Title

Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals

Permalink

https://escholarship.org/uc/item/6pj9k1bh

Journal

Water Resources Research, 60(6)

ISSN

0043-1397

Author

Vrugt, Jasper A

Publication Date

2024-06-01

DOI

10.1029/2023wr036710

Water Resources Research®

RESEARCH ARTICLE
10.1029/2023WR036710

Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals

Jasper A. Vrugt[1]

[1]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA

Key Points:

- Scoring rules of hydrograph functionals provide an information-theoretically principled means for watershed model training, evaluation, and diagnostics
- We present *strictly proper* (divergence) scores for flood frequency analysis, flow duration, and recession curves
- Propriety and elicitability offer useful working paradigms for metric development of hydrograph functionals

Correspondence to:

J. A. Vrugt,
jasper@uci.edu

Citation:

Vrugt, J. A. (2024). Distribution-based model evaluation and diagnostics: Elicitability, propriety, and scoring rules for hydrograph functionals. *Water Resources Research*, *60*, e2023WR036710. https://doi.org/10.1029/2023WR036710

Received 13 NOV 2023
Accepted 15 MAR 2024

Abstract

Distribution forecasts P over future quantities or events are routinely made in hydrology but usually traded for a (likelihood-weighted) mean or median prediction to accommodate error measures or *scoring functions* such as the mean absolute error or mean squared error. Case in point is the so-called KG efficiency (KGE) of Gupta et al. (2009, https://doi.org/10.1016/j.jhydrol.2009.08.003) and improvements thereof (Lamontagne et al., 2020, https://doi.org/10.1029/2020wr027101), which have rapidly gained popularity among hydrologists as alternative *scoring functions* to the commonly used Nash and Sutcliffe (1970, https://doi.org/10.1016/0022-1694(70)90255-6) efficiency, but are equally exclusive in how they quantify model performance using only single-valued output of the quantities of interest. This point-valued mapping necessarily implies a loss of information about model performance. This paper advocates the use of probabilistic watershed model training, evaluation and diagnostics. Distribution evaluation opens a mature literature on *scoring rules* whose strong statistical underpinning provides, as we will demonstrate, the theory, context and guidelines necessary for the development of robust information-theoretically principled metrics for watershed signatures. These so-called hydrograph functionals are scalar-valued mappings of major behavioral watershed functions embodied in a *strictly proper* scoring rule. We discuss past developments that led to the current state-of-the-art of distribution evaluation in hydrology and review scoring rules for dichotomous and categorical events, quantiles (intervals) and density forecasts. We are particularly concerned with elicitable functionals and scoring rule propriety, discuss the decomposition of scoring rules into a sharpness, reliability and entropy term and present diagnostically appealing *strictly proper* divergence scores of hydrograph functionals for flood frequency analysis, flow duration and recession curves. The usefulness and power of distribution-based model evaluation and diagnostics by means of scoring rules is demonstrated on simple illustrative problems and discharge distributions simulated with watershed models using random sampling and Bayesian model averaging. The presented theory (a) enables a more complete evaluation of distribution forecasts, (b) offers a statistically principled means for watershed model training, evaluation, diagnostics and selection using hydrograph functionals and/or extreme events and (c) provides a universal framework for metric development of watershed signatures, promoting metric standardization and reproducibility.

Plain Language Summary

The past decades have witnessed an unbridled growth in goodness-of-fit metrics of hydrologic models. These metrics may satisfy the needs of hydrologists but lack conforming theory and principles. This state of affairs (a) elicits improper model training and evaluation, (b) provokes and supports misguided inferences, (c) impedes statistically-principled uncertainty quantification, metric standardization and development of universal model benchmarks and (d) obfuscates determination of whether the model has finished learning. What is more, most hydrologic model evaluation metrics in use today are rather exclusive in how they quantify model performance using only single-valued simulated output of the quantities of interest. Predictive distributions derived from (quasi)-Bayesian methods or ensembles are usually traded for a (likelihood-weighted) mean or median prediction to accommodate error measures (scoring functions) such as the mean absolute error. This implies a large loss of information. This paper develops a distribution-based approach to hydrologic model evaluation and diagnostics. Distribution evaluation opens the necessary theory and guidelines for development of robust information-theoretically principled metrics of watershed signatures. These so-called hydrograph functionals are scalar-valued mappings of major behavioral watershed functions embodied in a *strictly proper* scoring rule. The hydrograph functionals offer a statistically principled means for hydrologic model evaluation, diagnostics and selection.

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 behavior 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 & Blöschl, 2001). Scientists use error metrics to quantify the goodness-of-fit or accuracy of model predictions (Krause et al., 2005). The need for error metrics and error quantification is generally understood, but the methods and metrics used in practice vary widely (Jackson et al., 2019; Reich et al., 2016). We follow Gneiting (2011) and use the terminology of a scoring function for an error measure on a general sample space Ω.

Definition 1. *A scoring function is any real-valued function $s : \Omega \times \Omega \rightarrow \mathbb{R}$ 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 pervasive squared error, $s_{\mathrm{SE}}(y, \omega) = (\omega - y)^2$, absolute error $s_{\mathrm{AE}}(y, \omega) = |\omega - y|$, absolute percentage error, $s_{\mathrm{APE}}(y, \omega) = |(\omega - y)/\omega|$, and relative error, $s_{\mathrm{RE}}(y, \omega) = |(\omega - y)/y|$, measure the performance of a point forecast y, where $|\cdot|$ is the absolute value operator and $\epsilon = \omega - y$ is the so-called residual. Smaller values of the scoring functions are preferred. If the scoring function is the squared error, $s_{\mathrm{SE}}(y, \omega)$, the optimal point forecast is the mean of the predictive distribution. In the case of the absolute error, $s_{\mathrm{AE}}(y, \omega)$ the Bayes rule is any median of the predictive distribution (Gneiting, 2011). But in simulation mode, hydrologic models generate a time series of forecasts rather than a scalar prediction. For a sequence of observation-forecast pairs (ω_t, y_t); $t = (1, \ldots, n)$ we resort to the average score (Ehm et al., 2016), $\bar{s}_{\mathrm{XX}}(\mathbf{y},\boldsymbol{\omega}) = \frac{1}{n}\sum_{t=1}^{n} s_{\mathrm{XX}}(y_t,\omega_t)$, a well-known example of which is the Nash and Sutcliffe (1970) efficiency, $\bar{s}_{\mathrm{NSE}}(\mathbf{y},\boldsymbol{\omega}) = 1 - \sum_{t=1}^{n}(\omega_t - y_t)^2/(\omega_t - m_\omega)^2$, where m_ω is the sample mean of the verifying data $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^\top$ and $\mathbf{y} = (y_1,\ldots,y_n)^\top$. This extension to a data record adds significant complexity to model performance evaluation (Jackson et al., 2019; Reich et al., 2016). For the purpose of this discussion, we classify the research on model evaluation into two groups including (a) theory-based methods and (b) empirical methods. Theory-based methods quantify model performance using scoring functions of the residuals of simulated watershed behavior. Empirical methods exploit hydrological context and theory and quantify model performance by comparing observed and simulated values of major watershed functions.

Theory-based methods are rooted in regression analysis and quantify model performance using classical residual-based measures of the goodness-of-fit of observed and simulated watershed behavior. This includes the use of (a) formal loss and likelihood functions and/or summary metrics derived from rigorous application of first principles to the assumed statistical properties of the residuals within the context of weighted and/or generalized least squares (Kavetski et al., 2006a, 2006b; Stedinger & Tasker, 1985; Tasker, 1980), maximum likelihood and Bayesian estimation (Ammann et al., 2019; Bates & Campbell, 2001; Kuczera, 1983; Scharnagl et al., 2015; Schoups & Vrugt, 2010; Sorooshian & Dracup, 1980; Vrugt et al., 2022) approximate Bayesian computation (Nott et al., 2012; Sadegh & Vrugt, 2013; Vrugt & Sadegh, 2013), information theoretical principled approaches (Lu et al., 2011; Neuman, 2003; Pachepsky et al., 2016; Schöniger et al., 2014; Volpi et al., 2017; Weijs, Schoups, et al., 2010; Ye et al., 2008), and sensitivity analysis (Gao et al., 2023), (b) pseudo loss and quasi-likelihood functions within the context of model training (Gupta et al., 2009; Knoben, Freer, Fowler, et al., 2019; Lamontagne et al., 2020; Nash & Sutcliffe, 1970; Pool et al., 2018; Schwemmle et al., 2021), informal Bayesian approaches (Beven & Binley, 1992; Beven & Freer, 2001; Freer et al., 1996) and multi-criteria model calibration (Boyle et al., 2000; Gupta et al., 1998), and (c) tolerable ranges of the simulated output within the context of limits of acceptability (Beven, 2006; Vrugt & Beven, 2018), regional sensitivity analysis (Spear et al., 2020; Spear & Hornberger, 1980), dynamic identifiability analysis (Wagener et al., 2003) and the parameter identification method based on the localization of information (Vrugt et al., 2002). These residual-based approaches have proven particularly useful for hydrologic model training and uncertainty quantification but do not guarantee an accurate representation of the major behavioral functions of a watershed (Vrugt & Sadegh, 2013; Yilmaz et al., 2008). Also, regression-based performance metrics fail to relate the information in the data to modeled processes in a diagnostic manner for the purpose of learning and scientific discovery (Gupta et al., 2008).

Empirical methods are much more geared toward hypothesis testing and scientific learning by exploiting hydrological context and theory in the evaluation of watershed model performance (Yilmaz et al., 2008). These methods do not work with the residuals of simulated and measured watershed behavior but rather quantify model performance by evaluating major behavioral functions of watershed behavior. This includes measures of watershed behavior such as the runoff ratio (Sawicz et al., 2011), baseflow index (Eckhardt, 2005), streamflow elasticity (Sankarasubramanian et al., 2001), flashiness index (Baker et al., 2004), rising limb density (Shamir et al., 2005), flow duration curve (FDC) (McMillan et al., 2017; Sadegh et al., 2016; Searcy, 1959; Vogel & Fennessey, 1994; Yadav et al., 2007), rate of runoff recession (Pool et al., 2017), mean fall and rise rates (Olden & Poff, 2003), streamflow variability (Jowett & Duncan, 1990), skewness of daily flows (Clausen & Biggs, 2000), proportion of zero flows (Olden & Poff, 2003) and frequency of low (Olden & Poff, 2003) and high (Clausen & Biggs, 2000) flows. Shamir et al. (2005) laid the foundation of this empirical approach in their work on hydrograph indices and this approach has matured further into what is now known as model diagnostics (Gupta et al., 2008). Watershed signatures are not commonly used for model training as it is not clear how to turn the FDC and numeral hydrograph descriptors such as the runoff ratio, baseflow index and streamflow elasticity in a proper loss function (Sadegh et al., 2015; Vrugt & Sadegh, 2013). Note that one may discern a third group of model evaluation methods which are theory-based but have a diagnostic intent. In this hybrid group are 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).

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 (Resin, 2023). Formally speaking, a functional is a mapping $T : \mathcal{P} \rightarrow \mathbb{R}$, where $\mathcal{P}$ is a collection of functions. Usually, functionals are single valued such as the mean $\mu_x = T_{\text{mean}}(P) = \int x\text{d}P(x)$, variance, $\sigma_x^2 = T_{\text{var}}(P) = \int (x - \mu_x)^2 \text{d}P(x)$, and median, $T_{\text{med}}(P) = P^{-1}\left(\frac{1}{2}\right)$, where $P \in \mathcal{P}$ is the cumulative distribution function (CDF) of quantity x, $\text{d}P(x) = p(x)\text{d}x$ and $p(x)$ is its probability density function (PDF). In this context, we coin the term *hydrograph functional* for a scalar-valued mapping T of a major behavioral function of the watershed. Thus, hydrograph functionals are real numbers, which quantify the most important characteristics of the catchment's response to rainfall. Functionals that incentivize a truthful description are called *elicitable* in decision-theory (Fissler et al., 2021; Gneiting, 2011; Roccioletti, 2015). This term was coined by Lambert et al. (2008), but its roots go back decades to the work of Savage (1971) and Osband (1985). The mean (expectation), median and quantiles of a distribution are elicitable but its mode and variance are not (Brehmer & Strokorb, 2019; Gneiting, 2011; Heinrich, 2014). Discrepancies between measured and simulated functionals are symptoms of model malfunctioning, and if able to relate functionals to a specific process, will provide guidance on model improvement (Gupta et al., 2008; Westerberg et al., 2011; Yilmaz et al., 2008).

