb{R}$ which quantifies the reward to the forecaster of a point forecast $x \in \Omega$ and verifying observation $\omega \in \Omega$. Then, according to definition 2, a *scoring rule* $S(P, \omega)$ is any extended real-valued and $\mathscr{P}$-quasi-integrable function $S : \mathscr{P} \times \Omega \to \overline{\mathbb{R}} \equiv [-\infty, \infty]$ for all $P \in \mathscr{P}$ which measures the reward when forecast P is issued and the observation $\omega \in \Omega$ materializes. The function $\mathcal{S}(P, Q) : \mathscr{P} \times \mathscr{P} \to \overline{\mathbb{R}}$ is equal to the expected value of $S(P, \omega)$ under the *true distribution* Q of ω.

5.2 Theory

The expected score of the probabilistic forecast P under the *true* distribution $Q \in \mathscr{P}$ is defined as follows

$$\mathcal{S}(P, Q) = \mathbb{E}_{\omega \overset{\mathcal{D}}{\sim} Q}\big[S(P, \omega)\big] = \int_{\Omega} S(P, \omega) \mathrm{d}Q(\omega) \tag{20}$$

where the symbol $\overset{\mathcal{D}}{\sim}$ means "*distributed according to*". Note that the order of the arguments has reversed with respect to the convention used in information theory. The *probabilistic forecast* P precedes the *true distribution* Q. This is in line with the statistical forecasting literature (*Gneiting and Raftery*, 2007; *Bröcker*, 2009). It is our convention that a higher score indicates a good forecast and, thus, our scoring

rules $S(P,Q)$ are positively oriented and defined as reward functions. Then, a scoring rule S is said to be *proper* relative to $\mathscr{P}$ if

$$\mathcal{S}(P,Q) \leq \mathcal{S}(Q,Q) \quad \text{for all } P,Q \in \mathscr{P}, \tag{21}$$

and is considered *strictly proper* if the above condition holds with equality if and only if $P = Q$. This implies that a *strictly proper* score rule is a sufficient condition, whereas a *proper* score rule is a necessary but not sufficient condition. In plain words, if $\mathcal{S}(P,Q)$ is a *strictly proper* score rule, then the larger its value, the closer the distribution of P will be to that of Q. This is not true for *proper* scoring rules, which can attain a maximum score even if $P \neq Q$ (*Vrugt et al.*, 2022). Based on early recommendations by *Brier* (1950) and *Shuford et al.* (1966), we restrict attention to the class of *proper* scoring rules. This includes *strictly proper* scoring rules.

A scoring rule $S : \mathscr{P} \times \Omega \to \overline{\mathbb{R}}$ is *regular* relative to the class $\mathscr{P}$ if $\mathcal{S}(P,Q)$ is real-valued for all $P,Q \in \mathscr{P}$, except possibly that $\mathcal{S}(P,Q) = \infty$ if $P \neq Q$. If $\mathcal{S}$ is *regular* and *proper*, then the excess score

$$d(P,Q) = \mathcal{S}(Q,Q) - \mathcal{S}(P,Q), \quad P,Q \in \mathscr{P}, \tag{22}$$

measures the discrepancy of the *probabilistic forecast* $P \in \mathscr{P}$ from the *true distribution* $Q \in \mathscr{P}$. This is a *divergence function* alike the relative entropy in Equations (1) and (11) and equal to a measure of the difference between two points defined in terms of a continuously-differentiable expected score function, $H(P) : \mathscr{P} \to \mathbb{R}$. For positively oriented *proper* scoring rules $H(P)$ is equal to the pointwise supremum (least upper bound) over the convex class of probability measures Q on $\mathscr{P}$ (*Gneiting and Raftery*, 2007)

$$\begin{aligned} H(P) &= \sup_{Q \in \mathscr{P}} \mathcal{S}(Q,P), \qquad P \in \mathscr{P}, \\ &= \mathcal{S}(P,P). \end{aligned} \tag{23}$$

and is convex on $\mathscr{P}$ since $\mathcal{S}(Q,P)$ is linear in P (*Rockafellar*, 1970). The statement holds with proper replaced by strictly proper, and convex replaced by strictly convex. If the sample space is finite and the expected score function $H(P)$ smooth, then $d(P,Q)$ equals the *Bregman* (1967) divergence associated with the convex function $H(P)$.

To understand the relationship between $H(P)$, $\mathcal{S}(P,Q)$ and $d(P,Q)$ Figure 5 displays the entropy function of the *strictly proper* quadratic score for a binary event (*rain* or *no rain*)

$$\begin{aligned} H_{\text{QS}}(p) &= p^2 + (1-p)^2 \\ &= 2p^2 - 2p + 1 \\ \implies H_{\text{QS}}(p) &\mapsto p(p-1), \end{aligned} \tag{24}$$

where $p \in [0, 1]$ is the quoted *probability* for *rain* to occur and the symbol $\mapsto$ signifies affine transformation. This is equivalent to a sample space $\Omega = \{1, 0\}$ with probability $1 - p$ that no $(= 0)$ precipitation was observed. Entropy functions and score divergences of (*strictly*) *proper* scoring rules are invariant to such geometric transformation. The expected score function of the quadratic score is strictly convex on p

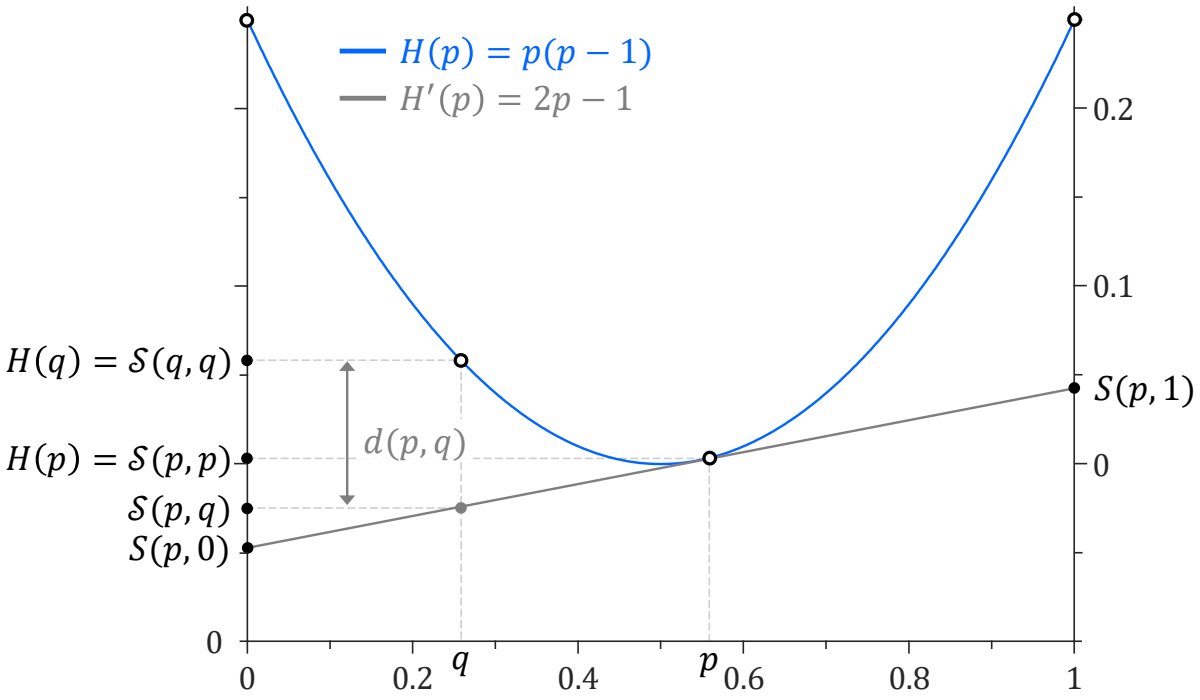

Figure 5: Generalized entropy function $H(p) = p(p-1)$ (blue curve) of the quadratic score for a dichotomous event $\Omega = \{1, 0\}$ with *probability forecast* $(p, 1-p)$ and *true probability* $(q, 1-q)$ with $p, q \in [0, 1]$. We present the values of the quadratic scoring rule $\mathcal{S}(p, q)$ at p and q (solid black dots) and display the so-called Bregman divergence, $d(p, q)$. For any probability forecast, $p \in [0, 1]$, the expected score, $\mathcal{S}(p, q) = qS(p, 1) + (1-q)S(p, 0)$, equals the ordinate of the tangent to H at p (solid gray line) when evaluated at $q \in [0, 1]$. In particular, the scores, $S(p, 0) = H(p) - pH'(p)$ and $S(p, 1) = H(p) + (1-p)H'(p)$, equal the tangent at $q = 0$ and $q = 1$, respectively. The divergence, $d_{\mathrm{QS}}(p, q) = \mathcal{S}(q, q) - \mathcal{S}(p, q)$, is equal to the difference between $H(q)$ and the tangent at p when evaluated at q (Adapted after Fig. 1 of *Gneiting and Raftery* (2007) and Fig. 8 of *Buja et al.* (2005)).

and satisfies continuity. The $H(p)$ function is referred to in the statistical literature as the *information measure* or *(generalized) entropy function* associated with the scoring rule S (*Grünwald and Dawid*, 2004; *Buja et al.*, 2005). This is the maximally achievable utility. Some authors refer instead to $-H(p)$ as the entropy function (*Bröcker*, 2009; *Dawid and Musio*, 2014) or *coherent uncertainty function* (*Dawid and Sebastiani*, 1999). According to Figure 5 and Equation (22), the score divergence $d(p, q)$ equals the difference of the value of H at point q and the first-order Taylor expansion of H around point p evaluated at point q

$$d(p, q) = \mathcal{S}(q, q) - \mathcal{S}(p, q)$$

$$
\begin{aligned}
&= H(q) - \big(H(p) - H'(p)(p-q)\big) \\
&= H(q) - H(p) + H'(p)(p-q),
\end{aligned} \tag{25}
$$

where $H'(p) = \mathrm{d}H(p)/\mathrm{d}p$ is the derivative of the entropy function with respect to p. The squared Euclidean distance $d(p,q) = \|p-q\|_2^2 = (p-q)^2$ is the canonical example of a Bregman distance, generated by the convex function $H(p) = p^2$. Indeed, according to Equation (25) the generalized entropy function $H(p) = p^2$ has divergence function

$$
\begin{aligned}
d(p,q) &= H(q) - H(p) + H'(p)(p-q) \\
&= q^2 - p^2 + 2p(p-q) \\
&= q^2 + p^2 - 2pq \\
&= (p-q)^2
\end{aligned} \tag{26}
$$

which is equal to the well-known Brier score (*Brier*, 1950).

5.3 Scoring Rules for Categorical Forecasts

For a categorical forecast of a finite number of m mutually exclusive and collectively exhaustive events $\Omega = \{1, \ldots, m\}$ the *distribution forecast* is a probability vector $\mathbf{p} = (p_1, \ldots, p_m)^\top$ issued on the convex class $\mathscr{P} = \mathscr{P}_m$ defined in Equation (9). Then Equation (25) may be written as

$$
d_{\mathrm{QS}}(\mathbf{p}, \mathbf{q}) = H(\mathbf{q}) - H(\mathbf{p}) + \big\langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \big\rangle \tag{27}
$$

where the gradient $\nabla H(\mathbf{p}) : \mathbb{R}^m \to \mathbb{R}^m$ at $\mathbf{p} \in \mathscr{P}_m$ is a vector-valued function

$$
\nabla H(\mathbf{p}) = \frac{\partial H(\mathbf{p})}{\partial \mathbf{p}} = \left(\frac{\partial H(\mathbf{p})}{\partial p_1}, \frac{\partial H(\mathbf{p})}{\partial p_2}, \ldots, \frac{\partial H(\mathbf{p})}{\partial p_m} \right)^\top \tag{28}
$$

and $\big\langle \mathbf{a}, \mathbf{b} \big\rangle$ denotes the inner product of the m-vectors $\mathbf{a}$ and $\mathbf{b}$. Furthermore, a *regular* scoring rule of a categorical forecast is *proper* if and only if (*McCarthy*, 1956; *Savage*, 1971)

$$
S(\mathbf{p}, j) = H(\mathbf{p}) - \big\langle \nabla H(\mathbf{p}), \mathbf{p} \big\rangle + \nabla H_j(\mathbf{p}) \quad \text{for } j = (1, \ldots, m) \tag{29}
$$

and reduces to a pair of functions, $S(p,1) : p \in [0,1] \to \overline{\mathbb{R}}$ and $S(p,0) : p \in [0,1] \to \overline{\mathbb{R}}$, for the binary event of *rain* or *no rain*. For a probability quote p the reward of the forecaster will equal to $S(p,1)$ if rainfall materializes and $S(p,0)$ otherwise. The expected score in Equation (20) then equals $\mathcal{S}(p,q) = qS(p,1) + (1-q)S(p,0)$, where q is the true probability of *rain*. For any two probability assignments $\mathbf{p}$ and $\mathbf{q}$ with *true probabilities* $\mathbf{q} = (q_1, \ldots, q_m)^\top$ constrained to the probability simplex,

Table 2: Quadratic, logarithmic and pseudospherical scoring rules for categorical variables: Entropy function, scoring rule, expected score function and divergence function for a *distribution forecast* $\mathbf{p} = (p_1, \ldots, p_m)^\top$ on the convex class $\mathscr{P} = \mathscr{P}_m$ of $m \geq 2$ mutually exclusive and collectively exhaustive events, $\Omega = \{1, \ldots, m\}$. The m-vector $\mathbf{q} = (q_1, \ldots, q_m)^\top$ lists the true event probabilities.

Name Score	Entropy function $H(\mathbf{p})$	Scoring rule $S(\mathbf{p}, j)$	Expectation $\mathcal{S}(\mathbf{p}, \mathbf{q})$	Divergence $d(\mathbf{p}, \mathbf{q})$
Quadratic[a,b]	$\sum_{k=1}^{m} p_k^2$	$2p_j - \sum_{k=1}^{m} p_k^2$	$2\sum_{k=1}^{m} p_k q_k - \sum_{k=1}^{m} p_k^2$	$\sum_{k=1}^{m} (p_k - q_k)^2$
Logarithmic[c]	$\sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k)$	$\log_{\mathfrak{b}}(p_j)$	$\sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(p_k)$	$\sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{q_k}{p_k}\right)$
Pseudospherical[d,e]	$\lVert\mathbf{p}\rVert_\eta^1$	$\frac{p_j^{\eta-1}}{\lVert\mathbf{p}\rVert_\eta^{\eta-1}}$	$\frac{\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle}{\lVert\mathbf{p}\rVert_\eta^{\eta-1}}$	$\lVert\mathbf{q}\rVert_\eta^1 - \frac{\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle}{\lVert\mathbf{p}\rVert_\eta^{\eta-1}}$

[a] Also known as proper linear score. Equals *Brier* (1950) score for a binary event, $\Omega = \{1, 0\}$
[b] Limiting case ($\eta \to 1$) of the pseudospherical score if suitably scaled
[c] Remains *strictly proper* under any logarithmic base $\mathfrak{b} > 1$.
[d] $\lVert\mathbf{p}\rVert_b^a = \left(\sum_{k=1}^{m} p_k^b\right)^{a/b}$ and $\langle \mathbf{p}, \mathbf{q}\rangle = \sum_{k=1}^{m} p_k q_k$ is the dot product of *forecast* and *true* probabilities
[e] For $\eta = 2$ the PSS reduces to the spherical score of *Friedman* (1983)

$\left\{\mathbf{q} \in \mathbb{R}_+^{m\times 1} : \mathbf{q}^\top \mathbf{1}_m = 1\right\}$ and $\mathbb{R}_+ = [0, \infty)$, the binary definition of the expected score generalizes to (*Bröcker*, 2009)

$$\mathcal{S}(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{m} q_j S(\mathbf{p}, j), \tag{30}$$

The interpretation of the scoring function $\mathcal{S}(\mathbf{p}, \mathbf{q})$ is that if ω is a random variable of distribution Q, then $\mathcal{S}(\mathbf{p}, \mathbf{q})$ is the mathematical expectation of the score of the assignment $\mathbf{p}$ in forecasting ω.

Table 2 presents the entropy functions $H(\mathbf{p})$ of three commonly used *strictly proper* categorical scoring rules on $\mathscr{P}_m$. Next, we use Equations (27), (29) and (30) to derive the mathematical expressions of the scoring rule, $S(\mathbf{p}, j)$, expected score function, $\mathcal{S}(\mathbf{p}, \mathbf{q})$, and score divergence, $d(\mathbf{p}, \mathbf{q})$, of the quadratic, logarithmic and (pseudo)spherical rule presented in Table 2. By definition, the probabilities p and q are dimensionless, and, consequently, all functions are unitless, except for the entropy function $H(\mathbf{p})$ of the logarithmic score which has units of information and, thus, bits if $\mathfrak{b} = 2$.

5.3.1 Quadratic Score

For a categorical forecast of m events with true probabilities, $p_1, \ldots, p_m$ issued on the probability simplex $P \in \mathscr{P}_m$ the entropy function of the quadratic score becomes $H(\mathbf{p}) = \sum_{k=1}^{m} p_k^2$ and equals an affine transformation of the Gini-index $G(\mathbf{p}) = \sum_{k=1}^{m} p_k(1 - p_k)$ (*Gini*, 1909). The gradient $\nabla H(\mathbf{p}) =$

$(2p_1, \ldots, 2p_m)^\top$ and Equation (27) leads to the divergence function

$$\begin{aligned} d_{\mathrm{QS}}(\mathbf{p}, \mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \big\langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \big\rangle \\ &= \sum_{k=1}^{m} q_k^2 - \sum_{k=1}^{m} p_k^2 + \big\langle (2p_1, \ldots, 2p_m), (p_1 - q_1, \ldots, p_m - q_m) \big\rangle \\ &= \sum_{k=1}^{m} q_k^2 - \sum_{k=1}^{m} p_k^2 + 2\big\langle \mathbf{p}, \mathbf{p} \big\rangle - 2\big\langle \mathbf{p}, \mathbf{q} \big\rangle \\ &= \sum_{k=1}^{m} q_k^2 + \sum_{k=1}^{m} p_k^2 - 2\sum_{k=1}^{m} p_k q_k \\ &= \sum_{k=1}^{m} (p_k - q_k)^2 \end{aligned} \tag{31}$$

The scoring rule of the quadratic score is now equal to

$$\begin{aligned} S_{\mathrm{QS}}(\mathbf{p}, j) &= H(\mathbf{p}) - \big\langle \nabla H(\mathbf{p}), \mathbf{p} \big\rangle + \nabla H_j(\mathbf{p}) \\ &= \sum_{k=1}^{m} p_k^2 - 2\big\langle \mathbf{p}, \mathbf{p} \big\rangle + 2p_j \\ &= 2p_j - \sum_{k=1}^{m} p_k^2 \end{aligned} \tag{32}$$

The quadratic score is also known as the proper linear score or *Brier* (1950) scoring rule

$$S_{\mathrm{BS}}(\mathbf{p}, j) = -\sum_{k=1}^{m} (\delta_{jk} - p_k)^2 = 2p_j - \sum_{k=1}^{m} p_k^2 - 1, \tag{33}$$

where the Kronecker symbol δ_{jk} equals one when the j^{th} event materializes ($j = k$) and $\delta_{jk} = 0$ otherwise.

With true probabilities, $\mathbf{q} = (q_1, \ldots, q_m)^\top$, the expected score of the quadratic rule equals

$$\begin{aligned} \mathcal{S}_{\mathrm{QS}}(\mathbf{p}, \mathbf{q}) &= \sum_{j=1}^{m} q_j S_{\mathrm{QS}}(\mathbf{p}, j) = \sum_{j=1}^{m} q_j \left(2p_j - \sum_{k=1}^{m} p_j^2 \right) = 2\sum_{j=1}^{m} p_j q_j - \sum_{j=1}^{m} q_j \sum_{k=1}^{m} p_k^2 \\ &= 2\sum_{j=1}^{m} p_j q_j - \sum_{k=1}^{m} p_k^2 \end{aligned} \tag{34}$$

and we can confirm the divergence function of the quadratic score in Table 2 and Equation (31)

$$\begin{aligned} d_{\mathrm{QS}}(\mathbf{p}, \mathbf{q}) &= \mathcal{S}_{\mathrm{QS}}(\mathbf{q}, \mathbf{q}) - \mathcal{S}_{\mathrm{QS}}(\mathbf{p}, \mathbf{q}) \\ &= 2\sum_{j=1}^{m} q_j q_j - \sum_{k=1}^{m} q_k^2 - 2\sum_{j=1}^{m} p_j q_j + \sum_{k=1}^{m} p_k^2 \\ &= \sum_{k=1}^{m} q_k^2 + \sum_{k=1}^{m} p_k^2 - 2\sum_{k=1}^{m} p_k q_k \\ &= \sum_{k=1}^{m} (p_k - q_k)^2. \end{aligned}$$

5.3.2 Logarithmic Score

The logarithmic score of *Good* (1952) is a linear equivalent of the relative entropy or KL divergence (*Lai et al.*, 2011) and also known in the statistical literature as the predictive deviance (*Knorr-Held and Rainer*, 2001) and ignorance score (*Roulston and Smith*, 2002). The entropy function of the logarithmic score $H(\mathbf{p}) = \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k)$ is equal to negative Shannon entropy $-\mathbb{H}(P)$ with gradient function $\nabla H(\mathbf{p}) = \big(\log_{\mathfrak{b}}(p_1) + 1, \ldots, \log_{\mathfrak{b}}(p_m) + 1\big)^{\top}$. The divergence function $d_{\mathrm{LS}}(\mathbf{p}, \mathbf{q})$ of the logarithmic score may now be derived from Equation (27) to yield

$$
\begin{aligned}
d_{\mathrm{LS}}(\mathbf{p}, \mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \big\langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \big\rangle \\
&= \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(q_k) - \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k) + \big\langle \big(\log_{\mathfrak{b}}(p_1) + 1, \ldots, \log_{\mathfrak{b}}(p_m) + 1\big), (p_1 - q_1, \ldots, p_m - q_m) \big\rangle \\
&= \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(q_k) - \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k) + \big\langle \log_{\mathfrak{b}}(\mathbf{p}), \mathbf{p} \big\rangle + \big\langle \mathbf{1}_m, \mathbf{p} \big\rangle - \big\langle \log_{\mathfrak{b}}(\mathbf{p}), \mathbf{q} \big\rangle - \big\langle \mathbf{1}_m, \mathbf{q} \big\rangle \\
&= \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(q_k) - \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(p_k) + \sum_{k=1}^{m} p_k - \sum_{k=1}^{m} q_k \\
&= \sum_{k=1}^{m} q_k \big(\log_{\mathfrak{b}}(q_k) - \log_{\mathfrak{b}}(p_k)\big) \\
&= \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\Big(\frac{q_k}{p_k}\Big), \qquad (35)
\end{aligned}
$$

and the scoring rule of the logarithmic score is now equal to

$$
\begin{aligned}
S_{\mathrm{LS}}(\mathbf{p}, j) &= H(\mathbf{p}) - \big\langle \nabla H(\mathbf{p}), \mathbf{p} \big\rangle + \nabla H_j(\mathbf{p}) \\
&= \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k) - \big\langle \big(\log_{\mathfrak{b}}(p_1) + 1, \ldots, \log_{\mathfrak{b}}(p_1) + 1\big), (p_1, \ldots, p_m) \big\rangle + \log_{\mathfrak{b}}(p_j) + 1 \\
&= \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k) - \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k) - \sum_{k=1}^{m} p_k + \log_{\mathfrak{b}}(p_j) + 1 \\
&= \log_{\mathfrak{b}}(p_j). \qquad (36)
\end{aligned}
$$

Thus the logarithmic score has negative Shannon entropy as its generalized entropy function and the reverse KL-divergence, $d_{\mathrm{LS}}(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(q_k/p_k)$, as its associated score divergence. The logarithmic score benefits a strong mathematical-statistical underpinning and links fundamental aspects from statistical theory, decision theory and information theory. The order of the arguments of the score divergence is reversed with respect to the convention used in information theory, but is in line with that of probabilistic forecasting (*Gneiting and Raftery*, 2007; *Bröcker*, 2009). The logarithmic score is equal to the logarithmic value of the probability p_j assigned to the materialized event, j, and, thus, only

strictly proper locally. *Roulston and Smith* (2002) provide an information-theoretic perspective on the logarithmic score and advocate using the so-called ignorance score, $\mathcal{S}_{\mathrm{IS}}(\mathbf{p}, j) = -\log_{\mathfrak{b}}(p_j)$.

With true probabilities, $\mathbf{q} = (q_1, \ldots, q_m)^\top$, the expected score of the logarithmic rule becomes

$$\mathcal{S}_{\mathrm{LS}}(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{m} q_j S_{\mathrm{LS}}(\mathbf{p}, j) = \sum_{j=1}^{m} q_j \log_{\mathfrak{b}}(p_j), \tag{37}$$

and we can confirm the logarithmic divergence score of Table 2 and Equation (35)

$$\begin{aligned} d_{\mathrm{LS}}(\mathbf{p}, \mathbf{q}) &= \mathcal{S}_{\mathrm{LS}}(\mathbf{q}, \mathbf{q}) - \mathcal{S}_{\mathrm{LS}}(\mathbf{p}, \mathbf{q}) \\ &= \sum_{j=1}^{m} q_j \log_{\mathfrak{b}}(q_j) - \sum_{j=1}^{m} q_j \log_{\mathfrak{b}}(p_j) \\ &= \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{q_k}{p_k}\right), \end{aligned}$$

which again is equal to the relative entropy $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p})$ from P to Q.

5.3.3 Pseudospherical Score

The entropy function of the pseudospherical score, $H(\mathbf{p}) = \|\mathbf{p}\|_\eta^1$ is equal to the ℓ_η-norm of the forecast probabilities $\mathbf{p} = (p_1, \ldots, p_m)^\top$

$$\|\mathbf{p}\|_\eta^1 = \left(p_1^\eta + \ldots + p_m^\eta\right)^{1/\eta}, \tag{38}$$

which for $\eta = 2$ reduces to the Euclidean norm, $\sqrt{\mathbf{p}^\top \mathbf{p}}$. The $m \times 1$-vector of partial derivatives $\nabla H(\mathbf{p})$ now equals

$$\nabla H(\mathbf{p}) = \left(\frac{p_1^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}}, \frac{p_2^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}}, \ldots, \frac{p_m^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}} \right)^\top, \tag{39}$$

where $\|\mathbf{p}\|_\eta^{\eta-1} = \left(\sum_{k=1}^m p_k^\eta\right)^{(\eta-1)/\eta}$. The gradient vector of the pseudospherical score may thus be written as the scalar-vector product, $\nabla H(\mathbf{p}) = \|\mathbf{p}\|_\eta^{1-\eta}\, \mathbf{p}^{\eta-1}$. According to Equation (27) the score divergence $d_{\mathrm{PSS}}(\mathbf{p}, \mathbf{q})$ is now equal to

$$\begin{aligned} d_{\mathrm{PSS}}(\mathbf{p}, \mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \left\langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \right\rangle \\ &= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta} \left\langle \mathbf{p}^{\eta-1}, \mathbf{p} - \mathbf{q} \right\rangle \\ &= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta} \left\langle \mathbf{p}^{\eta-1}, \mathbf{p} \right\rangle - \|\mathbf{p}\|_\eta^{1-\eta} \left\langle \mathbf{p}^{\eta-1}, \mathbf{q} \right\rangle \\ &= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta} \|\mathbf{p}\|_\eta^\eta - \|\mathbf{p}\|_\eta^{1-\eta} \left\langle \mathbf{p}^{\eta-1}, \mathbf{q} \right\rangle \\ &= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta} \left\langle \mathbf{p}^{\eta-1}, \mathbf{q} \right\rangle, \end{aligned} \tag{40}$$

for $\mathbf{p}, \mathbf{q} \in \mathscr{P}_m$ and $\eta > 1$. The scoring rule of the pseudospherical score now equals

$$\begin{aligned} S_{\text{PSS}}(\mathbf{p}, j) &= H(\mathbf{p}) - \big\langle \nabla H(\mathbf{p}), \mathbf{p} \big\rangle + \nabla H_j(\mathbf{p}) \\ &= \|\mathbf{p}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta} \big\langle \mathbf{p}^{\eta-1}, \mathbf{p} \big\rangle + \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\ &= \|\mathbf{p}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta} \|\mathbf{p}\|_\eta^\eta + \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\ &= \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1}. \end{aligned} \tag{41}$$

With true probabilities, $\mathbf{q} = (q_1, \ldots, q_m)^\top$, the expected pseudospherical score equals

$$\begin{aligned} \mathcal{S}_{\text{PSS}}(\mathbf{p}, \mathbf{q}) &= \sum_{j=1}^m q_j S_{\text{PSS}}(\mathbf{p}, j) \\ &= \sum_{j=1}^m q_j \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\ &= \|\mathbf{p}\|_\eta^{1-\eta} \big\langle \mathbf{p}^{\eta-1}, \mathbf{q} \big\rangle, \end{aligned} \tag{42}$$

and we can confirm the divergence function of the pseudospherical score in Table 2 and Equation (40)

$$\begin{aligned} d_{\text{PSS}}(\mathbf{p}, \mathbf{q}) &= \mathcal{S}_{\text{PSS}}(\mathbf{q}, \mathbf{q}) - \mathcal{S}_{\text{PSS}}(\mathbf{p}, \mathbf{q}) \\ &= \|\mathbf{q}\|_\eta^{1-\eta} \big\langle \mathbf{q}^{\eta-1}, \mathbf{q} \big\rangle - \|\mathbf{p}\|_\eta^{1-\eta} \big\langle \mathbf{p}^{\eta-1}, \mathbf{q} \big\rangle \\ &= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta} \big\langle \mathbf{p}^{\eta-1}, \mathbf{q} \big\rangle. \end{aligned}$$

For $\eta = 2$ the pseudospherical rule in Equation (42) reduces to the well known spherical scoring rule

$$\mathcal{S}_{\text{SS}}(\mathbf{p}, \mathbf{q}) = \|\mathbf{p}\|_2^{-1} \big\langle \mathbf{p}, \mathbf{q} \big\rangle, \tag{43}$$

with associated divergence function

$$d_{\text{SS}}(\mathbf{p}, \mathbf{q}) = \|\mathbf{q}\|_2^1 - \|\mathbf{p}\|_2^{-1} \big\langle \mathbf{p}, \mathbf{q} \big\rangle. \tag{44}$$

We look in more detail at the scoring rules and consider a binary event of *rain* and *no rain* with *probability forecast* $(p, 1-p)$ on $\Omega = \{1, 0\}$. Table 3 presents mathematical expressions of the scoring rules for the probability forecast p of this dichotomous event. When rain materializes, $j = 1$, the score is equal to $S(p, 1)$, otherwise $j = 0$ and the reward is equal to $S(p, 0)$. If we now treat the true forecast probability of the first event, $q = q_1$, then the expected score, $\mathcal{S}(p, q) = qS(p, 1) + (1-q)S(p, 0)$. Figure 6 displays the expected score of the quadratic (green), logarithmic (red) and spherical (blue) rules as function of $p \in [0, 1]$ and $q = q_1$ (different graphs). We used $\mathfrak{b} = 2$ for the logarithmic score to yield $\mathcal{S}_{\text{LS}}(p, q) = q \log_2(p) + (1-q) \log_2(1-p)$ in units of bits.

Table 3: Strictly proper scoring rules for a dichotomous event (*rain* and *no rain*) with *probability forecast* $\mathbf{p} = (p, 1-p)$ on $\Omega = \{1, 0\}$ with $p \in [0, 1]$.

Scoring rule	$S(p, 1)$	$S(p, 0)$
Brier	$-p^2 + 2p - 1$	$-p^2$
Quadratic	$4p - 2p^2 - 1$	$1 - 2p^2$
Logarithmic	$\log_{\mathfrak{b}}(p)$	$\log_{\mathfrak{b}}(1-p)$
Spherical	$p(2p^2 - 2p + 1)^{-1/2}$	$(1-p)(2p^2 - 2p + 1)^{-1/2}$
Pseudospherical	$p^{\eta-1}\big(p^\eta + (1-p)^\eta\big)^{(1-\eta)/\eta}$	$(1-p)^{\eta-1}\big(p^\eta + (1-p)^\eta\big)^{(1-\eta)/\eta}$

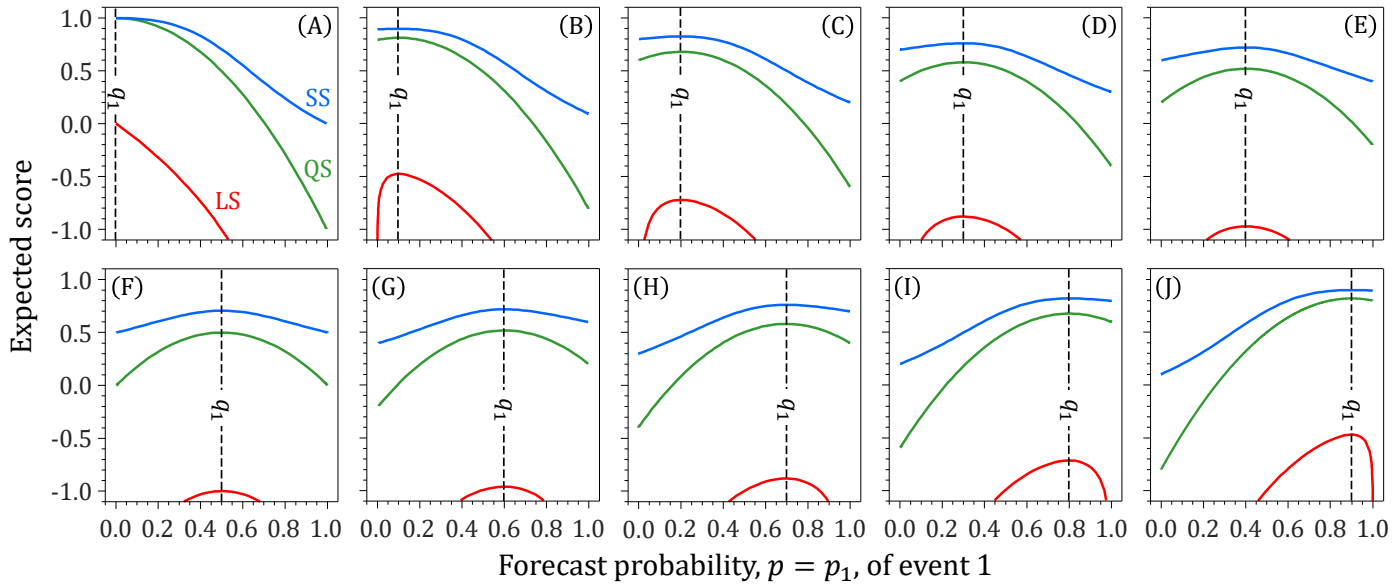

Figure 6: Binary event, $\Omega = \{1, 0\}$: Expected value of the quadratic (green), logarithmic (red) and spherical (blue) scoring rules as function of the *true probability* $q = q_1$ of the first event, (a) $q = 0$, (b) $q = 0.1$, (c) $q = 0.2$, (d) $q = 0.3$, (e) $q = 0.4$, (f) $q = 0.5$, (g) $q = 0.6$, (h) $q = 0.7$, (i) $q = 0.8$, and (j) $q = 0.9$.

The three scoring rules differ in their response to the quoted forecast probability $p = p_1$ of the *rain* event. The colored lines do not intersect and have a dissimilar functional shape, magnitude and range. But despite these differences, the three scoring rules have one property in common. The expected scores of QS, LS and SS are always maximized at the true *rain* probability $q = q_1$. In other words, the forecaster's reward is largest when he/she quotes $p = q$. This is exactly what (strict) propriety of the scoring rules implies and prompts a forecaster to be honest and report the *true* probabilities. The expected score decreases with increasing distance between the quoted and true *rain* probabilities p and q, respectively. The rate of decline is largest for the LS followed by the QS and SS. Note that magnitude differences are

inconsequential as *strictly proper* scoring rules S remain *strictly proper* under affine transformation

$$S^*(\mathbf{p}, j) = cS(\mathbf{p}, j) + \hbar(j), \tag{45}$$

where $c \neq 0$ is a constant and $\hbar(j)$ is a $\mathscr{P}$-integrable function (*Gneiting and Raftery*, 2007). If $c < 0$ the orientation of $S^*(\mathbf{p}, j)$ changes from a reward to a loss function.

5.4 Numerical examples

While it is generally agreed upon that scoring rules must at least be proper to adequately quantify the accuracy of probabilistic forecasts (*Winkler et al.*, 1996; *Gneiting and Ranjan*, 2011), the question of which ones to use in a specific applications remains largely open (*Gneiting and Raftery*, 2007; *Alexander et al.*, 2022). For the time being, we restrict our attention to the three categorical scoring rules of Table 2.

5.4.1 Case Study I: Simple Illustration

We revisit the distribution forecasts of Fig. 4a and turn the PDF of the *true* forecast distribution Q with continuous sample space $\Omega = [0, 6]$ into a PMF using $m = 60$ equally spaced values, $\omega_i = (6i - 3)/m$, where $i = (1, \ldots, m)$. The probability of each value (event) is determined from the CDF of Q and make up the m-vector $\mathbf{q} = (q_1, \ldots, q_m)^\top$ of *true* probabilities with unit sum. Similarly, we yield the probability assignment $\mathbf{p} = (p_1, \ldots, p_m)^\top$ for each distribution forecast, $P_1, \ldots, P_5$. Table 4 lists the generalized entropy, $H(\mathbf{p})$, expectation, $\mathcal{S}(\mathbf{p}, \mathbf{q})$ and divergence, $d(\mathbf{p}, \mathbf{q})$, of the *strictly proper* quadratic, logarithmic and spherical scoring rules for $P_1, \ldots, P_5$.

The tabulated values of $H(\mathbf{p})$, $\mathcal{S}(\mathbf{p}, \mathbf{q})$ and $d(\mathbf{p}, \mathbf{q})$ vary among the scoring rules and distribution forecasts and confirm several earlier points, (i) the expected score $\mathcal{S}(\mathbf{p}, \mathbf{q})$ of each scoring rule is maximized when the forecaster quotes the *true* probabilities, (ii) the logarithmic score is unbounded and operates on the extended real-line $\overline{\mathbb{R}}$ as it applies an indefinitely large penalty to P_1 and P_5 for each realized event *a priori* thought impossible by their respective uniform and generalized Pareto distribution forecasts, (iii) the QS, LS and SS divergence scores $d(\mathbf{p}, \mathbf{q})$ are strictly positive and zero only when $P = Q$, and (iv) *strictly proper* scoring rules do not necessarily yield the same ranking of the distribution forecasts. This justifies the use of multiple *strictly proper* scoring rules *Vrugt et al.* (2022). As a reminder, we reiterate that the logarithmic score (i) has negative Shannon entropy $-\mathbb{H}(\mathbf{p})$ in Equation (10) as its

Table 4: Entropy, $H(\mathbf{p})$, expectation, $\mathcal{S}(\mathbf{p},\mathbf{q})$ and divergence, $d(\mathbf{p},\mathbf{q})$ of the *strictly proper* categorical scoring rules of Table 2 for distribution forecasts $P_1,\ldots,P_5$ displayed in Fig. 4a using $m = 60$ discrete values, $\Omega = \frac{1}{20}\{1,3,5,\ldots,117,119\}$. Column R lists the rank of the distribution forecasts. The bottom row presents the values for a perfect distribution forecast, $P = Q$.

Forecast	Quadratic Score				Logarithmic Score, $\mathfrak{b} = 2$				Spherical Score			
	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$	$R^{\S}$	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$	R	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$	R
Eq.		(34)	(31)			(37)	(35)			(43)	(44)	
P_1	0.026	0.022	0.005	5	-5.248	-∞	∞	4	0.162	0.150	0.016	5
P_2	0.024	0.025	0.003	4	-5.522	-5.493	0.102	2	0.156	0.158	0.009	4
P_3	0.026	0.027	0.001	1	-5.483	-5.438	0.046	1	0.161	0.163	0.003	1
P_4	0.028	0.026	0.001	2	-5.389	-5.542	0.150	3	0.167	0.162	0.004	2
P_5	0.029	0.025	0.003	3	-5.314	-∞	∞	5	0.172	0.159	0.007	3
Q	0.028	0.028	0.000		-5.392	-5.392	0.000		0.166	0.166	0.000	

§ Rank of each distribution forecast obtained from sorting $d(\mathbf{p},\mathbf{q})$ in ascending order

entropy function and (ii) relative entropy $d_{\mathrm{KL}}(\mathbf{q},\mathbf{p})$ in Equation (11) as its divergence score $d_{\mathrm{LS}}(\mathbf{p},\mathbf{q})$ but is sometimes criticized for its unboundedness.

The QS, LS and SS may not give the exact same ranking of the distribution forecasts, but they are unanimous in their assessment of P_3 as the best forecast of the *true* distribution Q. This conclusion is supported by visual inspection of the distribution forecasts with the lognormal distribution forecast P_4 (yellow) as a close second contender. The results further demonstrate that (i) the entropy $H(\mathbf{p})$ cannot be used as sole determinant of the accuracy of a forecast distribution. This is implicit as the entropy is a function of the forecast distribution only, and (ii) outcomes with a zero-probability do not count in the computation of the entropy of the logarithmic score in accordance with the limit, $\lim_{x\downarrow 0} x\log_{\mathfrak{b}}(x) = 0$ for $\mathfrak{b} > 0$. This explains the elevated values of $H(\mathbf{p})$ for P_1 and P_5 under the logarithmic score.

5.4.2 Case Study II: Rainfall data

Next, we illustrate the application of the scoring rules to 24-h forecasts of daily precipitation probability from the Finnish Meteorological Institute for the city of Tampere in south-central Finland. The original data set of $n = 346$ daily rainfall forecasts is presented in Appendix E. *Hughes and Topp* (2015) used the individual daily forecasts to provide a diagrammatic interpretation of the Brier scoring rule in Equation (33) and associated score divergences. We simplify this analysis and report in Table 5 the entropy, expected score and score divergence of the quadratic, logarithmic and spherical scoring rules for the normalized probabilities of *true* and *forecasted* rainfall listed in the right block of Table E.1. The bottom

row presents the function values when the forecaster quotes the *true* rainfall probabilities, $\mathbf{p} = \mathbf{q}$. The

Table 5: Entropy, $H(\mathbf{p})$, expectation, $\mathcal{S}(\mathbf{p}, \mathbf{q})$ and divergence, $d(\mathbf{p}, \mathbf{q})$ of the *strictly proper* categorical scoring rules of Table 2 for the *true* and *forecasted* rainfall probabilities of Table E.1.

Forecast	Quadratic Score			Logarithmic Score, $\mathfrak{b} = 2$			Spherical Score		
	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p}, \mathbf{q})$	$d(\mathbf{p}, \mathbf{q})$	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p}, \mathbf{q})$	$d(\mathbf{p}, \mathbf{q})$	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p}, \mathbf{q})$	$d(\mathbf{p}, \mathbf{q})$
Eq.		(34)	(31)		(37)	(35)		(43)	(44)
$\mathbf{p}$	0.1241	0.1034	0.0043	-3.1557	-3.3505	0.0451	0.3523	0.3229	0.0052
$\mathbf{q}$	0.1077	0.1077	0.0000	-3.3054	-3.3054	0.0000	0.3282	0.3282	0.0000

794

795 tabulated data confirm our earlier points and are mainly listed for benchmark purposes.

796 6 Scoring rules for density forecasts

797 The task of determining whether the *probabilistic forecast* P matches the *true distribution* Q appears
798 difficult, perhaps hopeless, because Q is never observed, even after the fact. But as *Diebold et al.* (1998)
799 realized early on, the challenges posed by these subtleties are not insurmountable. The scoring rules for
800 categorical variables can be generalized to density forecasts to assist in forecast verification of continuous
801 variables.

802 Scoring rules for density forecasts are defined up to a so-called Lebesgue measure μ. The Lebesgue
803 integral plays an important role in probability theory but this topic is hardly taught in mathematics
804 courses. To explain the Lebesgue measure, please consider Figure 7 which displays an example Lebesgue
805 density of the standard normal distribution. The Riemann integral partitions the domain of a function
806 into a collection of small intervals and bars are constructed to meet the height of the graph. Then in $\mathbb{R}^2$
807 the resulting rectangles make up the area under the graph. The Lebesgue integral also uses rectangles,
808 but these rectangles are formed by partitioning the function's range (also called codomain) into different
809 intervals. For each horizontal slice, a rectangle is drawn with height of the corresponding function
810 value and width equal to the length of all intervals on the real line $\mathbb{R}$ (e.g. sample space) where the
811 function reaches approximately this height. This horizontal slicing of the codomain leads to much more
812 complicated sets of ω values, certainly for multivariate densities. Thus, the Lebesgue definition extends
813 integral calculation to a much broader class of functions. Now, the Lebesgue measure μ is equal to the
814 width of each slice, which, in turn, is the sum of the widths of all rectangles with the same height. For

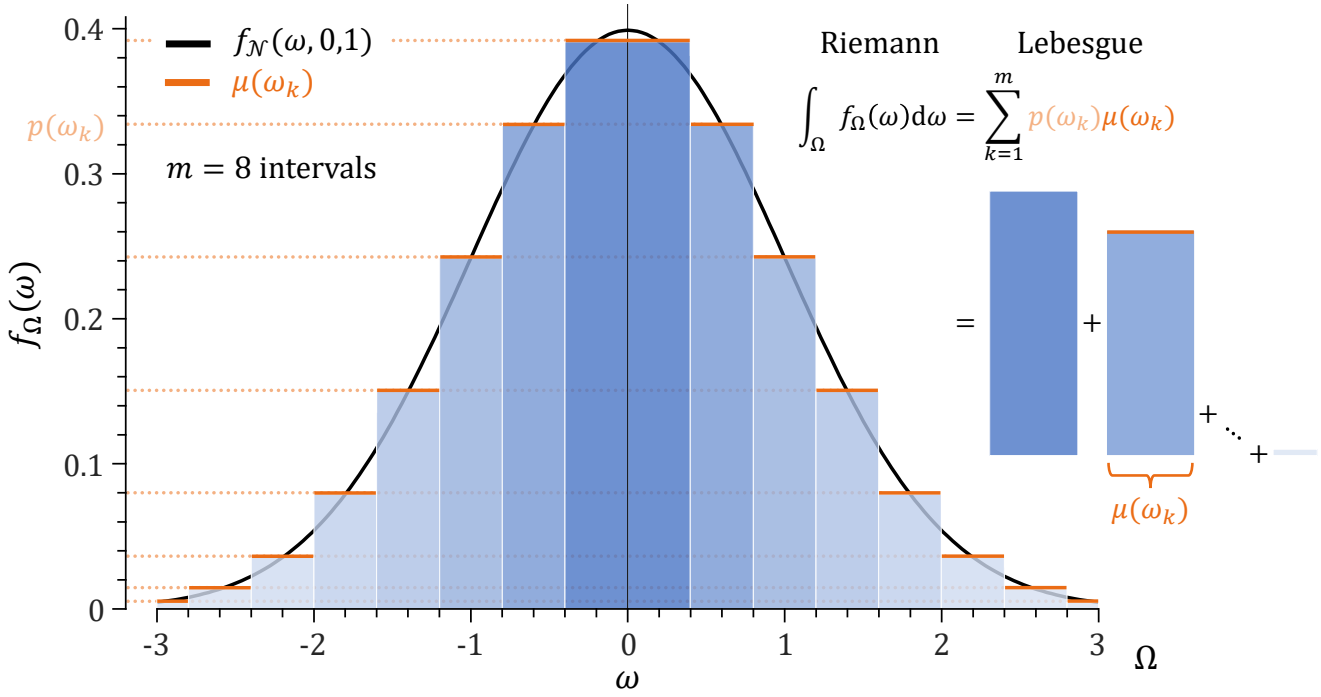

Figure 7: Illustration of the standard normal Lebesgue density on sample space $\Omega = [-3, 3]$ with range (codomain) of $\mathcal{N}(0, 1)$ partitioned into $m = 8$ small intervals. We use color coding to discern the intervals of ω. The Lebesgue measure $\mu(\omega_k)$ is equal to the length of each color coded interval containing ω_k. The Lebesgue density $f(\omega)$ is constant in each interval $\omega_1, \ldots, \omega_m$ of ω values. The sum of the areas of the rectangles is equal to the Lebesgue integral.

univariate densities, we can divide each interval of ω in non-overlapping bins with Lebesgue measure the bin width. This representation of the Lebesgue density is almost equal to a PMF with each bin made up of different points (univariate case) or sets (bivariate case and higher).

6.1 Univariate forecasts

We follow the formal measure-theoretic definition of *Gneiting and Raftery* (2007). Let μ be a nonnegative, countably additive set function on the measurable space (Ω, Σ) and let $\mathcal{L}_\eta(\Omega)$ with $\eta \in [1, \infty)$ denote the class of probability measures $f : \Omega \to \mathbb{R}$ on (Ω, Σ) that are absolutely continuous with respect to the measure μ on Σ and have an integral

$$\|f\|_\eta \equiv \left(\int_\Omega f(\omega)^\eta \mu(\mathrm{d}\omega) \right)^{1/\eta} < \infty, \tag{46}$$

equal to the ℓ_η-norm of the density f. The μ-density f_P of the probabilistic forecast $P \in \mathcal{L}_\eta$ is called a predictive density or density forecast. The above norm is invariant to changes in the *true distribution* Q that leave the probability of ω unchanged and induces a nonnegative metric (divergence)

$d(P,Q) = \|f_P - f_Q\|_\eta$ which is zero only if $P = Q$.

In most applications, the probabilistic forecast P will consist of a m-member ensemble $\mathbf{y} = (y_1, \ldots, y_m)^\top$ at given space and time coordinates. The approximate density of ω in a ε-neighborhood of a point y in $\mathbb{R}^m$ as

$$f(y) = \lim_{\varepsilon \to 0} \frac{\mu\big(\omega \cap B_\varepsilon(y)\big)}{\mu\big(B_\varepsilon(y)\big)} \tag{47}$$

where $B_\varepsilon(y) = \{y \in \mathbb{R} : |\omega - y| < \varepsilon\}$ is the closed ball (sphere) of radius ε centered at point y. If possible, we compute the unique density of ω according to *Bernardo* (1979) (P. 689) $f(\omega) = \lim_{\varepsilon \to 0} P\big(B_\varepsilon(\omega)\big)/\mu\big(B_\varepsilon(\omega)\big)$. The norm $\|f\|_\eta$ can be expressed as a sequence and/or vector by using the counting measure

$$\|f_P\|_\eta = \left[\sum_{k=1}^{m} \left(\frac{f_P(x_k)}{\sum_{j=1}^{m} f_P(x_j)}\right)^\eta\right]^{1/\eta}, \tag{48}$$

where $(x_1, \ldots, x_m)$ are m events for which we must evaluate the predictive density. The reciprocal of the denominator equals the Lebesgue measure (e.g. histogram bin width) of each event x_k and assumes that the x's are evenly distributed on the real line, $\mathbb{R}$. This is a pragmatic definition as we estimate the predictive density f_P of the probabilistic forecast P via kernel smoothing. This method returns a m-vector of equally-spaced x values along with estimates of their probability density $f_P(x)$ so that $\int_{\mathbb{R}} f_P(x)\mathrm{d}x = 1$.

6.1.1 Quadratic, Logarithmic and (Pseudo)spherical scoring rules

Scoring rules for the density forecast f assign a numerical score based on the predictive distribution P and on the event or value ω that materializes. In analogy to Equation (32), the quadratic or *Brier* (1950) score becomes

$$S_{\mathrm{QS}}(P,\omega) = 2f_P(\omega) - \|f_P\|_2^2\,, \tag{49}$$

where $\|f_P\|_2^2$ equals the sum of the squared normalized densities of the forecast distribution P. This scoring is *strictly proper* relative to the class $\mathcal{L}_2$ and has entropy function, $H(P) = \|f_P\|_2^2$ and divergence function, $d_{\mathrm{QS}}(P,Q) = \|f_P - f_Q\|^2$, where f_Q signifies the density of the true distribution Q. The logarithmic score of Equation (36) equals

$$S_{\mathrm{LS}}(P,\omega) = \log_{\mathfrak{b}}\big(f_P(\omega)\big), \tag{50}$$

and is *strictly proper* relative to the class $\mathcal{L}_1$ of probability measures with entropy function the negative Shannon entropy $H(P) = -\mathbb{H}(P)$ and divergence score the reverse KL-divergence. The logarithmic score

is a limiting case of the pseudospherical score for $\eta \to 1$

$$S_{\text{PSS}}(P, \omega) = \frac{f_P(\omega)^{\eta-1}}{\|f_P\|_\eta^{\eta-1}}, \tag{51}$$

which is *strictly proper* relative to the class $\mathcal{L}_\eta$. The strict convexity of the associated entropy function, $H(P) = \|f_P\|_\eta$, and the nonnegativity of the divergence function are straightforward consequences of the Hölder and Minkowski inequalities. For $\eta = 2$, we yield the spherical score

$$S_{\text{SS}}(P, \omega) = \frac{f_P(\omega)}{\|f_P\|_2}, \tag{52}$$

is *strictly proper* relative to the class $\mathcal{L}_2$ of probability measures.

6.1.2 Linear scoring rule: impropriety

To demonstrate the importance of scoring rule propriety we briefly discuss the intuitively appealing but improper linear score

$$S_{\text{LinS}}(P, \omega) = f_P(\omega), \tag{53}$$

Let $f_P = \frac{1}{2\epsilon}$ and $f_Q = \frac{1}{\sqrt{2\pi}} \exp(-\epsilon^2/2)$ denote the Lebesgue densities of the uniform *forecast distribution* P and standard normal *true distribution* Q on the interval $[-\epsilon, \epsilon]$. According to Equation (20), the expected score of the probabilistic forecasts P or Q under the true distribution Q becomes

$$\begin{aligned}
\mathcal{S}_{\text{LinS}}(P, Q) &= \int_{-\epsilon}^{\epsilon} S_{\text{LinS}}(P, \omega) \mathrm{d}Q(\omega) & \mathcal{S}_{\text{LinS}}(Q, Q) &= \int_{-\epsilon}^{\epsilon} S_{\text{LinS}}(Q, \omega) \mathrm{d}Q(\omega) \\
&= \int_{-\epsilon}^{\epsilon} \frac{1}{2\epsilon} \frac{1}{\sqrt{2\pi}} \exp(-\omega^2/2) \mathrm{d}\omega & &= \int_{-\epsilon}^{\epsilon} \Big(\frac{1}{\sqrt{2\pi}} \exp(-\omega^2/2) \Big)^2 \mathrm{d}\omega \\
&= \frac{1}{2\epsilon} \frac{1}{\sqrt{2\pi}} \int_{-\epsilon}^{\epsilon} \exp(-\omega^2/2) \mathrm{d}\omega & &= \frac{1}{2\pi} \int_{-\epsilon}^{\epsilon} \exp(-\omega^2) \mathrm{d}\omega \\
&= \frac{1}{2\epsilon} \frac{1}{\sqrt{2\pi}} \Big| \sqrt{2\pi} \operatorname{erf}\big(\omega/\sqrt{2}\,\big) \Big|_0^{\epsilon} & &= \frac{1}{2\pi} \Big| \sqrt{\pi} \operatorname{erf}(\omega) \Big|_0^{\epsilon} \\
&= \frac{1}{2\epsilon} \operatorname{erf}\big(\epsilon/\sqrt{2}\,\big) & &= \frac{1}{2\sqrt{\pi}} \operatorname{erf}(\epsilon),
\end{aligned} \tag{54}$$

and the score divergence $d_{\text{LinS}}(P, Q)$ is equal to

$$\begin{aligned}
d_{\text{LinS}}(P, Q) &= \mathcal{S}_{\text{LinS}}(Q, Q) - \mathcal{S}_{\text{LinS}}(P, Q) \\
&= \frac{1}{2\sqrt{\pi}} \operatorname{erf}(\epsilon) - \frac{1}{2\epsilon} \operatorname{erf}\big(\epsilon/\sqrt{2}\,\big).
\end{aligned} \tag{55}$$

For an infinitesimal interval $\epsilon \to 0$ we yield $d_{\text{LinS}}(P, Q) = -(2\pi)^{-1/2}$. Figure 8 displays the divergence of the linear score as function of the Lebesgue measure $0 < \epsilon \leq 3$. The score divergence is negative for small values of ϵ and changes sign at the root, $\epsilon = 1.6221$, of Equation (55). This demonstrates that the score divergence of the linear score $S_{\text{LinS}}(P, \omega) = f_P(\omega)$ does not have a proper zero point. Thus, $S_{\text{LinS}}(P, \omega)$ is an improper scoring rule.

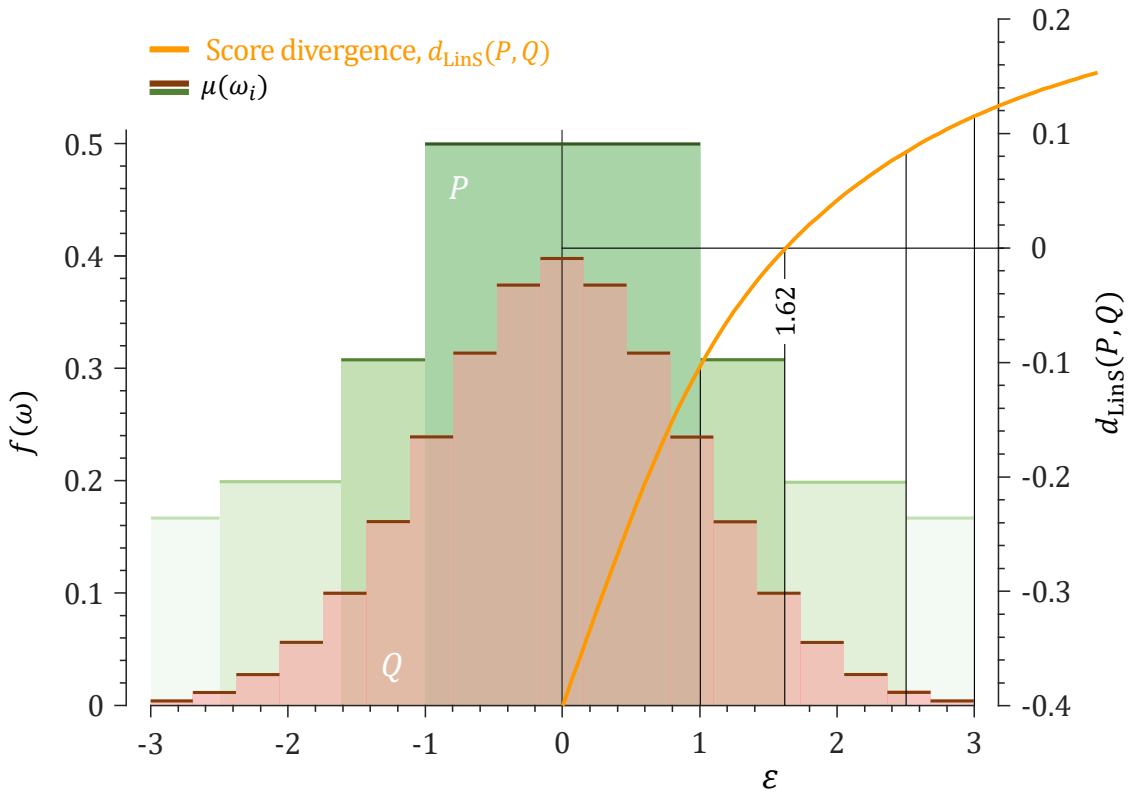

Figure 8: Score divergence $d_{\text{LinS}}(P, Q)$ of the linear scoring rule $S_{\text{LinS}}(P, \omega) = f_P(\omega)$ for a uniform probabilistic forecast P under a standard Gaussian true distribution Q and symmetric interval $[-\epsilon, \epsilon]$ with $\epsilon \in (0, 3]$. The horizontal green and blue lines correspond to the Lebesgue measure $\mu(\omega_i)$ or the length of the interval containing event ω_i.

6.1.3 Continuous Ranked Probability Score

The aforementioned scores are not particularly sensitive to distance in that they do not receive credit when assigning high probabilities to values near but not equal to the materialized outcome. The continuous ranked probability score or CRPS of Figure 9 addresses this deficiency. This scoring rule has found widespread application in the atmospheric sciences and is equal to the integral of the squared distance between the CDF, F_P, of the distribution forecast P and the empirical CDF of the observation (*Matheson and Winkler*, 1976; *Hersbach*, 2000)

$$\begin{aligned} S_{\text{CRPS}}(P, \omega) &= -\int_{-\infty}^{\infty} \big(F_P(z) - \mathbb{1}\{\omega \leq z\}\big)^2 \, \mathrm{d}z \\ &= -\int_{-\infty}^{\omega} F_P^2(z) \mathrm{d}z - \int_{\omega}^{\infty} \big(F_P(z) - 1\big)^2 \, \mathrm{d}z, \end{aligned} \tag{56}$$

where the indicator function $\mathbb{1}\{a\}$ returns 1 if a is true and zero otherwise, and the minus sign reverses the orientation to a reward function. The CRPS is *strictly proper* relative to the subclass $\mathscr{P}_1 \in \mathscr{P}$ of Borel probability measures that have finite first moment. The CRPS may also be written as integral of

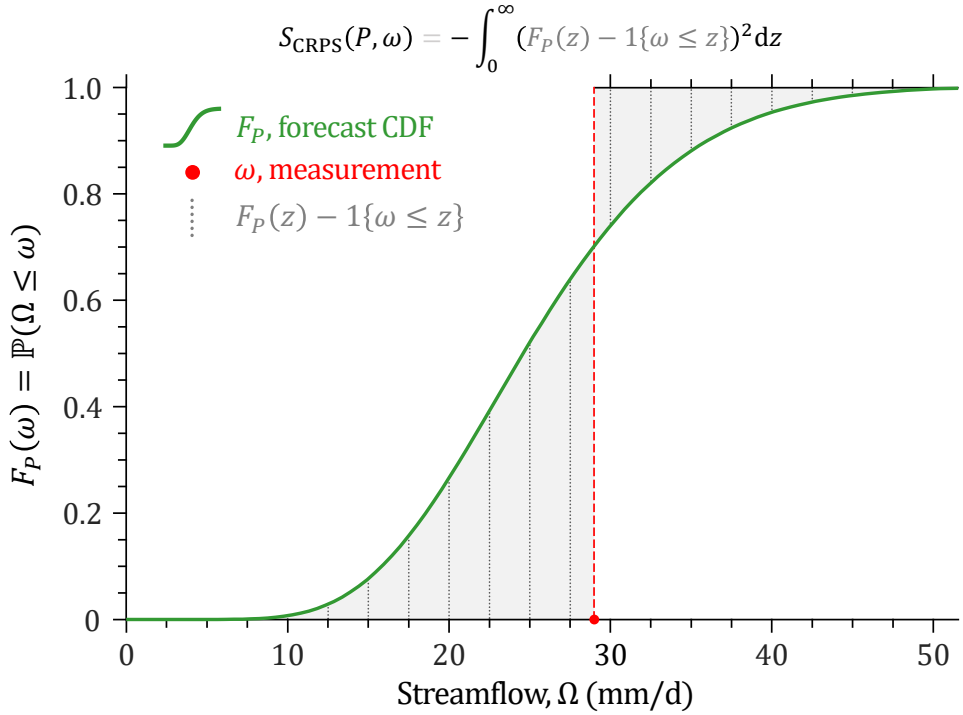

Figure 9: Graphical explanation of the continuous ranked probability scoring rule for a hypothetical streamflow forecast CDF F_P (black line) and verifying discharge measurement ω (red dot). The CRPS is equal to the integral of the squared differences (= gray dotted lines) of F_P and the Heaviside step function, $\mathbb{1}\{\omega \le z\}$.

the Brier probability score of the induced distribution forecast $F_P(z) = \int_{-\infty}^{z} f_P(t)\,\mathrm{d}t$ of the binary event $\{\omega \le z\}$ over all thresholds $z \in \mathbb{R}$

$$S_{\mathrm{CRPS}}(P,\omega) = \int_{-\infty}^{\infty} S_{\mathrm{BS}}\big(F_P(z), \mathbb{1}\{\omega \le z\}\big)\,\mathrm{d}z, \tag{57}$$

where

$$S_{\mathrm{BS}}\big(F_P(z), \mathbb{1}\{\omega \le z\}\big) = -\big(F_P(z) - \mathbb{1}\{\omega \le z\}\big)^2. \tag{58}$$

We can also write the CRPS using the $\tau \in [0,1]$-quantile forecast $y_\tau = F_P^{-1}(\tau)$ of P as follows (*Laio and Tamea*, 2007; *Gneiting and Ranjan*, 2011; *Grushka-Cockayne et al.*, 2017)

$$S_{\mathrm{CRPS}}(P,\omega) = -2\int_0^1 \big(\mathbb{1}\{\omega < y_\tau\} - \tau\big)\big(y_\tau - \omega\big)\mathrm{d}\tau, \tag{59}$$

with integrand the piecewise linear quantile score (*Friederichs and Hense*, 2007; *Bracher et al.*, 2021)

$$S^\tau_{\mathrm{QNT}}(P,\omega) = -2\big(\mathbb{1}\{\omega < y_\tau\} - \tau\big)\big(y_\tau - \omega\big), \tag{60}$$

which is also known as the pinball-loss, tick-loss or check-loss function. The equivalence of Equations (56) and (59) can be established via a change of variables (*Laio and Tamea*, 2007). A more friendly

formulation of Equation (59) is (see Appendix F)

$$S_{\mathrm{CRPS}}(P,\omega) = \omega\big(1 - 2F_P(\omega)\big) + 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau, \tag{61}$$

and promotes closed-form solutions of the CRPS for parametric distribution forecasts P with analytic expressions of their CDF, F_P, and quantile function, F_P^{-1} (*Jordan*, 2016; *Villez*, 2017). Appendix G presents such a derivation for a normal distribution forecast $P = \mathcal{N}(\mu_P, \sigma_P^2)$ to yield

$$S_{\mathrm{CRPS}}\big(\mathcal{N}(\mu_P,\sigma_P^2),\omega\big) = \frac{\sigma_P}{\sqrt{\pi}} - 2\sigma_P^2 f_{\mathcal{N}}(\omega,\mu_P,\sigma_P^2) - (\omega-\mu_P)\big(2F_{\mathcal{N}}(\omega,\mu_P,\sigma_P^2) - 1\big), \tag{62}$$

where $f_{\mathcal{N}}(x,\mu,\sigma^2)$ and $F_{\mathcal{N}}(x,\mu,\sigma^2)$ are the normal PDF and CDF, respectively.

Nonparametric distribution forecasts do not admit a closed-form expression for the CRPS and, thus, we must numerically solve for the integral of $S_{\mathrm{CRPS}}(P,\omega)$ using quadrature rules (*Staël von Holstein*, 1970; *Unger*, 1985). We can also resort to Lemma 2.2 of *Baringhaus and Franz* (2004) and use the convenient kernel representation of the CRPS

$$S_{\mathrm{CRPS}}(P,\omega) = \frac{1}{2}\mathbb{E}_P\big[|y-y^*|\big] - \mathbb{E}_P\big[|y-\omega|\big], \tag{63}$$

where y and y^* are samples of the forecast distribution P. For an ensemble forecast of m members, $\mathbf{y} = (y_1,\ldots,y_m)^\top$, Equation (63) simplifies to (*Grimit et al.*, 2006)

$$S_{\mathrm{CRPS}}(P,\omega) = \frac{1}{2m^2}\sum_{i=1}^m\sum_{j=1}^m |y_i - y_j| - \frac{1}{m}\sum_{i=1}^m |y_i-\omega|. \tag{64}$$

For a point forecast ($m = 1$) the first term equals zero and the CRPS reduces to the negative absolute error, $s_{\mathrm{NAE}}(y,\omega) = -|y-\omega|$. Thus, the CRPS is an extension of the absolute error to a distribution forecast. The computational complexity $\mathcal{O}(m^2)$ of Equation (64) can be reduced to a total of $\mathcal{O}\big(m\log_{\mathfrak{b}}(m)\big)$ operations if the CRPS is evaluated in terms of the quantile loss function of Equation (59) using the sorted ensemble members (*Murphy*, 1970; *Hersbach*, 2000; *Laio and Tamea*, 2007)

$$S_{\mathrm{CRPS}}(P,\omega) = -\frac{2}{m^2}\sum_{i=1}^m |y_i-\omega|\Big(m\mathbb{1}\{\omega\le y_i\} - i + \tfrac{1}{2}\Big). \tag{65}$$

The entropy function or information measure of the CRPS

$$H(P) = -\int_{-\infty}^{\infty} F_P(z)\big(1-F_P(z)\big)\mathrm{d}z = -\frac{1}{2}\mathbb{E}_P\big[|y-y^*|\big], \tag{66}$$

coincides with the negative selectivity function of *Matheron* (1984) and the *Gini* (1909) index. The CRPS divergence function

$$d_{\mathrm{CRPS}}(P,Q) = \int_{-\infty}^{\infty}\big(F_P(z) - F_Q(z)\big)^2\mathrm{d}z, \tag{67}$$

is symmetric by virtue of the quadratic term.

Table 6: Summary of *strictly proper* scoring rules for a density forecast f_P and verifying observation ω. The numerical form assumes that the forecast distribution P is a m-member ensemble $(y_1, \ldots, y_m)^\top$.

Score Name	XX	Scoring rule, $S_{\mathrm{XX}}(P, \omega)$		Note
		Analytic	Numerical	
Quadratic	QS	$2f_P(\omega) - \int_{-\infty}^{\infty} f_P^2(y)\,\mathrm{d}y$	$2f_P(\omega) - \sum_{k=1}^{m}\left(\frac{f_P(y_k)}{\sum_{j=1}^{m} f_P(y_j)}\right)^2$	a
Logarithmic	LS	$\log_{\mathfrak{b}}\big(f_P(\omega)\big)$	$\log_{\mathfrak{b}}\big(f_P(\omega)\big)$	a
Cnt. Rnk. Prb.	CRPS	$-\int_{-\infty}^{\infty}\big(F_P(z) - \mathbb{1}\{\omega \le z\}\big)^2\mathrm{d}z$	$\frac{1}{2m^2}\sum_{i=1}^{m}\sum_{k=1}^{m}\lvert y_i - y_k\rvert - \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \omega\rvert$	b
Spherical	SS	$\dfrac{f_P(\omega)}{\left(\int_{-\infty}^{\infty} f_P^2(y)\,\mathrm{d}y\right)^{1/2}}$	$f_P(\omega)\left[\sum_{k=1}^{m}\left(\frac{f_P(y_k)}{\sum_{j=1}^{m} f_P(y_j)}\right)^2\right]^{-1/2}$	a
Energy	ES	$\frac{1}{2}\mathbb{E}_P\big[\lvert y - y^*\rvert^\eta\big] - \mathbb{E}_P\big[\lvert y - \omega\rvert^\eta\big]$	$\frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lvert y_i - y_j\rvert^\eta - \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \omega\rvert^\eta$	b,c

[a] $f_P(x)$ signifies the empirical density of P at x. Determined from eCDF using kernel smoothing
[b] y_i and y_j are independent draws from the distribution forecast P
[c] Index $\eta \in (0, 2)$; For $\eta = 1$, we yield $S_{\mathrm{CRPS}}(P, \omega)$ and $\eta \to 2$ leads to $S_{\mathrm{SE}}(P, \omega) = -\lvert\mu_P - \omega\rvert^2$

6.1.4 Energy Score

Gneiting and Raftery (2007) proposed a generalization of the CRPS the so-called energy score

$$S_{\mathrm{ES}}(P, \omega) = \tfrac{1}{2}\mathbb{E}_P\big[\lvert y - y^*\rvert^\eta\big] - \mathbb{E}_P\big[\lvert y - \omega\rvert^\eta\big], \tag{68}$$

where the index $\eta \in (0, 2)$ and y and y^* are independent copies of the forecast distribution, $P \in \mathscr{P}_\eta$. This is a strictly proper score (*Székely*, 2003) and reduces to the CRPS for $\eta = 1$ and the negative squared error $S_{\mathrm{SE}}(P, \omega) = -\lvert\mu_P - \omega\rvert^2$ in the limit of $\eta \to 2$ (*Gneiting and Raftery*, 2007), where μ_P is the mean of the distribution forecast P. For an ensemble forecast of m values $\mathbf{y} = (y_1, \ldots, y_m)^\top$, the energy score may be computed as follows

$$S_{\mathrm{ES}}(P, \omega) = \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lvert y_i - y_j\rvert^\eta - \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \omega\rvert^\eta. \tag{69}$$

Table 6 summarizes the different *strictly proper* scoring rules for a distribution forecast P with density forecast p defined up to the μ-measure zero. All of them reward leptokurtic forecast distributions with probability mass that centers on the verifying observation, ω. As a reminder, the logarithmic score LS is the only tabulated scoring rule which ignores model predicted probabilities of all non-realized outcomes, thus, is strictly local only and highly sensitive to low probability events.

6.2 Numerical Examples

We use three different case studies to illustrate the computation of the scoring rules for univariate density forecasts. The first study provides a simple graphic illustration of the computation of the scoring rules. The second and third case study illustrate the application of scoring rules to discharge forecasts obtained from Bayesian Model Averaging and Generalized Likelihood Uncertainty Estimation.

6.2.1 Case Study III: Graphical Illustration

Suppose the forecast distribution P of the discharge Ω in mm/d is exactly described by a gamma distribution $P = \mathcal{G}(a_P, b_P)$ (see Figure 10) with dimensionless shape parameter $a_P = 3$ and scale parameter $b_P = 1$ mm/d. We slide the verifying observation ω from left to right across the distribution P and display the corresponding values of the quadratic (blue), logarithmic (red), spherical (green) and continuous ranked probability (yellow) scores of Table 6 on the y-axis. The different scoring rules exhibit a characteristic concave shape and provide the largest reward (smallest loss) in the high probability density region of the forecast distribution. Outside this region the different scoring rules decline in value with increasing distance from their maximum reward. The maximum of the quadratic, logarithmic and spherical scoring rules coincides exactly with the peak of the gamma distribution at $\omega = 2$. The maximum reward of the CRPS is well removed from the peak (mode) of the forecast distribution and concentrates on the median of P at about $\omega = 2.67$. The overall functional shape of the four scoring rules is rather similar but with appreciable differences in the curvature of QS, LS, SS and CRPS. The logarithmic score responds most strongly to changes in the density of the distribution forecast P, whereas the quadratic score displays a much more damped response to changes in the predictive density. Now imagine that $\omega = 4.0$ mm/d is the discharge which materializes (vertical dashed line) at a future time. According to the gamma distribution forecast $P = \mathcal{G}(3, 1)$ we yield $f_{\mathcal{G}}(\omega, 3, 1) = 0.1465$ (black diamond) with values of the scoring rules (colored dots) equal to $S_{\text{QS}} = 0.2912$, $S_{\text{LS}} = -1.9206$ nats, $S_{\text{SS}} = 3.375$ and $S_{\text{CRPS}} = -0.7585$ mm/d. As byproduct of another derivation, Appendix I presents a closed-form expression for the CRPS of a gamma distribution forecast. Equation (I.25) matches the yellow line.

The strong similarities between the QS, SS and LS do not come as a surprise. They belong to a limited class of *strictly proper* scoring rules on $\mathcal{L}_1(\Omega)$ and/or $\mathcal{L}_2(\Omega)$ and, thus, are expected to be related. Indeed, under some regularity conditions, *Bernardo* (1979) has shown that every proper local scoring rule is

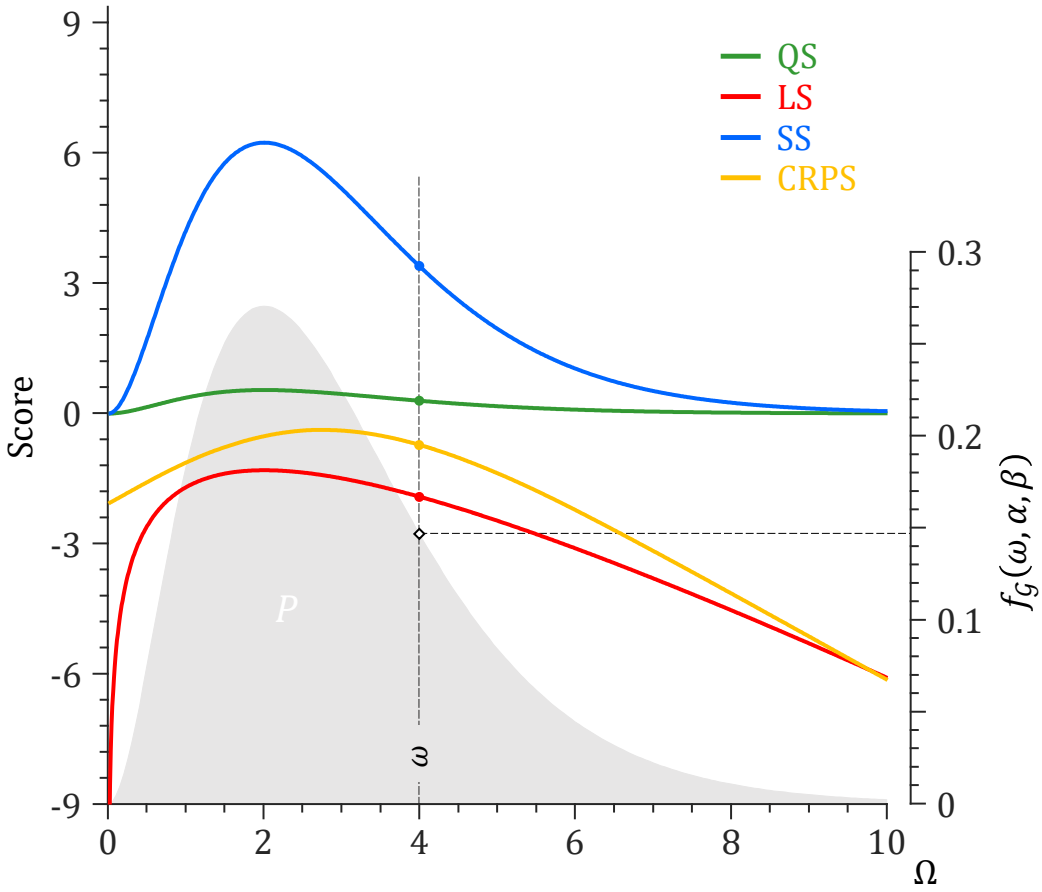

Figure 10: Hypothetical distribution forecast P of the discharge Ω (mm/d) and traces of the quadratic (green), logarithmic (red), spherical (blue) and continuous ranked probability (yellow) scoring rules across the streamflow distribution computed using their numerical definitions in Table 6. The probabilistic forecast is a gamma distribution $P = \mathcal{G}(a, b)$ with shape and scale parameters $a = 3$ (-) and $b = 1$ mm/d, respectively, and PDF $f_{\mathcal{G}}(\omega, a, b) = \Gamma^{-1}(a) b^{-a} \omega^{a-1} \exp(-\omega/b)$, where $\Gamma(x)$ is the gamma function.

equal to an affine transformation (45) of the logarithmic score. In principle, we only require one *strictly proper* scoring rule, though there are benefits in using multiple different scoring rules for verification of probabilistic model forecasts and/or multi-model ensemble prediction systems.

6.2.2 Case Study IV: Simple Illustration

We revisit the distribution forecasts of Fig. 4a and compute the *strictly proper* scoring rules of Table 6 using a collection of $n = 10^4$ observations $\omega_1, \ldots, \omega_n$ drawn at random from the *true distribution* Q. Table 7 documents the outcome of this analysis and lists mean values of the quadratic, logarithmic, spherical and continuous ranked probability scores for $P_1, \ldots, P_5$. Albeit different, these five distribution forecasts were deemed equivalent according to the coefficient of variation or C_v performance metric. The

Table 7: Mean values of the quadratic, logarithmic, spherical and continuous ranked probability scoring rules for the distribution forecasts $P_1, \ldots, P_5$ portrayed in Fig. 4a. The last column reports the mean scores for a perfect distribution forecast, $P = Q$.

Score	Fig. 4a					
	P_1 red	P_2 blue	P_3 green	P_4 yellow	P_5 purple	$P = Q$ gray
QS	0.484	0.467	0.501	0.513	0.525	0.534
LS	-6.520	-2.422	-2.190	-2.260	-2.479	-2.132
SS	6.197	6.148	6.390	6.332	6.454	6.608
CRPS	-0.649	-0.653	-0.663	-0.666	-0.642	-0.638

four scoring rules display a considerable variation among the distribution forecasts and assign different rewards to $P_1, \ldots, P_5$. This testifies to their complete evaluation of the forecast distribution. The QS, LS, SS and CRPS assign their highest scores to P_3 or the green forecast distribution in Fig. 4a. All four scoring rules agree that this forecast distribution is the best approximation of the *true distribution* Q, a conclusion that is supported by a visual comparison of $P_1, \ldots, P_5$ and Q. The last column lists the values of the scoring rules when the forecaster quotes $P = Q$. This is the maximum value each scoring rule can attain and equal to a Monte Carlo estimate of $\mathcal{S}(Q, Q)$ in Equation (21). Thus, if we subtract each column $P_1, \ldots, P_5$ from this last column, we yield estimates of the scores divergences $d_{\mathrm{QS}}(P, Q)$, $d_{\mathrm{LS}}(P, Q)$, $d_{\mathrm{SS}}(P, Q)$ and $d_{\mathrm{CRPS}}(P, Q)$ for each distribution forecast.

6.2.3 Case Study V: Bayesian Model Averaging

We now illustrate the scoring rules by application to density forecasts of river discharge from a multi-model ensemble of $K = 8$ conceptual hydrologic models of the Leaf River watershed (1950 km^2) located north of Collins, Mississippi. This ensemble is described in *Vrugt and Robinson* (2007) and interested readers are referred to this publication for more details. Figure 11 displays the discharge forecasts for a short but representative period of the training data set. The interval of the daily discharge forecasts generally envelops the measured discharge. Note that some models issue a negative forecast (days 1-40, 240-260 and 350-400) as a result of linear bias-correction as recommended by *Raftery et al.* (2005).