While the diagnostic 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. Case in point is the KG efficiency or KG efficiency (KGE) of Gupta et al. (2009) and refinements thereof (Lamontagne et al., 2020), which have quickly gained popularity with hydrologists as alternative scoring functions for the Nash and Sutcliffe (1970) efficiency, $\overline{s}_{\text{NSE}}(\mathbf{y}, \boldsymbol{\omega})$, but are equally exclusive in how they quantify model performance using only single-valued output $\mathbf{y} = (y_1, \ldots, y_n)^\top$ of the quantity of interest at $t = 1, \ldots, n$. Also, the NSE and KGE are inconsistent scoring functions as the model efficiency is not an elicitable data functional. The NSE directs the model to track the data as closely and consistently as possible. This directive is ambiguous and does not help determine the model's success in learning watershed behavior. The KGE is much more explicit about what it expects the model to do. It should match two data functionals (mean and variance) and maximize the correlation coefficient of measured and simulated data. Scoring functions should be (strictly) consistent for hydrograph functionals, in the sense that they optimize the expected score when following the directive. Thus, the scoring function and forecasting (simulation) task must be carefully matched (Gneiting, 2011).

More than a decade ago, Guttorp (2011) formulated a vision of how climate models should be evaluated against data (P. 820), "*…Climate models are di?cult 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 con?dence intervals.*" This change from point to distributional evaluation is supported by information-theoretic arguments (Weijs, Schoups, et al., 2010, p. 2545), "*…models should preferably be explicitly probabilistic and calibrated to maximize the information they provide.*", computer hardware and software advances and inspires a paradigm change in hydrologic model evaluation. Distribution forecasts express diversity in the form of a probability distribution over future quantities or events (Dawid, 1984) and contain information about model behavior, robustness, sensitivity and uncertainty that is not available in single-valued model output. Such forecasts are routinely made in epidemiology (Alkema et al., 2007), finance (Duffie & Pan, 1997; Groen et al., 2013), macroeconomics (Garratt et al., 2003; Granger, 2005), medicine (Hood et al., 2004), meteorology (Tracton & Kalnay, 1993) and hydrology (Thielen et al., 2008; Welles et al., 2007) and support influenza (Cheng et al., 2020), stock-market Nti et al. (2020), weather and climate (Gneiting et al., 2005; Palmer, 2002), seismic hazard (T. Jordan et al., 2011) and flood risk (Cloke & Pappenberger, 2009; Krzysztofowicz, 2001) prediction. Yet, predictive distributions P of (quasi)-Bayesian methods (Beven & Binley, 1992; Kavetski et al., 2006a; Kuczera & Parent, 1998; Schoups & Vrugt, 2010; Vrugt, 2016; Vrugt et al., 2003, 2022) are usually traded for some set-valued mapping $P \rightarrow T(P) \subseteq \Omega$ in hydrologic model evaluation with the (likelihood-weighted) mean or median prediction of P as key examples. This point-valued mapping is usually given in by a lack of knowledge of how to properly evaluate simulation distributions against data but necessarily implies a loss of information about model performance. This information loss is minimal when the modeled distributions are Gaussian and scoring functions such as the mean squared error, $\bar{s}_{\mathrm{SE}}(\mathbf{y},\boldsymbol{\omega})$, and mean absolute error, $\bar{s}_{\mathrm{AE}}(\mathbf{y},\boldsymbol{\omega})$, will do their job. The information loss is more colossal when simulated distributions deviate from normality in the face of epistemic and input data errors. 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.

Forecast verification is an active field of research in the climate, atmospheric and ocean sciences and is concerned with evaluating the predictive power of prognostic model forecasts (Jolliffe & Stephenson, 2011; Murphy & Katz, 1985; Storch & Zwiers, 1999). 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 & Raftery, 2007).

Definition 2. *A scoring rule is any extended real-valued function* $S : \mathcal{P} \times \Omega \rightarrow \overline{\mathbb{R}} \equiv [-\infty, \infty]$ *such that* $S(P, \omega)$ *is* $\mathcal{P}$*-quasi-integrable for all* $P \in \mathcal{P}$ *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 & 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; Fenicia et al., 2011; Schoups et al., 2010).

Weijs, Schoups, et al. (2010) 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 (Girons Lopez et al., 2021; Laio & Tamea, 2007; Vrugt et al., 2006). 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 is concerned with the basic question of how we should evaluate predictive (simulation) distributions of observed quantities. This is of crucial importance in yielding an accurate description of the probability distribution of predictands conditioned by deterministic model output, for example, using the Bluecat method of Koutsoyiannis and Montanari (2022). 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 scoring rule propriety, 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 point-valued 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 knowledge of the distribution of each verifying observation. Sections 4–7 discuss the more common and realistic situation in which we do not have knowledge of the probability distribution $Q \in \mathcal{P}$ that materializes with the event $\omega \in \Omega$. Section 4 illustrates the incompleteness of common metrics used in the hydrologic literature for evaluating distribution forecasts. This is followed by Section 5, which discusses scoring rules for distribution forecasts of categorical (discrete) variables and their extension in Section 6 to continuous variables. In this section we present diagnostically appealing divergence scores for hydrograph flow duration and recession curves. Section 7 revisits the decomposition of *strictly proper* scoring rules into an uncertainty, sharpness (resolution) and reliability term. The different sections are permeated with simple illustrative examples and case studies of the rainfall-discharge transformation. The penultimate Section 8 presents a brief outlook on the use of scoring rules and functions for diagnostic model evaluation, sensitivity analysis, Bayesian model selection, the prediction of extreme events and flood frequency analysis. To this end, we present closed-form expressions for the CRPS and logarithmic score (LS) for the Pearson type III distribution of annual maxima discharges. Section 9 concludes this paper with a summary of our main findings. To the extent possible, mathematical derivations and computational details have been deferred to Appendices. Those discouraged by our statistical treatment of this topic are directed to the case studies and the `ScoringRules` toolbox in MATLAB.

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 helps quantify 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 of probability theory will suffice to revise these subjective probabilities (=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 materialized events. Before we proceed any further, we first expose our treatment of probability and clarify the notation used.

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 $\mathcal{P}$ be a convex class of probability measures on (Ω, Σ). A *probabilistic forecast* is a set function $P \in \mathcal{P}$ 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, $P(\Omega) = 1$. Similarly, the *true forecast* $Q \in \mathcal{P}$ assigns probabilities $Q \in [0, 1]$ to all events $\Sigma \subseteq \Omega$, with unit sum of all probabilities, $Q(\Omega) = 1$. This measure theoretic treatment of probability allows us to simultaneously treat discrete and continuous probability distributions.

In the past two decades, probabilistic forecasting methods have found their way into hydrological practice. The topic of forecast verification has not yet received a systematic treatment in the hydrologic literature (but with some exceptions, for example Laio and Tamea (2007)) and, as a result, relevant methods are often used haphazardly. Throughout this paper, x and y are random variables, say, next day's peak discharge and $\omega \in \Omega$ is the measured (materialized) value. The forecaster's task is to quote a distribution $P \in \mathcal{P}$ which characterizes the uncertainty of x or y. 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. We follow McCarthy (1956) and assume that the forecaster cannot control the events predicted beyond experimentation, data collection and modeling. To demonstrate the requirements of a meaningful evaluation of distribution forecasts we have to be comprehensive in our statistical treatment of this topic. We spare the main text from lengthy mathematical derivations and defer such technicalities to appendices. The same holds for the description of models, data and computational procedures.

An adequate mathematical notation is crucially important as it shapes how we think, facilitates understanding and communication and streamlines problem solving (Holton, 2013). We use a lowercase italic font (a) for scalars, a lowercase bold font ($\mathbf{a}$) for vectors and an uppercase bold font ($\mathbf{A}$) for matrices. 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 has a normal distribution with 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 PDF with a lowercase f and a 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. The vertical bar "|" denotes conditional expectation. Thus, $p(x|\boldsymbol{\omega})$ is the conditional PDF of x given data $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^\top$ 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, 1959; Kullback & Leibler, 1951). Divergence is a physical measure of information gain in communication theory. For the time being, we adapt the notation 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

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \int_{x\in\Omega}\Big(Q(x)\log_{\mathfrak{b}}\big(Q(x)\big) - Q(x)\log_{\mathfrak{b}}\big(P(x)\big)\Big)\mathrm{d}x \\ &= \underbrace{\int_{x\in\Omega} Q(x)\log_{\mathfrak{b}}\big(Q(x)\big)\mathrm{d}x}_{-\mathbb{H}(Q)} - \underbrace{\int_{x\in\Omega} Q(x)\log_{\mathfrak{b}}\big(P(x)\big)\mathrm{d}x}_{-\mathbb{H}(Q,P)} = \mathbb{H}(Q,P) - \mathbb{H}(Q). \end{aligned} \tag{2}$$

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, 1948b). 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 given in Appendix A. If $P = Q$ and our distribution *forecast* P matches exactly the *true* distribution Q then $\mathbb{H}(Q,P) = \mathbb{H}(Q)$ and $d_{\mathrm{KL}}(Q, P)$ is zero. Thus, $d_{\mathrm{KL}}(Q, P) = 0$ if and only if $P = Q$. Hence, 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 $\mathcal{P}$ and follow a known statistical distribution then it may be possible to derive analytic expressions for the relative entropy $d_{\mathrm{KL}}(Q, P)$ (see e.g. Bouhlel & Dziri, 2019). Appendix B presents such derivations of the relative entropy for cases when the *probabilistic forecast* and *true distribution* are univariate normal, triangular and uniform, respectively. We also consider the case of a multivariate normal *probabilistic forecast* $P = \mathcal{N}_\zeta(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ and *true distribution* $Q = \mathcal{N}_\zeta(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ with means, $\boldsymbol{\mu}_Q, \boldsymbol{\mu}_P \in \mathbb{R}^{\zeta \times 1}$, and non-singular $\zeta \times \zeta$ covariance matrices $\boldsymbol{\Sigma}_Q$ and $\boldsymbol{\Sigma}_P$, respectively. The relative entropy in units of nats becomes (see Appendix B2)

$$d_{\mathrm{KL}}\Big(\mathcal{N}_\zeta(\boldsymbol{\mu}_Q,\boldsymbol{\Sigma}_Q),\mathcal{N}_\zeta(\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)\Big) = \frac{1}{2}\Big[\log_e\Big(\Big|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P\Big|\Big) - \zeta + \mathrm{tr}\big(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\big) + \big(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P\big)^\top \boldsymbol{\Sigma}_P^{-1}\big(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P\big)\Big] \tag{3}$$

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 $\zeta \times \zeta$ matrix, $\mathbf{A} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0$. Equation 3 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 3 reduces to

$$d_{\mathrm{KL}}\Big(\mathcal{N}_1(\mu_Q,\sigma_Q^2),\mathcal{N}_1(\mu_P,\sigma_P^2)\Big) = \frac{1}{2}\log_e\left(\frac{\sigma_P^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + (\mu_Q - \mu_P)^2 - \sigma_P^2}{2\sigma_P^2}. \tag{4}$$

The analytic expressions of the relative entropy in Appendix B confirm that $d_{\mathrm{KL}}(Q, P)$ 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 from P to Q does not equal its counterpart $d_{\mathrm{KL}}(P, Q)$ from Q to P. To convey this 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. In Appendix B3 we further show that $d_{\mathrm{KL}}(Q, P)$ does not satisfy the fourth and last axiom, the so-called triangle inequality, $d_{\mathrm{KL}}(Q, P) \leq d_{\mathrm{KL}}(Q, R) + d_{\mathrm{KL}}(R, P)$ of a metric d in space $\mathcal{M}$. Thus, relative entropy $d_{\mathrm{KL}}(Q, P)$ is not a metric or distance function in an Euclidean space with its usual physical or metaphorical notion of length but should be thought of as a divergence of two distributions. This distance is better known as the Kullback and Leibler (1951) divergence, hence our use of $d_{\mathrm{KL}}(Q, P)$. Due to its favorable properties (Liese & Vajda, 2006) the KL divergence $d_{\mathrm{KL}}(Q, P)$ is often used as measure of distance between two distributions (de Punder et al., 2023). The terminology of reverse KL-divergence is in use for its counterpart, $d_{\mathrm{KL}}(P, Q)$, or the relative entropy of P with respect to Q.

The fact that $d_{\mathrm{KL}}(Q, P)$ is strictly nonnegative and zero only when $P = Q$ suggests its interpretation as a so-called Bregman (1967) *divergence* of Q and P. The term Bregman *distance* is also used, even though $d_{\mathrm{KL}}(P, Q)$ and $d_{\mathrm{KL}}(Q, P)$ are not necessarily equal. We point forward to Figure 3 for a graphical representation of the Bregman divergence but refrain from detailed comments until Section 5 on scoring rules. In general, Bregman divergences play a key role in scientific forecast evaluation as they guarantee *strict propriety* of the scoring rules and associated divergence functions. *Strict propriety* implies that $d_{\mathrm{KL}}(Q, P) > 0$ if $P \neq Q$ and $d_{\mathrm{KL}}(Q, P) = 0$ if and only if $P = Q$. Propriety, "*…conformity to conventionally accepted standards of behavior or morals*" follows from Jensen's inequality (the secant line of a convex function lies above the graph of the function) and incentives a forecaster to be honest and volunteer $P = Q$ rather than any $P \neq Q$. In other words, only *strictly proper* scoring rules will lead us to the *true distribution*, $Q \in \mathcal{P}$. The minimum divergence $d(P, Q) = 0$ is achieved when $P = Q$, and this minimum is unique. *Proper* scoring rules also yield a minimum at $P = Q$ but this minimum may not be unique. The J-divergence named in honor of Sir Harold Jeffreys (Jeffreys, 1946) is a symmetrized variant of the Kullback and Leibler (1951) divergence

$$d_{\mathrm{J}}(P,Q) = d_{\mathrm{KL}}(Q,P) + d_{\mathrm{KL}}(P,Q) = \int_{x \in \Omega} (Q(x) - P(x)) \left(\frac{Q(x)}{P(x)}\right) \mathrm{d}x \tag{5}$$

and commonly used in pattern recognition and computer vision (Chang et al., 2009; Zheng & You, 2013). Divergences are sometimes called divergence functions or discrepancy functions or validation metrics (Liu et al., 2011), even though our example in Appendix B3 has shown that they may not necessarily satisfy the requirements of a metric in mathematical sense (Thorarinsdottir et al., 2013).

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 $\mathcal{P}= \mathcal{P}_m$

$$\mathcal{P}_m = \left\{\mathbf{p} = (p_1,\ldots,p_m)^\top : p_1 + \cdots + p_m = 1 \text{ and } p_k \geq 0 \text{ for all } k\right\}. \tag{6}$$

It is further assumed that the vector $\mathbf{q} = (q_1,\ldots,q_m)^\top$ reports the *true probabilities* of the m events, $\left\{\mathbf{q} \in \mathbb{R}_+^{m\times 1} : \mathbf{1}^\top \mathbf{q} = 1\right\}$, 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

$$d_{\mathrm{KL}}(\mathbf{q},\mathbf{p}) = \mathbb{H}(\mathbf{q},\mathbf{p}) - \mathbb{H}(\mathbf{q}), \tag{7}$$

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

$$d_{\mathrm{KL}}(\mathbf{q},\mathbf{p}) = \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{q_k}{p_k}\right), \tag{8}$$

which is equivalent to

$$d_{\mathrm{KL}}(\mathbf{q},\mathbf{p}) = -\sum_{k=1}^{m} q_k \log_{\mathfrak{b}}\left(\frac{p_k}{q_k}\right). \tag{9}$$

Appendix C demonstrates the use of the KL-divergence for two discrete distributions.

In the context of Bayesian inference, $d_{\mathrm{KL}}(Q, P)$ may be used as a measure of the information gained by revising one's beliefs from the prior probability distribution P to the posterior probability distribution Q (Scharnagl et al., 2010). This is equivalent to the amount of information lost when P is used to approximate Q (Burnham & Anderson, 2002). For example, if $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)^\top$ is a d-vector of unknown coefficients of some mathematical model (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. The relative entropy

$$d_{\mathrm{KL}}(p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}),p(\boldsymbol{\theta}|\mathcal{H})) = \mathbb{H}(p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}),p(\boldsymbol{\theta}|\mathcal{H})) - \mathbb{H}(p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})), \tag{10}$$

measures the added message length (bits) that an original code with 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 are the posterior and prior

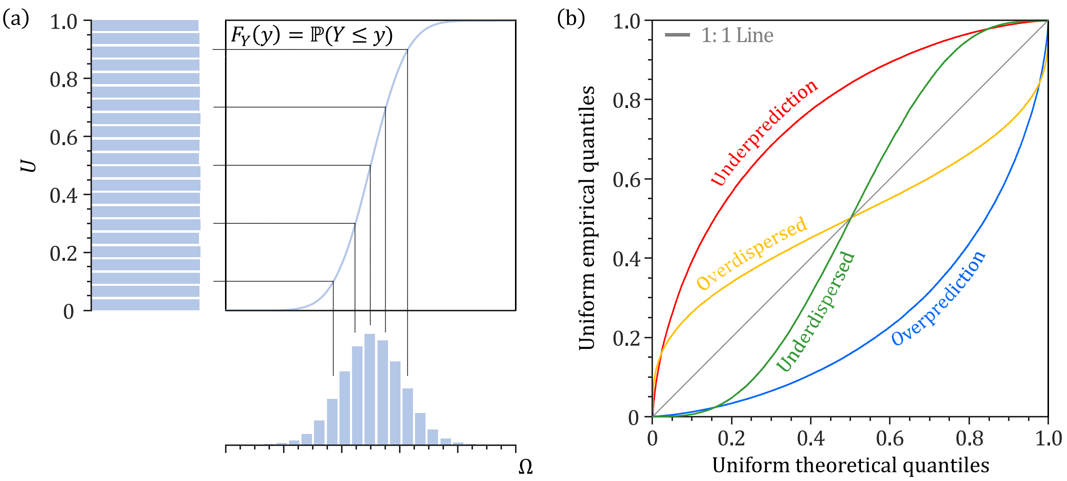

Figure 1. (a) Illustration of the probability integral transform and (b) interpretation of the so-called quantile-quantile plot (adapted from Laio and Tamea (2007)).

distributions, respectively. Thus, $d_{\mathrm{KL}}(p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}),p(\boldsymbol{\theta}|\mathcal{H}))$, is equal to the information gain about $\boldsymbol{\theta}$ 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, $\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = 1$.

4. Insufficient Scoring Rules for Density Forecasts

In most practical situations we do not have knowledge of the underlying distribution $Q \in \mathcal{P}$ which materializes with the measurement $\omega \in \Omega$. Then, we must find ways to evaluate the distribution forecast, $P \in \mathcal{P}$, 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. 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 1a). Let Y be a real-valued continuous random variable on a sample space Ω with CDF, $F_Y(y) = \mathbb{P}(Y \leq y)$. Then, random variable $U = F_Y(y)$ has a standard uniform distribution. Thus, if $\omega_1, \ldots, \omega_n$ are samples of Y (dependent or not) then $u_i = F_Y(\omega_i)$; $i = (1, \ldots, n)$ will be uniformly distributed on the unit interval. Hence, the probability integral transform reduces the assessment of F_Y to the question whether the sequence of u's behaves as a random sample of $\mathcal{U}[0,1]$. Figure 1b illustrates the consequences of using an incorrect distribution for Y on the relationship between the theoretical quantiles of the standard uniform distribution and the quantiles of the empirical distribution function of sampled data, $\omega_1, \ldots, \omega_n$. This so-called predictive quantile-quantile (Q-Q) plot (Casella & Berger, 2002; Dawid, 1984) is a common verification tool for probabilistic forecasts of meteorological (Gneiting & Raftery, 2007) and hydrologic (Laio & Tamea, 2007; Renard et al., 2011; Thyer et al., 2009) variables. This graph diagnoses 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 & 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_1 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

Table 1
Time-Averaged Performance Measures, $\overline{M}(\mathbf{P},\boldsymbol{\omega})$, of Distribution Forecasts $\boldsymbol{P} = \{P_1, \ldots, P_n\}$ and Verifying Observations $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^\top$

Performance measure	Symbol	$\overline{M}(\mathbf{P},\boldsymbol{\omega})$	Miscellaneous	Reference
Reliability[a]	R_1	$\frac{2}{n}\sum_{t=1}^{n}\lvert u'_t - \frac{t}{n}\rvert$	$u_t = F_{P_t}(\omega_t)$	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]	C	$\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\{l_t \leq \omega_t \leq u_t\}$		Dunsmore (1968)
Width[d]	W	$\frac{1}{n}\sum_{t=1}^{n}(u_t - l_t)$	$l_t = F_{P_t}^{-1}\left(\frac{\alpha}{2}\right)$ $u_t = F_{P_t}^{-1}\left(1 - \frac{\alpha}{2}\right)$	Raftery et al. (2005)

Note. F_P and F_P^{-1} are the cumulative distribution function (CDF) and inverse CDF of P. [a] $u'_1,\ldots,u'_n$ are ordered values of u. [b] Uses sample mean, m_P, and sample standard deviation, s_P, for an ensemble forecast. [c] The indicator function $\mathbb{1}\{a\}$ returns 1 if a is true and zero otherwise. [d] Lower l_t and upper u_t endpoints of $100(1-\alpha)\%$ prediction interval at $\alpha \in (0, 1)$ significance level.

$$R_1 = \frac{2}{n}\sum_{t=1}^{n}\left|u'_t - \frac{t}{n}\right|, \tag{11}$$

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 11 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 (Evin et al., 2013; Koutsoyiannis & Montanari, 2022; Renard et al., 2011; Thyer et al., 2009), 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_1 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 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 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 2). Specifically, the width W and coefficient of

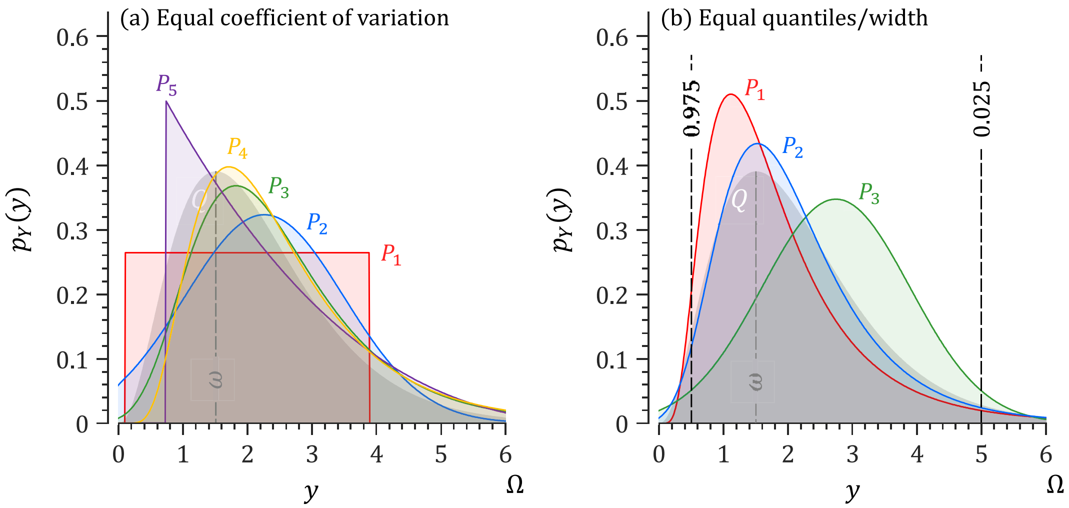

Figure 2. Hypothetical *true distribution* $Q = \mathcal{G}(3.36, 0.64)$ (gray) and *probabilistic forecasts* P (in color) with an equal (a) coefficient of variation, $C_\text{v} = 0.546$, and (b) width, $W = 4.5$, using lower and upper quantiles of 0.5 and 5.0 at $\alpha = 0.05$. The peak of the *true 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 $P_5 = \mathcal{GP}(-0.28, 2, 0.73)$.