Let β_k, y_{kt} and $f_k(y|y_{kt}, \psi_k)$ denote the weight, streamflow prediction and conditional density of the kth model of the ensemble at time t. The density of the multi-model forecast distribution P_t at $t = (1, \ldots, n)$

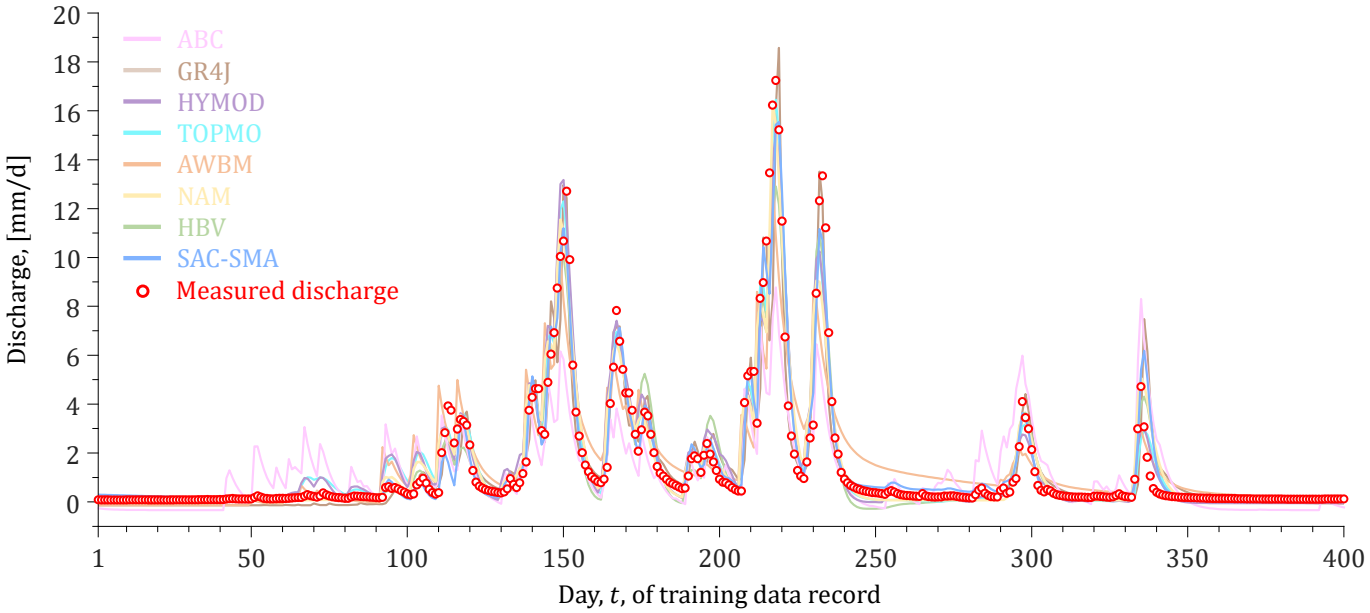

Figure 11: Streamflow predictions of the ABC (3), GR4J (4), HYMOD (5), TOPMO (6), AWBM (8), NAM (9), HBV (9) and SAC-SMA (13) conceptual watershed models for a small period of the $n = 3,000$ day training data record. The number between parenthesis lists the number of calibration parameters. The solid red circles correspond to the daily measured discharges.

is now equal to a mixture distribution of the models' conditional PDFs

$$f_{P_t}(y|\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{k=1}^{K} \beta_k f_k(y|y_{kt}, \psi_k), \tag{70}$$

centered on the weighted-average forecast

$$\mu_{P_t} = \sum_{k=1}^{K} \beta_k y_{kt} \tag{71}$$

with weights, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top$, constrained to the probability simplex, $\Delta^{K-1} = \{\boldsymbol{\beta} \in \mathbb{R}^K : \beta_1 + \cdots + \beta_K = 1;\ \beta_k \geq 0 \text{ for } k = 1, \ldots, K\}$ and shape parameters ψ_k of each model's predictive density assembled in the array $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$. Equations (70) and (71) are also known in the literature as the BMA forecast density and BMA model forecast, respectively (*Raftery et al.*, 2005). Table 8 presents mathematical formulations of the conditional PDFs, $f_k(y|y_{kt}, \psi_k);\ k = (1, \ldots, K)$, used to construct the forecast density.

Table 8: Formulation, coefficients and unknown parameters of the 1. generalized normal, 2. lognormal, 3. gamma and 4. Weibull predictive PDFs used in the BMA forecast density of Equation (70).

	Formulation	Coefficients	Unknown parameters
1.	$f_k(y\|\mu_{kt}, s_k^2, \tau_k) = \frac{\tau_k}{2s_k\Gamma(\tau_k^{-1})} \exp\left[-\left(\frac{\|y-\mu_{kt}\|}{s_k}\right)^{\tau_k}\right]$	$\mu_{kt} = y_{kt}$	$\boldsymbol{\psi} = (s_1^2, \ldots, s_K^2, \tau_1, \ldots, \tau_K)^\top$ §†
2.	$f_k(y\|\mu_{kt}, s_k^2) = \frac{1}{ys_k\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log(y)-\mu_{kt}}{s_k}\right)^2\right]$	$\mu_{kt} = \log(y_{kt}) - \frac{1}{2}s_k^2$	$\boldsymbol{\psi} = (s_1^2, \ldots, s_K^2)^\top$ §*
3.	$f_k(y\|a_{kt}, b_{kt}) = \frac{1}{\Gamma(a_{kt})b_{kt}^{a_{kt}}} y^{a_{kt}-1} \exp\left(-\frac{y}{b_{kt}}\right)$	$a_{kt} = \frac{\|y_{kt}\|^2}{s_k^2}$, $b_{kt} = \frac{s_k^2}{\|y_{kt}\|}$	$\boldsymbol{\psi} = (s_1^2, \ldots, s_K^2)^\top$ §†
4.	$f_k(y\|\lambda_{kt}, s_k^2) = \frac{s_k}{\lambda_{kt}}\left(\frac{y}{\lambda_{kt}}\right)^{s_k-1} \exp\left[-\left(\frac{y}{\lambda_{kt}}\right)^{s_k}\right]$	$\lambda_{kt} = \frac{\|y_{kt}\|}{\Gamma(1+1/s_k)}$	$\boldsymbol{\psi} = (s_1^2, \ldots, s_K^2)^\top$ ‖¶

§ We use a (i) group, $s_1^2 = \ldots = s_K^2$, and (ii) sole, $s_1^2, \ldots, s_K^2$, variance for the $K = 8$ models
† The variance σ_k^2 of $f_k(y|\cdot)$ is equal to s_k^2
* The variance of the lognormal density is equal to $\sigma_k^2 = \big(\exp(s_k^2) - 1\big)\exp(2\mu_{kt} + s_k^2)$
‖ We use s_k^2 for the shape parameter, thus, evaluate a group and model dependent (= sole) shape parameter
¶ The variance of the Weibull density is equal to $\sigma_k^2 = \lambda_{kt}^2[\Gamma(1+2/s_k) - \Gamma^2(1+1/s_k)]$

The coefficients of the conditional PDFs are defined so that the means of the generalized normal, lognormal, gamma and Weibull distributions coincide with the respective model forecasts, y_{kt}, where $k = (1, \ldots, K)$. The shape parameter τ_k determines the kurtosis of the generalized normal density. A value of $\tau_k = 2$ results in a normal distribution (albeit with variance $\sigma_k^2/2$), a value of $\tau_k = 1$ equates to a Laplace (double-exponential) distribution, and $\tau_k \to \infty$ converges to a uniform density on $[y_{kt} - \sigma_k, y_{kt} + \sigma_k]$ and a zero density outside this interval. Thus, the larger the value of τ_k the less peaked the conditional PDF of the kth model will be and the more tight its associated prediction intervals around y_{kt}.

If we assume that the models' forecast errors are independent, then the d-vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$ of weights $\boldsymbol{\beta}$ and shape parameters $\boldsymbol{\psi}$ of the conditional PDFs of Table 8 can be determined from maximization of the BMA log-likelihood function, $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega}) = \sum_{t=1}^{n} \log\big(f_{P_t}(\omega_t|\boldsymbol{\beta}, \boldsymbol{\psi})\big)$, using MCMC simulation with the DREAM algorithm (*Vrugt et al.*, 2008) and weights constrained to the probability simplex. Although the model ensemble does not satisfy the independence assumption, this should not affect much the estimates of the weights $\boldsymbol{\beta}$ and shape parameters $\boldsymbol{\psi}$, because we are estimating the conditional distribution for a scalar observation given K forecasts, rather than for several observations simultaneously (*Raftery et al.*, 2005).

The variance of the BMA forecast density $f_{P_t}(y|\boldsymbol{\beta}, \boldsymbol{\psi})$ in Equation (70) is equal to

$$\sigma_{P_t}^2 = \sum_{k=1}^{K} \beta_k(\sigma_k^2 + y_{kt}^2) - \mu_{P_t}^2 \tag{72}$$

where y_{kt} and σ_k^2 denote the mean and variance of the $f_k(y|y_{kt},\psi_k)$'s at time t. As the conditional PDFs of Table 8 admit a closed-form solution for their variances, $\sigma_k^2;\ k=(1,\ldots,K)$, the coefficient of variation of the BMA forecast density at time t is exactly defined; $C_{\mathrm{v},t}=\sigma_{P_t}/\mu_{P_t}$. Lower and upper endpoints of the $\gamma=100(1-\alpha)\%$ prediction interval of the BMA mixture density can be derived from the CDF

$$F_{P_t}(y|\boldsymbol{\beta},\boldsymbol{\psi})=\sum_{k=1}^{K}\beta_k F_k(y|y_{kt},\psi_k), \tag{73}$$

so that $F_{P_t}(l_t|\boldsymbol{\beta},\boldsymbol{\psi})=\alpha/2$ and $F_{P_t}(u_t|\boldsymbol{\beta},\boldsymbol{\psi})=1-\alpha/2$. At each time t, we solve for the lower and upper predictive quantiles at different α values using an iterative root finding procedure. If we evaluate the CDF in Equation (73) at each verifying observation and sort the resulting values in ascending order, then the reliability R_{l} of the BMA forecast distribution is easily computed using Equation (19). We use the `MODELAVG` toolbox of *Vrugt* (2018) in MATLAB to determine maximum likelihood values of the BMA model parameters for the conditional PDFs of Table 8 along with performance metrics, scoring rules and discharge prediction intervals of the BMA forecast density.

Figure 12 presents traces of the weighted-average BMA forecast (black line) and associated 50, 75 and 95% prediction intervals (gray regions) using the (a) normal ($\tau=2$), (b) lognormal, (c) generalized normal, (d) gamma and (e) Weibull predictive PDFs of Table 8 with a sole forecast variance, $\sigma_k^2;\ k=(1,\ldots,K)$ for each watershed model. The solid red squares depict the measured daily discharges. This weighted-average BMA forecasts describe the discharge observations quite well and appear relatively unaffected by the choice of the models' predictive distribution. This does not hold, however for the BMA forecast density. Indeed, we observe large differences in the spread of the 50, 75 and 95% prediction intervals of the BMA mixture distribution of Equation (70) among the conditional PDFs of Table 8. These differences are particularly well visible at the higher flows, for example between days 2,740 - 2,750, but also manifest themselves at lower flows, particularly during the dry period between days 2,800 and 2,870 of the training record. The symmetry of the normal and generalized normal distributions translates into BMA prediction interval with an approximately constant spread, independent of flow level. As a result, the prediction uncertainty of the BMA model is relatively small for the driven part of the hydrograph and comparatively large at the lower flows. The lognormal, gamma and Weibull distributions are defined only for $y\geq 0$ and, thus, are expected to yield a more accurate and fluent description of streamflow behavior. Indeed, the BMA prediction intervals of these three skewed distributions are comparatively large at the highest flows but their spread declines rapidly with discharge magnitude and collapses to a very small region around

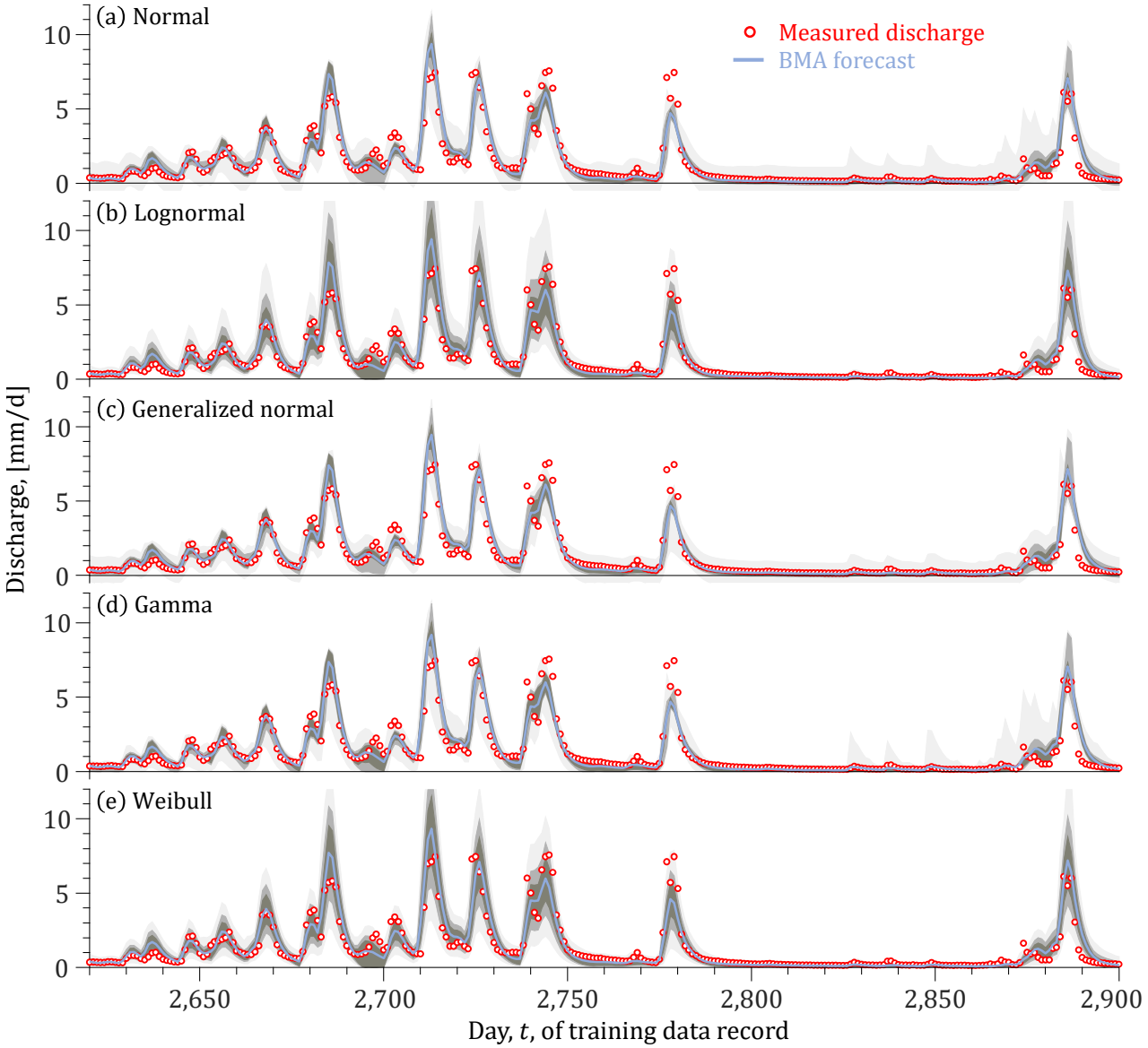

Figure 12: Weighted average BMA forecast (solid blue line) and 50% (dim gray), 75% (medium gray) and 95% (light gray) quantiles of the BMA forecast distribution for a representative 270-day period of the training data set using the (a) normal, (b) lognormal, (c) generalized normal, (d) gamma and (e) Weibull distributions with a model-dependent (= sole) forecast variance. The red circles are the daily discharge observations.

the mean forecast (solid blue line) during baseflow. The gamma distribution produces much sharper BMA density forests during peak flows than the lognormal and Weibull distributions but this is at the expense of an overly dispersed BMA forecast distribution for small rainfall events in the long dry period between days 2,800 and 2,870. This will deteriorate the average spread and performance of the gamma distribution. The normal and generalized normal distributions display similar jagged BMA prediction intervals in the long recession period.

The BMA forecast density offers an excellent opportunity for application of the scoring rules. For each

scoring rule, we compute a (time-averaged) mean score

$$\overline{S}_{\mathrm{XX}}(\mathbf{P}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{t=1}^{n} S_{\mathrm{XX}}(P_t, \omega_t), \tag{74}$$

where $\mathbf{P} = \{P_1, \ldots, P_n\}$ is the collection of probabilistic forecasts derived from the BMA model. Table 9 reports the mean values of the quadratic, logarithmic, spherical, continuous ranked probability and energy scoring rules for the BMA forecast density of Equation (70) with a normal, lognormal, generalized normal, gamma and Weibull predictive PDF using a constant *group* or *sole* forecast variance. We also list the values of the performance metrics of Table 1 and report the log-likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega}) = \sum_{t=1}^{n}\left\{\log_e\left(\sum_{k=1}^{K}\beta_k f_k(\omega_t|y_{kt}, \psi_k)\right)\right\}, \tag{75}$$

Root Mean Square Error (RMSE)

$$s_{\mathrm{RMSE}}(\boldsymbol{\mu}_P, \boldsymbol{\omega}) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\omega_t - \mu_{P_t})^2}, \tag{76}$$

coefficient of determination (R^2)

$$s_{R^2}(\boldsymbol{\mu}_P, \boldsymbol{\omega}) = 1 - \frac{\sum_{t=1}^{n}(\omega_t - \mu_{P_t})^2}{\sum_{t=1}^{n}(\omega_t - m_\omega)^2}, \tag{77}$$

and KG efficiency

$$s_{\mathrm{KGE}}(\boldsymbol{\mu}_P, \boldsymbol{\omega}) = 1 - \sqrt{\big(a-1\big)^2 + \big(b-1\big)^2 + (r-1)^2}, \tag{78}$$

of the weighted-average BMA forecast of Equation (71) where $a = m_{\mu_P}/m_\omega$ and $b = s_{\mu_P}/s_\omega$ are the unitless ratios of the sample means and sample standard deviations, respectively, of the weighted-average BMA discharge forecasts, $\boldsymbol{\mu}_P = (\mu_{P_1}, \ldots, \mu_{P_n})^\top$, and verifying measurements, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, respectively, and r is the sample Pearson correlation coefficient of the n-data pairs, $\{(\omega_1, \mu_{P_1}), \ldots, (\omega_n, \mu_{P_n})\}$. The tabulated data highlight several important findings.

1. The time-averaged values of the *strictly proper* scoring rules of the BMA forecast density differ substantially between the predictive PDFs of Table 8. This is particularly true for the QS, LS and SS, whereas the CRPS and ES demonstrate considerably less variation.

2. The use of a model-specific forecast variance generally increases the values of the scoring rules for the BMA forecast density but this improvement is marginal for the lognormal, generalized normal and Weibull distributions.

Table 9: Time-averaged values of the *strictly proper* scoring rules of Table 6 for the BMA density forecast $f_{P_t}(y|\boldsymbol{\beta},\boldsymbol{\psi})$; $t = (1,\ldots,n)$ of Equation (70) using the normal ($\tau = 2$), lognormal, generalized normal, gamma and Weibull predictive PDFs with a group or model-dependent (= sole) constant variance. We also list the performance metrics, R_l, C_v, C and W of Table 1 and report the log-likelihood, $\ell(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\omega})$, RMSE, R^2 and KG efficiency of the weighted-average BMA forecast. The bottom row lists the number d of unknown BMA parameters.

	Normal		Lognormal		Gen. normal		Gamma		Weibull	
	group	sole	group	sole	group	sole	group	sole	group	sole
QS	1.345	2.862	3.697	3.732	2.864	2.922	0.995	2.657	3.603	3.616
LS	-1.162	-0.291	0.145	0.153	-0.265	-0.257	-1.426	-0.408	0.103	0.108
SS	15.28	7.695	33.85	29.79	10.33	11.87	8.018	11.13	36.91	36.97
CRPS	-0.265	-0.238	-0.220	-0.220	-0.236	-0.236	-0.261	-0.237	-0.221	-0.220
ES	-0.502	-0.519	-0.513	-0.525	-0.516	-0.514	-0.563	-0.528	-0.520	-0.520
R_l	0.229	0.105	0.135	0.128	0.104	0.098	0.234	0.128	0.129	0.127
C_v	0.789	-3.917	0.772	0.859	1.849	1.784	2.189	-2.111	0.800	0.867
C	0.961	0.962	0.972	0.971	0.951	0.951	0.958	0.956	0.967	0.969
W	2.214	2.557	1.843	1.829	2.015	2.047	1.959	1.513	1.661	1.698
$\ell(\boldsymbol{\theta}\|\boldsymbol{\omega})$	-2416	-604.8	300.7	318.0	-550.1	-534.5	-2966	-847.8	215.2	223.8
RMSE	0.708	0.716	0.721	0.724	0.719	0.715	0.750	0.726	0.722	0.725
R^2	0.866	0.863	0.861	0.860	0.862	0.864	0.850	0.860	0.861	0.860
KGE	0.868	0.863	0.869	0.865	0.862	0.865	0.824	0.854	0.867	0.862
d	9	16	9	16	17	24	9	16	9	16

3. The lognormal and Weibull distributions maximize the overall performance of the BMA forecast density according to the *strictly formal* scoring rules. This confirms the advantages of using a conditional PDF commensurate with the skewed distribution of the discharge data. Perhaps rather surprisingly, but this advantage does not hold for the gamma distribution. While the spread of the BMA forecast density of this distribution is much smaller on average than the lognormal and Weibull distributions, the scoring rules of the gamma distribution suffer primarily on days with an underdispersed BMA forecast density (e.g. see days 2,740-2,750 and around 2,780) and to a lesser extent on days with overly dispersed BMA prediction intervals in the dry period (see Fig. 12d).

4. The reliability, R_l, varies quite a bit among the different conditional PDFs of the BMA forecast density. This measure of the statistical consistency of the BMA distribution forecasts and the observed discharge data does not prefer the lognormal and Weibull distributions but rather assigns the lowest value of about 0.10 to the generalized normal distribution. But as was illustrated in Fig. 4, the reliability is not a sufficient condition, hence, may support an erroneous ranking of the BMA

models.

5. The coefficient of variation C_v exhibits a large variation among the different conditional PDFs of the BMA forecast density with values ranging between -3.9 and 2.2 for the normal and gamma distributions, respectively. Time-averaging obfuscates interpretation of this measure of relative dispersion of the BMA forecast distribution as its value may be corrupted by large C_v values on days with near-zero discharges. This may at least explain in part why the C_v correlates poorly with the average spread W of the 95% BMA prediction intervals. The listed values of the C_v are exact for the BMA mixture density of Equation (70), but are otherwise difficult to estimate accurately, particularly for skewed discharge distributions with a mean close to zero.

6. All distributions of Table 8 achieve an adequate coverage of $C \approx 0.95$ at significance level $\alpha = 0.05$. This testifies to the appropriate form of the likelihood function and the DREAM algorithm's ability in successfully locating the maximum likelihood values of the BMA weights and shape parameters. The coverage, however, provides an incomplete evaluation of density forecasts as pointed out in Figure 4. We can turn the coverage into a *proper* scoring rule as will be discussed in the context of quantile prediction.

7. The tabulated data highlight the limitations of the conjectured sharpness principle of *Gneiting et al.* (2007). The gamma distribution with a model-dependent (sole) forecast variance results in the smallest spread (width) of the 95% BMA prediction intervals, yet, is inferior to almost all other conditional PDFs according to the *strictly proper* scoring rules and log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ of the time-averaged BMA forecast. Sharpness is a measure of the concentration (dispersion) of the BMA forecast density, hence is an *improper* scoring rule as it is a property of the predictive distribution only.

8. The RMSE, R^2 and KG efficiency of the weighted-average BMA forecast of Equation (71) appear insensitive to the choice of forecast variance (group or sole) and conditional PDF. These summary metrics are unresponsive to the large differences in the BMA forecast density shown in Fig. 12 between the predictive PDFs of Table 8. The log-likelihood on the contrary is much more responsive to changes in the BMA forecast distribution with values of $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ ranging between values of -2,966 and 318 for the different conditional PDFs. The log-likelihood demonstrates the advantages

of using a model-dependent (sole) forecast variance and confirms the superiority of the lognormal and Weibull distributions. In fact, the log-likelihood exhibits a perfect 1 : 1 relationship with the logarithmic scoring rule and an almost perfect linear relationship with the QS and CRPS (see Table 10).

Table 10: Pearson correlation coefficients of the log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ of the weighted-average BMA forecast and time-averaged values of the QS, LS, SS, CRPS and ES *strictly proper* scoring rules, improper performance metrics of Table 1 and RMSE, R^2 and KGE scoring functions.

	QS	LS	SS	CRPS	ES	R_{l}	C_{v}	C	W	RMSE	R^2	KGE
$\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$	0.998	1.000	0.690	0.976	0.361	0.809	0.596	0.560	-0.394	-0.351	0.360	0.666

Thus, the blatant insensitivity of the RMSE, R^2 and KG efficiency to the BMA forecast distribution is not a necessary consequence of these metrics acting only on the weighted-average BMA forecast.

We have learned that widely used metrics such as the RMSE, R^2 and KGE are woefully inadequate for evaluating distribution forecasts. This is a concerning finding as these metrics are commonly used by researchers and practitioners to evaluate model performance. The *strictly proper* scoring rules equip the hydrologist with an arsenal of "measures" for robustly evaluating distribution forecasts.

To provide insights into the temporal dynamics of the scoring rules, please consider Figure 13 which presents a time series plot of the (a) BMA forecast distribution and traces of the quadratic (green), logarithmic (red), spherical (blue), continuous ranked probability (yellow) and energy (cyan) scoring rules. The scoring rules vary dynamically across the hydrograph and display a rapid succession of small and occasionally larger fluctuations in a pattern independent of flow level. We briefly summarize the most important differences between the traces of the scoring rules. The LS appears most responsive to the BMA forecast distribution with relatively large day-to-day variations in the value of this scoring rule. The trace of the QS is almost similar to that of the LS but with a noticeably smaller amplitude during the first 150 days of the displayed record. The strong agreement in the temporal behavior of these two scoring rules confirms the similarities in the mathematical definition of the LS and QS. The SS is seemingly unresponsive to the rapid succession of storm events in the first half of the record as a result of the large values this scoring rule attains towards the end of the so-called dry period. The CRPS and to a lesser extent the ES show most variation during the wet period when streamflow is intermediate to high but are almost unresponsive during periods without rainfall when discharge is low.

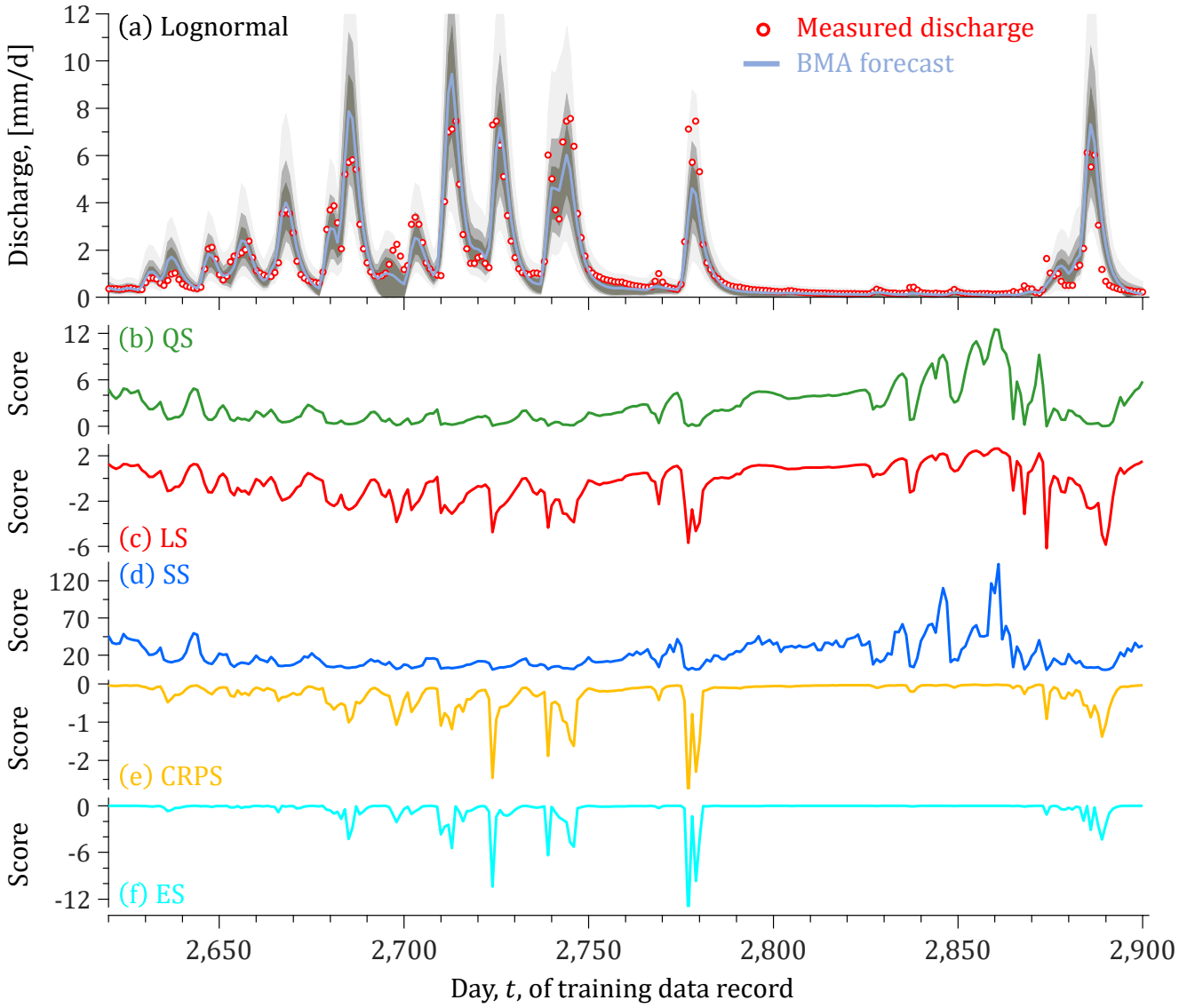

Figure 13: (a) 50%, 75% and 95% prediction intervals of the BMA forecast density for the lognormal distribution of Table 8 with a constant model-dependent (= sole) forecast variance and corresponding traces of the (b) quadratic, (c) logarithmic, (d) spherical, (e) continuous ranked probability and (f) energy scoring rules.

The scoring rules show important similarities and differences. The QS, LS and SS attain their largest values at the lowest flow levels when the discharge observation is contained within the high probability density (HPD) region of the BMA forecast density and this distribution is relatively tight. This comes at no surprise as a narrow forecast distribution will return, on average, higher probability densities than a much more dispersed forecast distribution. The CRPS and ES respond quite differently and do not peak when the BMA forecast distribution is accurate and narrow but instead retain values near zero. What all scoring rules have in common is that they achieve the lowest values when the discharge measurements materialize outside the HPD region and in the tails of the BMA forecast distribution. In principle, any

of the *strictly proper* rules will suffice in evaluating the quality of the BMA forecast distribution and ranking the conditional PDFs. But the question which one(s) to use remains largely open (*Gneiting and Raftery*, 2007; *Alexander et al.*, 2022).

6.2.4 Case Study VI: Generalized Likelihood Uncertainty Estimation

Before we move on to scoring rules for multi-variable forecasts, we illustrate the application of scoring rules for model training and evaluation using application of the Generalized Likelihood Uncertainty Estimation (GLUE) method of *Beven and Binley* (1992) to the HYdrologic MODEL (HYMOD, *Boyle* 2001), Hydrologic model (Hmodel, *Schoups et al.* 2010) and Sacramento Soil Moisture Accounting (SAC-SMA) model (*Burnash et al.*, 1973). Details of these models can be found in the cited references and numerous other publications. Readers are referred to Appendix H for a schematic description of each model along with mathematical formulations of the fluxes between the control volumes and a table with unknown parameters. We simulate the rainfall-discharge relationship for the 3,000-day training record of the Leaf River watershed using daily estimates of areal average rainfall and potential evapotranspiration. The model equations are solved using a mass-conservative second-order integration method with adaptive time stepping. This guarantees a robust and accurate numerical solution of the simulated fluxes and state variables. A one-year spin-up period mitigates the impact of state variable initialization.

We draw at random $m = 25,000$ parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ from the models' prior parameter ranges using Latin hypercube sampling. Next, we evaluate each parameter vector, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id})^\top$ and compute the quality-of-fit of its simulated daily discharge record, $\mathbf{y}(\boldsymbol{\theta}_i) = \big(y_1(\boldsymbol{\theta}_i), \ldots, y_n(\boldsymbol{\theta}_i)\big)^\top$, using the normal log-likelihood

$$\ell(\boldsymbol{\theta}_i|\boldsymbol{\omega}) = -\frac{n}{2}\log\left(\sum_{t=1}^{n}\big(\omega_t - y_t(\boldsymbol{\theta}_i)\big)^2\right), \tag{79}$$

where $e_t(\boldsymbol{\theta}) = \omega_t - y_t(\boldsymbol{\theta})$ signifies the tth residual. Then, we sort the simulations of the parameter vectors in decreasing likelihood. The result is a $n \times m$ matrix $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ where the $n \times 1$-vector $\mathbf{y}_1 = (y_{11}, \ldots, y_{n1})^\top$ maximizes the log-likelihood of Equation (79) and $\mathbf{y}_m = (y_{1m}, \ldots, y_{nm})^\top$ returns the smallest value of $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. Then, for any sufficiently large ensemble size the simulated values $y_{t1}, \ldots, y_{tK}$ at time t make up a discrete forecast distribution P_t whose PDF f_{P_t} may be approximated using a normal kernel function with bandwidth proportional to the number of points K.

Figure 14 displays the relationship between the ensemble size K and the time-averaged value of the

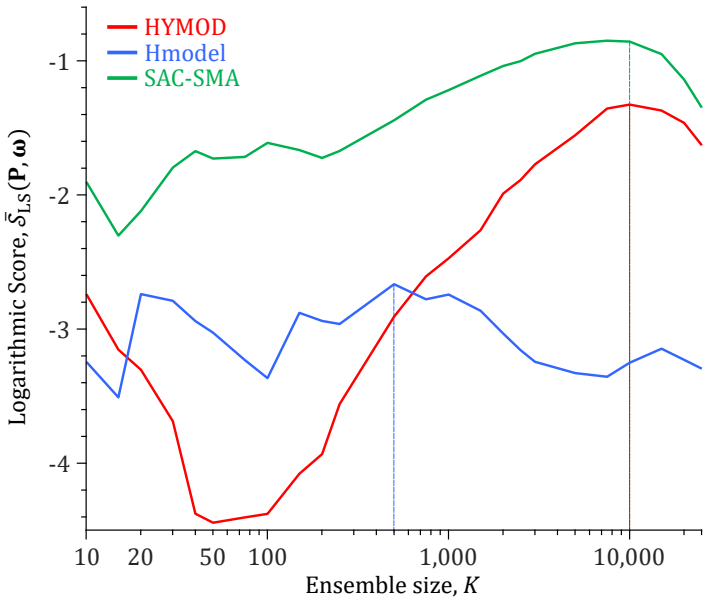

Figure 14: Relationship between the ensemble size, K, and the time-averaged value of the logarithmic scoring rule for HYMOD (red line), Hmodel (blue line) and SAC-SMA (green line) using daily discharge data from the Leaf River watershed. The dashed vertical lines point out the optimum size of the ensemble.

logarithmic scoring rule, $\overline{S}_{\mathrm{LS}}(\mathbf{P}, \boldsymbol{\omega})$. As a reminder, an ensemble of size K uses the first K simulations $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ of $\mathbf{Y}$. The three models display quite a different $\overline{S}_{\mathrm{LS}}(K)$ relationship. The logarithmic score of HYMOD and the SAC-SMA model displays an obvious dependence on the ensemble size, while the $\overline{S}_{\mathrm{LS}}(K)$ curve of the Hmodel meanders up and down the graph in a much more unpredictable manner. The logarithmic score is comparatively small for small ensemble sizes but steadily increases to attain a maximum value at $K = 500$ (Hmodel) and $K = 10,000$ for HYMOD and the SAC-SMA model. Beyond this optimum the value of the LS degrades either steadily (HYMOD and SAC-SMA) or intermittently (Hmodel) with ensemble size. Secondly, the optimum ensemble size of Hmodel is substantially smaller than that of HYMOD and the SAC-SMA model. This could have several causes. The default parameter ranges of Hmodel may be too wide for the Leaf River resulting in a proportionally large number of inferior simulations in the GLUE sample whose inclusion in the ensemble can only deteriorate the forecast distribution. But the small ensemble of Hmodel may also signal model inadequacy. Thirdly, the SAC-SMA model consistently achieves the largest values of the LS followed by HYMOD and Hmodel. The optimum ensemble of HYMOD attains a substantially larger LS value than Hmodel, which in turn has two more adjustable parameters. The structure and/or process descriptions of the 7-parameter Hmodel (Fig. H.2) can only be inferior to that of the 5-parameter HYMOD (Fig. H.1) in describing the rainfall-discharge relationship of the Leaf River. Note that the $\overline{S}_{\mathrm{LS}}(K)$ curves hardly differ between

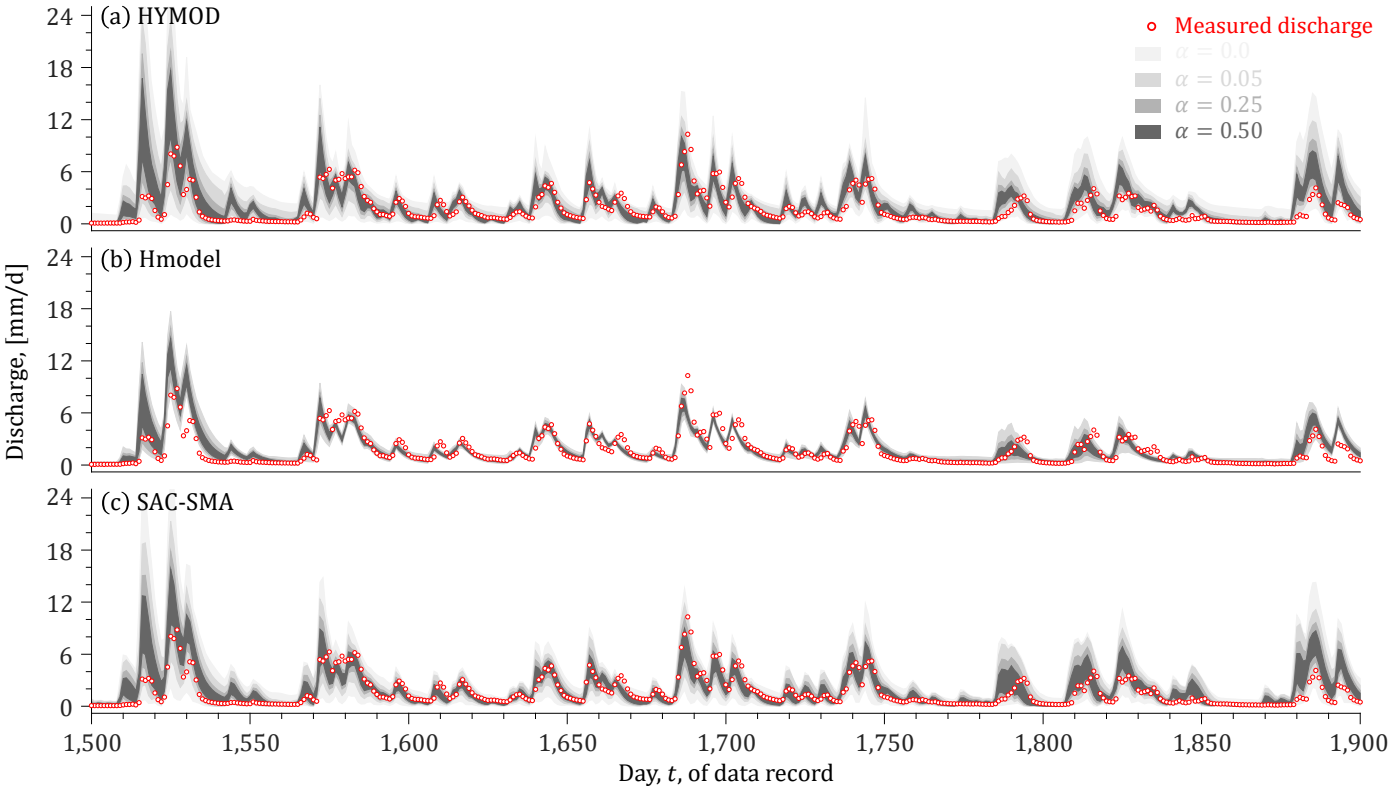

Figure 15: The 50, 75, 95 and 100% simulation uncertainty intervals of the (a) HYMOD, (b) Hmodel and (c) SAC-SMA conceptual hydrologic models for the optimum ensemble according to the logarithmic scoring rule.

different GLUE trials.

Next, we display the simulation uncertainty of the optimum ensemble (Figure 15) for (a) HYMOD, (b) Hmodel and (c) the SAC-SMA model. Table 11 reports different summary statistics of the so-obtained forecast distributions. The three graphs demonstrate several important findings.

1. The different models describe the discharge observations quite well but with the exception of some storm events, for example, on days $1,510$ - $1,530$ and $1,880$ - $1,900$ of the training data record. All three models significantly overestimate the measured discharge. Upon visual inspection of the simulated discharge records, we conclude that this discrepancy is not a consequence of an inaccurate initial state prior to the onset of the storm event but rather signals faulty rainfall estimates.