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 Figures 2a and 2b. A shift of the distribution forecasts to 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 (see 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 the selection of a single best distribution forecast among the rank one solutions remains inherently subjective. Furthermore, the performance metrics do not guarantee a complete evaluation of distribution forecasts.

5. Scoring Rules

Scoring rules are indispensable in our search for the *true distribution* Q, but have not yet entered mainstream use in hydrology. In terms of elicitation, scoring rules 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 probabilistic forecasts, reward probability assessors for forecasting jobs, and rank competing forecast procedures (Gneiting & Raftery, 2007). Meteorologists refer to this task as *forecast verification*. We briefly review underlying theory of scoring rules and demonstrate their application to categorical forecasts.

5.1. Theory

Per Definition 2, a *scoring rule* $S(P, \omega)$ measures the reward when distribution forecast $P \in \mathcal{P}$ is issued and the observation $\omega \in \Omega$ materializes. The expected score $\mathcal{S}(P, Q) : \mathcal{P} \times \mathcal{P} \rightarrow \overline{\mathbb{R}}$ of *probabilistic forecast* P under the *true distribution* $Q \in \mathcal{P}$ is defined by

$$\mathcal{S}(P, Q) = \mathbb{E}_{\omega \sim Q}[S(P, \omega)] = \int_\Omega S(P, \omega) \mathrm{d}Q(\omega) \tag{12}$$

and equal to the expected value of $S(P, \omega)$ under the *true distribution* Q of ω. Note that the order of the arguments of $\mathcal{S}(P,Q)$ has reversed with respect to the convention used in information theory. In line with the statistical forecasting literature, the *probabilistic forecast* P precedes the *true distribution* Q (Bröcker, 2009; Gneiting & Raftery, 2007). A higher score suggests a better forecast and, thus, our scoring rules $S(P, Q)$ are positively oriented and defined as reward functions which the forecaster aims to maximize. Then, a scoring rule S is said to be *proper* relative to $\mathcal{P}$ if

$$\mathcal{S}(P,Q) \le \mathcal{S}(Q,Q) \quad \text{for all } P,Q \in \mathcal{P}, \tag{13}$$

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

$$d(P,Q) = \mathcal{S}(Q,Q) - \mathcal{S}(P,Q), \quad P,Q \in \mathcal{P}, \tag{14}$$

measures the difference of the *probabilistic forecast* $P \in \mathcal{P}$ from the *true distribution* $Q \in \mathcal{P}$. This is a *divergence function* alike the relative entropy in Equations 1 and 8 and equal to a measure of difference between two points defined in terms of a continuously-differentiable expected score function, $H(P) : \mathcal{P} \to \mathbb{R}$. For positively oriented *proper* scoring rules, $H(P)$ is the pointwise supremum (least upper bound) over the convex class of probability measures Q on $\mathcal{P}$ (Gneiting & Raftery, 2007)

$$H(P) = \sup_{Q \in \mathcal{P}} \mathcal{S}(Q,P) = \mathcal{S}(P,P), \qquad P \in \mathcal{P}, \tag{15}$$

and is convex on $\mathcal{P}$ 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 Ω is finite and $H(P)$ smooth, then $d(P, Q)$ is the Bregman (1967) distance associated with convex function $H(P)$.

To understand the relationship between $H(P)$, $\mathcal{S}(P,Q)$ and $d(P, Q)$, Figure 3 displays the entropy function of the *strictly proper* quadratic score (QS) for a binary event (*rain* or *no rain*)

$$\begin{aligned} H_{\text{QS}}(p) &= p^2 + (1-p)^2 = 2p(p-1) + 1 \\ \Rightarrow H_{\text{QS}}(p) &\mapsto p(p-1), \end{aligned} \tag{16}$$

where $p \in [0, 1]$ is the quoted *probability* for *rain* and $\mapsto$ is an affine transformation, of which more later. The expected score function of the QS is strictly convex on p 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 (Buja et al., 2005; Grünwald & Dawid, 2004). This is the maximally achievable utility. Some authors refer instead to $-H(p)$ as the entropy function (Bröcker, 2009; Dawid & Musio, 2014) or *coherent uncertainty function* (Dawid & Sebastiani, 1999). According to Figure 3, 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

$$d(p,q) = \mathcal{S}(q,q) - \mathcal{S}(p,q) = H(q) - H(p) + H\prime(p)(p-q), \tag{17}$$

where $H'(p) = \mathrm{d}H(p)/\mathrm{d}p$ is the derivative of the entropy function with respect to p. Thus, the entropy function $H(p) = p^2$ has divergence function $d(p, q) = (p-q)^2$. This squared Euclidean distance is equal to the well-known (Brier, 1950) score.

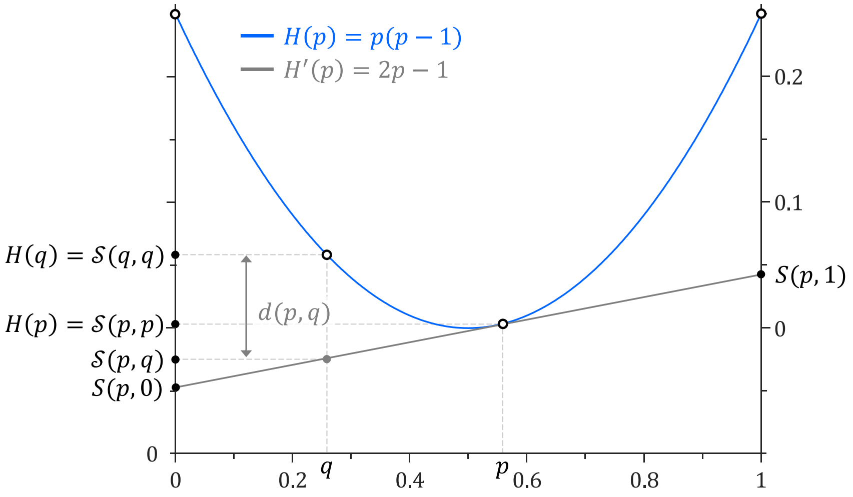

Figure 3. Generalized entropy function $H(p) = p(p-1)$ (blue curve) of the quadratic score for a dichotomous event $\Omega = \{1, 0\}$ with *probability forecast* $(p, 1-p)$ and *true probability* $(q, 1-q)$ with $p, q \in [0, 1]$. We present the values of the quadratic scoring rule $\mathcal{S}(p,q)$ 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, $d_{\text{QS}}(p,q) = \mathcal{S}(q,q) - \mathcal{S}(p,q)$, is equal to the difference between $H(q)$ and the tangent at p when evaluated at q (Adapted after Figure 1 of Gneiting and Raftery (2007) and Figure 8 of Buja et al. (2005)).

5.2. 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 $\mathcal{P} = \mathcal{P}_m$ defined in Equation 6. Then Equation 17 may be written as

$$d_{\text{QS}}(\mathbf{p},\mathbf{q}) = H(\mathbf{q}) - H(\mathbf{p}) + \langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \rangle \tag{18}$$

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

$$\nabla H(\mathbf{p}) = \frac{\partial H(\mathbf{p})}{\partial \mathbf{p}} = \left(\frac{\partial H(\mathbf{p})}{\partial p_1}, \frac{\partial H(\mathbf{p})}{\partial p_2}, \ldots, \frac{\partial H(\mathbf{p})}{\partial p_m} \right)^\top \tag{19}$$

and $\langle \mathbf{a}, \mathbf{b} \rangle$ 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)

$$S(\mathbf{p},j) = H(\mathbf{p}) - \langle \nabla H(\mathbf{p}), \mathbf{p} \rangle + \nabla H_j(\mathbf{p}) \quad \text{for} \quad j = (1,\ldots,m) \tag{20}$$

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 a binary event (*rain* or not). 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 of Equation 12 then equals $\mathcal{S}(p,q) = qS(p,1) + (1-q)S(p,0)$, where q is the true *rain* probability. For any two assignments $\mathbf{p}$ and $\mathbf{q}$ with *true probabilities* $\mathbf{q} = (q_1,\ldots,q_m)^\top$ constrained to the probability simplex, $\left\{\mathbf{q} \in \mathbb{R}_+^{m\times 1} : \mathbf{q}^\top \mathbf{1}_m = 1\right\}$ and $\mathbb{R}_+ = [0,\infty)$, the binary definition of the expected score generalizes to (Bröcker, 2009)

$$\mathcal{S}(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{m} q_j S(\mathbf{p},j), \tag{21}$$

Table 2
Entropy Function, Scoring Rule, Expected Score, and Score Divergence of Quadratic, Logarithmic, and Pseudospherical Scoring Rules for a Categorical Distribution Forecast $\mathbf{p} = (p_1,\ldots,p_m)^\top$ *on the Convex Class* $\mathcal{P} = \mathcal{P}_m$ *of* $m \geq 2$ *Mutually Exclusive and Collectively Exhaustive Events,* $\Omega = \{1, \ldots, m\}$

Score name	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]	$\sum_{k=1}^{m} p_k^2$	$2p_j - \sum_{k=1}^{m} p_k^2$	$2\sum_{k=1}^{m} p_k q_k - \sum_{k=1}^{m} p_k^2$	$\sum_{k=1}^{m} (p_k - q_k)^2$
Logarithmic[b,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]	$\lVert \mathbf{p} \rVert_\eta^1$	$\frac{p_j^{\eta-1}}{\lVert \mathbf{p} \rVert_\eta^{\eta-1}}$	$\frac{\sum_{k=1}^{m} p_k^{\eta-1} q_k}{\lVert \mathbf{p} \rVert_\eta^{\eta-1}}$	$\lVert \mathbf{q} \rVert_\eta^1 - \frac{\sum_{k=1}^{m} p_k^{\eta-1} q_k}{\lVert \mathbf{p} \rVert_\eta^{\eta-1}}$

Note. The m-vector $\mathbf{q} = (q_1,\ldots,q_m)^\top$ lists the true event probabilities. [a]Also known as proper linear score. Equals Brier (1950) score for a binary event, $\Omega = \{1, 0\}$. [b]Remains *strictly proper* under any logarithmic base $\mathfrak{b} > 1$. [c]Affine transformation of the pseudospherical score for $\eta \to 1$. [d]The η-norm $\lVert \mathbf{p} \rVert_\eta = \left(\sum_{k=1}^{m} p_k^\eta\right)^{1/\eta}$ raised to the power 1 or $\eta - 1$. [e]Reduces to the spherical score for $\eta = 2$ (Good, 1971; Roby, 1964).

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.

Thus, for any hypothesized generalized entropy function $H(\mathbf{p})$ whose graph is cup-shaped, we can use Equations 18, 20, and 21 to derive mathematical expressions of its corresponding scoring rule, $S(\mathbf{p}, j)$, expected score function, $\mathcal{S}(\mathbf{p},\mathbf{q})$, and score divergence, $d(\mathbf{p}, \mathbf{q})$. Appendix D presents such derivations for the entropy functions $H(\mathbf{p})$ of the quadratic, logarithmic and (pseudo)spherical scoring rules. These Equations are listed in Table 2. By definition, p and q are dimensionless and, consequently, all functions are unitless except for the entropy function $H(\mathbf{p})$ of the LS which has units of information and, thus, bits if $\mathfrak{b} = 2$. It is important to keep in mind that in practice we can only evaluate the scoring rule $S(\mathbf{p}, j)$ in the absence of knowledge of the true probabilities $\mathbf{q}$.

The LS of Good (1952) is also known as the predictive deviance (Knorr-Held & Rainer, 2001) or ignorance score (Roulston & Smith, 2002) and links fundamental aspects from statistical, decision and information theory. The generalized entropy function of the LS is equal to negative Shannon entropy $-\mathbb{H}(P)$ and its score divergence, $d_{\text{LS}}(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{m} q_k \log_{\mathfrak{b}}(q_k/p_k)$, is the reverse KL-divergence (Gneiting & Raftery, 2007). Note that the LS is equal to the logarithmic probability assigned to the materialized event and, thus, is only *strictly proper* locally.

We look in more detail at the scoring rules of Table 2 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 *strictly proper* categorical scoring rules of Table 2 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 $S(p, 0)$. If $q = q_1$ is the *true* event probability, then the expected score equals $\mathcal{S}(p,q) = qS(p,1) + (1 - q)S(p,0)$ according to Equation 21. Figure 4 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). For the LS we used $\mathfrak{b} = 2$ to yield $\mathcal{S}_{\text{LS}}(p,q) = q\log_2(p) + (1 - q)\log_2(1 - p)$ in units of bits.

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 values of the QS, LS, and

Table 3
Strictly Proper Scoring Rules $S(p, j)$ *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	$p^{\eta-1}(p^\eta + (1 - p)^\eta)^{(1-\eta)/\eta}$	$(1 - p)^{\eta-1}(p^\eta + (1 - p)^\eta)^{(1-\eta)/\eta}$

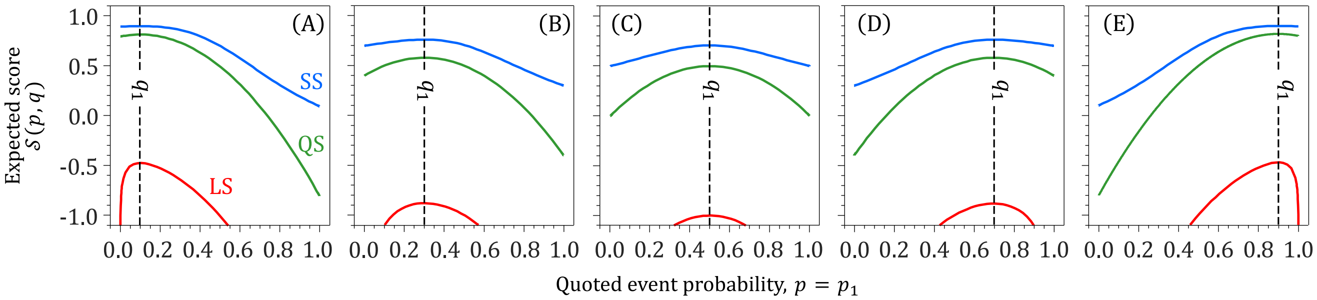

Figure 4. Binary event, $\Omega = \{1, 0\}$: Expected value of quadratic (green), logarithmic (red) and spherical (blue) scoring rules as function of quoted probability $p = p_1$ of the first event using true probabilities (a) $q = 0.1$, (b) $q = 0.3$, (c) $q = 0.5$, (d) $q = 0.7$, and (e) $q = 0.9$.

spherical score (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 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. The rate of decline is largest for the LS followed by the QS and SS. The magnitude differences of the three scoring rules are not relevant as *strictly proper* scoring rules S remain *strictly proper* under affine transformation

$$S^{*}(\mathbf{p},j) = cS(\mathbf{p},j) + \hbar(j), \tag{22}$$

where constant c is nonzero and $\hbar(j)$ is a $\mathcal{P}$-integrable function (Gneiting & Raftery, 2007). See Equation 16. If $c < 0$ the orientation of $S^{*}(\mathbf{p}, j)$ changes from a reward to a loss function.

5.3. Numerical Examples

While it is generally agreed upon that scoring rules must at least be proper to accurately quantify the quality of probabilistic forecasts (Gneiting & Ranjan, 2011; Winkler et al., 1996), the question which ones to use in practical application is unresolved (Alexander et al., 2022). For the time being, we restrict our attention to the three categorical scoring rules of Table 2.

5.3.1. Case Study I: Simple Illustration

We revisit the distribution forecasts of Figure 2a and turn the PDF of the *true* forecast distribution Q with continuous sample space $\Omega = [0, 6]$ into a probability mass function (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, (a) the expected score $\mathcal{S}(\mathbf{p},\mathbf{q})$ and score divergence $d(\mathbf{p}, \mathbf{q})$ are maximized and minimized, respectively, when the forecaster quotes the *true* probabilities, (b) the LS 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, (c) the QS, LS, and SS divergence scores $d(\mathbf{p}, \mathbf{q})$ are strictly positive and zero only when $P = Q$, and (d) *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). The LS is sometimes criticized for its unboundedness and (reminder) has negative Shannon entropy $-\mathbb{H}(\mathbf{p})$ (7) and relative entropy $d_{\mathrm{KL}}(\mathbf{q}, \mathbf{p})$ (8) as its entropy function $H(\mathbf{p})$ and score divergence $d(\mathbf{p}, \mathbf{q})$, respectively.

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

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 $P_1 = \mathcal{U}(0.11,3.89)$, $P_2 = \mathcal{N}(2.26,1.52)$, $P_3 = \mathcal{GEV}(0.04,1,1.86)$; $P_4 = \mathcal{LN}(0.80,0.26)$ and $P_5 = \mathcal{GP}(-0.28,2,0.73)$ Displayed in Figure 2a Using m = 60 Discrete Values, $\Omega = \frac{1}{20}\{1,3,5,\ldots,117,119\}$

	Quadratic score				Logarithmic score, $\mathfrak{b} = 2$				Spherical score			
		$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$			$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$			$\mathcal{S}(\mathbf{p},\mathbf{q})$	$d(\mathbf{p},\mathbf{q})$	
Fcst	$H(\mathbf{p})$	Equation D4	Equation D1	R[a]	$H(\mathbf{p})$	Equation D7	Equation D5	R	$H(\mathbf{p})$	Equation D13	Equation D14	R
P_1	0.026	0.022	0.005	5	−5.25	$-\infty$	∞	4	0.162	0.150	0.016	5
P_2	0.024	0.025	0.003	4	−5.52	−5.49	0.102	2	0.156	0.158	0.009	4
P_3	0.026	0.027	0.001	1	−5.48	−5.44	0.046	1	0.161	0.163	0.003	1
P_4	0.028	0.026	0.001	2	−5.39	−5.54	0.150	3	0.167	0.162	0.004	2
P_5	0.029	0.025	0.003	3	−5.31	$-\infty$	∞	5	0.172	0.159	0.007	3
Q	0.028	0.028	0.000		−5.39	−5.39	0.000		0.166	0.166	0.000	

Note. The bottom row presents the values for a perfect distribution forecast, $P = Q$. [a]Rank R of each distribution forecast obtained from sorting $d(\mathbf{p},\mathbf{q})$ in ascending order.

inspection of the distribution forecasts with the lognormal distribution forecast P_4 (yellow) as a close contender. The results further demonstrate that (a) 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 (b) outcomes with a zero-probability do not count in the entropy of the LS 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 LS.

5.3.2. Case Study II: Rainfall Data

Appendix E presents the application of the categorical scoring rules of Table 2 to 24-hr forecasts of precipitation probability in south-central Finland. This study is included for benchmarking purposes.

6. Scoring Rules for Density (Probabilistic) Forecasts

The task of determining whether the *probabilistic forecast* P matches the *true distribution* Q appears difficult, perhaps hopeless, because Q is never observed, even after the fact. But as Diebold et al. (1998) realized early on, the challenges posed by these subtleties are not insurmountable. Scoring rules for categorical variables can be generalized to density forecasts to assist in forecast verification of continuous variables. This involves use of a so-called Lebesgue (1902) measure μ and is explained in Appendix F. For now, please consider the Lebesgue measure to be equal to the histogram bin width.

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 $f : \Omega \rightarrow \mathbb{R}$ on (Ω, Σ) that are absolutely continuous with respect to the measure μ on Σ and have an integral

$$\|f\|_\eta \equiv \left(\int_\Omega f(\omega)^\eta \mu(\mathrm{d}\omega)\right)^{1/\eta} < \infty, \tag{23}$$

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) $d(P,Q) = \left\| f_P - f_Q \right\|_\eta$ which is zero only if $P = Q$. In general, the more compact predictive density P is, the larger will be its L_η-norm. For $\eta = 1$, we yield that $L_1 = 1$.

In most applications, the probabilistic forecast P will consist of a m-member ensemble $\mathbf{y} = (y_1,\dots,y_m)^\top$ at given space and time coordinates. The empirical CDF (eCDF), F_P, of the forecast distribution P can be construed from the ensemble

$$F_P(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{y_j \le x\}. \tag{24}$$

and is assumed continuous and strictly positive on Ω. Usually the ensemble size m is on the order of a few hundred members or less and we must use kernel dressing to translate the m discrete outcomes of P into a density forecast f_P (Bröcker & Smith, 2008). Such methods guarantee that f_P will integrate to one over its support Ω, thus, $\int_\Omega f_P(x)\mathrm{d}x = 1$. If the samples $x_1, \dots, x_M$ ($M \gg m$) are evenly distributed on the real line $\mathbb{R}$, then the η-norm of density forecast f_P simplifies to

$$\|f_P\|_\eta = \left[\Delta x \sum_{k=1}^{M} f_P^{\eta}(x_i)\right]^{1/\eta}, \tag{25}$$

where $\Delta x = x_2 - x_1$ is equal to the Lebesgue measure of each event x_i.

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 value ω that materializes. In analogy to Equation D2, the QS becomes

$$S_{\mathrm{QS}}(P,\omega) = 2f_P(\omega) - \|f_P\|_2^2, \tag{26}$$

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 is the density of the true distribution Q. The power scoring rule, $S_{\mathrm{PS}}(P,\omega) = \eta f_P(\omega)^{\eta-1} - (\eta-1)\|f_P\|_\eta^\eta$, is a generalization of the QS to an arbitrary positive power $\eta > 1$. For $\eta \to 1$ we yield the LS

$$S_{\mathrm{LS}}(P,\omega) = \log_b\left(f_P(\omega)\right), \tag{27}$$

and is *strictly proper* relative to the class $\mathcal{L}_1$ of probability measures. The LS has negative Shannon entropy $-\mathbb{H}(P)$ as its entropy function $H(P)$ and the reverse KL-divergence as its divergence score (e.g., Table 2). The pseudospherical score

$$S_{\mathrm{PSS}}(P,\omega) = \frac{f_P(\omega)^{\eta-1}}{\|f_P\|_\eta^{\eta-1}}, \tag{28}$$

is *strictly proper* relative to the class $\mathcal{L}_\eta$. Strict convexity of its entropy function $H(P) = \|f_P\|_\eta$ and nonnegativity of its divergence function are implied by the Hölder and Minkowski inequalities. For $\eta \to 1$ we yield a multiple of the LS and $\eta = 2$ results in the SS

$$S_{\mathrm{SS}}(P,\omega) = \frac{f_P(\omega)}{\|f_P\|_2}, \tag{29}$$

which is *strictly proper* relative to class $\mathcal{L}_2$ of probability measures. The expressions of the logarithmic and SSs in Equations 27 and 29 may inspire the use of a linear score, $S_{\mathrm{LinS}}(P, \omega) = f_P(\omega)$. This score may seem intuitively appealing but is improper as shown in Appendix G and Table 6.

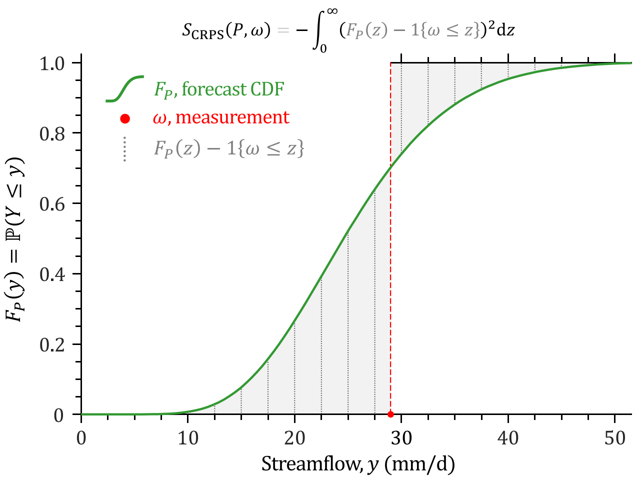

Figure 5. Graphical explanation of the continuous ranked probability scoring rule for a hypothetical streamflow forecast cumulative distribution function, F_P (black line), and verifying measurement ω (red dot). The continuous ranked probability score is the integral of squared differences (=gray dotted lines) of F_P and the Heaviside step function, $\mathbb{1}\{\omega \le z\}$.