2. The forecast distributions of HYMOD and SAC-SMA appear quite similar with intervals of the simulated discharge that are in close agreement with each other and much wider on average than those obtained from Hmodel. This is confirmed by the tabulated values of the width W of the 95% prediction intervals, which equal 2.401, 1.549 and 2.386 mm/d for HYMOD, Hmodel and

SAC-SMA, respectively.

3. The relatively sharp forecast distributions of Hmodel are at the expense of an insufficient coverage of its prediction intervals (see Table 11). The prediction intervals of the SAC-SMA model are remarkably accurate and envelop about 44.3, 67.4 and 91.3% of the discharge observations for $\gamma = 50$, $\gamma = 75$ and $\gamma = 95\%$, respectively.

4. The forecast distributions of HYMOD and the SAC-SMA model do not only have a reasonable coverage but also are also remarkably reliable with tabulated R_{l} values of 0.895 and 0.942, respectively. These values are close to one (optimal) and suggest a well dispersed ensemble.

Table 11: Time-averaged values of the moments (skew and kurtosis) and performance metrics (R_{l}, C_{v}, C and W) of the distribution forecasts simulated by HYMOD, Hmodel and the SAC-SMA model and RMSE, R^2 and KGE scoring functions of the mean functional of the forecasts. The coverage C corresponds to significance levels $\alpha = 0.5$, 0.25 and 0.05 of Fig. 15 and d equals the number of unknown parameters.

Model	d	Skew	Kurt.	R_{l}	C_{v}	Coverage C at γ			W	RMSE	R^2	KGE
						0.50	0.75	0.95				
HYMOD	5	1.485	10.60	0.105	0.796	0.385	0.615	0.855	2.401	1.590	0.695	0.679
Hmodel	7	0.003	2.854	0.558	0.365	0.214	0.446	0.778	1.549	1.555	0.708	0.611
SAC-SMA	13	1.042	6.016	0.058	0.739	0.443	0.674	0.913	2.386	1.375	0.772	0.708

Altogether, the results demonstrate that the GLUE methodology coupled with a normal likelihood and logarithmic scoring rule can provide a remarkably proficient forecast ensemble. This offers a simple method for obtaining probabilistic estimates of the simulated state variables and discharge. Parameter calibration may further enhance the ensemble.

6.2.5 Case Study VII: The Flow Duration Curve

One obvious hydrologic application of the scoring rules is the flow duration curve (FDC). This signature catchment characteristic relates the exceedance probability of streamflow, $\mathbb{P}(X > x)$, to its magnitude, x, and plays a critical role in (among others) flood frequency analysis, hydrologic model diagnostics, water quality management and the design of hydroelectric power plants (*Sadegh et al.*, 2016). The FDC is known as the survival function $S_X(x)$ in statistics and the reliability function $R_X(x)$ in engineering. We

adopt the nomenclature of reliability

$$\begin{aligned} R_X(x) = \mathbb{P}(X > x) &= \int_x^{\infty} f_X(t)\,\mathrm{d}t \\ &= 1 - \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \\ &= 1 - F_X(x), \end{aligned} \tag{80}$$

and reconfirm that the FDC is the complement of the streamflow CDF, $F_X(x)$ (*Vogel and Fennessey*, 1994). If we enter the above relationship, $F_X(x) = 1 - R_X(x)$, in the CRPS divergence function of Equation (67)

$$d_{\mathrm{CRPS}}(P,Q) = \int_0^{\infty} \Big(\big(1 - R_P(z)\big) - \big(1 - R_Q(z)\big) \Big)^2 \mathrm{d}z, \tag{81}$$

we yield the divergence function of the supposed continuous ranked exceedance probability score

$$d_{\mathrm{CREPS}}(P,Q) = \int_0^{\infty} \big(R_Q(z) - R_P(z) \big)^2 \,\mathrm{d}z = d_{\mathrm{CRPS}}(P,Q). \tag{82}$$

We do not derive an expression for the actual scoring rule, $S_{\mathrm{CREPS}}(P,\omega)$, of the CREPS divergence. This score will yield the same rankings of the simulated flow duration curves. Furthermore, $d_{\mathrm{CREPS}}(P,Q)$ is nonnegative and zero only when $R_P = R_Q$. We draw inspiration from *Thorarinsdottir et al.* (2013) and decompose the CREPS divergence into a term that summarizes the variability between the reliability functions of P and Q, and two other terms that measure the variability within the FDCs of P and Q

$$d_{\mathrm{CREPS}}(P,Q) = \mathbb{E}_{P,Q}\big[|y - \omega|\big] - \tfrac{1}{2}\big(\mathbb{E}_P\big[|y - y^*|\big] - \mathbb{E}_Q\big[|\omega - \omega^*|\big]\big), \tag{83}$$

where (ω, ω^*) and (y, y^*) are independent copies of the measured and simulated discharge records, respectively. For a measured $\omega_1, \ldots, \omega_n$ and simulated $y_1, \ldots, y_n$ streamflow time series, we use a Monte Carlo estimate of Equation (83)

$$\begin{aligned} d_{\mathrm{CREPS}}(P,Q) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - \omega_j| - \frac{1}{2}\frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - y_j| - \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_i - \omega_j| \right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - \omega_j| - \frac{1}{2}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \big\{ |y_i - y_j| - |\omega_i - \omega_j| \big\}, \end{aligned} \tag{84}$$

The above function is easy to compute and satisfies all the properties of a score divergence deemed desirable by *Ferson et al.* (2008). The $d_{\mathrm{CREPS}}(P,Q)$ function is (i) mathematically well behaved and understood; $d_{\mathrm{CREPS}}(P,Q) > 0$ unless $R_P = R_Q$ then $d_{\mathrm{CREPS}}(P,Q) = 0$, (ii) expressed in physical units (discharge, mm/d), (iii) sensitive to all moments of the FDC, not just mean and variance, and (iv) equal to the absolute error $d_{\mathrm{CREPS}}(\delta_y, \delta_\omega) = |\delta_y - \delta_\omega|$ between two point measures, δ_y and δ_ω.

We revisit the GLUE ensemble and compute Equation (84) for each of the $m = 25,000$ simulated discharge records of the SAC-SMA model for the Leaf River basin. Figure 16 compares the measured FDC of the 3,000-day training data record against simulated reliability functions of the fifty best ensemble members with lowest values of $d_{\text{CREPS}}(P, Q)$ in Equation (84). To benchmark the CREPS divergence, we also present scatter diagrams of (b) $d_{\text{CREPS}}(P, Q)$ and the mean taxicab distance, $\bar{d}_{\text{T}}(P, Q) = \frac{1}{n}\sum_{t=1}^{n} |\omega'_t - y'_t(\boldsymbol{\theta}_i)|$, where $\omega'_1, \ldots, \omega'_n$ and $y'_1, \ldots, y'_n$ are ordered point measures of the measured and simulated discharge records, respectively, and (c) $d_{\text{CREPS}}(P, Q)$ and the log-likelihood $\ell(\boldsymbol{\theta}_i|\boldsymbol{\omega})$ in Equation (79) of the simulated discharge record. Each gray square corresponds to a different member of the GLUE ensemble. The simulated FDCs of the SAC-SMA model are in close agreement with the actual reliability function

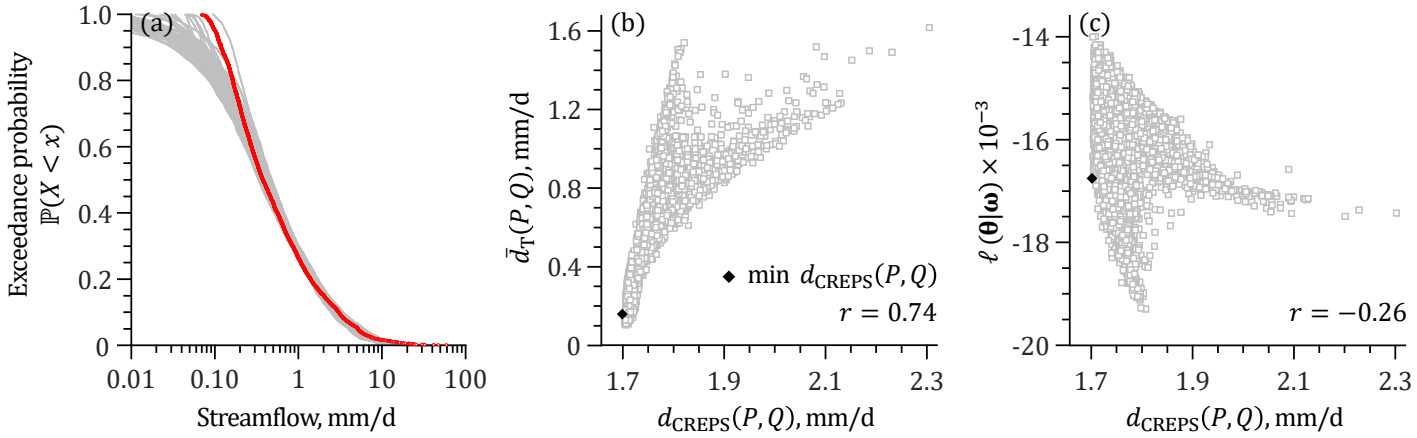

Figure 16: Preliminary results of the cumulative ranked exceedance probability divergence score: (a) measured FDC (red dots) and SAC-SMA model simulated reliability functions (gray lines) of the $K = 50$ discharge records of the ensemble with lowest values of the CREPS divergence and (b,c) bivariate scatter plots of the CREPS divergence and (b) mean taxicab distance, $\bar{d}_{\text{T}}(P, Q)$ and (c) normal log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ of Equation (79) using all discharge records of the GLUE ensemble. The black triangle corresponds to the minimum CREPS divergence.

of the Leaf River. The logarithmic scale emphasizes the discrepancies for streamflows smaller than about 0.1 mm/d. This apparent mismatch will disappear with a linear streamflow scale. Alternatively, we can apply a nonnegative weight function $\int_0^\infty w(z)\,\mathrm{d}z < \infty$ to the squared residuals of the measured and simulated FDCs

$$d_{\text{WCREPS}}(P, Q) = \int_0^\infty \big(R_Q(z) - R_P(z)\big)^2 w(z)\,\mathrm{d}z, \tag{85}$$

and emphasize description of the relative frequencies of the lowest streamflows. When $P, Q \in \mathscr{P}_1$ the above expression is a valid generalization of the CREPS divergence score (*Matheson and Winkler*, 1976). The CREPS divergence of the measured and simulated reliability functions correlates reasonably well with the taxicab distance of R_Q and R_P. But the CREPS divergence is not equivalent to a Manhattan

(and Euclidean) distance, and as a result, we find minimum values of $d_{\mathrm{CRPS}}(P,Q)$ over a range of taxicab distances. The CREPS divergence correlates poorly with the log-likelihood of the SAC-SMA simulated discharge records and attains a minimum value of about 1.7 mm/d well removed from the likelihood maximum of $\ell(\boldsymbol{\theta}|\boldsymbol{\theta}) \approx -14 \times 10^3$. This demonstrates once again that purely statistical metrics of the goodness of fit may compromise the ability of the model to reproduce hydrologically relevant signatures of catchment functioning. This testifies to the added value of process-based diagnostic measures embodied in a *strictly proper* score divergence.

6.3 Interval scoring rules

If so desired one can summarize probabilistic forecasts of a continuous quantity using predictive quantiles. Suppose the forecaster quotes quantiles $\mathbf{r} = (r_1, \ldots, r_z)^\top$ and x materializes, then he or she will be rewarded by the score $S(\mathbf{r}; x)$. Then, the expected score $\mathcal{S}(\mathbf{r}; P)$ under probability measure $P \in \mathscr{P}$ is equal to (*Gneiting and Raftery*, 2007)

$$\mathcal{S}(\mathbf{r}; P) = \int S(\mathbf{r}; x) \mathrm{d}P(x). \tag{86}$$

If $q_1, \ldots, q_z$ are the true quantiles for the class $\mathscr{P}$ of Borel probability measures on $\mathbb{R}$ then a scoring rule is proper if (*Cervera and Muñoz*, 1996)

$$\mathcal{S}(q_1, \ldots, q_k; P) \geq \mathcal{S}(r_1, \ldots, r_k; P) \tag{87}$$

for all real numbers $r_1, \ldots, r_k$ and $P \in \mathscr{P}$. *Gneiting and Raftery* (2007) present a general form of a scoring rule for quantiles. We focus our attention on the coverage C in Table 1 and formulate this insufficient performance metric as a *proper* scoring rule of the $100(1-\alpha)\%$ prediction interval

$$S_{\mathrm{IS}}^{\alpha}(P, \omega) = (l - u) - \frac{2}{\alpha}(l - \omega)\mathbb{1}\{\omega \leq l\} - \frac{2}{\alpha}(\omega - u)\mathbb{1}\{\omega \geq u\}, \tag{88}$$

where $l = F_P^{-1}(\alpha/2)$ and $u = F_P^{-1}(1 - \alpha/2)$ denote the lower and upper endpoints of the predictive quantiles at significance levels $\alpha/2$ and $1 - \alpha/2$, respectively. The interval score $S_{\mathrm{IS}}^{\alpha}(P, \omega)$ is positively oriented and incurs a penalty, the size of which depends on the significance level α, if the observation ω materializes outside the $[u, l]$ prediction interval. From the first term on the right-hand side of Equation (88) it is evident that $S_{\mathrm{IS}}^{\alpha}(P, \omega)$ rewards narrow prediction intervals.

We perform a simple numerical experiment to demonstrate that $S_{\mathrm{IS}}^{\alpha}(P, \omega)$ is a *proper* scoring rule. We draw at random $n = 10^4$ observations, $\omega_1, \ldots, \omega_n$, from a standard normal distribution. From tabulated

critical values, we expect that about 95% of these observations lie in the interval $-1.96 < \omega < 1.96$. Thus, if we fix $\alpha = 0.05$ in Equation (88) then $l = -1.96$ and $u = 1.96$ should maximize the expected value of $S^{\alpha}_{\text{IS}}(P, \omega)$. We verify this assertion in Figure 17 and plot the mean value of the interval score as function of the lower endpoint $l \in [-5, -0.1]$ using $u = 1.96$ and $\alpha = 0.05$ (red), $\alpha = 0.25$ (blue) and $\alpha = 0.50$ (green). The colored lines display a strong dependency of the mean interval score $S^{\alpha}_{\text{IS}}(P, \omega)$ on the choice

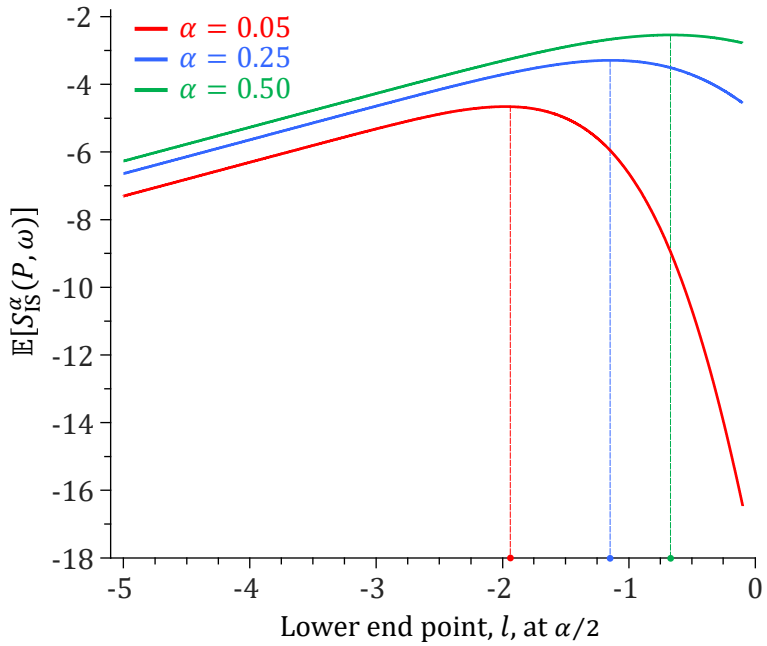

Figure 17: Traces of the expected value of the interval score $S^{\alpha}_{\text{IS}}(P, \omega)$ as function of the lower endpoint l of the $100(1-\alpha)\%$ prediction interval using $\alpha = 0.05$ (red), $\alpha = 0.25$ (blue) and $\alpha = 0.50$ (green). The colored dots are a projection of the maximum interval score on the x-axis.

of the lower endpoint $l = F_P^{-1}(\alpha/2)$. The interval score achieves its largest value, on average, when the forecaster quotes the *true* lower endpoints, $l = -1.96$, $l = -1.15$ and $l = -0.67$ of the $100(1-\alpha)\%$ prediction intervals of $\omega \sim \mathcal{N}(0, 1)$ at significance levels $\alpha = 0.05$, $\alpha = 0.25$ and $\alpha = 0.50$, respectively. This encourages the forecaster to be honest and volunteer his or her true beliefs. To guarantee an accurate description of the *true* forecast distribution one would need to compute $S^{\alpha}_{\text{IS}}(\mathbf{P}, \omega)$ at many different $\alpha \in (0, 1)$ values. But as in *Christoffersen* (1998) this is a daunting task, most certainly in the presence of trade-offs in the interval score of different prediction intervals. When confronted with a time series of forecasts $P_1, \ldots, P_n$ we work with the time-averaged interval score

$$\overline{S}^{\alpha}_{\text{IS}}(\mathbf{P}, \boldsymbol{\omega}) = \frac{1}{n} \sum_{t=1}^{n} S^{\alpha}_{\text{IS}}(P_t, \omega_t). \tag{89}$$

6.4 Multivariate Forecasts

Up until now, we have considered forecast distributions of only a single variable of interest, say, discharge, and verifying observations measured at different times. The overall skill score, $\overline{S}(\mathbf{P}, \boldsymbol{\omega})$, is then a time-averaged mean score as defined in Equation (74). We can expand this approach to multi-variable forecasts using marginalization by treating the distribution of each variable, say discharge, soil moisture content and respective aspects of stream water chemistry, separately. But for such multi-variable forecasts, $P \in \mathscr{P} \in \mathbb{R}^b$, we can also resort to multivariate scoring rules such as the energy score of *Gneiting and Raftery* (2007)

$$S_{\mathrm{ES}}(P, \boldsymbol{\omega}) = \tfrac{1}{2}\mathbb{E}_P\big[\|\mathbf{y} - \mathbf{y}^*\|_2^{\eta}\big] - \mathbb{E}_P\big[\|\mathbf{y} - \boldsymbol{\omega}\|_2^{\eta}\big], \tag{90}$$

where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^b$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_b)^\top$, $\eta \in (0, 2)$, and $\mathbf{y} = (y_1, \ldots, y_b)^\top$ and $\mathbf{y}^* = (y_1^*, \ldots, y_b^*)^\top$ are independent copies of the b-variate distribution, $P \in \mathscr{P}_\eta$. For $b = 1$, the above expression reduces to $S_{\mathrm{ES}}(P, \omega)$ in Equation (68). The multivariate form of the energy score is a *strictly proper* score (*Székely*, 2003) and reduces to

$$S_{\mathrm{SE}}(P, \boldsymbol{\omega}) = -\,\|\boldsymbol{\mu}_P - \boldsymbol{\omega}\|_2^2\,, \tag{91}$$

for $\eta \to 2$, where $\boldsymbol{\mu}_P = (\mu_{1P}, \ldots, \mu_{bP})^\top$ signifies the mean of the forecast distribution. The squared error can also be written as a vector-inner product, $S_{\mathrm{SE}}(P, \boldsymbol{\omega}) = -(\boldsymbol{\mu}_P - \boldsymbol{\omega})^\top(\boldsymbol{\mu}_P - \boldsymbol{\omega})$. Analogous to the numerical definition of the CRPS in Equation (64), the ES may be approximated using a large collection $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ of m samples of the forecast distribution P (*Grimit et al.*, 2006)

$$S_{\mathrm{ES}}(P, \boldsymbol{\omega}) = \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{y}_i - \mathbf{y}_j\|_2^{\eta} - \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{y}_i - \boldsymbol{\omega}\|_2^{\eta}\,. \tag{92}$$

Dawid (1998) and *Dawid and Sebastiani* (1999) studied scoring rules that depend only on the mean, $\boldsymbol{\mu}_P \in \mathbb{R}^{b\times 1}$, and the dispersion or covariance matrix, $\boldsymbol{\Sigma}_P \in \mathbb{R}^{b\times b}$ of the forecast distribution P. Their divergence function, $d_{\mathrm{DDS}}(P, Q)$, in Equation (5) is linked to the *proper* scoring rule (*Dawid and Sebastiani*, 1999)

$$S_{\mathrm{DSS}}(P, \boldsymbol{\omega}) = -\log_e(|\boldsymbol{\Sigma}_P|) - (\boldsymbol{\omega} - \boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_P), \tag{93}$$

and generalized entropy function

$$H(P) = -\log_e(|\boldsymbol{\Sigma}_P|) - b. \tag{94}$$

Note that the DSS is simply equal to the unnormalized log-likelihood of a multivariate normal density. For a univariate forecast, $b = 1$, the DSS simplifies to $S_{\mathrm{DSS}}(P, \omega) = -\log_e(\sigma_P^2) - (\omega - \mu_P)^2/\sigma_P^2$, where, again, μ_P and σ_P^2 are the mean and variance of the forecast distribution.

A multivariate forecast does not necessarily have to involve multiple different variables. It can also consist of a single variable but forecasted at many different sites. Suppose $i, j \in (1, \ldots, b)$ are single indices for the points in a square or rectangular grid made up of b sites. *Scheuerer and Hamill* (2015) investigate the accuracy of the forecast distribution P in describing the spatial structure of gridded measurements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_b)^\top$ of 80-meter wind speed forecasts. They introduced the so-called variogram score of order $\varsigma > 0$

$$S_{\mathrm{VS}}^{\varsigma}(P, \boldsymbol{\omega}) = -\sum_{i=1}^{b} \sum_{j=1}^{b} w_{ij} \big(|\omega_i - \omega_j|^{\varsigma} - \mathbb{E}_P \big[|y_i - y_j|^{\varsigma} \big] \big)^2, \tag{95}$$

where $w_{ij} \geq 0$ is a nonnegative weight attached to the $(i, j)^{\mathrm{th}}$ pair of sites and y_i and y_j are the i^{th} and the j^{th} elements (sites) of a random vector that is distributed according to P. The weights allows one to emphasize or downplay specific aspects of the distribution forecast. The variogram scoring rule is analogous to the squared error $S_{\mathrm{SE}}(P, \boldsymbol{\omega})$ of Equation (91) using residuals of the powered differences of pairs of measurement sites and pairs of forecast sites. For $\varsigma = 2$ the powered difference is known as the semi-variance. When the forecast distribution P is given in the form of a m-member ensemble $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ the second term of $S_{\mathrm{VS}}^{\varsigma}(P, \boldsymbol{\omega})$ can be approximated by

$$\mathbb{E}_P \big[|y_i - y_j|^{\varsigma} \big] = \frac{1}{m} \sum_{k=1}^{m} |y_{ik} - y_{jk}|^{\varsigma}. \tag{96}$$

The variogram scoring rule is more sensitive than existing scoring rules to multivariate site dependencies. But as the variogram score admits site differences, it is insensitive to the location of the distribution forecast, thus, is a *proper* and not *strictly proper* scoring rule. Forecast distributions that differ only in their mean will receive the same values of $S_{\mathrm{VS}}^{\varsigma}(P, \boldsymbol{\omega})$.

We could conveniently discard time as controlling variable of the rainfall-runoff transformation and treat the discharge time series of HYMOD, Hmodel and the SAC-SMA model as a multivariate forecast $\mathbf{y} = (y_1, \ldots, y_b)^\top$ with dimensionality b equal to the length $n = 3,000$ of the simulated records. The m-member ensemble $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ will produce values of $S_{\mathrm{ES}}(P, \boldsymbol{\omega})$, $S_{\mathrm{DSS}}(P, \boldsymbol{\omega})$ and $S_{\mathrm{VS}}^{\varsigma}(P, \boldsymbol{\omega})$ according to Equations (92), (94) and (95), respectively.

7 Decomposition of scoring rules for categorical forecasts

Strictly proper scoring rules condense the accuracy of the distribution forecast P to a single reward oriented value while retaining attractive statistical properties. This compression to a single reward value

simplifies tasks such as forecast verification, model evaluation and likelihood function selection (e.g. *Vrugt et al.* 2022), but makes it difficult to pinpoint exactly which attributes of the distribution forecast are deficient and in need of improvement. As a result, *strictly proper* scoring rules lack diagnostic power on how to improve the overall consistency, accuracy and precision of model forecasts. Thus, it may be beneficial to decompose scoring rules into their constituent parts. Decomposition of the expected loss into a calibration and refinement loss has facilitated the development of calibration methods (*Bella et al.*, 2013). The refinement loss consists of an uncertainty term and a resolution term (*DeGroot and Fienberg*, 1983; *Murphy*, 1973). *Kull and Flach* (2015) presents a decomposition of the logarithmic and Brier scoring rules into an epistemic and irreducible (aleatoric) loss term. Next, we review the decomposition of *strictly* proper scoring rules into an uncertainty, resolution and reliability term. These components relate directly to forecast attributes that are deemed desirable on grounds independent of the scoring rules themselves and provide an epistemological justification of measuring forecast quality by *strictly proper* scoring rules (*Bröcker*, 2009).

7.1 Theory

Let $\Omega = \{1, 0\}$ be the sample space of a binary event of *rain* or *no rain*. Let the quoted probability $p = p(\mathcal{D})$ of *rain* be a function of the data $\mathcal{D}$ available to the forecaster up to a certain lead time, where $p \in [0, 1]$. A forecasting scheme models the relationship between the data $\mathcal{D}$ and the quantity of interest or forecast (see *Murphy and Winkler* 1987; *Murphy* 1993, 1996 for a related discussion). Once we observe $\omega \in \Omega$, we can assign a score $S(p, \omega) : \mathscr{P}_2 \times \Omega \to \mathbb{R}$ to the prediction. Thus, ω takes on values of 0 (*no rain*) and 1 (*rain*). The law of total expectation states that if $S(p, \omega)$ and $\mathcal{D}$ are random variables on the same probability space then

$$\mathbb{E}[S(p, \omega)] = \mathbb{E}\big[\mathbb{E}[S(p, \omega)|\mathcal{D}]\big]. \tag{97}$$

Suppose we use the Brier or quadratic score, $S_{\mathrm{QS}}(p, \omega) = -\sum_{k=1}^{2}(\delta_{\omega k} - p_k)^2$, where $\delta_{\omega k} = 1$ if $\omega = k$ and $\delta_{\omega k} = 0$ otherwise. Then we can decompose the conditional expectation

$$\begin{aligned}
\mathbb{E}[S_{\mathrm{QS}}(p, \omega)|\mathcal{D}] &= \mathbb{E}\big[-\big(\omega - p(\mathcal{D})\big)^2|\mathcal{D}\big] \\
&= -\mathbb{E}\big[\big(\omega - \mathbb{E}[\omega|\mathcal{D}] + \mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D})\big)^2|\mathcal{D}\big] \\
&= -\operatorname{Var}[\omega|\mathcal{D}] - \big(\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D})\big)^2,
\end{aligned} \tag{98}$$

and insert Equation (98) into Equation (97) to yield

$$\begin{aligned}\mathbb{E}[S_{\mathrm{QS}}(p,\omega)] &= \mathbb{E}\big[-\operatorname{Var}[\omega|\mathcal{D}] - \big(\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D})\big)^2\big] \\ &= -\mathbb{E}\big[\operatorname{Var}[\omega|\mathcal{D}]\big] - \mathbb{E}\big[\big(\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D})\big)^2\big] \\ &= -\operatorname{Var}[\omega] + \operatorname{Var}\big[\mathbb{E}[\omega|\mathcal{D}]\big] - \mathbb{E}\big[\big(p(\mathcal{D}) - \mathbb{E}[\omega|\mathcal{D}]\big)^2\big], \end{aligned} \tag{99}$$

where $\mathbb{E}[\omega|\mathcal{D}]$ is simply equal to the conditional probability of *rain* and $p(\mathcal{D})$ equals the unconditional *rain* probability. *Bröcker* (2009) generalized the above decomposition to a generic *strictly proper* scoring rule, $S(\mathbf{p},\omega) : \mathscr{P}_m \times \Omega \to \mathbb{R}$ of a categorical forecast of $m \geq 2$ events to yield

$$\mathbb{E}[S(\mathbf{p},\omega)] = \underbrace{H(\overline{\mathbf{p}})}_{1} + \underbrace{\mathbb{E}[d(\overline{\mathbf{p}},\boldsymbol{\pi})]}_{2} - \underbrace{\mathbb{E}[d(\mathbf{p},\boldsymbol{\pi})]}_{3}, \tag{100}$$

where $\overline{\mathbf{p}} = (\overline{p}_1, \ldots, \overline{p}_m)^\top$ is the unconditional probability of ω (so-called climatology in meteorology) and $\pi_k = \mathbb{P}(\omega = k|\mathbf{p})$ signifies the conditional probability of observation ω for the probabilities $\mathbf{p}$ quoted, $k = (1, \ldots, m)$. Hence, $\boldsymbol{\pi}$ is a mapping ($m \times m$ matrix) which specifies for every $\omega \in \Omega$ a probability measure on $\mathscr{P}_m$. The three terms are nonnegative for *strictly proper* scoring rules and referred to as (1) uncertainty (of ω), (2) resolution (or sharpness) and (3) reliability (*Bröcker*, 2009). The entropy and resolution have a positive effect on $\mathbb{E}[S(\mathbf{p},\omega)]$, whereas the reliability decreases the expected value of the score. Note that the minus sign of $-\operatorname{Var}[\omega]$ in Equation (99) has disappeared from the uncertainty term as *Bröcker* (2009) uses $-H(\overline{\mathbf{p}})$ for the entropy function. Furthermore, the use of positively oriented scoring rules reverses the signs in front of the resolution and reliability terms.

The first term of the decomposition is uncertainty and quantifies our state of knowledge in the absence of an underlying theory to generate forecasts. This is equivalent to the expected score of the average event frequencies or climatology as forecast probabilities. This term is independent of the forecasting system and depends only on the statistics of the observations (*Christensen*, 2015). The second term, $\mathbb{E}[d(\overline{\mathbf{p}},\boldsymbol{\pi})]$ or resolution measures the mean divergence of the conditional event probabilities $\boldsymbol{\pi}$ from their average probabilities $\overline{\mathbf{p}}$ and, thus is a proxy for the variance of $\boldsymbol{\pi}$ over the sample Ω or data $\mathcal{D}$ space. At zero resolution $\boldsymbol{\pi}$ is always equal to the climatology (or prior mean) $\overline{\mathbf{p}}$ and, thus, the data $\mathcal{D}$ provides no useful information. Thus, in accordance with the sharpness principle of (*Gneiting and Raftery*, 2007) larger values of the resolution are preferred and reflect case-dependent probabilistic forecasts. The third and last term, reliability, measures the average deviation of the probabilistic forecast $\mathbf{p}$ from the conditional event probabilities $\boldsymbol{\pi}$. This is a measure of the statistical consistency of a forecast and evaluates whether

the quoted forecast probabilities are in agreement with the materialized event frequencies. Thus, the reliability penalizes poorly calibrated forecasts as $\mathbf{p}$ will not match $\boldsymbol{\pi}$.

Equation (100) is a generalisation of the well known decomposition of the Brier score of *Murphy* (1973)

$$\text{1. uncertainty} = \overline{p}(1-\overline{p}) \qquad \text{2. resolution} = \mathbb{E}[(\overline{p}-\pi_1)^2] \qquad \text{3. reliability} = \mathbb{E}[(p-\pi_1)^2], \tag{101}$$

where $\pi_1 = \pi$ is the conditional probability of *rain* given p, $\pi_1 = \mathbb{P}(x = 1|p)$. We refer to *Weijs et al.* (2010b) for a reliability-resolution-uncertainty decomposition of the KL-divergence, the divergence of the logarithmic scoring rule.

7.2 Case Study VII: Discharge Forecast Ensemble

We illustrate the analytic decomposition of Equation (100) by application to the multi-model ensemble of discharge forecasts displayed in Figure 11. In doing so we must first convert discharge to a categorical variable with number of possible outcomes m equal to the ensemble size, $K = 8$. In this discrete sample space, $\Omega = \{1, \ldots, m\}$, the measured daily discharge ω_t at $t \geq 1$ coincides with the "best" discharge forecast among the watershed models. The index $\hbar$ of this "best" forecast at each time t is the index of the minimum entry of the m-vector of absolute residuals

$$\hbar = \underset{\hbar \in (1,\ldots,K)}{\arg\min} |\omega_t - y_{\hbar t}|, \tag{102}$$

between the measured and forecasted discharges. Suppose the vector $(\hbar_1, \ldots, \hbar_n)^\top$ stores the indices of the best models for our training record of $n = 3,000$ days, then the event frequencies $\overline{\mathbf{p}} = (\overline{p}_1, \ldots, \overline{p}_m)$ may be computed using

$$\overline{p}_k = \mathbb{P}(\omega = y_k) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hbar_t = k\}, \tag{103}$$

and the conditional forecast probabilities equal

$$\begin{aligned} \pi_{jk} = \mathbb{P}(\omega = y_j|y_k) &= \frac{\sum_{t=2}^{n} \mathbb{1}\{\hbar_t = j|\hbar_{t-1} = k\}}{\sum_{t=1}^{n} \mathbb{1}\{\hbar_t = k\}} \\ &= \frac{\sum_{t=2}^{n} \mathbb{1}\{\hbar_t = j|\hbar_{t-1} = k\}}{n\overline{p}_k}, \end{aligned} \tag{104}$$

where $j, k = (1, \ldots, K)$. Table 12 reports the unconditional and conditional probabilities of the watershed models.

Table 12: Unconditional, $\overline{\mathbf{p}}$, and conditional, $\boldsymbol{\pi}$, probabilities of the watershed models estimated from the 3,000-day training data record: $\pi_{jk} = \mathbb{P}(\omega = y_j | y_k)$ is the probability of y_j given that y_k is the best forecast in the ensemble at the previous time.

	Model	ABC	GR4J	HYMOD	TOPMO	AWBM	NAM	HBV	SAC-SMA
$\overline{\mathbf{p}}$		0.064	0.148	0.088	0.101	0.080	0.142	0.175	0.203
$\boldsymbol{\pi}$	ABC	**0.267**	0.050	0.075	0.056	0.046	0.056	0.048	0.035
	GR4J	0.157	**0.534**	0.098	0.076	0.087	0.033	0.061	0.100
	HYMOD	0.099	0.063	**0.343**	0.102	0.083	0.035	0.057	0.051
	TOPMO	0.099	0.070	0.094	**0.383**	0.054	0.059	0.061	0.069
	AWBM	0.037	0.043	0.053	0.063	**0.442**	0.052	0.033	0.059
	NAM	0.099	0.045	0.087	0.106	0.062	**0.609**	0.038	0.061
	HBV	0.105	0.086	0.094	0.102	0.033	0.061	**0.608**	0.094
	SAC-SMA	0.136	0.110	0.155	0.112	0.192	0.094	0.094	**0.531**
	$\sum_{j=1}^{8} \pi_{jk}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

The unconditional forecast probabilities tend to increase with model complexity and is largest for the SAC-SMA model. The conditional forecast probabilities are largest on the main diagonal (in bold) confirming that a model's probability is largest conditional on it having the best forecast in the ensemble. Note that each column of $\boldsymbol{\pi}$ sums to unity.

Table 13 presents the expected value of the quadratic score, entropy, resolution and reliability of the BMA mixture distribution for the PDFs of Table 8 with a constant group and sole forecast variance, respectively. In each case, we set the forecast probabilities $\mathbf{p} = (p_1, \ldots, p_K)^\top$ equal to the maximum likelihood values of the weights $\beta_1, \ldots, \beta_K$ of the BMA mixture distribution.

Table 13: Time-averaged expected value of the *strictly proper* quadratic scoring rule (see Table 6), entropy, resolution and reliability for the BMA density forecasts of Equation (70) using the normal, lognormal, generalized normal, gamma and Weibull distributions with a constant group and sole variance, respectively. The bottom row completes the decomposition of Equation (100).

	Normal[a]		Lognormal		Gen. Normal		Gamma		Weibull	
	group	sole	group	sole	group	sole	group	sole	group	sole
$\mathbb{E}[S_{\mathrm{QS}}(\mathbf{p}, \omega)]$	0.027	0.100	0.031	0.033	0.108	0.108	0.073	0.116	0.066	0.071
$H(\overline{\mathbf{p}})$	0.142	0.142	0.142	0.142	0.142	0.142	0.142	0.142	0.142	0.142
$\mathbb{E}[d(\overline{\mathbf{p}}, \boldsymbol{\pi})]$	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162
$\mathbb{E}[d(\mathbf{p}, \boldsymbol{\pi})]$	0.277	0.204	0.272	0.271	0.196	0.196	0.231	0.188	0.238	0.233
Sum	0.027	0.100	0.031	0.033	0.108	0.108	0.073	0.116	0.066	0.071

[a] We fix $\tau = 2$ in the generalized normal density of Table 8

The tabulated data confirm the decomposition of Equation (100). The expected value of the quadratic

scoring rule (top row) is indeed equal to the entropy, $H(\overline{\mathbf{p}})$, plus the resolution, $\mathbb{E}[d(\overline{\mathbf{p}}, \boldsymbol{\pi})]$, minus the reliability, $\mathbb{E}[d(\mathbf{p}, \boldsymbol{\pi})]$. The first two terms of this decomposition depend only on the unconditional and conditional event frequencies, $\overline{\mathbf{p}}$ and $\boldsymbol{\pi}$, respectively, hence, do not differ among the component functions of the BMA forecast distribution. With exception of the generalized normal density, the use of a model-dependent forecast variance increases forecast reliability. The BMA model probabilities $\mathbf{p} = (p_1, \ldots, p_K)^\top$ derived from the gamma distribution maximize the expected value of the quadratic scoring rule. This conclusion contradicts earlier findings reported in Table 9 which linked the gamma distribution to the lowest value of the quadratic score. Thus, the use of the maximum likelihood solution of the gamma distribution leads to diametrically opposed conclusions about model adequacy. This testifies to the premise of this paper that only a complete evaluation of the forecast density provides an honest and fair assessment of model adequacy.

8 Outlook

Scoring rules lie at the heart of statistical theory and practice and guarantee a more robust and complete evaluation of hydrologic models. But scoring rules may satisfy other purposes as well. We briefly describe a few avenues for future work. For obvious reasons, this discussion is succinct and incomplete.

8.0.1 Localized Scoring Rules: Extreme Events

One may only be interested in evaluating certain aspects of the forecast distribution, for example, the probability of extreme events in the lower and/or upper tail of the distribution. The CREPS divergence $d_{\mathrm{WCREPS}}(P, Q)$ in Equation (85) allows differential weighting of the flow levels in the description of the flow duration curve. If all weights are strictly positive, then this localization will not sacrifice the strict propriety of the CREPS divergence. The nonnegative weight function $w_{ij} = \max(0, 1 - \frac{1}{9}|i - j|^2)$ used by *Scheuerer and Hamill* (2015) in the application of the $S^{\varsigma}_{\mathrm{VS}}(P, \boldsymbol{\omega})$ favors an accurate probabilistic description of the powered differences between nearby sites over such differences from distant sites. But as all sites with $|i - j| \geq 3$ attain a zero weight, the above formulation of the variogram score is locally proper at best. No localized scoring rule can differentiate between predictive densities on sets with a zero weight, thus, satisfy propriety only for the domain of interest, say extreme events (*Diks et al.*, 2011;

de Punder et al., 2022). A simple work-around is to assign a minimum nominal weight to each set. The weight function also enables users to incorporate soft information and/or personal judgment in model evaluation.