6.1.2. 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 CRPS of Figure 5 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 (Hersbach, 2000; Matheson & Winkler, 1976)

$$\begin{aligned} S_{\mathrm{CRPS}}(P,\omega) &= -\int_{-\infty}^{\infty} (F_P(z) - \mathbb{1}\{\omega \le z\})^2 \,\mathrm{d}z \\ &= -\int_{-\infty}^{\omega} F_P^2(z)\mathrm{d}z - \int_{\omega}^{\infty} (F_P(z) - 1)^2 \,\mathrm{d}z, \end{aligned} \tag{30}$$

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 $\mathcal{P}_1 \in \mathcal{P}$ of Borel probability measures that have finite first moment. The CRPS is equal to the Brier probability score of distribution forecast $F_P(z) = \int_{-\infty}^{z} f_P(t)\mathrm{d}t$ of binary event $\{\omega \le z\}$ over all thresholds $z \in \mathbb{R}$

$$S_{\mathrm{CRPS}}(P,\omega) = \int_{-\infty}^{\infty} S_{\mathrm{BS}}\,(F_P(z), \mathbb{1}\{\omega \le z\})\,\mathrm{d}z, \tag{31}$$

where $S_{\mathrm{BS}}\,(F_P(z), \mathbb{1}\{\omega \le z\}) = -(F_P(z) - \mathbb{1}\{\omega \le z\})^2$. The CRPS can also be written using the $\tau \in [0, 1]$-quantile forecast $y_\tau = F_P^{-1}(\tau)$ of P (Grushka-Cockayne et al., 2017; Laio & Tamea, 2007)

$$S_{\mathrm{CRPS}}(P,\omega) = -2\int_0^1 (1\{\omega < y_\tau\} - \tau)(y_\tau - \omega)\mathrm{d}\tau, \tag{32}$$

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

$$S_{\mathrm{QNT}}^{\tau}(P,\omega) \;=\; (1\{\omega < y_\tau\} - \tau)(\omega - y_\tau), \tag{33}$$

known also in negative orientation as the pinball-loss, tick-loss or check-loss function and used by Tyralis and Papacharalampous (2021) for hydrologic model calibration. Laio and Tamea (2007) proof the equivalence of Equations 30 and 32. The more friendly form of Equation 32 (see Appendix H)

$$S_{\mathrm{CRPS}}(P,\omega) = \omega(1 - 2F_P(\omega)) + 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau, \tag{34}$$

simplifies closed-form solutions of the CRPS for parametric distribution forecasts P (A. Jordan, 2016; Villez, 2017). Appendix I1 presents such derivation for a normal distribution forecast $P = \mathcal{N}(\mu_P, \sigma_P^2)$

$$S_{\mathrm{CRPS}}\big(\mathcal{N}(\mu_P, \sigma_P^2), \omega\big) = \frac{\sigma_P}{\sqrt{\pi}} - 2\sigma_P^2 f_{\mathcal{N}}(\omega, \mu_P, \sigma_P^2) \;-\; (\omega - \mu_P)\big(2F_{\mathcal{N}}(\omega, \mu_P, \sigma_P^2) - 1\big), \tag{35}$$

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 evaluate the integral of Equation 34 using Monte Carlo techniques, for example, 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 (Gneiting & Raftery, 2005)

$$S_{\mathrm{CRPS}}(P,\omega) = \frac{1}{2}\mathbb{E}_P[|y - y^*|] - \mathbb{E}_P[|y - \omega|], \tag{36}$$

where y and y^* are samples of forecast distribution P. This form shows that $S_{\mathrm{CRPS}}(P, \omega)$ has units of ω. For a m-member ensemble forecast, $\mathbf{y} = (y_1,\ldots,y_m)^\top$, Equation 36 equals (Grimit et al., 2006)

$$S_{\mathrm{CRPS}}(P,\omega) = \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|y_i - y_j| - \frac{1}{m}\sum_{i=1}^{m}|y_i - \omega|. \tag{37}$$

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 a generalization of the absolute residual to a distribution forecast. The computational complexity $\mathcal{O}(m^2)$ of Equation 37 reduces to a total of $\mathcal{O}(m\log_b(m))$ operations if the CRPS is evaluated in terms of the quantile form of Equation 32 using the sorted ensemble members (Hersbach, 2000; Laio & Tamea, 2007; Murphy, 1970)

$$S_{\mathrm{CRPS}}(P,\omega) = -\frac{2}{m^2}\sum_{i=1}^{m}|y_i - \omega|\left(m\mathbb{1}\{\omega \le y_i\} - i + \frac{1}{2}\right). \tag{38}$$

The generalized entropy function or information measure of the CRPS

$$H(P) = -\int_{-\infty}^{\infty} F_P(z)(1 - F_P(z))\,\mathrm{d}z = -\frac{1}{2}\mathbb{E}_P[|y - y^*|], \tag{39}$$

is the negative selectivity function (Gneiting & Raftery, 2007; Matheron, 1984) and intimately related to the Gini (1909) coefficient G, a measure of the inequality degree in income and wealth distribution. The CRPS divergence function

$$d_{\mathrm{CRPS}}(P,Q) = \int_{-\infty}^{\infty}\left(F_P(z) - F_Q(z)\right)^2\mathrm{d}z, \tag{40}$$

is symmetric by virtue of the quadratic term and reminiscent of the Cramér-von Mises distance between an empirical and given CDF (Cramér, 1928; Von Mises, 1928) or two sample CDFs (Anderson, 1962).

6.1.3. Energy Score

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

$$S_{\mathrm{ES}}(P,\omega) = \frac{1}{2}\mathbb{E}_P[|y - y^*|^\eta] - \mathbb{E}_P[|y - \omega|^\eta], \tag{41}$$

where $\eta \in (0, 2)$ and y and y^* are independent copies of $P \in \mathcal{P}_\eta$. This is a *strictly proper* score (Székely, 2003) and reduces to the CRPS for $\eta = 1$ and the negative squared error $S_{\mathrm{SE}}(P, \omega) = -|\mu_P - \omega|^2$ for $\eta \to 2$ (Gneiting & Raftery, 2007), where μ_P is the mean of 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

$$S_{\mathrm{ES}}(P,\omega) = \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|y_i - y_j|^\eta - \frac{1}{m}\sum_{i=1}^{m}|y_i - \omega|^\eta. \tag{42}$$

Table 5 summarizes the different *strictly proper* scoring rules for a distribution forecast P with density forecast p defined up to the μ-measure zero. Functional analysis of the numerical expressions of the scoring rules provides

Table 5
Summary of Strictly Proper Scoring Rules for a Density Forecast f_P and Verifying Observation ω

		Scoring rule, $S_{\mathrm{XX}}(P, \omega)$		
Score name	XX	Analytic	Numerical	Note
Quadratic	QS	$2f_P(\omega) - \int_{-\infty}^{\infty} f_P^2(y)\mathrm{d}y$	$2f_P(\omega) - \lVert f_P \rVert_2^2$	[a]
Logarithmic	LS	$\log_{\mathfrak{b}}\left(f_P(\omega)\right)$	$\log_{\mathfrak{b}}\left(f_P(\omega)\right)$	[a]
Cnt. Rnk. Prb.	CRPS	$-\int_{-\infty}^{\infty}\left(F_P(z) - \mathbb{1}\{\omega \leq z\}\right)^2\mathrm{d}z$	$\frac{1}{2m^2}\sum_{i=1}^{m}\sum_{k=1}^{m}\lvert y_i - y_k\rvert - \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \omega\rvert$	[b]
Spherical	SS	$f_P(\omega)\left(\int_{-\infty}^{\infty} f_P^2(y)\mathrm{d}y\right)^{-1/2}$	$f_P(\omega)\lVert f_P \rVert_2^{-1}$	[a]
Energy	ES	$\frac{1}{2}\mathbb{E}_P[\lvert y - y^*\rvert^{\eta}] - \mathbb{E}_P[\lvert y - \omega\rvert^{\eta}]$	$\frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lvert y_i - y_j\rvert^{\eta} - \frac{1}{m}\sum_{i=1}^{m}\lvert y_i - \omega\rvert^{\eta}$	[b,c]

Note. The numerical form assumes that the forecast distribution P is a m-member ensemble $(y_1,\ldots,y_m)^\top$. [a]$f_P(x)$ is the empirical density of P at x. Determined from eCDF in Equation 24 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 $\eta \rightarrow 2$ leads to $S_{\mathrm{SE}}(P, \omega) = -\lvert\mu_P - \omega\rvert^2$.

insights into how the QS, LS, CRPS, SS, and ES respond to a particular distribution forecast. In general, the more concentrated the probability mass of P is around ω, the larger the reward. Jose et al. (2009) provide a generalization of the sensitivity of scoring rules to distance. As a reminder, the LS 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 demonstrate the power and usefulness of scoring rules by application to a variety of different studies using two illustrative examples, distribution forecasts of discharge obtained from Bayesian model averaging and the analysis of model simulated hydrograph recession and flow duration curves.

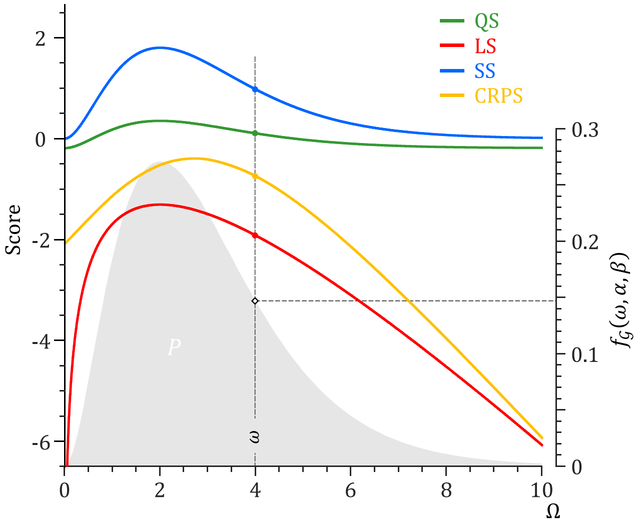

Figure 6. Gamma distribution forecast $P = \mathcal{G}(a,b)$ of discharge (mm/d) for $a = 3$ and $b = 1$ mm/d and traces of the quadratic (green), logarithmic (red), spherical (blue) and continuous ranked probability (yellow) scoring rules for a hypothetical measurement $\omega \in [0, 10]$. We use the numerical form of the scoring rules listed in Table 5. The right y-axis is the probability density function of the discharge distribution, $f_{\mathcal{G}}(\omega,a,b) = \Gamma^{-1}(a)b^{-a}\omega^{a-1}\exp(-\omega/b)$, where $\Gamma(x)$ is the gamma function.

6.2.1. Case Study III: Graphical Illustration

Suppose the forecast distribution P of discharge (in mm/d) is exactly described by a gamma distribution $P = \mathcal{G}(a_P,b_P)$ (see Figure 6) 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 5 on the y-axis. The 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 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 QS, LS, SS, and CRPS is rather similar but their curvatures are noticeably different. The LS responds strongest to the density of distribution forecast P, whereas the response and thus curvature of the QS is much more damped. Suppose $\omega = 4.0$ mm/d (vertical dashed line) materializes at a future time. According to the gamma discharge forecast $P = \mathcal{G}(3,1)$, we yield $f_{\mathcal{G}}(\omega,3,1) = 0.1465$ (black diamond) and the scoring rules (colored dots) attain values of $S_{\mathrm{QS}} = 0.106$, $S_{\mathrm{LS}} = -1.921$ nats, $S_{\mathrm{SS}} = 0.975$, and $S_{\mathrm{CRPS}} = -0.758$ mm/d, respectively. Appendix I2 presents a closed-form expression for the CRPS of $P = \mathcal{G}(a,b)$. Equation I37 is in perfect agreement with the yellow line.

Table 6
Mean Values of the Quadratic, Logarithmic, Spherical, Continuous Ranked Probability, and Linear Scoring Rules for Distribution Forecasts $P_1 = \mathcal{U}(0.11, 3.89)$, $P_2 = \mathcal{N}(2.26, 1.52)$, $P_3 = \mathcal{GEV}(0.04, 1, 1.86)$, $P_4 = \mathcal{LN}(0.80, 0.26)$ *and* $P_5 = \mathcal{GP}(-0.28, 2, 0.73)$ *Displayed in Figure 2a*

Score	P_1 red	P_2 blue	P_3 green	P_4 yellow	P_5 purple	$P = Q$ gray
QS	0.219	0.243	0.257	0.255	0.212	0.269
LS	$-\infty$	−2.302	−2.202	−2.309	$-\infty$	−2.138
SS	0.470	0.493	0.507	0.505	0.481	0.519
CRPS	−0.658	−0.661	−0.669	−0.673	−0.665	−0.645
LinS	0.242	0.236	0.252	0.259	0.299	0.270

Note. The last column reports the mean scores for a perfect distribution forecast, $P = Q$.

The strong similarities between 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 equal to an affine transformation (e.g., Equation 22) of the LS. In principle, we only require one *strictly proper* scoring rule, though multiple different scoring rules can aid in singling out the most adequate distribution forecast (Vrugt et al., 2022).

6.2.2. Case Study IV: Simple Illustration

We revisit the distribution forecasts P_1, …, P_5 of Figure 2a with an equal coefficient of variation (C_v) and compute the *strictly proper* scoring rules of Table 5 and *improper* linear score, $S_\mathrm{LinS}(P, \omega) = f_P(\omega)$, using $n = 10^4$ observations ω_1, …, ω_n from the *true distribution* Q. Table 6 documents the outcome of this analysis. The scoring rules display a considerable variation between the distribution forecasts of Figure 2a and assign different rewards to P_1, …, P_5. The QS, LS and SS single out P_3 (green) as best distribution forecast, whereas the CRPS and *improper* linear score reward most the uniform and generalized Pareto distribution forecasts, P_1 and P_5, respectively. This shows again that *strictly proper* scoring rules may not always give the exact same ranking of distribution forecasts. The last column lists the scoring rules for when the forecaster quotes $P = Q$. This is the maximum attainable value for each scoring rule and equals a Monte Carlo estimate of $\mathcal{S}(Q,Q)$ in Equation 13. This column confirms the improperness of the linear score (Winkler, 1969). LinS awards a higher score to P_5 than to the true distribution. If we subtract each column P_1, …, P_5 from this last column, we yield score divergences, $d_\mathrm{QS}(P, Q)$, $d_\mathrm{LS}(P, Q)$, $d_\mathrm{SS}(P, Q)$, and $d_\mathrm{CRPS}(P, Q)$ for P_1, …, P_5.

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 (1,950 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 7 displays the discharge forecasts for a short but representative period of the $n = 3{,}000$ day training record. The discharge ensemble generally envelops the measured streamflows. Some models issue negative forecasts as a result of linear bias-correction as recommended by Raftery et al. (2005).

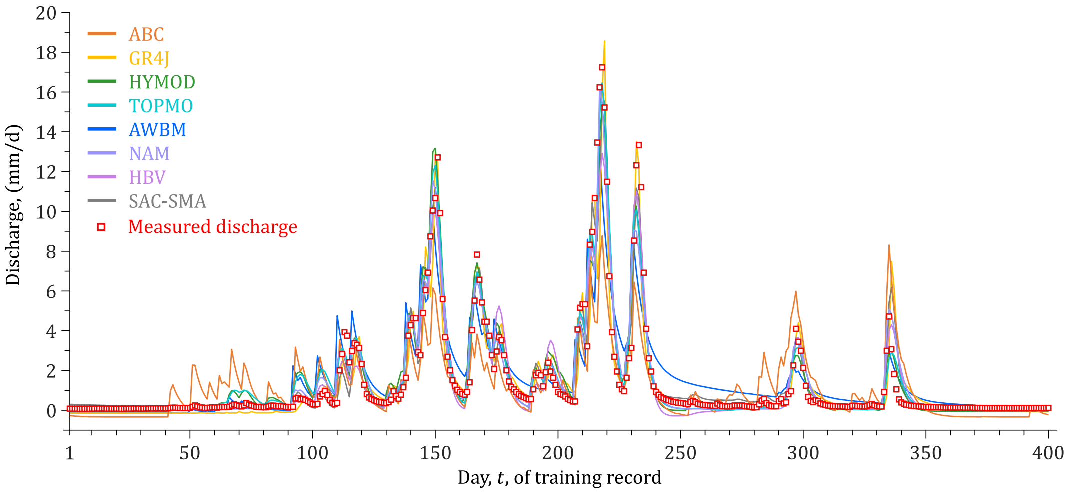

Figure 7. Streamflow simulations of the ABC, GR4J, HYMOD, TOPMO, AWBM, NAM, HBV, and SAC-SMA conceptual watershed models for a short but representative period of the $n = 3,\ 000$ day training record. The solid red circles are daily measured streamflows.

Table 7
Parametric Expressions, Settings, and Shape Parameters of (1) $\mathcal{G}$eneralized $\mathcal{N}$ormal, (2) $\mathcal{L}$og $\mathcal{N}$ormal, (3) $\mathcal{T}$runcated $\mathcal{N}$ormal, (4) $\mathcal{G}$amma, and (5) $\mathcal{G}$eneralized $\mathcal{E}$xtreme $\mathcal{V}$alue PDFs of BMA Density

Formulation	Settings	Shape parameters, $\boldsymbol{\psi}$
1. $f_{\mathcal{GN},k}(y\vert\mu_k,s_k^2,\tau_k) = \frac{\tau_k}{2s_k\Gamma(\tau_k^{-1})}\exp\left[-\frac{\vert y-\mu_k\vert^{\tau_k}}{s_k^{\tau_k}}\right]$	$\mu_k = y_k$	a: $c, \tau_1, \ldots, \tau_K$ b: $c_1, \ldots, c_K, \tau_1, \ldots, \tau_K$[a]
2. $f_{\mathcal{LN},k}(y\vert\mu_k,v_k^2) = \frac{1}{y\cdot v_k\sqrt{2\pi}}\exp\left[-\frac{\left(\log_e(y)-\mu_k\right)^2}{2v_k}\right]$	$\mu_k = \log_e\left(y_k^2\left(s_k^2+y_k^2\right)^{-1/2}\right)$ $v_k^2 = \log_e(s_k^2 y_k^{-2} + 1)$	a: c b: $c_1, \ldots, c_K$[a]
3. $f_{\mathcal{TN},k}(y\vert\mu_k,s_k^2) = \frac{1}{s_k\sqrt{2\pi}}\left[\frac{\exp\left[-\frac{1}{2}\left(y-\mu_k\right)^2/s_k^2\right]}{\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left[-\mu_k/(s_k\sqrt{2})\right]}\right]$	$\mu_k = y_k$	a: c b: $c_1, \ldots, c_K$[a,b]
4. $f_{\mathcal{G},k}(y\vert a_k,b_k) = \frac{1}{\Gamma(a_k)b_k^{a_k}}y^{(a_k-1)}\exp\left[-\frac{y}{b_k}\right]$	$\mu_k = \vert y_k\vert$ $a_k = y_k^2/s_k^2,\ b_k = s_k^2/\vert y_k\vert$	a: c b: $c_1, \ldots, c_K$[a]
5. $f_{\mathcal{GEV},k}(\bar{y}\vert\xi_k) = \left(1+\xi_k\bar{y}\right)^{-1-\frac{1}{\xi_k}}\exp\left[-\left(1+\xi_k\bar{y}\right)^{-\frac{1}{\xi_k}}\right]$	$\mu_k = y_k + \left(1-g_1(\xi_k)\right)s_k/\xi_k$ $\bar{y} = (y_k-\mu_k)/s_k$	a: $c, \xi_1, \ldots, \xi_K$ b: $c_1, \ldots, c_K, \xi_1, \ldots, \xi_K$[a,c]

[a] With a nonconstant a: group, $s_k^2 = \left(c\cdot y_k\right)^2$, or b: single, $s_k^2 = \left(c_k\cdot y_k\right)^2$, forecast variance; $k = 1, \ldots, K$. [b] Mode of truncated normal is set equal to deterministic forecast, y_k. [c] The function $g_a(\xi) = \Gamma(1-a\xi)$. Note that $f_k(y\vert\mu_k,s_k^2,\xi_k) = s_k^{-1}f_k(\bar{y}\vert\xi_k)$.

Let β_k, y_{kt}, and $f_k(y\vert y_{kt}, \psi_k)$ denote the weight, streamflow prediction and conditional density of the kth model of the ensemble, $k = (1, \ldots, K)$, at time t. The density of the multi-model forecast distribution P_t at $t = (1, \ldots, n)$ now equals a mixture distribution of the models' conditional PDFs

$$f_{P_t}(y\vert\boldsymbol{\beta},\boldsymbol{\psi}) = \sum_{k=1}^{K}\beta_k f_k(y\vert y_{kt},\psi_k), \tag{43}$$

centered on the weighted-average forecast

$$\mu_{P_t} = \sum_{k=1}^{K}\beta_k y_{kt} \tag{44}$$

with weights $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)^\top$ on the probability simplex, $\Delta^{K-1} = \left\{\boldsymbol{\beta}\in\mathbb{R}^K : \beta_1+\cdots+\beta_K = 1;\ \beta_k\geq 0 \text{ for } k = 1,\ldots,K\right\}$ and shape parameters ψ_k of each model's predictive PDF assembled in the array $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_K)^\top$. Equations 43 and 44 are known as the BMA forecast density and BMA model forecast, respectively (Raftery et al., 2005). Table 7 lists parametric expressions of the component PDFs $f_k(y\vert y_{kt}, \psi_k)$; $k = (1, \ldots, K)$ of the BMA mixture density of Equation 43. The location parameters of the generalized normal, lognormal, gamma and GEV conditional PDFs are defined so that their means coincide exactly with the model's deterministic forecasts y_{kt} at each time t, where $k = 1, \ldots, K$. We cannot center the truncated normal PDF this way and instead set y_{kt} equal to the mode of this distribution. The shape parameter τ_k of the generalized normal distribution determines its kurtosis. A value of $\tau_k = 2$ results in a normal distribution (albeit with variance $\sigma_k^2/2$), a value of $\tau_k = 1$ gives a Laplace distribution and $\tau_k \to \infty$ converges to a uniform PDF on $[y_{kt} - \sigma_k, y_{kt} + \sigma_k]$ with a zero density outside this interval. Thus, the larger the value of τ_k, the more peaked the PDF of the kth model will be and the tighter its prediction intervals around y_{kt}. The truncated normal distribution is bounded at zero and assumes an infinite upper streamflow bound. The shape parameter ξ of the GEV distribution controls its tail behavior. For $\xi = 0$, $\xi > 0$, and $\xi < 0$ we yield the Gumbel (unbounded), Fréchet (lower tail bounded) and reversed Weibull (upper tail bounded) distributions, respectively. The maximum likelihood values of the BMA weights $\hat{\boldsymbol{\beta}}$ and shape parameters $\hat{\boldsymbol{\psi}}$ of the component PDFs of Table 7 are determined for the n = 3,000 day training record using MCMC simulation with the DREAM algorithm (Vrugt, 2016; Vrugt et al., 2008) as part of the `MODELAVG` toolbox of Vrugt (2018). This software package returns the performance metrics, scoring functions, scoring rules and prediction intervals of the BMA distribution forecasts for the calibration and 2,000 day evaluation periods. Details of their computation appear in Appendix J.