8.0.2 Flood frequency analysis

Flood frequency analysis usually involves the fitting of the parameters of some known probability distribution to a training data record of log-transformed annual maxima discharges, $\omega_1, \ldots, \omega_n$. The marginal likelihood (Bayes factors) of the estimated model parameters will convey which assumption about the mean of the distribution (constant or time-dependent) receives most support from the measured data record (*Luke et al.*, 2017). This is one of many examples which may benefit from the application of scoring rules for hypothesis testing of stationary and nonstationary flood frequency models. Suppose we choose the Pearson type III distribution $\mathcal{P}_{\text{III}}(\mu, \sigma^2, \rho)$ with mean μ, variance σ^2 and skewness ρ for the log-transformed annual maxima discharges. If we reparameterize the location, shape and scale parameters of the PIII distribution to $\xi = \mu - 2\sigma/\rho$, $a = 4/\rho^2$ and $b = \frac{1}{2}\sigma|\rho|$, respectively, then the PDF of P simplifies to (*Hosking and Wallis*, 1997; *Tegos et al.*, 2022)

$$f_P(x, \xi, a, b) = \frac{|x - \xi|^{a-1}}{b^a \Gamma(a)} \exp\big(-b^{-1}|x - \xi|\big), \tag{105}$$

where $x, \xi, a, b \in \mathbb{R}$, $a > 0$ and $b > 0$. If $\rho > 0$ then $x \in (\xi, \infty)$, otherwise for $\rho < 0$ we yield $x \in (-\infty, \xi)$. The logarithmic score, $S_{\text{LS}}(P, \omega) = \log_{\mathfrak{b}}\big(f_P(\omega)\big)$, of distribution forecast $P = \mathcal{P}_{\text{III}}(\xi, a, b)$ for the materialized outcome ω_t is now equal to

$$S_{\text{LS}}\big(\mathcal{P}_{\text{III}}(\xi, a, b), \omega_t\big) = (a-1)\log_e(|\omega_t - \xi|) - a\log_e(b) - \log_e\big(\Gamma(a)\big) - b^{-1}|\omega_t - \xi|, \tag{106}$$

with Euler's number as logical choice for the base of the score. The model that achieves the highest value of the *strictly proper* logarithmic score is preferred by the data record. The QS, PSS and SS are readily computed as well but an analytic expression for the CRPS of a PIII distribution forecast $P = \mathcal{P}_{\text{III}}(\xi, a, b)$ is more involved (see Appendix I) and results in

$$\begin{aligned} S_{\text{CRPS}}\big(\mathcal{P}_{\text{III}}(\xi, a, b), \omega\big) = 2 \times \frac{4^{-a} b}{B(a, a)} - ab + |\omega - \xi| + 2abF_{\mathcal{G}}(|\omega - \xi|, a+1, b) \\ - 2|\omega - \xi| F_{\mathcal{G}}(|\omega - \xi|, a, b), \end{aligned} \tag{107}$$

where $B(u, v) = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ is the beta function of the first kind and $F_{\mathcal{G}}(z, a, b)$ is the CDF of the gamma distribution $\mathcal{G}(a, b)$ with shape and scale parameters, $a > 0$ and $b > 0$, respectively. The

first term in the CRPS of $P = \mathcal{P}_{\text{III}}(\xi, a, b)$ is equal to the first term, $\frac{1}{2}\mathbb{E}_P\big[|y - y^*|\big]$, of Equation (63) and related to the well-known concentration index G of Corrado Gini for measuring the degree of inequality in the distribution of income and wealth (*McDonald and Jensen*, 1979; *Scheuerer and Möller*, 2015)

$$\frac{1}{2}\mathbb{E}_P\big[|y - y^*|\big] = 2 \times \frac{4^{-a}b}{B(a,a)} = abG = ab\frac{\Gamma(a+\frac{1}{2})}{\sqrt{\pi}\Gamma(a+1)}. \tag{108}$$

As $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we yield the numerically more robust $\frac{ab}{\pi}B\big(a+\frac{1}{2},\frac{1}{2}\big)$ for the first term of Equation (107). For completeness, Appendix J presents an analytic form of the CRPS for a generalized extreme value distribution forecast $P = \mathcal{GEV}(\mu, \sigma^2, \xi)$ (*Friederichs and Thorarinsdottir*, 2012).

8.0.3 Bayesian model selection

We revisit Bayes theorem in Equation (15) and look in more detail at the normalization constant $p(\mathcal{D}|\mathcal{H})$

$$p(\mathcal{D}|\mathcal{H}) = \int_{\boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta}|\mathcal{H})\, \mathrm{d}\boldsymbol{\theta}. \tag{109}$$

where $L(\boldsymbol{\theta}|\mathcal{D}, \mathcal{H}) \equiv p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H})$ is the likelihood under hypothesis $\mathcal{H}$ and $p(\boldsymbol{\theta}|\mathcal{H})$ denotes the prior parameter distribution. The normalizing constant is also referred to as the integrated or marginal likelihood. Now for two competing hypothesis, $\mathcal{H}_0$ and $\mathcal{H}_1$, of a finite or countable class of statistical models with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in a d-dimensional Euclidean space, we wish to determine which hypothesis (possibly with parameter values) is most supported by a sample $\mathcal{D} = (\omega_1, \ldots, \omega_n)^\top$ of materialized outcomes of $\boldsymbol{\Omega}$. The Bayes factor $B_{1,0}$ for $\mathcal{H}_1$ against $\mathcal{H}_0$ (*Jeffreys*, 1939; *Kass and Raftery*, 1995)

$$B_{1,0} = \frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_0)}, \tag{110}$$

summarizes the evidence provided by the data $\mathcal{D}$ in favor of hypothesis $\mathcal{H}_1$, represented by a watershed model, as opposed to the null hypothesis, $\mathcal{H}_0$. *Good* (1952) established a simple relationship between the logarithmic score and the logarithmic value of the Bayes factor

$$\log_{\mathfrak{b}}(B_{1,0}) = \log_{\mathfrak{b}}\left(\frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_0)}\right) = S_{\text{LS}}(\mathcal{H}_1, \mathcal{D}) - S_{\text{LS}}(\mathcal{H}_0, \mathcal{D}), \tag{111}$$

where $\log_{\mathfrak{b}}(B_{1,0})$ is also referred as the *weight of evidence*. We can use this identity to compute the values of the Bayes factors for the competing distribution forecasts of Fig. 4a (see Table 14). The listed Bayes factors confirm that P_3 the best predictive density among the distribution forecasts unless, of course, we set $P = Q$. The evidence for P_3 is strong as a result of the large sample of ten-thousand observations.

Table 14: Mean values of the Bayes factors $B_{i,j}$ of the distribution forecasts $P_1, \ldots, P_5$ displayed in Fig. 4a from application of Equation (111) to the sum of the logarithmic scores $S_{\mathrm{LS}}(P, \omega)$ in units of bits ($\mathfrak{b} = e$) of the $n = 10^4$ outcomes $\omega_1, \ldots, \omega_n$. The last row and column correspond to a perfect distribution forecast, $P = Q$. *Kass and Raftery* (1995) (P. 777) categorize the Bayes factors and present descriptive statements on the strength of the evidence; $B_{i,j} = 0$ favors in strongest possible terms the null hypothesis, P_j, and vice-versa $B_{i,j} > 75$ means that there is decisive evidence against P_j.

Forecast	$\mathcal{S}_{\mathrm{LS}}(P, Q)$	P_1 red	P_2 blue	P_3 green	P_4 yellow	P_5 purple	$P = Q$ gray
P_1	-6.520		0.000	0.000	0.000	0.000	0.000
P_2	-2.422	> 75		0.000	0.000	> 75	0.000
P_3	-2.190	> 75	> 75		> 75	> 75	0.000
P_4	-2.260	> 75	> 75	0.000		> 75	0.000
P_5	-2.479	> 75	0.000	0.000	0.000		0.000
$P = Q$	-2.132	> 75	> 75	> 75	> 75	> 75	

For this same reason, we only need to look at the time-averaged values of the logarithmic score $\overline{S}_{\mathrm{LS}}(P, \boldsymbol{\omega})$ in Table 9 to realize that there is overwhelming evidence for the lognormal density.

When the data come in a particular sequence, such as time order, we may develop a more intuitive understanding of the above identity if we look at the predictive density of ω_t given past observations $\mathcal{D}^{t-1} = (\omega_1, \ldots, \omega_{t-1})^\top$

$$p(\omega_t|\mathcal{D}^{t-1}, \mathcal{H}) = \int_{\boldsymbol{\Theta}} p(\omega_t|\boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta}|\mathcal{D}^{t-1}, \mathcal{H}) \, \mathrm{d}\boldsymbol{\theta}. \tag{112}$$

where $p(\omega_t|\boldsymbol{\theta}, \mathcal{H})$ is the predictive density of w_t given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $p(\boldsymbol{\theta}|\mathcal{D}^{t-1}, \mathcal{H})$ signifies the posterior distribution of the parameters and $t = (1, \ldots, n)$. The logarithmic score for ω_t is now equal to

$$S_{\mathrm{LS}}(\mathcal{H}, \omega_t) = \log_{\mathfrak{b}}\big(p(\omega_t|\mathcal{D}^{t-1}, \mathcal{H})\big), \tag{113}$$

and we can write

$$S_{\mathrm{LS}}(\mathcal{H}, \mathcal{D}) = \sum_{t=1}^{n} \log_{\mathfrak{b}}\big(p(\omega_t|\mathcal{D}^{t-1}, \mathcal{H})\big), \tag{114}$$

for the total score in Equation (111). The above scoring rule is asymptotically equivalent to the Bayes information criterion (*Dawid*, 1984). Thus, the logarithmic score may not only be used to evaluate, contrast and rank different distribution forecasts but has broader application to model selection.

8.0.4 Synthesis with model diagnostics

There is an urgent need for *proper* scoring rules of hydrologic signatures in support of diagnostic model evaluation. We proposed steps in this direction for the flow duration curve with the CREPS divergence

score but are in need of a much larger family of *strictly proper* signature scoring rules. The relevant number and form of diagnostic signatures is catchment and theory dependent. Let us restrict attention to numeral descriptors of the stream hydrograph, examples of which are the baseflow index, runoff ratio, rising limb density (*Shamir et al.*, 2005) and flashiness index (*Baker et al.*, 2004). Time-averaged values of these indices are frequently used for hydrologic model evaluation, but this projection of the hydrograph to a handful of points implies a significant loss of information. But this loss does not have to be as colossal if we work instead with distribution functions of the hydrologic signatures. We can implement the moving-block bootstrap of *Kunsch* (1989) and obtain frequency distributions of the hydrologic signatures by shifting a window of constant width, say 365 days, by one or more days through the streamflow record and/or hyetograph. The choice of increment exerts control on the smoothness of the signatures' distribution functions. A distribution-based approach to model diagnostics arguably is much more robust and complete in that it, (i) compares measured and simulated signature distributions and not just their mean values, (ii) acknowledges temporal variability of the signatures, and (iii) implicitly accounts for signature uncertainty. The CRPS divergence function in Equation (67) may be used to quantify the statistical distance between measured and simulated signature distributions. Confidence intervals can be derived from the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (*Dvoretzky et al.*, 1956) and its extension by *Naaman* (2021) to multivariate distributions.

The methodology described above evaluates each signature separately, thus, is invariant to the multivariate relationships between the signatures. Suppose F_Q and F_P are r-variate CDFs of the *true* and *simulated* signatures and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_r)^\top$ is a signature-valued vector drawn at random from Q. We can generalize distribution-based model diagnostics to the joint distribution of the signatures using the multivariate form of the CRPS

$$S_{\mathrm{MCRPS}}(P, \boldsymbol{\omega}) = -\int_{\Omega} \big(F_P(\mathbf{u}) - \mathbb{1}\{\boldsymbol{\omega} \leq \mathbf{u}\}\big)^2 \mathrm{d}\mathbf{u}, \tag{115}$$

and associated divergence function

$$d_{\mathrm{MCRPS}}(P, Q) = \int_{\Omega} \big(F_P(\mathbf{u}) - F_Q(\mathbf{u})\big)^2 \mathrm{d}\mathbf{u}, \tag{116}$$

where $\mathbf{u} \in \Omega \subseteq \mathbb{R}^r$. By expanding the integrand of Equation (115) we yield a term $\int_{\Omega} \mathbb{1}\{\boldsymbol{\omega} \leq \mathbf{u}\}^2 \mathrm{d}\mathbf{u}$, which depends only on $\boldsymbol{\omega}$ and not F_P. Thus,

$$S^*_{\mathrm{MCRPS}}(P, \boldsymbol{\omega}) = -\int_{\Omega} F_P^2(\mathbf{u}) \mathrm{d}\mathbf{u} + 2\int_{\Omega} F_P(\mathbf{u}) \mathbb{1}\{\boldsymbol{\omega} \leq \mathbf{u}\} \mathrm{d}\mathbf{u}, \tag{117}$$

is an affine transformation of the MCRPS scoring rule (see e.g. *Meng et al.* 2022). If the distribution P is made up of m samples $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\} \in \mathbb{R}^{r \times m}$ of signature values, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})^\top \in \mathbb{R}^{r \times 1}$, then the multivariate eCDF at point $\mathbf{u} = (u_1, \ldots, u_r)^\top$ is equal to

$$F_P(\mathbf{u}) = F_P(\mathbf{u}|\mathbf{Y}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{y_{i1} \leq u_1, \ldots, y_{ir} \leq u_r\}. \tag{118}$$

The above formulation can be used in Equation (116) for the signatures of the simulated and measured discharge records. *Langrené and Warin* (2021) presents two different algorithms for accurate and CPU-efficient estimation of the multivariate eCDF requiring only $\mathcal{O}(m)$ operations if samples are sorted. Note that the MCRPS divergence score can be used as a loss function for watershed model parameter estimation so as to maximize the agreement between the measured and simulated joint signature distributions.

The multivariate form of the CRPS divergence can also be used for individual signatures. For example, *Brutsaert and Nieber* (1977), hereafter referred to as BN77, made simplifying assumptions about the catchment water balance in a recession period to arrive at the following relationship between the time rate of change in discharge $\mathrm{d}y/\mathrm{d}t$ (mm/d^2) and discharge y (mm/d)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -ay^b, \tag{119}$$

where a (d$^{-1/b}$) and b (-) are unknown recession constants that depend on watershed characteristics. BN77 suggest using a $\log_{\mathfrak{b}} - \log_{\mathfrak{b}}$ graph of $-\mathrm{d}y/\mathrm{d}t$ versus y to estimate a and b. This graphical interpretation of recession hydrographs is not without practical problems and has been subject to active debate in the hydrologic literature (*Rupp and Selker*, 2006; *Kirchner*, 2009; *Thomas et al.*, 2013; *Roques et al.*, 2017; *Tashie et al.*, 2020). But, we do not need to know the values of a and b for a meaningful model evaluation and/or calibration (e.g. *Jepsen et al.* 2016). We can devise a much stronger test of model performance by comparing directly measured Q and simulated P distributions of the $\log_{10}(y)$ and $\log_{10}(-\mathrm{d}y/\mathrm{d}t)$ relationship. The divergence of P and Q can be measured using the bivariate form of $d_{\mathrm{MCRPS}}(P, Q)$ in Equation (116). The mass-conservative numerical solver of the watershed models will help remedy artefacts of BN77 analysis as a result of a constant time step between successive discharge observations.

To illustrate our idea, we revisit the BMA ensemble and compute the divergence of measured and simulated $\log_{10}(y)$ and $\log_{10}(-\mathrm{d}y/\mathrm{d}t)$ point clouds of the $K = 8$ conceptual watershed models (see Figure 18). BN77 scatterplot analysis demonstrates that the recession curves of (a) NAM, (b) GR4J, (c) HYMOD and (d) TOPMO are inconsistent with the measured y and $-\mathrm{d}y/\mathrm{d}t$ point cloud. This can be

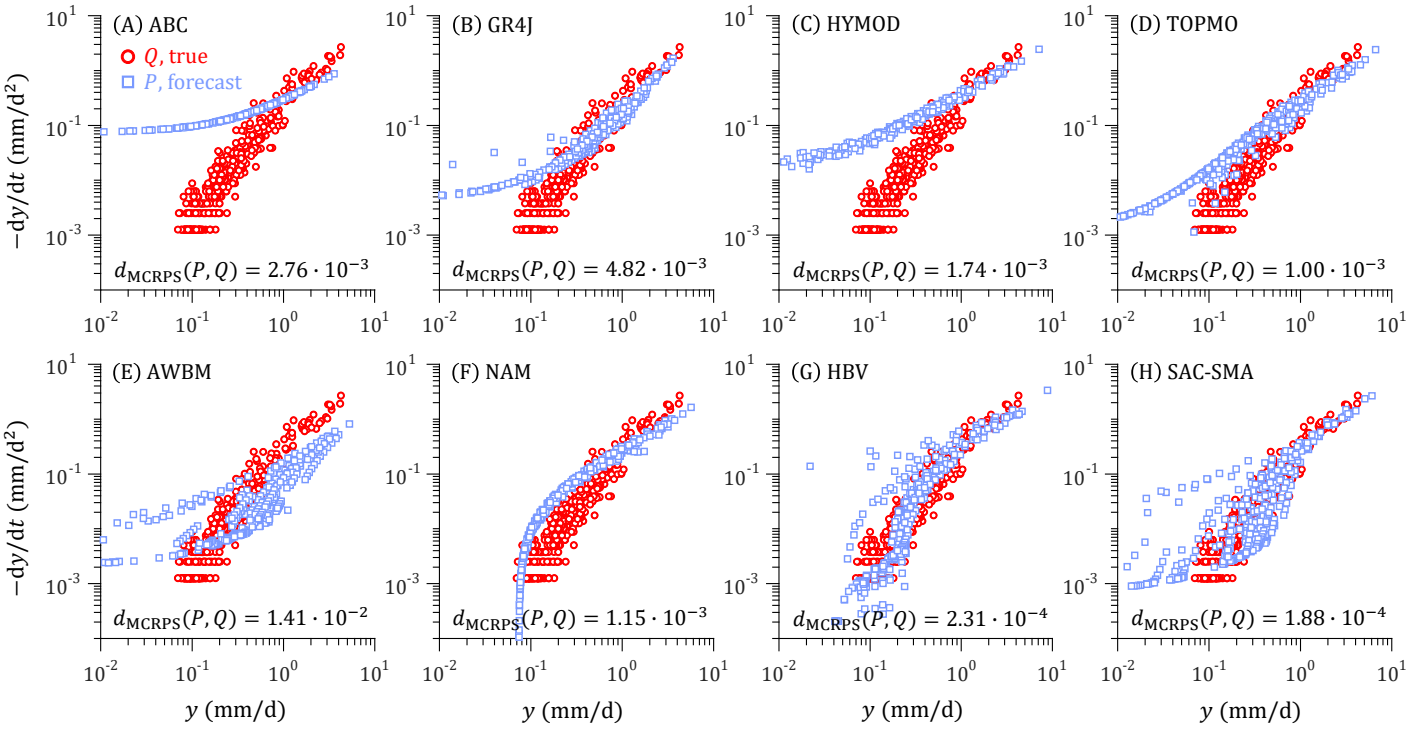

Figure 18: Scatter plots (blue squares) of the $\log_{10}(y) - \log_{10}(\mathrm{d}y/\mathrm{d}t)$ relationship for the different models of the BMA discharge ensemble. The red dots correspond to the measured discharge record.

attributed in part to linear bias-correction, but is reason enough to disqualify these four models from the BMA ensemble. The recession curves of the HBV and SAC-SMA models display the closest match to the measured y and $-\mathrm{d}y/\mathrm{d}t$ point cloud. The MCRPS score divergence, $d_{\mathrm{MCRPS}}(P, Q)$, confirms our inherently subjective visual assessment of the watershed models and attains minimum values for the SAC-SMA model and HBV as runner-up.

We can reformulate signatures using existing scoring rules or develop specific scoring rules such as the CREPS divergence for the flow duration curve and MCRPS and CRPS for recession and flood frequency analysis. In general, the extraction of theory-based signatures from measurements of catchment behavior should go hand in hand with the development of adequate scoring rules.

8.0.5 Standardization of model evaluation metrics

The last decades have witnessed an unbridled growth in the number of performance measures used for hydrologic model evaluation. This proliferation is partly a result of the lack of conforming theory and principles for scoring functions. This state of affairs leads to misguided inferences. Elicitability and propriety offer two useful working paradigms for the development, application and evaluation of

hydrologic scoring functions and scoring rules. This will set much higher standards for metric development and reduces a model's susceptibility to misinformation and unfinished learning. Note that there is ample opportunity to extend hydrologic scoring functions and rules to high-dimensional data from ground-based sensor networks and satellites and the spectral domain.

8.0.6 Practical implementation

Most hydrologist will be familiar with goodness-of-fit metrics such as the coefficient of determination, RMSE, NSE and KGE that quantify the agreement (or distance) between point estimates of measured and simulated quantities. Scoring rules, on the contrary, evaluate the accuracy of a distribution forecast but may appear more difficult to compute. Therefore, we conclude this paper with a few words about their practical use. The case studies of this paper should explicate what is needed for computation of the different scoring rules. These requirements are met with a simple GLUE ensemble of simulated watershed behavior but are also satisfied with a more sophisticated Bayesian description of predictive densities obtained from posterior realizations of the DREAM algorithm (e.g. *Vrugt et al.* 2022). The `ScoringRules` toolbox in MATLAB automatically computes the different scoring rules from simulation distributions specified in the form of a m-member ensemble. This toolbox will be released through Zenodo upon acceptance.

9 Conclusions

This paper was inspired in part by sociocultural developments related to equity, diversity and inclusion and advocates the use of simulation distributions for hydrologic model evaluation and diagnostics. A simulation distribution coalesces the diversity of model responses across the (prior/posterior) parameter and/or input space and contains information about model functioning, behavior, robustness, sensitivity and uncertainty that is not available in single-valued model output. But simulation distributions demand a fundamentally different approach to model evaluation and diagnostics.

In this paper, we discussed past developments that led to the current state-of-the-art of simulation distribution evaluation in hydrology and explored the use of *strictly proper* scoring rules for a more robust and complete method of hydrologic model evaluation and diagnostics. *Strictly proper* scoring

rules condense a distribution forecast to a single reward value for the materialized outcome(s) and have a strong underpinning in statistical, decision and information theory. We reviewed scoring rules for dichotomous and categorical events, quantiles (intervals) and density forecasts, discussed the importance of scoring rule propriety and sufficiency, and addressed diagnostic aspects of *strictly proper* scoring rules by means of their analytic decomposition into an entropy (uncertainty), sharpness and reliability term. The power and usefulness of probabilistic (distribution) model evaluation was illustrated by application to a few illustrative examples, 24-h forecasts of daily rainfall, the survival function of discharge and discharge distributions derived from conceptual watershed models using Bayesian model averaging and the GLUE methodology. These studies and the ensuing discussion on future research work, demonstrated the use of scoring rules for diagnostic model evaluation, Bayesian model selection and extreme event prediction. Specifically, we pointed out the relationship between the logarithmic score and Bayes factors, introduced *strictly proper* divergence scores for the flow duration and recession curves and presented an analytic CRPS expression for flood frequency analysis with the Pearson type III distribution.

Our most important results are as follows.

1. Ubiquitous scoring functions such as the mean squared error, coefficient of determination and KG efficiency which quantify model performance using single-valued model output are insensitive to the underlying distribution of this output. The likelihood, on the contrary, is also computed from point-valued model output but preserves information about the underlying distribution of simulated quantities.

2. Distribution forecasts express diversity in the form of a probability distribution over quantities of interests and contain information about model behavior, robustness, sensitivity and uncertainty that is not available in single-valued model output. As we have witnessed in the BMA case study, this can lead to diametrically opposed conclusions about model adequacy.

3. Albeit diagnostically appealing but hydrologic forecast verification metrics such as the coefficient of variation, reliability, width and coverage measure only a few aspects of the statistical consistency between distribution forecasts and verifying observations. This incomplete evaluation does not guarantee honest forecasts.

4. Bregman divergences are of paramount importance in distribution evaluation as they ensure *strict*

propriety of the scoring rules and their associated divergence functions. *Strict propriety* implies that the score divergence $d(P, Q)$ of the probabilistic forecast P and true but unknown distribution Q is strictly positive and zero only when $P = Q$.

5. *Strictly proper* scoring rules such as the quadratic, logarithmic, spherical, continuous ranked probability and energy scores facilitate a more robust and complete method of hydrologic model evaluation. The larger their values at any time t, the closer is the simulated distribution to the unknown distribution Q of ω.

6. The logarithmic score has negative Shannon entropy as its generalized entropy function and the relative entropy (reverse KL divergence) as its score divergence.

7. The continuous ranked probability score is a generalization of the absolute error (residual) to a distribution forecast.

8. Propriety and elicitability offer two useful working paradigms for the development and application of hydrologic scoring functions and scoring rules. Within the context of model evaluation, information-theoretic principles set much higher and universal standards for metric development thereby promoting metric standardization, reproducibility, stability and comparative analysis across models and data sets. Within the context of model training, information-theoretic principles reduce a model's susceptibility to misinformation and unfinished learning.

9. Sharpness is an *improper* scoring rule as it is a property of the predictive distribution only.

10. *Strictly proper* scoring rules can be used to determine the optimum size of a GLUE ensemble.

11. The use of time-averaged functionals of the hydrograph such as the baseflow index, runoff ratio, rising limb density and flashiness index imply a large loss of information about the temporal behavior of watershed functioning.

12. Frequency distributions of numeral streamflow descriptors derived from a moving-block bootstrap method enable a more robust and complete model evaluation and diagnostics by means of *strictly proper* scoring rules. The signature frequency distributions may be treated separately or as a joint multivariate distribution.

13. The CREPS divergence measures in a single reward value the agreement of measured R_Q and simulated R_P flow duration curves. This function is strictly positive and zero only if $R_P = R_Q$, expressed in physical units of discharge, sensitive to all moments of the FDC and equal to the distance (absolute error) between two point measures of the FDC.

14. The bivariate form of the continuous ranked probability score offers a *strictly proper* scoring rule for BN77 recession analysis of the joint distribution of the time rate of change in discharge and discharge itself. The comparison of observed and simulated point clouds simplifies model diagnostics and selection.

15. Differential weighting provides an opportunity to hone in on specific aspects of the simulation distribution. These so-called local and/or censored scoring functions do not sacrifice local propriety and are specifically suited to evaluating a models' ability in simulating accurately the frequency of extreme events.

16. There is a pressing need for *proper* and *strictly proper* scoring rules of hydrologic signatures. The CREPS divergence of the flow duration curve is an important step in this direction but model diagnostics would benefit from decision-theoretically principled hydrograph functionals.

Last but not least, but currently, there is no delineation or field of science in hydrology and water resources that expects and appreciates mathematical and statistical rigor. By analogy with the fields of biometrics, econometrics and psychometrics, which grew out of the application of theory, mathematics, and statistical inference to biology, economy and psychology, respectively, we believe that this work would fit naturally under *hydrometrics*. Though, this requires a redefinition of hydrometry by the International Organization for Standardization *"... the science of monitoring water in natural water resources"* to include the application of theory, mathematics and statistical inference to study, model and predict the water distribution and movement. This would develop the new title of hydrometrician.

Data and Software Availability

The different case studies presented in this paper are part of the MATLAB toolbox `Scoring_Rules`. This toolbox will be made available for download from Zenodo upon acceptance. Data, models and

algorithms are organized in separate folders. The `MODELAVG` toolbox of (*Vrugt and Beven*, 2018) can be obtained from GitHub at `https://github.com/jaspervrugt/MODELAVG`. Hydrologic data from the CAMELS data set is described in *Newman et al.* (2015) and can be downloaded from `https://dx.doi.org/10.5065/D6MW2F4D`.

References

1. Alexander, C., M. Coulon, Y. Han, and X. Meng (2022), Evaluating the discrimination ability of proper multi-variate scoring rules, *Annals of Operations Research*, doi:10.1007/s10479-022-04611-9.

2. Ammann, L., F. Fenicia, and P. Reichert (2019), A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation, *Hydrology and Earth System Sciences*, *23*(4), 2147–2172, doi:10.5194/hess-23-2147-2019.

3. Baker, D. B., R. P. Richards, T. T. Loftus, and J. W. Kramer (2004), A new flashiness index: Characteristics and applications to midwestern rivers and streams, *JAWRA Journal of the American Water Resources Association*, *40*(2), 503–522, doi:10.1111/j.1752-1688.2004.tb01046.x.

4. Baringhaus, L., and C. Franz (2004), On a new multivariate two-sample test, *Journal of Multivariate Analysis*, *88*, 190–206, doi:10.1016/S0047-259X(03)00079-4.

5. Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, *37*(4), 937–947, doi:10.1029/2000WR900363.

6. Bayes, T. (1763), An essay toward solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., *Philosophical Transactions of the Royal Society of London*, *53*, 370–418, doi:10.1098/rstl.1763.0053.

7. Bella, A., C. Ferri, J. Hernández-Orallo, Ramírez-Quintana, and J. María (2013), On the effect of calibration in classifier combination, *Applied Intelligence*, *38*, 566–585, doi:10.1007/s10489-012-0388-2.

8. Bernardo, J. M. (1979), Expected information as expected utility, *The Annals of Statistics*, *7*, 686–690, doi:10.1214/aos/1176344689.

9. Beven, K. (2006), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

10. Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrological Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

11. Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*(1), 11–29, doi:10.1016/S0022-1694(01)00421-8.

12. Bouhlel, N., and A. Dziri (2019), Kullback–leibler divergence between multivariate generalized gaussian distributions, *IEEE Signal Processing Letters*, *26*(7), 1021–1025, doi:10.1109/LSP.2019.2915000.

13. Boyle, D. P. (2001), *Multicriteria calibration of hydrological models (PhD thesis)*, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ.

14. Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resources Research*, *36*(12), 3663–3674, doi:10.1029/2000WR900207.

15. Bracher, J., E. L. Ray, T. Gneiting, and N. G. Reich (2021), Evaluating epidemic forecasts in an interval format, *PLOS Computational Biology*, *17*(2), 1–15, doi:10.1371/journal.pcbi.1008618.

16. Bregman, L. (1967), The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217, doi:10.1016/0041-5553(67)90040-7.

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

18. Bröcker, J. (2009), Reliability, sufficiency, and the decomposition of proper scores, *Quarterly Journal of the Royal Meteorological Society*, *135*(643), 1512–1519, doi:10.1002/qj.456.

19. Brunetti, C., N. Linde, and J. A. Vrugt (2017), Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the south oyster bacterial transport site, virginia, usa, *Advances in Water Resources*, *102*, 127–141, doi:10.1016/j.advwatres.2017.02.006.

20. Brutsaert, W., and J. L. Nieber (1977), Regionalized drought flow hydrographs from a mature glaciated plateau, *Water Resources Research*, *13*(3), 637–643, doi:10.1029/WR013i003p00637.

21. Buja, A., W. Stuetzle, and Y. Shen (2005), Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications, *Tech. rep.*, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302.

22. Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, *Tech. rep.*, Joint Federal-State River Forecast Center: US Department of Commerce, National Weather Service and CA Department of Water Resources, Sacramento, CA.

23. Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2 ed., 488 pp., Springer, New York, doi:10.1007/b97636.

24. Casella, G., and R. L. Berger (2002), *Statistical Inference*, Duxbury Advanced Series, 2 ed., 660 pp., Duxbury, Pacific Grove, CA.

25. Cervera, J. L., and J. Muñoz (1996), Proper scoring rules for fractiles, in *Bayesian Statistics 5*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 513–519, Oxford University Press.

26. Chang, H., Y. Yao, A. Koschan, B. Abidi, and M. Abidi (2009), Improving face recognition via narrowband spectral range selection using jeffrey divergence, *IEEE Transactions on Information Forensics and Security*, *4*(1), 111–122, doi:10.1109/TIFS.2008.2012211.

27. Chen, L., Z. Shen, X. Yang, Q. Liao, and S. L. Yu (2014), An interval-deviation approach for hydrology and water quality model evaluation within an uncertainty framework, *Journal of Hydrology*, *509*, 207–214, doi:10.1016/j.jhydrol.2013.11.043.

28. Christensen, H. M. (2015), Decomposition of a new proper score for verification of ensemble forecasts, *Monthly Weather Review*, *143*(5), 1517–1532, doi:10.1175/MWR-D-14-00150.1.

29. Christoffersen, P. F. (1998), Evaluating interval forecasts, *International Economic Review*, *39*(4), 841–862.

30. Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (FUSE): A modular framework

to diagnose differences between hydrological models, *Water Resources Research*, *44*(12), doi: 10.1029/2007WR006735.

31. Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory*, Telecommunications and Signal Processing, 2 ed., 784 pp., John Wiley & Sons, Inc., Hoboken, NJ.

32. Cox, D. R., and P. A. W. Lewis (1966), *Statistical Analysis of Series of Events*, Methuen's Monographs on Applied Probability and Statistics (MMAPS), 285 pp., Methuen, London.

33. Csiszar, I. (1975), I-Divergence Geometry of Probability Distributions and Minimization Problems, *The Annals of Probability*, *3*(1), 146–158, doi:10.1214/aop/1176996454.

34. Dawid, A. P. (1984), Present position and potential developments: Some personal views: Statistical theory: The prequential approach, *Journal of the Royal Statistical Society. Series A (General)*, *147*(2), 278–292.

35. Dawid, A. P. (1998), Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design, *Tech. rep.*, Department of Statistical Science, University College London, London, UK.

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

37. Dawid, P., and P. Sebastiani (1999), Coherent dispersion criteria for optimal experimental design, *The Annals of Statistics*, *27*(1), 65–81, doi:10.1214/aos/1018031101.

38. de Finetti, B. (2017), *Theory of Probability: A Critical Introductory Treatment*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd, United Kingdom.

39. de Punder, R., C. Diks, R. Laeven, and D. van Dijk (2022), A general procedure for localising strictly proper scoring rules, *Tech. rep.*, Erasmus University Rotterdam, Tinbergen Institute, Rotterdam, The Netherlands.

40. DeGroot, M. H., and S. E. Fienberg (1983), The comparison and evaluation of forecasters, *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(1/2), 12–22.

41. Diebold, F. X., T. A. Gunther, and A. S. Tay (1998), Evaluating density forecasts with applications to financial risk management, *International Economic Review*, *39*(4), 863–883.

42. Diks, C., V. Panchenko, and D. van Dijk (2011), Likelihood-based scoring rules for comparing density forecasts in tails, *Journal of Econometrics*, *163*(2), 215–230, doi:10.1016/j.jeconom.2011.04.001.

43. Dimitriadis, T., T. Gneiting, and A. I. Jordan (2021), Stable reliability diagrams for probabilistic classifiers, *Proceedings of the National Academy of Sciences*, *118*(8), e2016191,118, doi:10.1073/pnas.2016191118.

44. Duffie, D., and J. Pan (1997), An overview of value at risk, *The Journal of Derivatives*, *4*(3), 7–49, doi:10.3905/jod.1997.407971.

45. Dunsmore, I. R. (1968), A Bayesian approach to calibration, *Journal of the Royal Statistical Society. Series B (Methodological)*, *30*(2), 396–405.

46. Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956), Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator, *The Annals of Mathematical Statistics*, *27*(3), 642–669, doi:10.1214/aoms/1177728174.

47. Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water Resources Research*, *49*(7), 4518–4524, doi:10.1002/wrcr.20284.

48. Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resources Research*, *50*(3), 2350–2375, doi:10.1002/2013WR014185.

49. Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. motivation and theoretical development, *Water Resources Research*, *47*(11), doi:10.1029/2010WR010174.

50. Ferson, S., W. L. Oberkampf, and L. Ginzburg (2008), Model validation and predictive capability for the thermal challenge problem, *Computer Methods in Applied Mechanics and Engineering*, *197*(29), 2408–2430, doi:10.1016/j.cma.2007.07.030, validation Challenge Workshop.

51. Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resources Research*, *32*(7), 2161–2173, doi:10.1029/95WR03723.

52. Friederichs, P., and A. Hense (2007), Statistical downscaling of extreme precipitation events using censored quantile regression, *Monthly Weather Review*, *135*(6), 2365–2378, doi:10.1175/MWR3403.1.

53. Friederichs, P., and T. L. Thorarinsdottir (2012), Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, *Environmetrics*, *23*(7), 579–594, doi:10.1002/env.2176.

54. Friedman, D. (1983), Effective scoring rules for probabilistic forecasts, *Management Science*, *29*(4), 447–454, doi:10.1287/mnsc.29.4.447.

55. Garratt, A., K. Lee, M. H. Pesaran, and Y. Shin (2003), Forecast uncertainties in macroeconomic modeling: An application to the u.k. economy, *Journal of the American Statistical Association*, *98*(464), 829–838.

56. Garthwaite, P. H., J. B. Kadane, and A. O'Hagan (2005), Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, *100*(470), 680–701, doi:10.1198/016214505000000105.

57. Gerstner, T., and M. Griebel (1998), Numerical integration using sparse grids, *Numerical Algorithms*, *18*, 209–232, doi:10.1023/A:1019129717644.

58. Gerstner, T., and M. Griebel (2003), Dimension-adaptive tensor-product quadrature, *Computing*, *71*(1), 65–87, doi:10.1007/s00607-003-0015-5.

59. Gini, C. (1909), Concentration and dependency ratios (in italian), *Rivista di Politica Economica*, *87*(8-9), 769–790.

60. Girons Lopez, M., L. Crochemore, and I. G. Pechlivanidis (2021), Benchmarking an operational hydrological model for providing seasonal forecasts in sweden, *Hydrology and Earth System Sciences*, *25*(3), 1189–1209, doi:10.5194/hess-25-1189-2021.

61. Gneiting, T. (2011), Making and evaluating point forecasts, *Journal of the American Statistical Association*, *106*(494), 746–762, doi:10.1198/jasa.2011.r10138.

62. Gneiting, T., and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, *102*(477), 359–378, doi:10.1198/016214506000001437.

63. Gneiting, T., and R. Ranjan (2011), Comparing density forecasts using threshold-and quantile-weighted scoring rules, *Journal of Business & Economic Statistics*, *29*(3), 411–422.

64. Gneiting, T., F. Balabdaoui, and A. E. Raftery (2005), Probabilistic forecasts, calibration, and sharpness, *Tech. rep.*, Department of Statistics, University of Washington, Seattle, WA, USA.

65. Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

66. Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero III (2013), Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, *Water Resources Research*, *49*(4), 2253–2273, doi:10.1002/wrcr.20161.

67. Good, I. J. (1952), Rational decisions, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *14*(1), 107–114.

68. Granger, C. W. J. (2005), Preface: Some thoughts on the future of forecasting, *Oxford Bulletin of Economics and Statistics*, *67*(s1), 707–711, doi:10.1111/j.1468-0084.2005.00138.x.

69. Grayson, R., and G. Blöschl (2001), *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, p. 416, Cambridge University Press.

70. Grimit, E. P., T. Gneiting, B. V. J., and N. A. Johnson (2006), The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification, *Quarterly Journal of the Royal Meteorological Society*, *132*(621C), 2925–2942, doi:10.1256/qj.05.235.

71. Grünwald, P. D., and A. P. Dawid (2004), Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, *The Annals of Statistics*, *32*(4), 1367 – 1433, doi:10.1214/009053604000000553.

72. Grushka-Cockayne, Y., K. C. Lichtendahl, V. R. R. Jose, and R. L. Winkler (2017), Quantile evaluation, sensitivity to bracketing, and sharing business payoffs, *Operations Research*, *65*(3), 712–728, doi:10.1287/opre.2017.1588.

73. Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resources Research*, *34*(4), 751–763, doi:10.1029/97WR03495.

74. Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, *Hydrological Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.

75. Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, *377*(1), 80–91, doi:10.1016/j.jhydrol.2009.08.003.

76. Guttorp, P. (2011), The role of statisticians in international science policy, *Environmetrics*, *22*(7), 817–825, doi:10.1002/env.1109.

77. Halton, J. H. (1960), On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals, *Numerische Mathematik*, *2*, 84–90, doi:10.1007/BF01386213.

78. Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Monthly Weather Review*, *129*(3), 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

79. Hammersley, J. M., and D. C. Handscomb (1960), *Monte Carlo Methods*, Monographs on Statistics and Applied Probability (MSAP), 1 ed., 178 pp., Springer, Dordrecht, doi:10.1007/978-94-009-5819-7.

80. Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather and Forecasting*, *15*(5), 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

81. Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-moments*, p. 224, Cambridge University Press, doi:10.1017/cbo9780511529443.

82. Hughes, G., and C. F. Topp (2015), Probabilistic forecasts: Scoring rules and their decomposition and diagrammatic representation via Bregman divergences, *Entropy*, *17*(8), 5450–5471, doi:10.3390/e17085450.

83. Hunt, V., S. Dixon-Fyle, S. Prince, and K. Dolan (2020), Diversity wins: How inclusion matters, *Tech. rep.*, McKinsey & Company.

84. Jakeman, J. D., and A. Narayan (2018), Generation and application of multivariate polynomial quadrature rules, *Computer Methods in Applied Mechanics and Engineering*, *338*, 134–161, doi:10.1016/j.cma.2018.04.009.

85. Jaynes, E. T. (1963), *Information Theory and Statistical Mechanics*, pp. 181–218, W. A. Benjamin, Inc.

86. Jeffreys, H. (1939), *The theory of probability*, p. 470, Oxford Classic Texts in the Physical Sciences, 3 ed., Oxford University Press.

87. Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*(1007), 453–461.

88. Jepsen, S. M., T. C. Harmon, and Y. Shi (2016), Watershed model calibration to the base flow recession curve with and without evapotranspiration effects, *Water Resources Research*, *52*(4), 2919–2933, doi:10.1002/2015WR017827.

89. Jolliffe, I. T., and D. B. Stephenson (2011), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, p. 296, 2 ed., Wiley Blackwell.

90. Jordan, A. (2016), Facets of forecast evaluation, Ph.D. thesis, Karlsruhe Institute of Technology, doi:10.5445/IR/1000063629.

91. Kass, R. E., and A. E. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, *90*(430), 773–795, doi:10.1080/01621459.1995.10476572.

92. Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*(3), doi:10.1029/2005WR004368.

93. Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resources Research*, *42*(3), doi:10.1029/2005WR004376.

94. Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resources Research*, *45*(2), doi:10.1029/2008WR006912.

95. Knoben, W. J. M., J. E. Freer, K. J. A. Fowler, M. C. Peel, and R. A. Woods (2019), Modular assessment of rainfall–runoff models toolbox (marrmot) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, *Geoscientific Model Development*, *12*(6), 2463–2480, doi:10.5194/gmd-12-2463-2019.

96. Knorr-Held, L., and E. Rainer (2001), Projections of lung cancer in West Germany: A case study in Bayesian prediction, *Biostatistics*, *2*(1), 109–129, doi:10.1093/biostatistics/2.1.109.

97. Kolmogorov, A. N. (1933), Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83–91.

98. Kuczera, G. (1983), Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, *Water Resources Research*, *19*(5), 1151–1162, doi:10.1029/WR019i005p01151.

99. Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm, *Journal of Hydrology*, *211*(1), 69–85, doi:10.1016/S0022-1694(98)00198-X.

100. Kull, M., and P. Flach (2015), Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration, in *Machine Learning and Knowledge Discovery in Databases*, edited by A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, and A. Jorge, pp. 68–85, Springer International Publishing, Cham.

101. Kullback, S. (1959), *Information Theory and Statistics*, John Wiley & Sons, New York.

102. Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Annals of Mathematical Statistics*, *22*(1), 79–86, doi:10.1214/aoms/1177729694.

103. Kunsch, H. R. (1989), The Jackknife and the Bootstrap for General Stationary Observations, *The Annals of Statistics*, *17*(3), 1217–1241, doi:10.1214/aos/1176347265.

104. Lai, T. L., S. T. Gross, and D. B. Shen (2011), Evaluating probability forecasts, *The Annals of Statistics*, *39*(5), 2356–2382.

105. Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrology and Earth System Sciences*, *11*(4), 1267–1277, doi:10.5194/hess-11-1267-2007.

106. Lamontagne, J. R., C. A. Barber, and R. M. Vogel (2020), Improved estimators of model performance efficiency for skewed hydrologic data, *Water Resources Research*, *56*(9), 1–25, doi:10.1029/2020wr027101.

107. Langrené, N., and X. Warin (2021), Fast multivariate empirical cumulative distribution function with connection to kernel density estimation, *Computational Statistics & Data Analysis*, *162*, 107,267, doi:10.1016/j.csda.2021.107267.

108. Liu, Y., W. Chen, P. Arendt, and H.-Z. Huang (2011), Toward a better understanding of model validation metrics, *Journal of Mechanical Design*, *133*, 071,005.

109. Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of bayesian model selection criteria and fisher information matrix on sample size, *Mathematical Geosciences*, *43*, 971–993, doi:10.1007/s11004-011-9359-0.

110. Luke, A., J. A. Vrugt, A. AghaKouchak, R. Matthew, and B. F. Sanders (2017), Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the united states, *Water Resources Research*, *53*(7), 5469–5494, doi:10.1002/2016WR019676.

111. Matheron, G. (1984), *The Selectivity of the Distributions and the "Second Principle of Geostatistics"*, p. 421–433, Springer, doi:10.1007/978-94-009-3699-7_24.

112. Matheson, J. E., and R. L. Winkler (1976), Scoring rules for continuous probability distributions, *Management Science*, *22*(10), 1087–1096, doi:10.1287/mnsc.22.10.1087.

113. McCarthy, J. (1956), Measures of the value of information, *Proceedings of the National Academy of Sciences*, *42*(9), 654–655, doi:10.1073/pnas.42.9.654.

114. McDonald, J. B., and B. C. Jensen (1979), An analysis of some properties of alternative measures of income inequality based on the gamma distribution function, *Journal of the American Statistical Association*, *74*(368), 856–860.

115. McInerney, D., M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera (2017), Improving probabilistic prediction of daily streamflow by identifying pareto optimal approaches for modeling heteroscedastic residual errors, *Water Resources Research*, *53*(3), 2199–2239, doi:10.1002/2016WR019168.

116. McInerney, D., D. Kavetski, M. Thyer, J. Lerat, and G. Kuczera (2019), Benefits of explicit treatment of zero flows in probabilistic hydrological modeling of ephemeral catchments, *Water Resources Research*, *55*(12), 11,035–11,060, doi:10.1029/2018WR024148.

117. Meng, X., J. W. Taylor, S. Ben Taieb, and S. Li (2022), Scores for multivariate distributions and level sets, *Tech. rep.*, University of Sussex, Falmer, Brighton, UK, BN1 9SN, doi:10.48550/arXiv.2002.09578.

118. Murphy, A. H. (1970), The ranked probability score and the probability score: A comparison, *Monthly Weather Review*, *98*(12), 917–924, doi:10.1175/1520-0493(1970)098<0917:TRPSAT>2.3.CO;2.

119. Murphy, A. H. (1973), A new vector partition of the probability score, *Journal of Applied Meteorology and Climatology*, *12*(4), 595 – 600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

120. Murphy, A. H. (1993), What is a good forecast? an essay on the nature of goodness in weather forecasting, *Weather and Forecasting*, *8*(2), 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

121. Murphy, A. H. (1996), General decompositions of mse-based skill scores: Measures of some basic aspects of forecast quality, *Monthly Weather Review*, *124*(10), 2353–2369, doi:10.1175/1520-0493(1996)124<2353:GDOMBS>2.0.CO;2.

122. Murphy, A. H., and R. W. Katz (1985), *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, p. 560, 1 ed., Westview Press.

123. Murphy, A. H., and R. L. Winkler (1987), A general framework for forecast verification, *Monthly Weather Review*, *115*(7), 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

124. Naaman, M. (2021), On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality, *Statistics & Probability Letters*, *173*, 109,088, doi:10.1016/j.spl.2021.109088.

125. Nash, J., and J. Sutcliffe (1970), River flow forecasting through conceptual models part I - A discussion of principles, *Journal of Hydrology*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

126. Neuman, S. P. (2003), Maximum likelihood bayesian averaging of uncertain model predictions, *Stochastic Environmental Research and Risk Assessment*, *17*, 291–305, doi:10.1007/s00477-003-0151-7.

127. Newman, A. J., et al. (2015), Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrology and Earth System Sciences*, *19*(1), 209–223, doi:10.5194/hess-19-209-2015.

128. Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conference Series in Applied Mathematics, 237 pp., Society for Industrial and Applied Mathematics, doi:10.1137/1.9781611970081.

129. Nott, D. J., L. Marshall, and J. Brown (2012), Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?, *Water Resources Research*, *48*(12), doi:10.1029/2011WR011128.

130. Pachepsky, Y. A., G. Martinez, F. Pan, T. Wagener, and T. Nicholson (2016), Evaluating hydrological model performance using information theory-based metrics, *Hydrology and Earth System Sciences Discussions*, *2016*, 1–24, doi:10.5194/hess-2016-46.

131. Palmer, T. N. (2002), The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, *Quarterly Journal of the Royal Meteorological Society*, *128*(581), 747–774, doi:10.1256/0035900021643593.

132. Pool, S., M. Vis, and J. Seibert (2018), Evaluating model performance: towards a non-parametric variant of the kling-gupta efficiency, *Hydrological Sciences Journal*, *63*(13-14), 1941–1953, doi: 10.1080/02626667.2018.1552002.

133. Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, *133*(5), 1155–1174, doi: 10.1175/MWR2906.1.

134. Ramsey, F. P. (1926), Truth and probability, in *The Foundations of Mathematics and other Logical Essays*, edited by R. B. Braithwaite, chap. 7, pp. 156–198, McMaster University Archive for the History of Economic Thought.

135. Rathinasamy, M., R. Khosa, J. Adamowski, S. ch, G. Partheepan, J. Anand, and B. Narsimlu (2014), Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models, *Water Resources Research*, *50*(12), 9721–9737, doi:10.1002/2013WR014650.

136. Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, *46*(5), doi:10.1029/2009WR008328.

137. Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resources Research*, *47*(11), doi:10.1029/2011WR010643.

138. Reusser, D. E., T. Blume, B. Schaefli, and E. Zehe (2009), Analysing the temporal dynamics of model performance for hydrological models, *Hydrology and Earth System Sciences*, *13*(7), 999–1018, doi:10.5194/hess-13-999-2009.

139. Rockafellar, R. T. (1970), *Convex Analysis*, Princeton Mathematical Series, 472 pp., Princeton University Press, Princeton, NJ.

140. Roques, C., D. E. Rupp, and J. S. Selker (2017), Improved streamflow recession parameter estimation with attention to calculation of - dQ/dt, *Advances in Water Resources*, *108*, 29–43, doi:10.1016/j.advwatres.2017.07.013.

141. Rosenblatt, M. (1952), Remarks on a multivariate transformation, *The Annals of Mathematical Statistics*, *23*(3), 470–472.

142. Roulston, M. S., and L. A. Smith (2002), Evaluating probabilistic forecasts using information theory, *Monthly Weather Review*, *130*(6), 1653 – 1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

143. Rupp, D. E., and J. S. Selker (2006), Information, artifacts, and noise in dq/dt − q recession analysis, *Advances in Water Resources*, *29*(2), 154–160, doi:10.1016/j.advwatres.2005.03.019, experimental Hydrology: A Bright Future.

144. Sadegh, M., and J. A. Vrugt (2013), Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation, *Hydrology and Earth System Sciences*, *17*(12), 4831–4850, doi:10.5194/hess-17-4831-2013.

145. Sadegh, M., J. Vrugt, H. Gupta, and C. Xu (2016), The soil water characteristic as new class of closed-form parametric expressions for the flow duration curve, *Journal of Hydrology*, *535*, 438–456, doi:10.1016/j.jhydrol.2016.01.027.

146. Savage, L. J. (1971), Elicitation of personal probabilities and expectations, *Journal of the American Statistical Association*, *66*(336), 783–801, doi:doi.org/10.2307/2284229.

147. Scharnagl, B., J. A. Vrugt, H. Vereecken, and M. Herbst (2010), Information content of incubation experiments for inverse estimation of pools in the rothamsted carbon model: a bayesian perspective, *Biogeosciences*, *7*(2), 763–776, doi:10.5194/bg-7-763-2010.

148. Scharnagl, B., S. C. Iden, W. Durner, H. Vereeken, and M. Herbst (2015), Inverse modelling of in situ soil water dynamics: accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals, *Hydrology and Earth System Sciences Discussions*, *12*, 2155–2199.

149. Scheuerer, M., and T. M. Hamill (2015), Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, *Monthly Weather Review*, *143*(4), 1321–1334, doi:10.1175/mwr-d-14-00269.1.

150. Scheuerer, M., and D. Möller (2015), Probabilistic wind speed forecasting on a grid based on ensemble model output statistics, *The Annals of Applied Statistics*, *9*(3), 1328–1349.

151. Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate bayesian model evidence, *Water Resources Research*, *50*(12), 9484–9513, doi:10.1002/2014WR016062.

152. Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resources Research*, *46*(10), doi:10.1029/2009WR008933.

153. Schoups, G., J. A. Vrugt, F. Fenicia, and N. C. van de Giesen (2010), Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, *Water Resources Research*, *46*(10), doi:10.1029/2009WR008648.

154. Schwemmle, R., D. Demand, and M. Weiler (2021), Technical note: Diagnostic efficiency – specific evaluation of model performance, *Hydrology and Earth System Sciences*, *25*, 2187–2198, doi:10.5194/hess-25-2187-2021.

155. Shamir, E., B. Imam, E. Morin, H. V. Gupta, and S. Sorooshian (2005), The role of hydrograph indices in parameter estimation of rainfall–runoff models, *Hydrological Processes*, *19*(11), 2187–2207, doi:10.1002/hyp.5676.

156. Shannon, C. E. (1948a), A mathematical theory of communication, *Bell System Technical Journal*, *27*(3), 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.

157. Shannon, C. E. (1948b), A mathematical theory of communication, *Bell System Technical Journal*, *27*(4), 623–656, doi:10.1002/j.1538-7305.1948.tb00917.x.

158. Shuford, E. H., A. Albert, and H. E. Massengill (1966), Admissible probability measurement procedures, *Psychometrika*, *31*, 125–145.

159. Smirnov, N. (1948), Table for estimating the goodness of fit of empirical distributions, *The Annals of Mathematical Statistics*, *19*(2), 279–281, doi:10.1214/aoms/1177730256.

160. Smolyak, S. A. (1963), Quadrature and interpolation formulas for tensor products of certain classes of functions, *Doklady Akademii Nauk SSSR*, *148*, 1042–1045.

161. Sobol', I. M., and B. V. Shukhman (1995), Integration with quasirandom sequences: Numerical experience, *International Journal of Modern Physics C*, *6*(2), 263–275, doi:10.1142/S0129183195000204.

162. Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.

163. Spear, R. C., and G. Hornberger (1980), Eutrophication in peel inlet-ii. identification of critical uncertainties via generalized sensitivity analysis, *Water Research*, *14*(1), 43–49.

164. Spear, R. C., Q. Cheng, and S. L. Wu (2020), An example of augmenting regional sensitivity analysis using machine learning software, *Water Resources Research*, *56*(4), 1–16, doi:10.1029/2019wr026379.

165. Staël von Holstein, C.-A. S. (1970), A family of strictly proper scoring rules which are sensitive to distance, *Journal of Applied Meteorology*, *9*, 360–364.

166. Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis: 1. ordinary, weighted, and generalized least squares compared, *Water Resources Research*, *21*(9), 1421–1432, doi:10.1029/wr021i009p01421.

167. Storch, H. v., and F. W. Zwiers (1999), *Statistical Analysis in Climate Research*, p. 484, 1 ed., Cambridge University Press, doi:10.1017/CBO9780511612336.

168. Székely, G. J. (2003), $\mathcal{E}$-statistics: The energy of statistical samples, *Tech. rep.*, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA.

169. Tashie, A., T. Pavelsky, and L. E. Band (2020), An empirical reevaluation of streamflow recession analysis at the continental scale, *Water Resources Research*, *56*(1), e2019WR025,448, doi:10.1029/2019WR025448.

170. Tasker, G. D. (1980), Hydrologic regression with weighted least squares, *Water Resources Research*, *16*(6), 1107–1113, doi:10.1029/wr016i006p01107.

171. Tegos, S. A., G. K. Karagiannidis, P. D. Diamantoulakis, and N. D. Chatzidiamantis (2022), New results for pearson type iii family of distributions and application in wireless power transfer, *IEEE Internet of Things Journal*, *9*(23), 24,038–24,050, doi:10.1109/JIOT.2022.3189220.

172. Thielen, J., J. Schaake, R. Hartman, and R. Buizza (2008), Aims, challenges and progress of the hydrological ensemble prediction experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 june 2007, *Atmospheric Science Letters*, *9*, 29–35, doi:10.1002/asl.168.

173. Thomas, B. F., R. M. Vogel, C. N. Kroll, and J. S. Famiglietti (2013), Estimation of the base flow recession constant under human interference, *Water Resources Research*, *49*(11), 7366–7379, doi:10.1002/wrcr.20532.

174. Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl (2013), Using proper divergence functions to evaluate climate models, *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534, doi:10.1137/130907550.

175. Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resources Research*, *45*(12), doi:10.1029/2008WR006825.

176. Tsyplakov, A. (2011), Evaluating density forecasts: A comment, *SSRN*, doi:10.2139/ssrn.1907799.

177. Unger, D. A. (1985), A method to estimate the continuous ranked probability score, in *Preprints of the Ninth Conference on Probability and Statistics in Atmospheric Sciences*, pp. 206–213, American Meteorological Society, Virginia Beach, Virginia, USA.

178. Villez, K. (2017), Analytical expressions to compute the Continuous Ranked Probability Score (CRPS), *Tech. Rep. 4*, Eawag, Aquatic Research, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland.

179. Vogel, R. M., and N. M. Fennessey (1994), Flow-duration curves. i: New interpretation and confidence intervals, *Journal of Water Resources Planning and Management*, *120*(4), 485–504, doi:10.1061/(ASCE)0733-9496(1994)120:4(485).

180. Volpi, E., G. Schoups, G. Firmani, and J. A. Vrugt (2017), Sworn testimony of the model evidence: Gaussian mixture importance (GAME) sampling, *Water Resources Research*, *53*(7), 6133–6158, doi:10.1002/2016WR020167.

181. Vrugt, J. A. (2016), Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and matlab implementation, *Environmental Modelling & Software*, *75*, 273–316, doi:10.1016/j.envsoft.2015.08.013.

182. Vrugt, J. A. (2018), MODELAVG: A MATLAB toolbox for postprocessing of model ensembles, *Tech. rep.*, University of California, Irvine.

183. Vrugt, J. A., and K. J. Beven (2018), Embracing equifinality with efficiency: Limits of acceptability sampling using the $\text{DREAM}_{(\text{LOA})}$ algorithm, *Journal of Hydrology*, *559*, 954–971, doi:10.1016/j.jhydrol.2018.02.026.

184. Vrugt, J. A., and D. Y. de Oliveira (2022), Confidence intervals of the kling-gupta efficiency, *Journal of Hydrology*, *612*, 127,968, doi:10.1016/j.jhydrol.2022.127968.

185. Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resources Research*, *43*(1), doi:10.1029/2005WR004838.

186. Vrugt, J. A., and M. Sadegh (2013a), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resources Research*, *49*(7), 4335–4345, doi:10.1002/wrcr.20354.

187. Vrugt, J. A., W. Bouten, H. V. Gupta, and S. Sorooshian (2002), Toward improved identifiability of hydrologic model parameters: The information content of experimental data, *Water Resources Research*, *38*(12), 48–1–48–13, doi:10.1029/2001WR001118.

188. Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resources Research*, *39*(8), doi:10.1029/2002WR001746.

189. Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson (2006), Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophysical Research Letters*, *33*(19), doi:10.1029/2006GL027126.

190. Vrugt, J. A., C. G. H. Diks, and M. P. Clark (2008), Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling, *Environmental Fluid Mechanics*, *8*(5), 579–595, doi:10.1007/s10652-008-9106-3.

191. Vrugt, J. A., D. Yumi de Oliveria, G. Schoups, and C. G. H. Diks (2022), On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking, *Journal of Hydrology*, *615*, 128,542, doi:10.1016/j.jhydrol.2022.128542.

192. Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, *Hydrological Processes*, *17*(2), 455–476, doi:10.1002/hyp.1135.

193. Weijs, S. V., G. Schoups, and N. van de Giesen (2010a), Why hydrological predictions should be evaluated using information theory, *Hydrology and Earth System Sciences*, *14*(12), 2545–2558, doi:10.5194/hess-14-2545-2010.

194. Weijs, S. V., R. van Nooijen, and N. van de Giesen (2010b), Kullback-leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition, *Monthly Weather Review*, *138*(9), 3387–3399, doi:10.1175/2010MWR3229.1.

195. Welles, E., S. Sorooshian, G. Carter, and B. Olsen (2007), Hydrologic verification: A call for action and collaboration, *Bulletin of the American Meteorological Society*, *88*(4), 503–512, doi:10.1175/BAMS-88-4-503.

196. Westerberg, I. K., J.-L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C.-Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrology and Earth System Sciences*, *15*(7), 2205–2227, doi:10.5194/hess-15-2205-2011.

197. Winkler, R. L., et al. (1996), Scoring rules and the evaluation of probabilities, *Test*, *5*(1), 1–60, doi:10.1007/BF02562681.

198. Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resources Research*, *44*(3), doi:10.1029/2008WR006803.

199. Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resources Research*, *44*(9), doi:10.1029/2007WR006716.

200. Zheng, J., and H. You (2013), A new model-independent method for change detection in multi-temporal SAR images based on radon transform and Jeffrey divergence, *IEEE Geoscience and Remote Sensing Letters*, *10*(1), 91–95, doi:10.1109/LGRS.2012.2193659.

Figure Captions

Figure 1: Probability density function of the uniform true distribution $Q = \mathcal{U}(a, b)$ and symmetric triangular forecast distribution $P = \mathcal{T}(a, b)$ of quantity $\Omega = [a, b]$.

Figure 2: Illustration of the computation of the relative entropy $d_{\mathrm{KL}}(Q, P)$ on a sample space $\Omega = \{0, 1, 2, 3, 4\}$: (a) *true distribution*, $Q = \mathcal{B}(n, p)$, and (b) *forecast distribution*, $P = \mathcal{U}_{\mathrm{d}}(n)$. According to data, Ω follows a binomial distribution with $n = 4$, $p = \frac{1}{2}$ and PMF, $f_{\mathcal{B}}(\omega, n, p) = c(n, \omega)p^{\omega}(1-p)^{n-\omega}$, where $c(a, b) = a!/b!(a-b)!$ denotes the binomial coefficient and ! is the factorial function. Theory predicts a discrete uniform distribution for ω with equal density, $f_{\mathcal{U}_{\mathrm{d}}}(\omega, n) = 1/n$, for all $n = 5$ values.

Figure 3: (a) Illustration of the probability integral transform and (b) interpretation of the so-called quantile-quantile plot (adapted from *Laio and Tamea* (2007)).

Figure 4: Hypothetical *true distribution* $Q = \mathcal{G}(3.36, 0.64)$ (gray) and different *probabilistic forecasts* P (in color) with an equal (a) coefficient of variation, $C_{\mathrm{v}} = 0.546$, and (b) width, $W = 4.5$, at $\alpha = 0.05$ using using lower and upper quantiles of 0.5 and 5.0, respectively. The peak of the *true distribution* is equal to the verifying observation, ω. The distribution forecasts of the left graph are used in later studies: $P_1 = \mathcal{U}(0.11, 3.89)$, $P_2 = \mathcal{N}(2.26, 1.52)$, $P_3 = \mathcal{GEV}(0.04, 1, 1.86)$; $P_4 = \mathcal{LN}(0.80, 0.26)$ and $P_5 = \mathcal{GP}(-0.28, 2, 0.73)$.

Figure 5: Generalized entropy function $H(p) = p(p-1)$ (blue curve) of the quadratic score for a dichotomous event $\Omega = \{1, 0\}$ with *probability forecast* $(p, 1-p)$ and *true probability* $(q, 1-q)$ with $p, q \in [0, 1]$. We present the values of the quadratic scoring rule $\mathcal{S}(p, q)$ at p and q (solid black dots) and display the so-called Bregman divergence, $d(p, q)$. For any probability forecast, $p \in [0, 1]$, the expected score, $\mathcal{S}(p, q) = qS(p, 1) + (1-q)S(p, 0)$, equals the ordinate of the tangent to H at p (solid gray line) when evaluated at $q \in [0, 1]$. In particular, the scores, $S(p, 0) = H(p) - pH'(p)$ and $S(p, 1) = H(p) + (1-p)H'(p)$, equal the tangent at $q = 0$ and $q = 1$, respectively. The divergence, $d_{\mathrm{QS}}(p, q) = \mathcal{S}(q, q) - \mathcal{S}(p, q)$, is equal to the difference between $H(q)$ and the tangent at p when evaluated at q (Adapted after Fig. 1 of *Gneiting and Raftery* (2007) and Fig. 8 of *Buja et al.* (2005)).

Figure 6: Binary event, $\Omega = \{1, 0\}$: Expected value of the quadratic (green), logarithmic (red) and spherical (blue) scoring rules as function of the *true probability* $q = q_1$ of the first event, (a) $q = 0$, (b) $q = 0.1$, (c) $q = 0.2$, (d) $q = 0.3$, (e) $q = 0.4$, (f) $q = 0.5$, (g) $q = 0.6$, (h) $q = 0.7$, (i) $q = 0.8$, and (j)

$q = 0.9$.

Figure 7: Illustration of the standard normal Lebesgue density on sample space $\Omega = [-3, 3]$ with range of $\mathcal{N}(0, 1)$ partitioned into 14 intervals. The Lebesgue measure $\mu(\omega_k)$ is equal to the length of the interval (= bin width) containing ω_k. The Lebesgue density $f(\omega)$ is constant in each bin $\omega_1, \ldots, \omega_m$ of ω values, where $m = 29$. The sum of the areas of the rectangles is equal to the Lebesgue integral.

Figure 8: Score divergence $d_{\text{LinS}}(P, Q)$ of the linear scoring rule $S_{\text{LinS}}(P, \omega) = f_P(\omega)$ for a uniform probabilistic forecast P under a standard Gaussian true distribution Q and symmetric interval $[-\varepsilon, \varepsilon]$ with $\varepsilon \in (0, 3]$.

Figure 9: Graphical explanation of the continuous ranked probability scoring rule for a hypothetical streamflow forecast CDF (F_P, black line) and verifying observation (ω, red dot).

Figure 10: Hypothetical distribution forecast P of the discharge Ω (mm/d) and traces of the quadratic (green), logarithmic (red), spherical (blue) and continuous ranked probability (yellow) scoring rules across the streamflow distribution computed using their numerical definitions in Table 6. The probabilistic forecast is a gamma distribution $P = \mathcal{G}(a, b)$ with shape and scale parameters $a = 3$ (-) and $b = 1$ mm/d, respectively, and PDF $f_{\mathcal{G}}(\omega, a, b) = \Gamma^{-1}(a) b^{-a} \omega^{a-1} \exp(-\omega/b)$, where $\Gamma(x)$ is the gamma function.

Figure 11: Streamflow predictions of the ABC (3), GR4J (4), HYMOD (5), TOPMO (6), AWBM (8), NAM (9), HBV (9) and SAC-SMA (13) conceptual watershed models for a small period of the $n = 3,000$ day training data record. The number between parenthesis lists the number of calibration parameters. The solid red circles correspond to the daily measured discharges.

Figure 12: Weighted average BMA forecast (solid blue line) and 50% (dim gray), 75% (medium gray) and 95% (light gray) quantiles of the BMA forecast distribution for a representative 270-day period of the training data set using the (a) normal, (b) lognormal, (c) generalized normal, (d) gamma and (e) Weibull distributions with a model-dependent (= sole) forecast variance. The red circles are the daily discharge observations.

Figure 13: (a) 50%, 75% and 95% prediction intervals of the BMA forecast density for the lognormal distribution of Table 8 with a constant model-dependent (= sole) forecast variance and corresponding traces of the (b) quadratic, (c) logarithmic, (d) spherical, (e) continuous ranked probability and (f) energy scoring rules.

Figure 14: Relationship between the ensemble size, K, and the time-averaged value of the logarithmic scoring rule for HYMOD (red line), Hmodel (blue line) and SAC-SMA (green line) using daily discharge data from the Leaf River watershed. The dashed vertical lines point out the optimum size of the ensemble.

Figure 15: The 50, 75, 95 and 100% simulation uncertainty intervals of the (a) HYMOD, (b) Hmodel and (c) SAC-SMA conceptual hydrologic models for the optimum ensemble according to the logarithmic scoring rule.

Figure 16: Preliminary results of the cumulative ranked exceedance probability divergence score: (a) measured FDC (red dots) and SAC-SMA model simulated reliability functions (gray lines) of the $K = 50$ discharge records of the ensemble with lowest values of the CREPS divergence and (b,c) bivariate scatter plots of the CREPS divergence and (b) mean taxicab distance, $\overline{d}_{\mathrm{T}}(P, Q)$ and (c) normal log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ of Equation (79) using all discharge records of the GLUE ensemble.

Figure 17: Traces of the expected value of the interval score $S_{\mathrm{IS}}^{\alpha}(P, \omega)$ as function of the lower endpoint l of the $100(1-\alpha)\%$ prediction interval using $\alpha = 0.05$ (red), $\alpha = 0.25$ (blue) and $\alpha = 0.50$ (green). The colored dots are a projection of the maximum interval score on the x-axis.

Figure 18: Scatter plots (blue squares) of the $\log_{10}(y)$ versus $\log_{10}(\mathrm{d}y/\mathrm{d}t)$ relationship for the different models of the BMA discharge ensemble. The red dots correspond to the measured discharge record.

Table Captions

Table 1: Time-averaged performance measures, $\overline{M}(\mathbf{P}, \boldsymbol{\omega})$, of distribution forecasts $\mathbf{P} = \{P_1, \ldots, P_n\}$ and verifying observations $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$.

Table 2: Quadratic, logarithmic and pseudospherical scoring rules for categorical variables: Entropy function, scoring rule, expected score function and divergence function for a *distribution forecast* $\mathbf{p} = (p_1, \ldots, p_m)^\top$ on the convex class $\mathscr{P} = \mathscr{P}_m$ of $m \geq 2$ mutually exclusive and collectively exhaustive events, $\Omega = \{1, \ldots, m\}$. The m-vector $\mathbf{q} = (q_1, \ldots, q_m)^\top$ lists the true event probabilities.

Table 3: Strictly proper scoring rules for a dichotomous event (*rain* and *no rain*) with *probability forecast* $\mathbf{p} = (p, 1-p)$ on $\Omega = \{1, 0\}$ with $p \in [0, 1]$.

Table 4: Entropy, $H(\mathbf{p})$, expectation, $\mathcal{S}(\mathbf{p}, \mathbf{q})$ and divergence, $d(\mathbf{p}, \mathbf{q})$ of the *strictly proper* categorical scoring rules of Table 2 for distribution forecasts $P_1, \ldots, P_5$ displayed in Fig. 4a using $m = 60$ discrete values, $\Omega = \frac{1}{20}\{1, 3, 5, \ldots, 117, 119\}$. Column R lists the rank of the distribution forecasts. The bottom row presents the values for a perfect distribution forecast, $P = Q$.

Table 5: Entropy, $H(\mathbf{p})$, expectation, $\mathcal{S}(\mathbf{p}, \mathbf{q})$ and divergence, $d(\mathbf{p}, \mathbf{q})$ of the *strictly proper* categorical scoring rules of Table 2 for the *true* and *forecasted* rainfall probabilities of Table E.1.

Table 6: Summary of *strictly proper* scoring rules for a density forecast f_P and verifying observation ω. The numerical form assumes that the forecast distribution P is a m-member ensemble $(y_1, \ldots, y_m)^\top$.

Table 7: Mean values of the quadratic, logarithmic, spherical and continuous ranked probability scoring rules for the distribution forecasts $P_1, \ldots, P_5$ portrayed in Fig. 4a. The last column reports the mean scores for a perfect distribution forecast, $P = Q$.

Table 8: Formulation, coefficients and unknown parameters of the 1. generalized normal, 2. lognormal, 3. gamma and 4. Weibull predictive PDFs used in the BMA forecast density of Equation (70).

Table 9: Time-averaged values of the *strictly proper* scoring rules of Table 6 for the BMA density forecast $f_{P_t}(y|\boldsymbol{\beta}, \boldsymbol{\psi})$; $t = (1, \ldots, n)$ of Equation (70) using the normal ($\tau = 2$), lognormal, generalized normal, gamma and Weibull predictive PDFs with a group or model-dependent (= sole) constant variance. We also list the performance metrics, R_{l}, C_{v}, C and W of Table 1 and report the log-likelihood, $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$, Root Mean Square Error (RMSE), coefficient of determination (R^2) and KG efficiency of the weighted-average BMA forecast. The bottom row lists the number d of unknown BMA parameters.

Table 10: Pearson correlation coefficients of the log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ of the weighted-average BMA forecast and time-averaged values of the QS, LS, SS, CRPS and ES *strictly proper* scoring rules, improper performance metrics of Table 1 and RMSE, R^2 and KGE scoring functions.

Table 11: Time-averaged values of the moments (skew and kurtosis) and performance metrics (R_{l}, C_{v}, C and W) of the distribution forecasts simulated by HYMOD, Hmodel and the SAC-SMA model and RMSE, R^2 and KGE scoring functions of the mean functional of the forecasts. The coverage C of the $\gamma = 1 - \alpha$ prediction intervals corresponds to the significance levels $\alpha = 0.5$, 0.75 and 0.95 of Fig. 15. The second column d lists the number of unknown parameters.

Table 12: Unconditional, $\overline{\mathbf{p}}$, and conditional, $\boldsymbol{\pi}$, probabilities of the watershed models estimated from the $3,000$-day training data record: $\pi_{jk} = \mathbb{P}(\omega = y_j|y_k)$ is the probability of y_j given that y_k is the best forecast in the ensemble at the previous time.

Table 13: Time-averaged expected value of the *strictly proper* quadratic scoring rule (see Table 6), entropy, resolution and reliability for the BMA density forecasts of Equation (70) using the normal, lognormal, generalized normal, gamma and Weibull distributions with a constant group and sole variance, respectively. The bottom row completes the decomposition of Equation (100).

Table 14: Mean values of the Bayes factors $B_{i,j}$ of the distribution forecasts $P_1, \ldots, P_5$ displayed in Fig. 4a from application of Equation (111) to the sum of the logarithmic scores $S_{\mathrm{LS}}(P, \omega)$ in units of bits ($\mathfrak{b} = e$) of the $n = 10^4$ outcomes $\omega_1, \ldots, \omega_n$. The last row and column correspond to a perfect distribution forecast, $P = Q$. *Kass and Raftery* (1995) (P. 777) categorize the Bayes factors and present descriptive statements on the strength of the evidence; $B_{i,j} = 0$ favors in strongest possible terms the null hypothesis, P_j, and vice-versa $B_{i,j} > 75$ means that there is decisive evidence against P_j.

Appendix A: On Gibbs' inequality

Gibbs' inequality

$$\mathbb{H}(Q,P) \geq \mathbb{H}(Q) \tag{A.1}$$

was presented by the American scientist Josiah Willard Gibbs (1839-1903) and states that the cross-entropy $\mathbb{H}(Q,P)$ of two probability distributions Q and P will always exceed the entropy $\mathbb{H}(Q) = \mathbb{H}(Q,Q)$ of distribution Q alone unless $P = Q$ then $\mathbb{H}(Q,P) = \mathbb{H}(Q)$. Different mathematical proofs exist of this inequality. For completeness, we present one of them in this Appendix.

Suppose Q and P are discrete probability distributions on a common sample space Ω. If x is a possible outcome then $q(x) \geq 0$ and $p(x) \geq 0$ denote the probability for $x \in \Omega$ with $\sum_{x\in\Omega} q(x) = 1$ and $\sum_{x\in\Omega} p(x) = 1$. Equation (A.1) may now be written in discretized form

$$\mathbb{H}(\mathbf{q},\mathbf{p}) \geq \mathbb{H}(\mathbf{q}) \tag{A.2}$$

where $\mathbf{q}$ and $\mathbf{p}$ are vectors with probabilities of Q and P for all events of Ω. Now we can write

$$\sum_{x\in\Omega} q(x)\log_{\mathfrak{b}}\left(\frac{1}{p(x)}\right) \geq \sum_{x\in\Omega} q(x)\log_{\mathfrak{b}}\left(\frac{1}{q(x)}\right), \tag{A.3}$$

which is equal to

$$\sum_{x\in\Omega} q(x)\log_{\mathfrak{b}}\big(q(x)\big) - \sum_{x\in\Omega} q(x)\log_{\mathfrak{b}}\big(p(x)\big) \geq 0, \tag{A.4}$$

and simplifies further to

$$\begin{aligned}
&\sum_{x\in\Omega} q(x)\Big(\log_{\mathfrak{b}}\big(q(x)\big) - \log_{\mathfrak{b}}\big(p(x)\big)\Big) \geq 0\\
&\sum_{x\in\Omega} q(x)\log_{\mathfrak{b}}\left(\frac{q(x)}{p(x)}\right) \geq 0\\
&\implies d_{\mathrm{KL}}(Q,P) \geq 0.
\end{aligned} \tag{A.5}$$

Thus, to prove Gibbs' inequality we need to demonstrate that the relative entropy, $d_{\mathrm{KL}}(Q,P)$, is nonnegative. We define the function $t(x) = p(x)/q(x)$. This function satisfies the following condition

$$\log_{\mathfrak{b}}\big(t(x)\big) \leq t(x) - 1 \tag{A.6}$$

for all $t(x) > 0$ and $b > 0$ with equality if and only if $t(x) = 1$ and, thus, $P = Q$. The above inequality may be written as follows

$$-\log_{\mathfrak{b}}\left(\frac{q(x)}{p(x)}\right) \leq \frac{p(x)}{q(x)} - 1, \tag{A.7}$$

and, thus, we yield

$$\log_{\mathfrak{b}}\left(\frac{q(x)}{p(x)}\right) \geq 1 - \frac{p(x)}{q(x)}. \tag{A.8}$$

We can multiply both sides of Equation (A.8) with $q(x)$

$$\begin{aligned}\sum_{x\,\in\,\Omega} q(x)\log_{\mathfrak{b}}\left(\frac{q(x)}{p(x)}\right) &\geq \sum_{x\,\in\,\Omega} q(x)\left(1 - \frac{p(x)}{q(x)}\right)\\ &\geq \sum_{x\,\in\,\Omega} q(x) - \sum_{x\,\in\,\Omega} p(x),\end{aligned} \tag{A.9}$$

to arrive at the following inequality

$$d_{\mathrm{KL}}(Q,P) \geq 1 - \sum_{x\,\in\,\Omega} p(x). \tag{A.10}$$

Thus, for Gibbs' inequality to hold we simply need to show that

$$1 - \sum_{x\,\in\,\Omega} p(x) \geq 0 \quad\Longleftrightarrow\quad \sum_{x\,\in\,\Omega} p(x) \leq 1. \tag{A.11}$$

In plain words, the sum of the probabilities of distribution P at the collection of outcomes $x \in \Omega$ cannot exceed unity. This condition will always be satisfied as, (i) all non-zero $q(x)$ values will sum to one and so do their $p(x)$ values, and (ii) all points $x \in \Omega$ at which $p(x) > 0$ but $q(x) = 0$ do not contribute to the relative entropy as $\lim_{q\downarrow 0} q\log(q) = 0$. Hence, the sum of the $p(x)$'s at which $q(x) > 0$ will almost surely be smaller than one unless $P = Q$ then $\mathbb{H}(Q,P) = \mathbb{H}(Q)$ and $d_{\mathrm{KL}}(Q,P) = 0$.

Appendix B: Relative Entropy Triangular and Uniform Distributions

In this Appendix, we present the derivation of the relative entropy of the triangular forecast distribution P and uniform true distribution Q displayed in Figure 1. If we enter the analytic expressions of the PDFs of P and Q into the integral of Equation (1) we yield

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \int_a^b \frac{1}{b-a}\log_{\mathfrak{b}}\left(\frac{(b-a)^2}{(b-a)\big(2(b-a)-2|a+b-2x|\big)}\right)\mathrm{d}x \qquad \text{(B.1)}\\
&= \frac{1}{b-a}\int_a^c \log_{\mathfrak{b}}\left(\frac{b-a}{2(b-a)-2(a+b-2x)}\right)\mathrm{d}x + \frac{1}{b-a}\int_c^b \log_{\mathfrak{b}}\left(\frac{b-a}{2(b-a)-2(2x-a-b)}\right)\mathrm{d}x\\
&= \frac{1}{b-a}\int_a^c \log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{x-a}\right)\mathrm{d}x + \frac{1}{b-a}\int_c^b \log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{b-x}\right)\mathrm{d}x\\
&= \frac{1}{b-a}\Bigg|(x-a)\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{a-b}{a-x}\right)+1\right]\Bigg|_a^c + \frac{1}{b-a}\Bigg|(x-b)\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{b-x}\right)+1\right]\Bigg|_c^b.
\end{aligned}
$$

At the midpoint c, we yield $x-a=\frac{1}{2}(b-a)$ and $x-b=\frac{1}{2}(a-b)$, thus, the expression above simplifies to

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \left(\frac{\frac{1}{2}(b-a)}{b-a}\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{a-b}{\frac{1}{2}(a-b)}\right)+1\right]-0\right)+\left(0-\frac{\frac{1}{2}(a-b)}{b-a}\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{\frac{1}{2}(b-a)}\right)+1\right]\right)\\
&= \frac{1}{2}\left[\log_{\mathfrak{b}}\left(\frac{1}{2}\right)+1\right]+\frac{1}{2}\left[\log_{\mathfrak{b}}\left(\frac{1}{2}\right)+1\right]\\
&= 1-\log_{\mathfrak{b}}(2). \qquad \text{(B.2)}
\end{aligned}
$$

We can swap the arguments Q and P and compute the relative entropy from Q to P to yield

$$
\begin{aligned}
d_{\mathrm{KL}}(P,Q) &= \int_a^b \frac{2(b-a)-2|a+b-2x|}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{(b-a)\big(2(b-a)-2|a+b-2x|\big)}{(b-a)^2}\right)\mathrm{d}x \qquad \text{(B.3)}\\
&= \int_a^c \frac{4(x-a)}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{4(x-a)}{b-a}\right)\mathrm{d}x + \int_c^b \frac{4(b-x)}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{4(b-x)}{b-a}\right)\mathrm{d}x\\
&= \left|\frac{(4a-4x)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{4(a-x)}{a-b}\right)-\frac{1}{2}\right]\right|_a^{\frac{a+b}{2}} + \left|-\frac{(4b-4x)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{4(x-b)}{a-b}\right)-\frac{1}{2}\right]\right|_{\frac{a+b}{2}}^b\\
&= \left(\frac{4(a-b)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{2(a-b)}{a-b}\right)-\frac{1}{2}\right]-0\right)+\left(0+\frac{4(b-a)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{2(a-b)}{a-b}\right)-\frac{1}{2}\right]-0\right)\\
&= \frac{1}{2}\left[\log_{\mathfrak{b}}(2)-\frac{1}{2}\right]+\frac{1}{2}\left[\log_{\mathfrak{b}}(2)-\frac{1}{2}\right]\\
&= \log_{\mathfrak{b}}(2)-\frac{1}{2}. \qquad \text{(B.4)}
\end{aligned}
$$

This concludes our derivation.

Appendix C: KL divergence of two multivariate normal distributions

Suppose that the *true* joint distribution Q of $\mathbf{x} = (x_1, \ldots, x_b)^\top$ and its probabilistic forecast, P, are each described by a multivariate normal distribution, $\mathcal{N}_b(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ and $\mathcal{N}_b(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$, respectively, with means, $\boldsymbol{\mu}_Q = [\,\mu_{Q,1}\ \ldots\ \mu_{Q,b}\,]^\top$ and $\boldsymbol{\mu}_P$, and $b \times b$ covariance matrices, $\boldsymbol{\Sigma}_Q$ and $\boldsymbol{\Sigma}_P$, respectively. The probability density at $\mathbf{x}$ is then equal to

$$f_{\mathrm{N}}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{b/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big), \tag{C.1}$$

where $|\cdot|$ is the determinant operator and the symbol $\top$ denotes transpose. We can use the above expression to derive a closed-form expression for the KL-divergence, $d_{\mathrm{KL}}(Q, P)$, of Q and P in $\mathbb{R}^b$. Indeed, we can write

$$\begin{aligned} d_{\mathrm{KL}}(Q, P) &= \mathbb{E}_Q\left[\log_e\left(\frac{Q(x)}{P(x)}\right)\right] \\ &= \mathbb{E}_Q\big[\log_e\big(f_{\mathrm{N}}(\mathbf{x}, \boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)\big) - \log_e\big(f_{\mathrm{N}}(\mathbf{x}, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)\big)\big]. \end{aligned} \tag{C.2}$$

To cancel the exponential function in Equation (C.1), the base of the logarithm must be fixed to Euler's number, $e = 2.7182818\ldots$, and as a result the relative entropy, $d_{\mathrm{KL}}(Q, P)$, has units of nats. If we admit to Equation (C.2) the normal density of Equation (C.1) we yield

$$\begin{aligned} d_{\mathrm{KL}}(Q, P) = \mathbb{E}_Q\Bigg[&\log_e\left(\frac{1}{(2\pi)^{b/2}|\boldsymbol{\Sigma}_Q|^{1/2}}\right) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q) \\ &- \left[\log_e\left(\frac{1}{(2\pi)^{b/2}|\boldsymbol{\Sigma}_P|^{1/2}}\right) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\right]\Bigg] \\ = \mathbb{E}_Q\Big[&-\tfrac{b}{2}\log_e(2\pi) - \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q|) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q) + \tfrac{b}{2}\log_e(2\pi) \\ &+ \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_P|) + \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\Big] \\ = \tfrac{1}{2}\mathbb{E}_Q\big[&\log_e(|\boldsymbol{\Sigma}_P|) - \log_e(|\boldsymbol{\Sigma}_Q|) + (\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P) \\ &- (\mathbf{x}-\boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\big]. \end{aligned} \tag{C.3}$$

The expected value of a constant, $\mathbb{E}[c]$, is equal to c itself and, thus, Equation (C.3) becomes

$$\begin{aligned} d_{\mathrm{KL}}(Q, P) &= \tfrac{1}{2}\log_e\left(\frac{|\boldsymbol{\Sigma}_P|}{|\boldsymbol{\Sigma}_Q|}\right) + \tfrac{1}{2}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P) - (\mathbf{x}-\boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\big] \\ &= \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) + \tfrac{1}{2}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\big] \\ &\quad - \tfrac{1}{2}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\big]. \end{aligned} \tag{C.4}$$

The vector-matrix-vector product, $(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, produces a scalar whose value depends on the entries of the vector $\mathbf{x}$. The covariance matrix, $\boldsymbol{\Sigma}$, is constant and can be taken out of the expectation

$$\begin{aligned}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] &= \mathbb{E}\big[\mathrm{tr}\big((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)\big]\\ &= \mathbb{E}\big[\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big)\big]\\ &= \mathrm{tr}\big(\mathbb{E}\big[\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big]\big)\\ &= \mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\mathbb{E}\big[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big]\big),\end{aligned} \tag{C.5}$$

where the trace function

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{b} a_{ii} = a_{11} + a_{22} + \ldots + a_{bb}, \tag{C.6}$$

computes the sum of the elements on the main diagonal of the $b \times b$ matrix, $\boldsymbol{\Sigma}^{-1}\mathbb{E}\big[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big]$. We can use Equation (C.5) to write Equation (C.4) as follows

$$\begin{aligned}d_{\mathrm{KL}}(Q,P) = \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) &+ \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_P)(\mathbf{x}-\boldsymbol{\mu}_P)^\top\big]\big)\\ &- \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_Q^{-1}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_Q)(\mathbf{x}-\boldsymbol{\mu}_Q)^\top\big]\big),\end{aligned} \tag{C.7}$$

The expected value of the vector outer product, $(\mathbf{x}-\boldsymbol{\mu}_Q)(\mathbf{x}-\boldsymbol{\mu}_Q)^\top$, with respect to Q is simply equal to the covariance matrix, $\boldsymbol{\Sigma}_Q$, of this distribution. Thus, we yield

$$\begin{aligned}d_{\mathrm{KL}}(Q,P) &= \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_Q\big) + \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q\big[(\mathbf{x}-\boldsymbol{\mu}_P)(\mathbf{x}-\boldsymbol{\mu}_P)^\top\big]\big)\\ &= \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \tfrac{1}{2}b + \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q\big[\mathbf{x}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}_P^\top - \boldsymbol{\mu}_P\mathbf{x}^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top\big]\big)\\ &= \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \tfrac{1}{2}b + \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q\big[\mathbf{x}\mathbf{x}^\top - 2\mathbf{x}\boldsymbol{\mu}_P^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top\big]\big),\end{aligned} \tag{C.8}$$

where the sum of the diagonal elements of the $b \times b$ identity matrix, $\mathbf{I}_b = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}$, equals the dimension, $b \in \mathbb{N}_+$, of the multivariate normal distribution and the $b \times b$ matrix, $-\mathbf{x}\boldsymbol{\mu}_P^\top - \boldsymbol{\mu}_P\mathbf{x}^\top$, is conveniently written as $-2\mathbf{x}\boldsymbol{\mu}_P^\top$. This formulation for the sum of the two vector outer products holds only for the main diagonal elements of $\mathbf{x}\boldsymbol{\mu}_P^\top - \boldsymbol{\mu}_P\mathbf{x}^\top$ on which the trace function operates.

We can generate a large collection of N points, $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$, from $Q \sim \mathcal{N}_b(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$, and compute numerically the expected (mean) value of the $b \times b$ matrix, $\mathbf{x}\mathbf{x}^\top - 2\mathbf{x}\boldsymbol{\mu}_P^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top$, between square brackets of Equation (C.8). With a little bit more effort, however, we can yield an analytic expression for the KL-divergence. The expected value of $\mathbf{x}$ under Q is equal to the mean, $\boldsymbol{\mu}_Q$, of this distribution. From the general definition of the covariance matrix, we can derive a simple expression for the expected value of the vector outer product, $\mathbf{x}\mathbf{x}^\top$, in Equation (C.8) as follows

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{x}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}\boldsymbol{\mu}^\top] - \mathbb{E}[\boldsymbol{\mu}\mathbf{x}^\top] + \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^\top] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbb{E}[\mathbf{x}^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&\implies \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \qquad \text{(C.9)}
\end{aligned}
$$

Thus, Equation (C.8) is equal to

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \tfrac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \tfrac{1}{2}b + \tfrac{1}{2}\operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\Sigma}_Q + \boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top - 2\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top)\big) \\
&= \tfrac{1}{2}\big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - b + \operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q + \boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top - 2\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top + \boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top\big)\big] \\
&= \tfrac{1}{2}\big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - b + \operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\big) + \operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top\big) - 2\operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top\big) \\
&\qquad + \operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top\big)\big]. \qquad \text{(C.10)}
\end{aligned}
$$

As corollary of Equation (C.5), we yield

$$
\operatorname{tr}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big) = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}), \qquad \text{(C.11)}
$$

and, consequently, Equation (C.10) may be written as follows

$$
d_{\mathrm{KL}}(Q,P) = \tfrac{1}{2}\big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - b + \operatorname{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\big) + \boldsymbol{\mu}_Q^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q - 2\boldsymbol{\mu}_Q^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P + \boldsymbol{\mu}_P^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\big]. \qquad \text{(C.12)}
$$

The three vector-matrix-vector products in the above expression can be factorized to yield

$$
d_{\mathrm{KL}}(Q,P) = \tfrac{1}{2}\big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - b + \operatorname{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + (\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P)\big]. \qquad \text{(C.13)}
$$

This concludes our derivation of the KL-divergence of two multivariate normal distributions, $Q \sim \mathcal{N}_b(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ and $P = \mathcal{N}_b(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ in $\mathbb{R}^b$ with $b \in \mathbb{N}_+$.

In the special case of $b = 1$ or two univariate normal distributions, $Q \sim \mathcal{N}(\mu_Q, \sigma_Q^2)$ and $P = \mathcal{N}(\mu_P, \sigma_P^2)$, the above expression simplifies to

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \tfrac{1}{2}\left[\log_e\left(\frac{\sigma_P^2}{\sigma_Q^2}\right) - 1 + \frac{\sigma_Q^2}{\sigma_P^2} + \frac{(\mu_Q-\mu_P)^2}{\sigma_P^2}\right] \\
&= \log_e\left(\frac{\sigma_P}{\sigma_Q}\right) + \frac{\sigma_Q^2 + (\mu_Q-\mu_P)^2 - \sigma_P^2}{2\sigma_P^2}. \qquad \text{(C.14)}
\end{aligned}
$$

The use of the natural logarithm in our derivation of Equations (C.13) and (C.14) affixes the unit of nats to $d_{\mathrm{KL}}(Q,P)$. To change units of the relative entropy, one should just divide $d_{\mathrm{KL}}(Q,P)$ by $\log_e(z)$. Then for $z = 2$ we yield $d_{\mathrm{KL}}(Q,P)$ in bits. Furthermore, we obtain the reverse KL-divergence $d_{\mathrm{KL}}(P,Q)$ by swapping arguments Q and P in the respective Equations.

Appendix D: KL divergence of normal and uniform distributions

We derive an analytic expression for the relative entropy $d_{\mathrm{KL}}(Q, P)$

$$d_{\mathrm{KL}}\big(Q, P\big) = \mathbb{H}(Q, P) - \mathbb{H}(Q), \tag{D.1}$$

of a normal *true distribution* $Q = \mathcal{N}(\mu_Q, \sigma_Q^2)$ and uniform *forecast distribution* $P = \mathcal{U}(a_P, b_P)$ on a bounded sample space $x \in [a_P, b_P]$ and $b_P > a_P$. The cross-entropy of Q and P is equal to

$$\begin{aligned}
\mathbb{H}(Q, P) &= -\int_{a_P}^{b_P} Q(x) \log_{\mathfrak{b}}\big(P(x)\big)\mathrm{d}x \\
&= -\int_{a_P}^{b_P} Q(x) \log_e\left(\frac{1}{b_P - a_P}\right)\mathrm{d}x \\
&= \log_e(b_P - a_P)\int_{a_P}^{b_P} Q(x)\mathrm{d}x \\
&= \log_e(b_P - a_P),
\end{aligned} \tag{D.2}$$

in units of nats. Note that the cross-entropy will attain an infinite value on the extended real line, $x \in \overline{\mathbb{R}}$ as $P(x) \to 0$. The entropy of the normal *true distribution* Q may be computed as follows

$$\begin{aligned}
\mathbb{H}(Q) &= -\int_{a_P}^{b_P} Q(x) \log_{\mathfrak{b}}\big(Q(x)\big)\mathrm{d}x \\
&= -\int_{a_P}^{b_P} Q(x) \log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu_Q)^2}{\sigma_Q^2}\right)\right]\mathrm{d}x \\
&= -\int_{a_P}^{b_P} Q(x)\left(\log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\right] - \frac{1}{2}\frac{(x-\mu_Q)^2}{\sigma_Q^2}\right)\mathrm{d}x \\
&= \log_e\big(2\pi\sigma_Q^2\big)^{1/2}\int_{a_P}^{b_P} Q(x)\mathrm{d}x + \frac{1}{2\sigma_Q^2}\int_{a_P}^{b_P}(x-\mu_Q)^2 Q(x)\mathrm{d}x \\
&= \tfrac{1}{2}\log_e(2\pi\sigma_Q^2) \times 1 + \frac{1}{2\sigma_Q^2} \times \sigma_Q^2 \\
&= \tfrac{1}{2}\log_e(2e\pi\sigma_Q^2).
\end{aligned} \tag{D.3}$$

The relative entropy is now equal to

$$\begin{aligned}
d_{\mathrm{KL}}\big(Q, P\big) &= \mathbb{H}(Q, P) - \mathbb{H}(Q) \\
&= \log_e(b_P - a_P) - \tfrac{1}{2}\log_e(2e\pi\sigma_Q^2),
\end{aligned} \tag{D.4}$$

in units of nats. If the interval of $P(x)$ exceeds by far that of $Q(x)$ and, thus, $b_P - a_P \gg \sigma_Q$ then $d_{\mathrm{KL}}\big(Q, P\big)$ may attain an infinite value. For any scalar $\beta \in \mathbb{R}_+$ and $\sigma_Q = (b_P - a_P)/\nu$ the relative entropy simplifies to $d_{\mathrm{KL}}\big(Q, P\big) = -\frac{1}{2}\log_e(2e\pi) + \log_e(\nu)$.

To resolve problems with the uniform distribution of P on an unbounded interval we could specify $P(x) \propto 1$ instead. Then the cross-entropy $\mathbb{H}(Q,P) = 0$ and the relative entropy $d_{\mathrm{KL}}(Q,P)$ reduces to the so-called differential entropy $\frac{1}{2}\log_e(2e\pi\sigma_Q^2)$ of the normal distribution Q.

We can follow a similar derivation for the reverse KL-divergence, $d_{\mathrm{KL}}(P,Q)$. The cross-entropy of P and Q in units of nats is equal to

$$\begin{aligned}
\mathbb{H}(P,Q) &= -\int_{a_P}^{b_P} P(x)\log_{\mathfrak{b}}\big(Q(x)\big)\mathrm{d}x \\
&= -\int_{a_P}^{b_P} P(x)\log_{\mathfrak{b}}\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu_Q)^2}{\sigma_Q^2}\right)\right]\mathrm{d}x \\
&= -\int_{a_P}^{b_P} P(x)\left(\log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\right] - \frac{1}{2}\frac{(x-\mu_Q)^2}{\sigma_Q^2}\right)\mathrm{d}x \\
&= \log_e\big(2\pi\sigma_Q^2\,\big)^{1/2}\int_{a_P}^{b_P} P(x)\mathrm{d}x + \frac{1}{2\sigma_Q^2}\int_{a_P}^{b_P}(x-\mu_Q)^2P(x)\mathrm{d}x \\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2) + \frac{1}{2(b_P-a_P)\sigma_Q^2}\left|-\frac{1}{3}(\mu_Q-x)^3\right|_{a_P}^{b_P} \\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2) + \frac{(\mu_Q-a_P)^3-(\mu_Q-b_P)^3}{6(b_P-a_P)\sigma_Q^2}. \qquad \text{(D.5)}
\end{aligned}$$

The entropy of P is equal to Equation (D.2) to yield

$$\begin{aligned}
\mathbb{H}(P) &= -\int_{a_P}^{b_P} P(x)\log_{\mathfrak{b}}\big(P(x)\big)\mathrm{d}x \\
&= -\int_{a_P}^{b_P} P(x)\log_e\left(\frac{1}{b_P-a_P}\right)\mathrm{d}x \\
&= \log_e(b_P-a_P)\int_{a_P}^{b_P} P(x)\mathrm{d}x \\
&= \log_e(b_P-a_P), \qquad \text{(D.6)}
\end{aligned}$$

in units of nats. Now we yield the following expression for the relative entropy $d_{\mathrm{KL}}(P,Q)$ in nats

$$\begin{aligned}
d_{\mathrm{KL}}\big(P,Q\big) &= \mathbb{H}(P,Q) - \mathbb{H}(P) \\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2) + \frac{(\mu_Q-a_P)^3-(\mu_Q-b_P)^3}{6(b_P-a_P)\sigma_Q^2} - \log_e(b_P-a_P). \qquad \text{(D.7)}
\end{aligned}$$

This confirms again that the relative entropy is not symmetric in Q and P.

Appendix E: Rainfall data

Table E.1 is taken from *Hughes and Topp* (2015) and summarizes a data set of $n = 346$ forecasts of 24-hour precipitation probability made by the Finnish Meteorological Institute during 2003 for the city of Tampere in south-central Finland. The left block presents the original data, and the right block lists the data used in our case study.

Table E.1: Rainfall data from Table 1 of *Hughes and Topp* (2015) for the city of Tampere, Finland.

a: Original data						b: Adapted data		
k	p_k	n_k	o_k	$\overline{o}_k$	n_k/n	k	p_k	q_k
1	0.05	46	1	0.0217	0.1329	1	0.1727	0.1359
2	0.1	55	1	0.0182	0.1590	2	0.1636	0.1364
3	0.2	59	5	0.0847	0.1705	3	0.1455	0.1272
4	0.3	41	5	0.1220	0.1185	4	0.1273	0.1220
5	0.4	19	4	0.2105	0.0549	5	0.1091	0.1097
6	0.5	22	8	0.3636	0.0636	6	0.0909	0.0884
7	0.6	22	6	0.2727	0.0636	7	0.0727	0.1011
8	0.7	34	16	0.4706	0.0983	8	0.0545	0.0736
9	0.8	24	16	0.6667	0.0694	9	0.0364	0.0463
10	0.9	11	8	0.7273	0.0318	10	0.0182	0.0379
11	0.95	13	11	0.8462	0.0376	11	0.0091	0.0214
	$\sum$	346	81		1.0000	$\sum$	1.000	1.000

Forecast probabilities of rainfall p_k; $k = (1, \ldots, m)$ were issued using $m = 11$ categories. The variable n_k lists the number of days for which the Finnish Meteorological Institute quoted p_k. Then, o_k, signifies the number of days on which measured rainfall depths for the city of Tampere exceeded ≥ 0.2 mm, otherwise a no-rain day was recorded. Next, the ratio $\overline{o}_k = o_k/n_k$ equals the *true* rainfall probability for each forecast category. Finally, n_k/n corresponds to the relative frequency of each forecast category. We refer readers to *Hughes and Topp* (2015) for a more detailed description of the data set and its use in a diagrammatic interpretation of the Brier scoring rule and related score divergence. The raw precipitation data can be found at `https://www.cawcr.gov.au/projects/verification/POP3/POP_3cat_2003.txt`.

The right block tabulates the data that was used in our case study. The forecast probabilities of the individual categories are normalized to sum to unity. This defines a m-vector $\mathbf{p} = (p_1, \ldots, p_m)^\top$ of rainfall probability forecasts. We apply a similar normalization to the $\overline{o}_k$'s to yield the vector $\mathbf{q} = (q_1, \ldots, q_m)^\top$ of *true* rainfall probabilities.

Appendix F: Quantile form of Continuous Ranked Probability Score

The CRPS of a distribution forecast P and verifying observation $\omega \in \Omega$ is equal to the integral of the quantile score function

$$S_{\mathrm{CRPS}}(P, \omega) = \int_0^1 S^{\tau}_{\mathrm{QNT}}(P, \omega)\, \mathrm{d}\tau, \tag{F.1}$$

and, thus, we yield

$$S_{\mathrm{CRPS}}(P, \omega) = -2 \int_0^1 (\mathbb{1}\{\omega < y_\tau\} - \tau)(y_\tau - \omega)\, \mathrm{d}\tau. \tag{F.2}$$

The above expression may be rearranged and written as follows

$$\begin{aligned}
S_{\mathrm{CRPS}}(P, \omega) &= 2 \int _0^1 \tau(y_\tau - \omega)\, \mathrm{d}\tau - 2 \int_0^1 \mathbb{1}\{\omega < y_\tau\}(y_\tau - \omega)\, \mathrm{d}\tau \\
&= 2 \int_0^1 \tau \big(F_P^{-1}(\tau) - \omega\big)\, \mathrm{d}\tau - 2 \int_0^1 \mathbb{1}\{F_P(\omega) < \tau\} \big(F_P^{-1}(\tau) - \omega\big)\, \mathrm{d}\tau \\
&= 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau - 2\omega \int_0^1 \tau\, \mathrm{d}\tau - 2 \int_{F_P(\omega)}^1 \big(F_P^{-1}(\tau) - \omega\big)\, \mathrm{d}\tau \\
&= 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau - 2\omega \int_0^1 \tau\, \mathrm{d}\tau - 2 \int_{F_P(\omega)}^1 F_P^{-1}(\tau)\, \mathrm{d}\tau + 2\omega \int_{F_P(\omega)}^1 \mathrm{d}\tau \\
&= 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau - 2\omega \Big[\tfrac{1}{2}\tau^2\Big]_0^1 - 2 \int_{F_P(\omega)}^1 F_P^{-1}(\tau)\, \mathrm{d}\tau + 2\omega \big[\tau\big]_{F_P(\omega)}^1 \\
&= 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau - \omega - 2 \int_{F_P(\omega)}^1 F_P^{-1}(\tau)\, \mathrm{d}\tau + 2\omega - 2\omega F_P(\omega) \\
&= 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau + \omega - 2 \int_{F_P(\omega)}^1 F_P^{-1}(\tau)\, \mathrm{d}\tau - 2\omega F_P(\omega),
\end{aligned} \tag{F.3}$$

and finally, we yield

$$S_{\mathrm{CRPS}}(P, \omega) = \omega\big(1 - 2F_P(\omega)\big) + 2 \int_0^1 \tau F_P^{-1}(\tau)\, \mathrm{d}\tau - 2 \int_{F_P(\omega)}^1 F_P^{-1}(\tau)\, \mathrm{d}\tau. \tag{F.4}$$

This concludes the derivation.

Appendix G: Continuous Ranked Probability Score for $\mathcal{N}(\mu, \sigma^2)$

The continuous ranked probability score (CRPS) is given by

$$S_{\text{CRPS}}(P, \omega) = -\int_{-\infty}^{\infty} \big(F_P(z) - \mathbb{1}\{z \geq \omega\}\big)^2 \mathrm{d}z, \tag{G.1}$$

where $F_P(z)$ denotes the cumulative distribution function of P and the indicator function, $\mathbb{1}\{a\}$, returns 1 if a is true and zero otherwise. If we make the convenient assumption that the probability measure is univariate normal, $P = \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function of P has a closed-form expression

$$F_P(x, \mu, \sigma^2) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \tag{G.2}$$

where $\text{erf}\,(x)$ is the error function for element x

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\mathrm{d}t, \tag{G.3}$$

and the CRPS becomes

$$S_{\text{CRPS}}\big(\mathcal{N}(\mu, \sigma^2), \omega\big) = -\int_{-\infty}^{\omega} \left(\frac{1}{2}\left[1 + \text{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]\right)^2 \mathrm{d}z - \int_{\omega}^{\infty} \left(\frac{1}{2}\left[1 + \text{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right] - 1\right)^2 \mathrm{d}z. \tag{G.4}$$

We use the symbolic toolbox in MATLAB to yield closed-form expressions for the two definite integrals in the above expression. We first write out the left integral

$$\begin{aligned}
-\int_{-\infty}^{\omega} \left(\frac{1}{2}\left[1 + \text{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]\right)^2 \mathrm{d}z = -\Bigg| &\frac{z}{4} + \frac{(\mu - z)}{2}\,\text{erf}\left(\frac{\mu - z}{\sigma\sqrt{2}}\right) - \frac{(\mu - z)}{4}\,\text{erf}\left(\frac{\mu - z}{\sigma\sqrt{2}}\right)^2 \\
&+ \frac{\sigma}{2\sqrt{\pi}}\,\text{erf}\left(\frac{\mu - z}{\sigma}\right) + \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(\mu - z)^2}{2\sigma^2}\right) \\
&- \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(\mu - z)^2}{2\sigma^2}\right) \text{erf}\left(\frac{\mu - z}{\sigma\sqrt{2}}\right)\Bigg|_{-\infty}^{\omega} \\
= \Bigg| &-\frac{z}{4} - \frac{(\mu - z)}{2} f(z, \mu, \sigma) + \frac{(\mu - z)}{4} f(z, \mu, \sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(z, \mu, \sigma) \\
&- \frac{\sigma}{\sqrt{2\pi}} h(z, \mu, \sigma) + \frac{\sigma}{\sqrt{2\pi}} f(z, \mu, \sigma) h(z, \mu, \sigma)\Bigg|_{-\infty}^{\omega},
\end{aligned} \tag{G.5}$$

where $f(z,\mu,\sigma) = \operatorname{erf}\big(\frac{1}{2}\sqrt{2}(\mu - z)/\sigma\big)$, $g(z,\mu,\sigma) = \operatorname{erf}\big((\mu - z)/\sigma\big)$ and $h(z,\mu,\sigma) = \exp\big(-\frac{1}{2}(\mu - z)^2/\sigma^2\big)$. We follow a similar recipe for the right integral of Equation (G.4) to yield

$$
\begin{aligned}
-\int_{\omega}^{\infty}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right]-1\right)^2 \mathrm{d}z &= -\Bigg|\frac{z}{4}-\frac{(\mu-z)}{2}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)-\frac{(\mu-z)}{4}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)^2 \\
&\quad+\frac{\sigma}{2\sqrt{\pi}}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right)-\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^2}{2\sigma^2}\right) \\
&\quad-\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^2}{2\sigma^2}\right)\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)\Bigg|_{\omega}^{\infty} \\
&= \Bigg|-\frac{z}{4}+\frac{(\mu-z)}{2}f(z,\mu,\sigma)+\frac{(\mu-z)}{4}f(z,\mu,\sigma)^2-\frac{\sigma}{2\sqrt{\pi}}g(z,\mu,\sigma) \\
&\quad+\frac{\sigma}{\sqrt{2\pi}}h(z,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\mu,\sigma)h(z,\mu,\sigma)\Bigg|_{\omega}^{\infty}. \qquad \text{(G.6)}
\end{aligned}
$$

Before admitting the integral bounds, we first perform limit analysis of the constituent functions

$$
\begin{aligned}
\lim_{z\to-\infty} f(z,\mu,\sigma) &= \lim_{z\to-\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right) = 1 \\
\lim_{z\to\infty} f(z,\mu,\sigma) &= \lim_{z\to\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right) = -1 \\
\lim_{z\to-\infty} g(z,\mu,\sigma) &= \lim_{z\to-\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right) = 1 \\
\lim_{z\to\infty} g(z,\mu,\sigma) &= \lim_{z\to\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right) = -1 \\
\lim_{z\to-\infty} h(z,\mu,\sigma) &= \lim_{z\to-\infty}\exp\left(-\frac{(\mu-z)^2}{2\sigma^2}\right) = 0 \\
\lim_{z\to\infty} h(z,\mu,\sigma) &= \lim_{z\to\infty}\exp\left(-\frac{(\mu-z)^2}{2\sigma^2}\right) = 0. \qquad \text{(G.7)}
\end{aligned}
$$

The left integral is now equal to

$$
\begin{aligned}
-\int_{-\infty}^{\omega}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]\right)^2 \mathrm{d}z &= \Bigg|-\frac{z}{4}-\frac{(\mu-z)}{2}f(z,\mu,\sigma)+\frac{(\mu-z)}{4}f(z,\mu,\sigma)^2-\frac{\sigma}{2\sqrt{\pi}}g(z,\mu,\sigma) \\
&\quad-\frac{\sigma}{\sqrt{2\pi}}h(z,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\mu,\sigma)h(z,\mu,\sigma)\Bigg|_{-\infty}^{\omega} \\
&= \left(-\frac{\omega}{4}-\frac{(\mu-\omega)}{2}f(\omega,\mu,\sigma)+\frac{(\mu-\omega)}{4}f(\omega,\mu,\sigma)^2-\frac{\sigma}{2\sqrt{\pi}}g(\omega,\mu,\sigma)\right. \\
&\quad\left.-\frac{\sigma}{\sqrt{2\pi}}h(\omega,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(\omega,\mu,\sigma)h(\omega,\mu,\sigma)\right) \\
&\quad-\lim_{z\to-\infty}\left(-\frac{z}{4}-\frac{(\mu-z)}{2}f(z,\mu,\sigma)+\frac{(\mu-z)}{4}f(z,\mu,\sigma)^2\right. \\
&\quad\left.-\frac{\sigma}{2\sqrt{\pi}}g(z,\mu,\sigma)-\frac{\sigma}{\sqrt{2\pi}}h(z,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\mu,\sigma)h(z,\mu,\sigma)\right) \\
&= \left(-\frac{\omega}{4}-\frac{(\mu-\omega)}{2}f(\omega,\mu,\sigma)+\frac{(\mu-\omega)}{4}f(\omega,\mu,\sigma)^2-\frac{\sigma}{2\sqrt{\pi}}g(\omega,\mu,\sigma)\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg) \\
&\quad - \lim_{z\to-\infty}\Bigg(-\frac{z}{4} - \frac{(\mu-z)}{2}\times 1 + \frac{(\mu-z)}{4}\times 1^2 - \frac{\sigma}{2\sqrt{\pi}}\times 1 \\
&\quad - \frac{\sigma}{\sqrt{2\pi}}\times 0 + \frac{\sigma}{\sqrt{2\pi}}\times 0\times 1\Bigg) \\
&= \Bigg(-\frac{\omega}{4} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) \\
&\quad - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg) \\
&\quad - \lim_{z\to-\infty}\Bigg(-\frac{z}{4} + \frac{z}{2} - \frac{\mu}{2} - \frac{z}{4} + \frac{\mu}{4} - \frac{\sigma}{2\sqrt{\pi}}\Bigg) \\
&= -\frac{\omega}{4} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) \\
&\quad - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma) + \frac{\mu}{4} + \frac{\sigma}{2\sqrt{\pi}}, \qquad \text{(G.8)}
\end{aligned}
$$

and the right integral becomes

$$
\begin{aligned}
-\int_\omega^\infty \left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]-1\right)^2 \mathrm{d}z &= \Bigg|-\frac{z}{4} + \frac{(\mu-z)}{2} f(z,\mu,\sigma) + \frac{(\mu-z)}{4} f(z,\mu,\sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(z,\mu,\sigma) \\
&\quad + \frac{\sigma}{\sqrt{2\pi}} h(z,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(z,\mu,\sigma) h(z,\mu,\sigma)\Bigg|_\omega^\infty \\
&= \lim_{z\to\infty}\Bigg(-\frac{z}{4} + \frac{(\mu-z)}{2} f(z,\mu,\sigma) + \frac{(\mu-z)}{4} f(z,\mu,\sigma)^2 \\
&\quad - \frac{\sigma}{2\sqrt{\pi}} g(z,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} h(z,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(z,\mu,\sigma) h(z,\mu,\sigma)\Bigg) \\
&\quad - \Bigg(-\frac{\omega}{4} + \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 \\
&\quad - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg) \\
&= \lim_{z\to\infty}\Bigg(-\frac{z}{4} + \frac{(\mu-z)}{2}\times(-1) + \frac{(\mu-z)}{4}\times(-1)^2 - \frac{\sigma}{2\sqrt{\pi}}\times(-1) \\
&\quad + \frac{\sigma}{\sqrt{2\pi}}\times 0 + \frac{\sigma}{\sqrt{2\pi}}\times 0\times(-1)\Bigg) - \Bigg(-\frac{\omega}{4} + \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) \\
&\quad + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) \\
&\quad + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg) \\
&= \lim_{z\to\infty}\Bigg(-\frac{z}{4} + \frac{z}{2} - \frac{\mu}{2} - \frac{z}{4} + \frac{\mu}{4} + \frac{\sigma}{2\sqrt{\pi}}\Bigg) \\
&\quad - \Bigg(-\frac{\omega}{4} + \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 \\
&\quad - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg)
\end{aligned}
$$

$$= -\frac{\mu}{4} + \frac{\sigma}{2\sqrt{\pi}} + \frac{\omega}{4} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) - \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2$$
$$+ \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) - \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma). \tag{G.9}$$

We can now add up Equations (G.8) and (G.9) to yield the CRPS of $P = \mathcal{N}(\mu, \sigma^2)$ in Equation (G.4)

$$\begin{aligned} S_{\text{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) &= \Bigg(\cancel{-\frac{\omega}{4}} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \cancel{\frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2} - \cancel{\frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma)} \\ &\quad - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \cancel{\frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)} + \cancel{\frac{\mu}{4}} + \frac{\sigma}{2\sqrt{\pi}} \Bigg) + \\ &\quad \Bigg(\cancel{-\frac{\mu}{4}} + \frac{\sigma}{2\sqrt{\pi}} + \cancel{\frac{\omega}{4}} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) - \cancel{\frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2} \\ &\quad \cancel{+ \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma)} - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) \cancel{- \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)} \Bigg) \\ &= -\frac{2(\mu-\omega)}{2} f(\omega,\mu,\sigma) - \frac{2\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{2\sigma}{2\sqrt{\pi}} \\ &= -(\mu-\omega)\,\text{erf}\left(\frac{\mu-\omega}{\sigma\sqrt{2}}\right) - \frac{\sigma\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{(\mu-\omega)^2}{2\sigma^2}\right) + \frac{\sigma}{\sqrt{\pi}}. \end{aligned} \tag{G.10}$$

We can manipulate this expression into a function of the normal PDF, $f_{\mathcal{N}}(x, \mu, \sigma^2)$, and normal CDF, $F_{\mathcal{N}}(x, \mu, \sigma^2)$, in Equation (G.2) as follows

$$\begin{aligned} S_{\text{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) &= -(\omega-\mu)\,\text{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right) - 2\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mu-\omega)^2}{2\sigma^2}\right) + \frac{\sigma}{\sqrt{\pi}} \\ &= -(\omega-\mu)\left[1 + \text{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right] + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}} \\ &= -2(\omega-\mu)\frac{1}{2}\left[1 + \text{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right] + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}} \\ &= -2(\omega-\mu) F_{\mathcal{N}}(\omega,\mu,\sigma^2) + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}}, \end{aligned} \tag{G.11}$$

which may be rearranged and simplified to

$$S_{\text{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) = \frac{\sigma}{\sqrt{\pi}} - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) - (\omega-\mu)\big(2F_{\mathcal{N}}(\omega,\mu,\sigma^2) - 1\big). \tag{G.12}$$

Note that the quantile form of the CRPS in Equation (61) would lead to an equivalent solution as above but in fewer steps. This concludes the derivation of the CRPS for a normal distribution forecast $P = \mathcal{N}(\mu, \sigma^2)$ and verifying observation $\omega \in \Omega$.

Appendix H: Description of Hydrologic models

In this Appendix we present a brief overview of the HYMOD, Hmodel and SAC-SMA conceptual watershed models. The models are coded in MATLAB and C++ and use a mass-conservative second-order integration method with adaptive time step. This guarantees a robust and accurate numerical solution of the simulated fluxes and state variables. Next, we discuss each of the models separately.

H.1 HYdrologic MODEL

The HYdrologic MODel (HYMOD) originates from the PhD thesis of *Boyle* (2001) and describes the rainfall-discharge relationship using five fictitious control volumes. These reservoirs simulate processes such as evaporation, percolation, river inflow and baseflow (see Figure H.1). Table H.1 presents the five

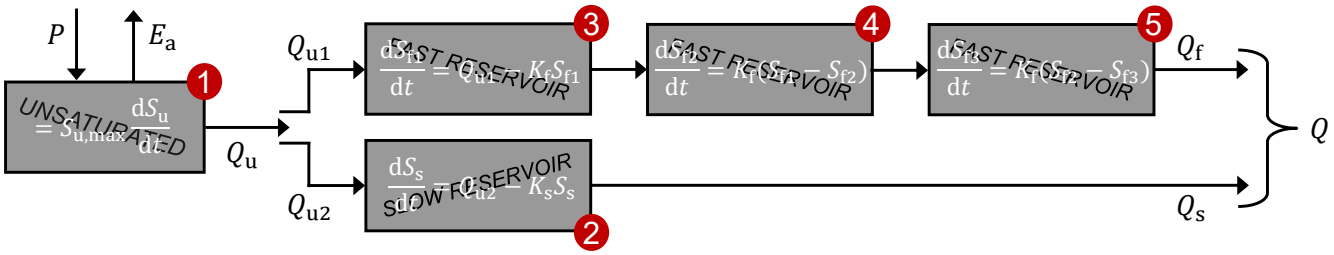

Figure H.1: Schematic illustration of the HYdrologic MODel of *Boyle* (2001). Grey boxes, labeled in red, correspond to fictitious control volumes of the watershed which govern the rainfall-runoff transformation. The state variables, S_u, S_s, S_{f1}, S_{f2} and S_{f3}, correspond to the water storage in each compartment. Arrows portray the fluxes into and out of the compartments, including precipitation, P, evaporation, E_a, precipitation converted into flow, Q_u, fast flow, Q_f, and baseflow, Q_s. The fluxes are computed as follows, $Q_u = P\left(1-(1-\overline{S}_u)^b\right)$, $E_a = E_p\overline{S}_u(1+c)/(\overline{S}_u+c)$, $Q_{u1} = aQ_u$, $Q_{u2} = (1-a)Q_u$, $Q_f = K_f S_{f3}$ and $Q_s = K_s S_s$, where E_p signifies the potential evapotranspiration, $c = 10^{-2}$, $\overline{S}_u = S_u/S_{u,max}$ and $S_{u,max}$, a, b, K_s and K_f are unknown parameters.

hymod parameters with their corresponding symbols, units, and lower and upper bounds.

H.2 Hydrologic model

The Hydrologic model (Hmodel) is a parsimonious conceptual watershed model originally developed by *Schoups et al.* (2010). This model transforms rainfall into runoff at the watershed outlet using an interception, unsaturated zone, fast and slow flow reservoir, respectively, which simulate interception,

Table H.1: Summary of hymod parameters and their symbols, units, and lower and upper bounds.

Parameter	Symbol	Units	Min.	Max.
Maximum storage unsaturated zone	$S_{\mathrm{u,max}}$	mm	50	1000
Spatial variability of soil moisture capacity	b	–	10^{-1}	10
Flow partitioning coefficient	a	–	0	1
Recession constant, slow reservoir	K_{s}	1/d	10^{-4}	1
Recession constant, fast reservoir	K_{f}	1/d	10^{-1}	5

throughfall, evaporation, surface runoff, percolation, fast streamflow and baseflow (see Figure H.2). Table

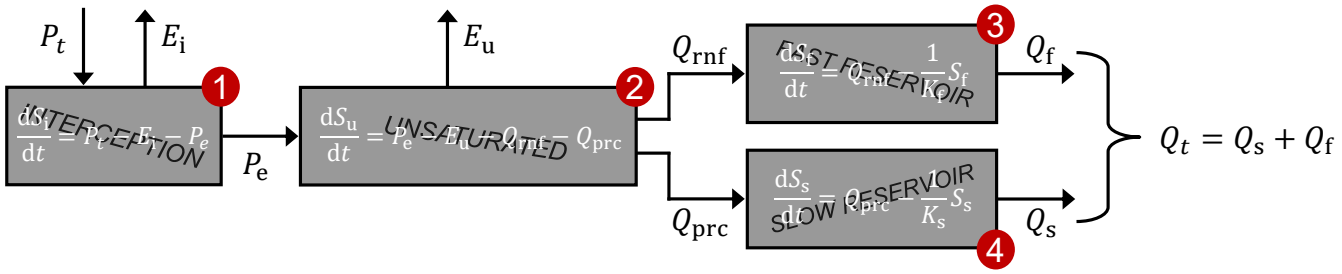

Figure H.2: Schematic illustration of the hmodel after *Schoups et al.* (2010). Grey boxes, labeled in red, correspond to fictitious control volumes of the watershed which govern the rainfall-runoff transformation. The state variables, S_{i}, S_{u}, S_{f} and S_{s}, correspond to the water storage in each compartment. Arrows portray the fluxes into and out of the compartments, including precipitation, P_t, interception evaporation, E_{i}, excess precipitation, P_{e}, evaporation, E_{u}, surface runoff, Q_{rnf}, percolation, Q_{prc}, fast flow, Q_{f} and baseflow, Q_{s}. The fluxes are computed as follows, $E_{\mathrm{i}} = E_{\mathrm{p}} f(\overline{S}_{\mathrm{i}}, \alpha_{\mathrm{i}})$, $P_{\mathrm{e}} = P f(\overline{S}_{\mathrm{i}}, \alpha_{\mathrm{p}})$, $E_{\mathrm{u}} = (E_{\mathrm{p}} - E_{\mathrm{i}}) f(\overline{S}_{\mathrm{u}}, \alpha_{\mathrm{e}})$, $Q_{\mathrm{rnf}} = P_{\mathrm{e}} f(\overline{S}_{\mathrm{u}}, \alpha_{\mathrm{f}})$, $Q_{\mathrm{prc}} = Q_{\mathrm{max}} f(\overline{S}_{\mathrm{u}}, \alpha_{\mathrm{f}})$, $Q_{\mathrm{f}} = S_{\mathrm{f}}/K_{\mathrm{f}}$ and $Q_{\mathrm{s}} = S_{\mathrm{s}}/K_{\mathrm{s}}$, where E_{p} signifies the potential evapotranspiration, the functions, $f(x, y) = \big(1 - g(-xy)\big)/\big(1 - g(-x)\big)$ and $g(c) = \exp\big(\min(c, 300)\big)$, protect against overflow, $\alpha_{\mathrm{i}} = 50$, $\alpha_{\mathrm{p}} = -50$, $\overline{S}_{\mathrm{i}} = S_{\mathrm{i}}/I_{\mathrm{max}}$, $\overline{S}_{\mathrm{u}} = S_{\mathrm{u}}/S_{\mathrm{max}}$ and I_{max}, S_{max}, Q_{max}, α_{e}, α_{f}, K_{f} and K_{s} are unknown parameters.

H.2 lists the seven hmodel parameters and their corresponding symbols, units and upper and lower bounds.

Table H.2: Description of Hmodel parameters, including symbols, units, lower and upper bounds.

Parameter	Symbol	Units	Min.	Max.
Maximum interception	I_{max}	mm	0.1	10
Soil water storage capacity	S_{max}	mm	10	1000
Maximum percolation rate	Q_{max}	mm/d	10^{-1}	100
Evaporation parameter	α_{e}	–	0	100
Runoff parameter	α_{f}	–	-10	10
Time constant, fast reservoir	K_{f}	d	10^{-1}	10
Time constant, slow reservoir	K_{s}	d	1	150

H.3 Sacramento Soil Moisture Accounting model

The Sacramento Soil Moisture Accounting (SAC-SMA) model is used by the National Weather Service River Forecast System (NWSRFS) for flood forecasting throughout the United States. The model converts areal average precipitation into streamflow (*Burnash et al.*, 1973). Our implementation follows *Clark et al.* (2008) and is presented in Figure H.3. Table H.3 presents the thirteen SAC-SMA model parameters with their corresponding symbols, units, and lower and upper bounds This concludes the description of the models.

Table H.3: Description of SAC-SMA parameters, including symbols, units, lower and upper bounds.

Parameter	Symbol	Units	Min.	Max.
Upper zone tension water maximum storage	$S^1_{\mathrm{T,max}}$	mm	50	500
Upper zone free water maximum storage	$S^1_{\mathrm{F,max}}$	mm	10	500
Lower zone tension water maximum storage	$S^2_{\mathrm{T,max}}$	mm	10	500
Lower zone free water primary maximum storage	$S^2_{\mathrm{FP,max}}$	mm	10	1000
Lower zone free water supplemental maximum storage	$S^2_{\mathrm{FS,max}}$	mm	10	1000
Percolation multiplier for the lower layer	α	-	1	250
Percolation exponent for the lower layer	ψ	-	1	5
Upper zone free water lateral depletion rate (interflow rate)	k_{i}	mm/d	10^{-2}	100
Fraction of percolation to tension storage in the lower layer	κ	-	0.05	0.95
Base flow depletion rate for primary reservoir	ν_{P}	d^{-1}	10^{-3}	0.25
Base flow depletion rate for secondary reservoir	ν_{S}	d^{-1}	10^{-3}	0.25
Maximum saturated area (fraction)	$A_{\mathrm{c,max}}$	-	0.05	0.95
Recession constant, routing	K_{f}	d^{-1}	10^{-1}	5

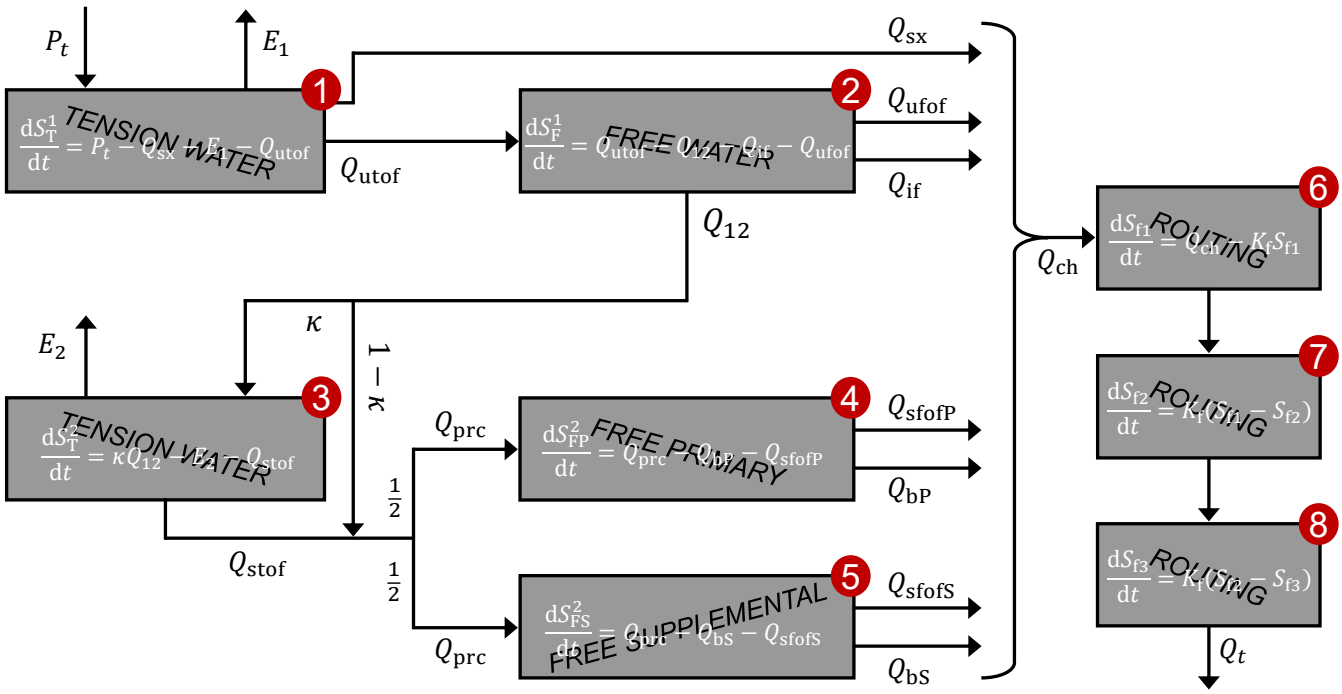

Figure H.3: Schematic illustration of the SAC-SMA model after *Burnash et al.* (1973) and *Clark et al.* (2008). Grey boxes, labeled in red, correspond to fictitious control volumes of the watershed which govern the rainfall-runoff transformation. The model has eight state variables, including the tension water content in the upper soil layer, S_{T}^1, the free water content in the upper soil layer, S_{F}^1, the tension water content in the lower soil layer, S_{T}^2, the free water content in the primary baseflow reservoir, S_{FP}^2, the free water content in the secondary baseflow reservoir, S_{FS}^2, and the water storage, S_{f1}, S_{f2} and S_{f3}, in the cascade of three linear reservoirs used for routing. Arrows portray the fluxes into and out of the compartments, including precipitation, P_t, evaporation from upper soil layer, E_1, overflow of water from tension storage in the upper soil layer, Q_{utof}, surface runoff, Q_{sx}, overflow of water from free storage in the upper soil layer, Q_{ufof}, interflow, Q_{if}, percolation of water from the upper to the lower layer, Q_{12}, evaporation from the lower soil layer, E_2, overflow of water from tension storage in the lower soil layer, Q_{stof}, water flow into primary and supplemental storage, Q_{prc}, overflow of water from primary base flow storage in the lower soil layer, Q_{sofP}, overflow of water from secondary base flow storage in the lower soil layer, Q_{sofS}, base flow from the primary reservoir, Q_{bP}, and baseflow from the secondary reservoir, Q_{bS}. These fluxes are computed as follows, $E_1 = E_{\mathrm{p}}(S_{\mathrm{T}}^1/S_{\mathrm{T,max}}^1)$, $Q_{\mathrm{sx}} = A_{\mathrm{c,max}}(S_{\mathrm{T}}^1/S_{\mathrm{T,max}}^1)P_t$, $Q_{\mathrm{utof}} = (P_t - Q_{\mathrm{sx}})f(S_{\mathrm{T}}^1, S_{\mathrm{T,max}}^1)$, $Q_{12} = Q_0 d_{\mathrm{lz}}(S_{\mathrm{F}}^1/S_{\mathrm{F,max}}^1)$, $Q_{\mathrm{ufof}} = Q_{\mathrm{utof}}f(S_{\mathrm{F}}^1, S_{\mathrm{F,max}}^1)$, $Q_{\mathrm{if}} = k_{\mathrm{i}}(S_{\mathrm{F}}^1/S_{\mathrm{F,max}}^1)$, $E_2 = (E_{\mathrm{p}} - E_1)(S_{\mathrm{T}}^2/S_{\mathrm{T,max}}^2)$, $Q_{\mathrm{stof}} = \kappa Q_{12} f(S_{\mathrm{T}}^2, S_{\mathrm{T,max}}^2)$, $Q_{\mathrm{prc}} = \frac{1}{2}(1-\kappa)Q_{12} + \frac{1}{2}Q_{\mathrm{stof}}$, $Q_{\mathrm{sofP}} = Q_{\mathrm{prc}}f(S_{\mathrm{FP}}^2, S_{\mathrm{FP,max}}^2)$, $Q_{\mathrm{sofS}} = Q_{\mathrm{prc}}f(S_{\mathrm{FS}}^2, S_{\mathrm{FS,max}}^2)$, $Q_{\mathrm{bP}} = \nu_{\mathrm{P}}S_{\mathrm{FP}}^2$, $Q_{\mathrm{bS}} = \nu_{\mathrm{S}}S_{\mathrm{FS}}^2$, where E_{p} signifies the potential evapotranspiration, $Q_0 = \nu_{\mathrm{P}}S_{\mathrm{FP,max}}^2 + \nu_{\mathrm{S}}S_{\mathrm{FS,max}}^2$, $d_{\mathrm{lz}} = 1+\alpha[(S_{\mathrm{T}}^2+S_{\mathrm{FP}}^2+S_{\mathrm{FS}}^2)/(S_{\mathrm{T,max}}^2+S_{\mathrm{FP,max}}^2+S_{\mathrm{FS,max}}^2)]^{\psi}$, the smoothing function $f(x,y) = 1/(1+\exp\{-[x-(y-\epsilon\rho y)]/\rho y\})$, with $\rho = 10^{-2}$ and $\epsilon = 5$, and $S_{\mathrm{F,max}}^1$, $S_{\mathrm{T,max}}^1$, $S_{\mathrm{FP,max}}^2$, $S_{\mathrm{FS,max}}^2$, $S_{\mathrm{T,max}}^2$, α, ψ, k_{i}, κ, ν_{P}, ν_{S} and $A_{\mathrm{c,max}}$ are unknown parameters. The channel inflow, $Q_{\mathrm{ch}} = Q_{\mathrm{sx}}+Q_{\mathrm{ufof}}+Q_{\mathrm{if}}+Q_{\mathrm{QstofP}}+Q_{\mathrm{bP}}+Q_{\mathrm{QstofS}}+Q_{\mathrm{bS}}$, is routed through three linear reservoirs with a common recession constant, K_{f}, to yield the streamflow, $Q_t = K_{\mathrm{f}}S_{\mathrm{f3}}$.

Appendix I: Continuous Ranked Probability Score of $\mathcal{P}_{\mathrm{III}}(\mu, \sigma^2, \rho)$

The continuous ranked probability score (CRPS) is equal to the integral of the quantile scores

$$S_{\mathrm{CRPS}}(P, \omega) = -\int_{-\infty}^{\infty} S^{\tau}\big(F_P(z) - \mathbb{1}\{z \geq \omega\}\big)^2 \mathrm{d}z, \tag{I.1}$$

where $F_P(z)$ denotes the cumulative distribution function of P and the indicator function, $\mathbb{1}\{a\}$, returns 1 if a is true and zero otherwise. Suppose that the distribution forecast P follows a univariate Pearson type III distribution $\mathcal{P}_{\mathrm{III}}(\mu, \sigma^2, \rho)$ with mean μ, variance σ^2 and skewness ρ. If we reparametrize the PIII distribution and define $\xi = \mu - 2\sigma/\rho$, $a = 4/\rho^2$ and $b = \frac{1}{2}\sigma|\rho|$ as location, shape and scale parameters, respectively, then the cumulative distribution function of $P = \mathcal{P}_{\mathrm{III}}(\xi, a, b)$ simplifies to (*Tegos et al.*, 2022)

$$F_P(x, \xi, a, b) = \begin{cases} \dfrac{1}{\Gamma(a)}\gamma\big(a, b^{-1}(x - \xi)\big) & \text{if } \rho > 0 \\ \dfrac{1}{\Gamma(a)}\Gamma\big(a, b^{-1}(\xi - x)\big) & \text{if } \rho < 0 \end{cases} \tag{I.2}$$

where $x, \xi, a, b \in \mathbb{R}$, $a > 0$, $b > 0$ and

$$\Gamma(a, q) = \int_q^{\infty} t^{a-1}\, \exp(-t)\, \mathrm{d}t \tag{I.3}$$

and

$$\gamma(a, q) = \int_0^{q} t^{a-1}\, \exp(-t)\, \mathrm{d}t \tag{I.4}$$

denote the upper and lower incomplete gamma functions, respectively. If $\rho > 0$ then $x \in (\xi, \infty)$ and if $\rho < 0$ then $x \in (-\infty, \xi)$. Next, we derive an analytic expression for the CRPS of $P = \mathcal{P}_{\mathrm{III}}(\xi, a, b)$ and verifying observation $\omega \in \Omega$.

I.1 Analytic expression for $S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi, a, b), \omega\big)_{\rho>0}$: positive skewness

Let us first assume that $\rho > 0$ and, thus, $x \in [\xi, \infty)$. In our derivation we work with $z = x - \xi$ and, thus, $z \in [0, \infty)$. This change of variables simplifies an analytic solution of the CRPS as we will demonstrate next. Equation (I.1) is now equal to

$$S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi, a, b), \omega\big)_{\rho>0} = -\int_0^{z_\omega} \left[\frac{1}{\Gamma(a)}\gamma(a, b^{-1}z)\right]^2 \mathrm{d}z - \int_{z_\omega}^{\infty} \left[\frac{1}{\Gamma(a)}\gamma(a, b^{-1}z) - 1\right]^2 \mathrm{d}z, \tag{I.5}$$

where $z_\omega = \omega - \xi$. We first derive an expression for the left integral using integration by parts

$$\int u \mathrm{d}v = uv - \int v \mathrm{d}u, \tag{I.6}$$

with $u = \gamma^2(a, b^{-1}z)$ and $\mathrm{d}v = 1\mathrm{d}z$. Then

$$\mathrm{d}u = \frac{2}{z}\exp(-b^{-1}z)(b^{-1}z)^a\gamma(a, b^{-1}z) \tag{I.7}$$

and $v = z$ to yield

$$\begin{aligned}-\frac{1}{\Gamma^2(a)}\int \gamma^2(a, b^{-1}z)\mathrm{d}z &= -\frac{1}{\Gamma^2(a)}\Big(z\gamma^2(a, b^{-1}z) - \int z\frac{2}{z}\exp(-b^{-1}z)(b^{-1}z)^a\gamma(a, b^{-1}z)\Big)\\ &= -\frac{1}{\Gamma^2(a)}\Big(z\gamma^2(a, b^{-1}z) - 2\int \exp(-b^{-1}z)(b^{-1}z)^a\gamma(a, b^{-1}z)\Big).\end{aligned} \tag{I.8}$$

The online calculator of Wolfram|Alpha returns a closed-form expression for the indefinite integral

$$\begin{aligned}\int \exp(-b^{-1}z)(b^{-1}z)^a\gamma(a, b^{-1}z)\mathrm{d}z &= \int \exp(-b^{-1}z)(b^{-1}z)^a\big(\Gamma(a) - \Gamma(a, b^{-1}z)\big)\mathrm{d}z\\ &= b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a, b^{-1}z) + \tfrac{1}{2}ab\Gamma^2(a, b^{-1}z)\\ &\quad - 4^{-a}b\Gamma(2a, 2b^{-1}z) - b\Gamma(a)\Gamma(a+1, b^{-1}z) + C.\end{aligned} \tag{I.9}$$

If we substitute Equation (I.9) into Equation (I.8)

$$\begin{aligned}-\frac{1}{\Gamma^2(a)}\int \gamma^2(a, z)\mathrm{d}z = -\frac{1}{\Gamma^2(a)}\Big(&z\gamma^2(a, b^{-1}z) - 2\big(b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a, b^{-1}z) + \tfrac{1}{2}ab\Gamma^2(a, b^{-1}z)\\ &- 4^{-a}b\Gamma(2a, 2b^{-1}z) - b\Gamma(a)\Gamma(a+1, b^{-1}z)\big)\Big) + C,\end{aligned} \tag{I.10}$$

and admit the integral limits

$$\begin{aligned}-\frac{1}{\Gamma^2(a)}\int_0^{z_\omega} \gamma^2(a, b^{-1}z)\mathrm{d}z = -\frac{1}{\Gamma^2(a)}\Big|&z\gamma^2(a, b^{-1}z) - 2\big(b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a, b^{-1}z)\\ &+ \tfrac{1}{2}ab\Gamma^2(a, b^{-1}z) - 4^{-a}b\Gamma(2a, 2b^{-1}z) - b\Gamma(a)\Gamma(a+1, b^{-1}z)\big)\Big|_0^{z_\omega},\end{aligned} \tag{I.11}$$

then we yield the following expression for the left integral of Equation (I.5)

$$\begin{aligned}-\frac{1}{\Gamma^2(a)}\int_0^{z_\omega} \gamma^2(a, b^{-1}z)\mathrm{d}z &= -\frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a, b^{-1}z_\omega) - 2\big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a, b^{-1}z_\omega)\\ &\quad + \tfrac{1}{2}ab\Gamma^2(a, b^{-1}z_\omega) - 4^{-a}b\Gamma(2a, 2b^{-1}z_\omega) - b\Gamma(a)\Gamma(a+1, b^{-1}z_\omega)\big)\Big)\\ &\quad + \frac{1}{\Gamma^2(a)}\Big(-2\big(\tfrac{1}{2}ab\Gamma^2(a) - 4^{-a}b\Gamma(2a) - b\Gamma(a)\Gamma(a+1)\big)\Big)\\ &= -\frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a, b^{-1}z_\omega) - 2\big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a, b^{-1}z_\omega)\\ &\quad + \tfrac{1}{2}ab\Gamma^2(a, b^{-1}z_\omega) - 4^{-a}b\Gamma(2a, 2b^{-1}z_\omega) - b\Gamma(a)\Gamma(a+1, b^{-1}z_\omega)\big)\Big)\\ &\quad + \frac{2}{\Gamma^2(a)}\Big(4^{-a}b\Gamma(2a) + b\Gamma(a)\Gamma(a+1)\Big) - ab.\end{aligned} \tag{I.12}$$

Next, we proceed with the right integral of Equation (I.5)

$$\begin{aligned}-\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\gamma(a, b^{-1}z) - 1\right]^2\mathrm{d}z &= -\int_{z_\omega}^{\infty}\left(\frac{1}{\Gamma^2(a)}\gamma^2(a, b^{-1}z) - \frac{2}{\Gamma(a)}\gamma(a, b^{-1}z) + 1\right)\mathrm{d}z\\ &= -\frac{1}{\Gamma^2(a)}\int_{z_\omega}^{\infty}\gamma^2(a, b^{-1}z)\mathrm{d}z + \frac{2}{\Gamma(a)}\int_{z_\omega}^{\infty}\gamma(a, b^{-1}z)\mathrm{d}z - \int_{z_\omega}^{\infty}\mathrm{d}z.\end{aligned} \tag{I.13}$$

The first of three integrals is equal to Equation (I.10), and the other two integrals yield

$$\frac{2}{\Gamma(a)}\int_{z_\omega}^{\infty}\gamma(a,b^{-1}z)\mathrm{d}z-\int_{z_\omega}^{\infty}\mathrm{d}z=\left|\frac{2}{\Gamma(a)}\Big(z\Gamma(a)-z\Gamma(a,b^{-1}z)+b\Gamma(a+1,b^{-1}z)\Big)-z\right|_{z_\omega}^{\infty}$$
$$=\left|z-\frac{2}{\Gamma(a)}\Big(z\Gamma(a,b^{-1}z)-b\Gamma(a+1,b^{-1}z)\Big)\right|_{z_\omega}^{\infty}. \tag{I.14}$$

Thus, Equation (I.13) becomes

$$\begin{aligned}
-\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)-1\right]^2\mathrm{d}z &= -\frac{1}{\Gamma^2(a)}\Big|z\gamma^2(a,b^{-1}z)-2\Big(b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a,b^{-1}z)\\
&\quad+\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z)-4^{-a}b\Gamma(2a,2b^{-1}z)-b\Gamma(a)\Gamma(a+1,b^{-1}z)\Big)\Big|_{z_\omega}^{\infty}\\
&\quad+\left|z-\frac{2}{\Gamma(a)}\big(z\Gamma(a,b^{-1}z)-b\Gamma(a+1,b^{-1}z)\big)\right|_{z_\omega}^{\infty}\\
&= -\frac{1}{\Gamma^2(a)}\Big|z\gamma^2(a,b^{-1}z)-2\Big(b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a,b^{-1}z)\\
&\quad+\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z)-4^{-a}b\Gamma(2a,2b^{-1}z)-b\Gamma(a)\Gamma(a+1,b^{-1}z)\Big)-z\Gamma^2(a)\\
&\quad+2\Gamma(a)\big(z\Gamma(a,b^{-1}z)-b\Gamma(a+1,b^{-1}z)\big)\Big|_{z_\omega}^{\infty}\\
&= \lim_{z\to\infty}-\frac{1}{\Gamma^2(a)}\Big(\cancel{z\gamma^2(a,b^{-1}z)}^{\Gamma^2(a)}-2\Big(\cancel{b\exp(-b^{-1}z)(b^{-1}z)^a\Gamma(a,b^{-1}z)}^{0}\\
&\quad+\cancel{\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z)}^{0}-4^{-a}\cancel{b\Gamma(2a,2b^{-1}z)}^{0}-b\Gamma(a)\cancel{\Gamma(a+1,b^{-1}z)}^{0}\Big)-z\Gamma^2(a)\\
&\quad+2\Gamma(a)\Big(\cancel{z\Gamma(a,b^{-1}z)}^{0}-\cancel{b\Gamma(a+1,b^{-1}z)}^{0}\Big)\Big)+\frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a,b^{-1}z_\omega)\\
&\quad-2\Big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a,b^{-1}z_\omega)+\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega)\\
&\quad-4^{-a}b\Gamma(2a,2b^{-1}z_\omega)-b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\Big)-z_\omega\Gamma^2(a)\\
&\quad+2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega)-b\Gamma(a+1,b^{-1}z_\omega)\big)\Big)\\
&= \frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a,b^{-1}z_\omega)-2\Big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a,b^{-1}z_\omega)\\
&\quad+\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega)-4^{-a}b\Gamma(2a,2b^{-1}z_\omega)-b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\Big)\\
&\quad-z_\omega\Gamma^2(a)+2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega)-b\Gamma(a+1,b^{-1}z_\omega)\big)\Big).
\end{aligned}\tag{I.15}$$

The sum of Equations (I.12) and (I.15) equals $S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho>0}$, whence we can write

$$\begin{aligned}
S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho>0} &= -\frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a,b^{-1}z_\omega)-2\Big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a,b^{-1}z_\omega)\\
&\quad+\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega)-4^{-a}b\Gamma(2a,2b^{-1}z_\omega)-b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\Big)\Big)\\
&\quad+\frac{2}{\Gamma^2(a)}\Big(4^{-a}b\Gamma(2a)+b\Gamma(a)\Gamma(a+1)\Big)-ab
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{\Gamma^2(a)}\Big(z_\omega\gamma^2(a,b^{-1}z_\omega) - 2\big(b\exp(-b^{-1}z_\omega)(b^{-1}z_\omega)^a\Gamma(a,b^{-1}z_\omega)\\
&+ \tfrac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega) - 4^{-a}b\Gamma(2a,2b^{-1}z_\omega) - b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\big)\\
&- z_\omega\Gamma^2(a) + 2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega) - b\Gamma(a+1,b^{-1}z_\omega)\big)\Big)\\
&= \frac{2}{\Gamma^2(a)}\Big(4^{-a}b\Gamma(2a) + b\Gamma(a)\Gamma(a+1)\Big) - ab - \frac{1}{\Gamma^2(a)}\Big(z_\omega\Gamma^2(a)\\
&- 2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega) - b\Gamma(a+1,b^{-1}z_\omega)\big)\Big)\\
&= 2\frac{4^{-a}b\Gamma(2a)}{\Gamma^2(a)} + 2b\frac{\Gamma(a+1)}{\Gamma(a)} - ab - z_\omega + 2\frac{z_\omega\Gamma(a,b^{-1}z_\omega)}{\Gamma(a)}\\
&- 2b\frac{\Gamma(a+1,b^{-1}z_\omega)}{\Gamma(a)}\\
&= 2\frac{4^{-a}b}{B(a,a)} + ab - z_\omega + 2z_\omega\big(1 - F_{\mathcal{G}}(z_\omega,a,b)\big)\\
&- 2ab\big(1 - F_{\mathcal{G}}(z_\omega,a+1,b)\big), \qquad \text{(I.16)}
\end{aligned}
$$

where $B(u,v) = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ is the beta function of the first kind and $F_{\mathcal{G}}(z,a,b)$ is the CDF of the gamma distribution

$$F_{\mathcal{G}}(z,a,b) = \frac{1}{\Gamma(a)}\gamma(a,b^{-1}z), \qquad \text{(I.17)}$$

with $z > 0$, unitless shape parameter $a > 0$ and scale parameter $b > 0$. We can rearrange Equation (I.16) to yield a final expression for the CRPS of a PIII distribution forecast $P = \mathcal{P}_{\text{III}}(\xi,a,b)$ with positive skewness, $\rho > 0$ and $z_\omega = \omega - \xi$

$$S_{\text{CRPS}}\big(\mathcal{P}_{\text{III}}(\xi,a,b),\omega\big)_{\rho>0} = 2\frac{4^{-a}b}{B(a,a)} + ab\big(2F_{\mathcal{G}}(z_\omega,a+1,b) - 1\big) + z_\omega\big(1 - 2F_{\mathcal{G}}(z_\omega,a,b)\big). \qquad \text{(I.18)}$$

I.2 Analytic expression for $S_{\text{CRPS}}\big(\mathcal{P}_{\text{III}}(\xi,a,b),\omega\big)_{\rho<0}$: negative skewness

Let us now assume that $\rho < 0$ and, thus, $x \in (-\infty,\xi]$. We define $z = \xi - x$ and, thus, $z \in [0,\infty)$. Equation (I.1) becomes

$$S_{\text{CRPS}}\big(\mathcal{P}_{\text{III}}(\xi,a,b),\omega\big)_{\rho<0} = -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\Gamma(a,b^{-1}z)\right]^2\mathrm{d}z - \int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\Gamma(a,b^{-1}z) - 1\right]^2\mathrm{d}z. \qquad \text{(I.19)}$$

We take advantage of the following identity

$$\gamma(a,b^{-1}z) + \Gamma(a,b^{-1}z) = \Gamma(a), \qquad \text{(I.20)}$$

to write Equation (I.19) in another form

$$
\begin{aligned}
S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho<0} &= -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\Big(\Gamma(a)-\gamma(a,b^{-1}z)\Big)\right]^2\mathrm{d}z \\
&\quad -\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\Big(\Gamma(a)-\gamma(a,b^{-1}z)\Big)-1\right]^2\mathrm{d}z \\
&= -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)-1\right]^2\mathrm{d}z-\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)\right]^2\mathrm{d}z, \qquad \text{(I.21)}
\end{aligned}
$$

which is identical to Equation (I.5) but with left and right integrals swapped. This confirms that

$$
S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho<0} = S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho>0}, \qquad \text{(I.22)}
$$

and, thus, we yield the following expression for the CRPS of a PIII distribution forecast $P=\mathcal{P}_{\mathrm{III}}(\xi,a,b)$ with $\rho\neq 0$ and $z_\omega=\xi-\omega$

$$
S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho<0} = 2\frac{4^{-a}b}{B(a,a)}+ab\big(2F_{\mathcal{G}}(z_\omega,a+1,b)-1\big)+z_\omega\big(1-2F_{\mathcal{G}}(z_\omega,a,b)\big), \qquad \text{(I.23)}
$$

I.3 Analytic expression for $S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)$: positive/negative skewness

We can combine the mathematical expressions of $S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho>0}$ and $S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big)_{\rho<0}$ into one general Equation for the CRPS of $P=\mathcal{P}_{\mathrm{III}}(\xi,a,b)$ and verifying observation $\omega\in\Omega$

$$
\begin{aligned}
S_{\mathrm{CRPS}}\big(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega\big) &= 2\frac{4^{-a}b}{B(a,a)}-ab+|\omega-\xi|+2abF_{\mathcal{G}}(|\omega-\xi|,a+1,b) \\
&\quad -2|\omega-\xi|F_{\mathcal{G}}(|\omega-\xi|,a,b), \qquad \text{(I.24)}
\end{aligned}
$$

where $z_\omega=|\omega-\xi|$. The first term in the above expression is equal to the first term, $\frac{1}{2}\mathbb{E}_P\big[|y-y^*|\big]$, of Equation (63) and the sum of all the remaining terms of Equation (I.24) equals $-\mathbb{E}_P\big[|y-\omega|\big]$.

We would be remiss not to address two well-known limiting cases of the PIII distribution. For $\xi=0$, thus, $\mu=2\sigma/\rho$, the PIII distribution $\mathcal{P}_{\mathrm{III}}(\xi,a,b)$ reduces to the gamma distribution $\mathcal{G}(a,b)$ and Equation (I.24) simplifies to

$$
S_{\mathrm{CRPS}}\big(\mathcal{G}(a,b),\omega\big) = 2\frac{4^{-a}b}{B(a,a)}-ab+|\omega|+2abF_{\mathcal{G}}(|\omega|,a+1,b)-2|\omega|F_{\mathcal{G}}(|\omega|,a,b). \qquad \text{(I.25)}
$$

This expression for the CRPS of $P=\mathcal{G}(a,b)$ matches the numerical estimates of the CRPS shown in Figure 10 using the solid yellow line. If $\rho=0$, then the PIII distribution $\mathcal{P}_{\mathrm{III}}(\mu,\sigma^2,0)$ simplifies to a normal distribution $\mathcal{N}(\mu,\sigma^2)$ and the CRPS can be computed using Equation (G.12). This concludes the derivation.

Appendix J: Continuous Ranked Probability Score of $\mathcal{GEV}(\mu, \sigma^2, \xi)$

We revisit the quantile form of the continuous ranked probability score (CRPS)

$$S_{\text{CRPS}}(P, \omega) = \omega\big(1 - 2F_P(\omega)\big) + 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau. \tag{J.1}$$

Suppose that the distribution forecast P follows a generalized extreme value distribution $\mathcal{GEV}(\mu, \sigma^2, \xi)$ with mean $\mu \in \mathbb{R}$, variance $\sigma^2 > 0$ and shape parameter $\xi \in \mathbb{R}$. The CDF of the GEV distribution equals

$$F_{\mathcal{GEV}}(x, \mu, \sigma^2, \xi) = \begin{cases} \exp\Big[-\exp\Big(-\dfrac{\xi}{\sigma}(x-\mu)\Big)\Big] & \text{if } \xi = 0 \\ \exp\Big[-\Big(1 + \dfrac{\xi}{\sigma}(x-\mu)\Big)^{-1/\xi}\Big] & \text{if } \xi < 0 \text{ and } x < -\frac{1}{\xi} \\ 1 & \text{if } \xi < 0 \text{ and } x \geq -\frac{1}{\xi} \\ 0 & \text{if } \xi > 0 \text{ and } x \leq -\frac{1}{\xi} \\ \exp\Big[-\Big(1 + \dfrac{\xi}{\sigma}(x-\mu)\Big)^{-1/\xi}\Big] & \text{if } \xi > 0 \text{ and } x > -\frac{1}{\xi}, \end{cases} \tag{J.2}$$

and its quantile function has the following explicit expression

$$F^{-1}_{\mathcal{GEV}}(x, \mu, \sigma^2, \xi) = \begin{cases} \mu - \sigma\log_e\big(-\log_e(\tau)\big) & \text{if } \xi = 0 \text{ and } \tau \in (0,1) \\ \mu + \dfrac{\sigma}{\xi}\Big(\big(-\log_e(\tau)\big)^{-\xi} - 1\Big) & \text{if } \xi \neq 0\ , \end{cases} \tag{J.3}$$

where $\tau \in [0, 1)$ if $\xi > 0$ and $\tau \in (0, 1]$ if $\xi < 0$. Next, we derive an analytic expression for the CRPS of $P = \mathcal{GEV}(\mu, \sigma^2, \xi)$ and verifying observation $\omega \in \Omega$. For $\xi \geq 1$ the CRPS of $P = \mathcal{GEV}(\mu, \sigma^2, \xi)$ is undefined. Thus, we restrict our attention to $\xi < 1$.

J.1 Analytic expression for $S_{\text{CRPS}}\big(\mathcal{GEV}(\mu, \sigma^2, \xi), \omega\big)_{\xi<1,\, \xi\neq 0}$

We first consider the case of a non-zero shape parameter, $\xi \neq 0$, of the GEV distribution. The indefinite form of the first integral of the CRPS in Equation (J.1) can be expressed analytically using integration by parts

$$\begin{aligned} \int \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \frac{\tau}{\xi}\Big(\mu\xi\tau + \sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) + C \\ &\quad - \int \frac{1}{\xi}\Big(\mu\xi\tau + \sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big)\,\mathrm{d}\tau \\ &= \frac{\tau}{\xi}\Big(\mu\xi\tau + \sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) \\ &\quad - \frac{1}{2\xi}\Big[\tau\Big(\mu\tau\xi + 2\sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) - 2^{\xi}\sigma\Gamma\big(1-\xi, -2\log_e(\tau)\big)\Big] + C, \end{aligned} \tag{J.4}$$

where

$$\Gamma(a, q) = \int_q^\infty t^{a-1} \exp(-t)\,\mathrm{d}t \tag{J.5}$$

denotes the upper incomplete gamma function. If we now admit the lower and upper limits of the quantiles, we yield

$$\begin{aligned}\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \Bigg[\frac{\tau}{\xi}\Big(\mu\xi\tau + \sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) \\ &\quad - \frac{1}{2\xi}\Big[\tau\Big(\mu\tau\xi + 2\sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) - 2^\xi\sigma\Gamma\big(1-\xi, -2\log_e(\tau)\big)\Big] + C\Bigg]_0^1 \\ &= \left(\mu + \frac{\sigma}{\xi}\Gamma(1-\xi) - \frac{\sigma}{\xi} - \frac{\mu}{2} - \frac{\sigma}{\xi}\Gamma(1-\xi) + \frac{\sigma}{2\xi} + \frac{2^\xi\sigma\Gamma(1-\xi)}{2\xi}\right) - \big(0\big) \\ &= \frac{\mu}{2} + \frac{\sigma}{2\xi}\big(2^\xi\Gamma(1-\xi) - 1\big).\end{aligned} \tag{J.6}$$

The second or right integral of the CRPS in Equation (J.1) results in

$$\begin{aligned}\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau &= \Bigg[\frac{1}{\xi}\Big(\mu\xi\tau + \sigma\Gamma\big(1-\xi, -\log_e(\tau)\big) - \sigma\tau\Big) + C\Bigg]_{F_P(\omega)}^1 \\ &= \left(\mu + \frac{\sigma}{\xi}\Gamma(1-\xi) - \frac{\sigma}{\xi}\right) - \Bigg(\mu F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) \\ &\quad + \frac{\sigma}{\xi}\Gamma\Big(1-\xi, -\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\big)\Big) - \frac{\sigma}{\xi}F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\Bigg) \\ &= \mu - \mu F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) + \frac{\sigma}{\xi}\Big[\Gamma(1-\xi) + F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) \\ &\quad - \Gamma\Big(1-\xi, -\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\big)\Big) - 1\Big].\end{aligned} \tag{J.7}$$

Next, we can insert the analytic expressions of the two integrals into Equation (J.1)

$$\begin{aligned}S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\omega, \mu, \sigma^2, \xi), \omega\big)_{\xi<1,\,\xi\neq0} &= \omega\big(1 - 2F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\big) + \mu + \frac{\sigma}{\xi}\big(2^\xi\Gamma(1-\xi) - 1\big) \\ &\quad - 2\mu + 2\mu F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) - \frac{2\sigma}{\xi}\Big[\Gamma(1-\xi) + F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) \\ &\quad - \Gamma\Big(1-\xi, -\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\big)\Big) - 1\Big].\end{aligned}$$

The above expression may be rearranged and rewritten to yield

$$\begin{aligned}S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\omega, \mu, \sigma^2, \xi), \omega\big)_{\xi<1,\,\xi\neq0} &= (\mu - \omega)\big(2F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi) - 1\big) + \frac{\sigma}{\xi}\Big(1 - (2 - 2^\xi)\Gamma(1-\xi)\Big) \\ &\quad + \frac{2\sigma}{\xi}\Big[\Gamma\Big(1-\xi, -\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\big)\Big) \\ &\quad - F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, \xi)\Big].\end{aligned} \tag{J.8}$$

J.2 Analytic expression for $S_{\text{CRPS}}\big(\mathcal{GEV}(\mu, \sigma^2, 0), \omega\big)$

Next, we consider $\xi = 0$. We use integration by parts

$$\int u \mathrm{d}v = uv - \int v \mathrm{d}u, \tag{J.9}$$

and yield the following expression for the first of two integrals of Equation (J.1)

$$\begin{aligned}
\int \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \tfrac{1}{2}\mu\tau^2 - \tfrac{1}{2}\sigma\tau^2 \log_e\big(-\log_e(\tau)\big) + C - \int -\frac{\sigma\tau^2}{2\tau\log_e(\tau)}\,\mathrm{d}\tau \\
&= \tfrac{1}{2}\mu\tau^2 - \tfrac{1}{2}\sigma\tau^2 \log_e\big(-\log_e(\tau)\big) + C + \tfrac{1}{2}\sigma\int \frac{\tau}{\log_e(\tau)}\,\mathrm{d}\tau \\
&= \tfrac{1}{2}\mu\tau^2 - \tfrac{1}{2}\sigma\tau^2 \log_e\big(-\log_e(\tau)\big) + \tfrac{1}{2}\sigma\,\mathrm{Ei}\big(2\log_e(\tau)\big) + C,
\end{aligned} \tag{J.10}$$

where $\mathrm{Ei}(x)$ is the exponential integral function. If we admit the integral limits, we yield

$$\begin{aligned}
\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \Big[\tfrac{1}{2}\mu\tau^2 - \tfrac{1}{2}\sigma\tau^2 \log_e\big(-\log_e(\tau)\big) + \tfrac{1}{2}\sigma\,\mathrm{Ei}\big(2\log_e(\tau)\big) + C\Big]_0^1 \\
&= \tfrac{1}{2}\mu - \tfrac{1}{2}\sigma \lim_{\tau\to 1^-}\Big(\tau^2\log_e\big(-\log_e(\tau)\big) - \mathrm{Ei}\big(2\log_e(\tau)\big)\Big) \\
&= \tfrac{1}{2}\mu + \tfrac{1}{2}\sigma\big(\gamma_{\mathrm{c}} + \log_e(2)\big),
\end{aligned} \tag{J.11}$$

where $\gamma_{\mathrm{c}} = 0.57721566\ldots$ is the Euler-Mascheroni constant. The right integral of Equation (J.1) becomes

$$\begin{aligned}
\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau &= \Big[\mu\tau + \sigma\,\mathrm{Li}(\tau) - \sigma\tau\log_e\big(-\log_e(\tau)\big)\Big]_{F_P(\omega)}^1 \\
&= \Big(\mu + \lim_{\tau\to 1^-}\Big(\sigma\,\mathrm{Li}(\tau) - \sigma\tau\log_e\big(-\log_e(\tau)\big)\Big)\Big) \\
&\quad - \Big(\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) + \sigma\,\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) \\
&\quad - \sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\Big)\Big) \\
&= \mu + \gamma_{\mathrm{c}}\sigma - \mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) - \sigma\,\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) \\
&\quad + \sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\Big).
\end{aligned} \tag{J.12}$$

where $\mathrm{Li}(x)$ signifies the logarithmic integral function. If we substitute Equations (J.11) and (J.12) into Equation (J.1), we yield the following expression for the CRPS of $P = \mathcal{GEV}(\mu, \sigma^2, 0)$

$$\begin{aligned}
S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\omega,\mu,\sigma^2,0),\omega\big) &= \omega\big(1 - 2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) + \mu + \sigma\big(\gamma_{\mathrm{c}} + \log_e(2)\big) - 2\mu \\
&\quad - 2\gamma_{\mathrm{c}}\sigma + 2\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) + 2\sigma\,\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) \\
&\quad - 2\sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\Big) \\
&= \omega\big(1 - 2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) - \mu - \gamma_{\mathrm{c}}\sigma + \sigma\log_e(2)
\end{aligned}$$

$$
\begin{aligned}
&\quad + 2\mu F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0) + 2\sigma \operatorname{Li}\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big) \\
&\quad - 2\sigma F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0) \log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big)\Big) \\
&= \omega - \mu - \gamma_{\mathrm{c}}\sigma + \sigma \log_e(2) + 2\sigma \operatorname{Li}\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big) \\
&\quad - 2\Big[\omega - \mu + \sigma \log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big)\Big)\Big] F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0). \qquad \text{(J.13)}
\end{aligned}
$$

Per the quantile function in Equation (J.3), we find that

$$
-\mu + \sigma \log_e\Big(-\log_e\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big)\Big) = -\omega, \qquad \text{(J.14)}
$$

and, thus, the last term amounts to zero. This results in the following expression for the CRPS of a GEV distribution forecast $P = \mathcal{GEV}(\mu, \sigma^2, 0)$ with shape parameter $\xi = 0$

$$
S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\omega, \mu, \sigma^2, 0), \omega\big) = \omega - \mu + \sigma \log_e(2) + 2\sigma \operatorname{Li}\big(F_{\mathcal{GEV}}(\omega, \mu, \sigma^2, 0)\big) - \gamma_{\mathrm{c}}\sigma. \qquad \text{(J.15)}
$$

This concludes the derivation.