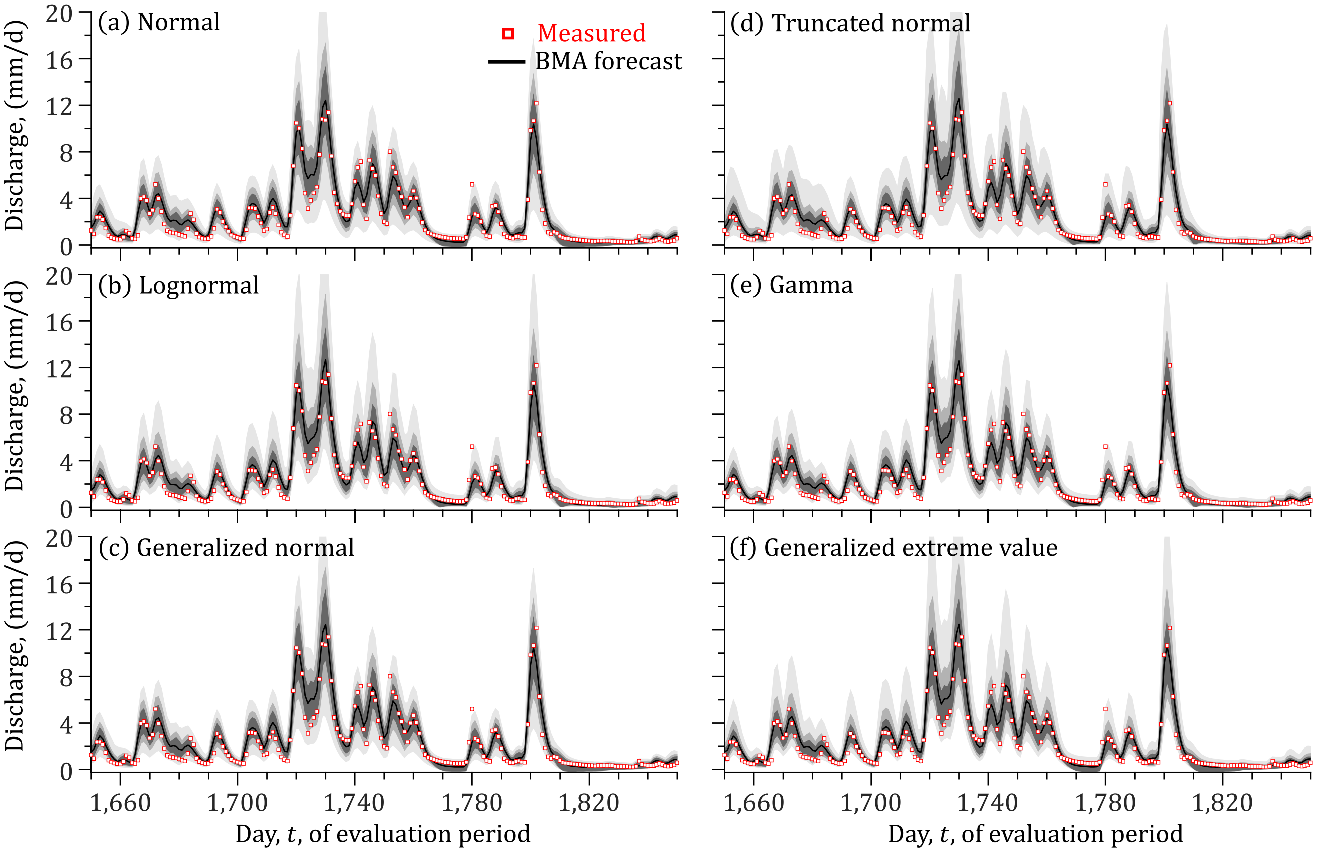

Figure 8. Weighted-average BMA forecast (black line) and 50% (dim gray), 75% (medium gray) and 95% (light gray) BMA prediction intervals for a 200-day evaluation period using the (a) normal, (b) lognormal, (c) generalized normal, (d) truncated normal, (e) gamma, and (e) generalized extreme value distribution with a nonconstant variance. Red squares are discharge measurements.

Figure 8 presents traces of the weighted-average BMA forecast (black line) and associated 50%, 75%, and 95% prediction intervals (gray regions) for a small portion of the evaluation period using the (a) normal ($\tau = 2$), (b) lognormal, (c) generalized normal, (d) truncated normal, (e) gamma, and (f) generalized extreme value PDFs of Table 7 with a nonconstant forecast variance. The BMA distribution forecasts of the different conditional PDFs describe the measured discharge record quite well. The 95% prediction intervals encapsulate the large majority of the discharge observations (red squares) and exhibit a variable spread in accordance with the use of a nonconstant forecast variance. The BMA prediction intervals are comparatively large for the peak flows but decrease rapidly in spread with flow level and collapse to a small region surrounding the mean forecast (solid black line) for the lowest streamflows. Intuitively, one may hypothesize that asymmetric and/or truncated component PDFs hold the greatest promise for describing the skewed discharge data but the different graphs do not seem to support this conjecture. At first sight, the BMA distribution forecasts of the lognormal, truncated normal, gamma and GEV conditional PDFs look very similar to those obtained from the symmetric normal and generalized normal distributions and have an about equal coverage (shown next). Thus, the symmetry of the normal and generalized normal distributions does not impair the BMA model's ability to describe the measured hydrograph. This testifies to the adaptability and flexibility of the BMA mixture distribution. On closer inspection, the normal and generalized normal distributions display the smallest 95% BMA prediction intervals of all conditional PDFs listed in Table 7. This is followed by the gamma, lognormal, truncated normal and GEV distributions (shown next). Their positive skew enlarges substantially BMA prediction uncertainty of the largest streamflows. Last but not least, the weighted-average BMA forecast (black line) appears largely unaffected by the choice of conditional PDF and even disappears behind the streamflow data toward the end of the 200-day evaluation period. Overall, the normal and generalized normal conditional PDFs yield the sharpest BMA distribution forecasts and, thus, appear to receive most support by the discharge data. But visual interpretation alone is not enough for judging the quality of distribution forecasts.

Table 8
Evaluation Period: Time-Averaged Values of the Strictly Proper Scoring Rules of Table 5 for the BMA Density Forecast $f_{P_t}(y|\boldsymbol{\beta},\boldsymbol{\psi});t = (1,\ldots,n)$ of Equation 43 Using the Conditional PDFs of Table 7 With a Nonconstant Group and Single Variance

		Normal		Lognormal		Gen. Normal		Trunc. Normal		Gamma		GEV	
		Group	Single	Group	Single	Group	Single	Group	Single	Group	Single	Group	Single
Scoring rules	QS	0.071	−0.131	0.156	0.245	0.143	−0.158	−0.126	0.026	0.133	0.086	−0.033	−0.037
	LS	0.000	0.011	0.124	0.125	0.014	0.026	0.008	0.020	0.095	0.092	0.141	0.132
	SS	1.100	1.114	1.178	1.179	1.117	1.123	1.100	1.133	1.156	1.153	1.188	1.182
	CRPS	−0.216	−0.216	−0.207	−0.207	−0.215	−0.216	−0.212	−0.220	−0.210	−0.209	−0.208	−0.206
	ES	−0.346	−0.353	−0.335	−0.333	−0.349	−0.351	−0.337	−0.403	−0.339	−0.337	−0.343	−0.343
Metrics	R_1	0.119	0.109	0.124	0.132	0.112	0.117	0.135	0.129	0.130	0.132	0.105	0.102
	C_v	0.558	0.573	0.602	0.627	0.564	0.556	0.580	0.692	0.588	0.588	0.675	1.005
	C	0.974	0.983	0.966	0.972	0.970	0.970	0.980	0.986	0.970	0.973	0.961	0.955
	W	1.713	1.819	1.959	1.956	1.698	1.722	1.718	2.235	1.843	1.821	2.036	2.032
Summary	$\ell(\hat{\boldsymbol{\Phi}}\|\boldsymbol{\omega})$	0.282	14.64	171.4	173.1	19.46	36.42	11.05	27.27	131.6	128.0	195.4	183.6
	RMSE	0.588	0.596	0.579	0.579	0.590	0.591	0.581	0.596	0.580	0.580	0.586	0.583
	NSE	0.877	0.874	0.881	0.881	0.876	0.876	0.880	0.874	0.880	0.880	0.878	0.879
	KGE	0.807	0.807	0.807	0.811	0.808	0.808	0.812	0.809	0.808	0.810	0.805	0.805
	d	9	16	9	16	17	24	9	16	9	16	17	24

Note. We also list the performance metrics, R_1, C_v, C, and W of Table 1 and report the BMA log-likelihood $\ell\big(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}}|\boldsymbol{\omega}\big)$ and RMSE, NSE, and KGE of the weighted-average BMA forecast. The bottom row lists the number d of BMA model parameters.

The BMA distribution forecasts offer an excellent application of the scoring rules. For each scoring rule, we compute a time-averaged mean score for an independent $n = 2{,}000$ day evaluation period

$$\bar{S}_{\mathrm{XX}}(\mathbf{P},\boldsymbol{\omega}) = \frac{1}{n}\sum_{t=1}^{n} S_{\mathrm{XX}}(P_t,\omega_t), \tag{45}$$

where $\mathbf{P} = \{P_1, \ldots, P_n\}$ is the collection of probabilistic forecasts derived from the BMA model. Table 8 reports the mean values of the quadratic, logarithmic, spherical, continuous ranked probability and energy scoring rules for the BMA mixture distribution of Equation 43 using the normal, lognormal, generalized normal, gamma, and GEV predictive PDFs with a nonconstant group and single variance. We also present the performance metrics of Table 1 and report the BMA log-likelihood $\ell\big(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}}|\boldsymbol{\omega}\big)$ and Root Mean Square Error (RMSE), NSE and KGE of the weighted-average BMA forecast of Equation 44 using the maximum likelihood BMA weights $\hat{\boldsymbol{\beta}}$ and shape parameters, $\hat{\boldsymbol{\psi}}$.

The tabulated data highlight several important findings.

1. The time-averaged values of the *strictly proper* scoring rules display only a small variation within and between the predictive PDFs of Table 7.
2. A common slope c for the models' discharge-variance relationships, $s_k^2 = (c \cdot y_k)^2$ suffices for maximizing the overall quality of the BMA distribution forecasts. A model-dependent slope, c_k; $k = 1, \ldots, K$, does not improve the values of the *strictly proper* scoring rules.
3. The scoring rules differ in their rankings of the conditional PDFs but are confident in their selection of the best predictive distributions for the BMA model. The LS and SS favor the GEV distribution, the QS and ES assign the largest rewards to the lognormal distribution and the CRPS is maximized by both conditional PDFs. This dichotomy in the "best" conditional PDF is a result of the strong resemblance in the maximum likelihood BMA mixture distributions of the lognormal and GEV PDFs at each t (not shown). Differences in the rankings of the conditional PDFs by each scoring rule highlights differences in how they evaluate distribution forecasts, and are impacted by structural errors of the BMA model. The preference of the scoring rules for asymmetric component PDFs testifies to the skewed distribution of measured streamflows.

Table 9
Pearson Correlation Coefficients of BMA Log-Likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ and Time-Averaged Values of Scoring Rules, Performance Metrics of Table 1 and Root Mean Square Error, NSE, and KGE Scoring Functions

	QS	LS	SS	CRPS	ES	R_1	C_v	C	W	RMSE	NSE	KGE
$\ell(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}}|\boldsymbol{\omega})$	0.447	1.000	0.637	0.825	−0.440	−0.360	−0.756	0.018	−0.367	0.771	−0.769	−0.758

4. The BMA likelihood confirms the findings of the scoring rules and advantages of using a skewed distribution for the models' predictive PDFs. The GEV distribution maximizes $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ followed by the lognormal, gamma, generalized normal, truncated normal and normal distributions.
5. The reliability R_1 of the BMA mixture distribution displays only small differences between the conditional PDFs. The variation in this measure of statistical consistency of the BMA distribution forecasts and materialized discharge events is reminiscent of some of the scoring rules. The smallest value of R_1 is obtained by the GEV distribution followed by the normal, generalized normal, lognormal, gamma and truncated normal distributions.
6. The coefficient of variation C_v of the BMA mixture distribution is remarkably similar for the conditional PDFs of Table 7 with exception of the GEV distribution which attains largest C_v values. The use of a model-dependent slope of the discharge-variance relationship tends to increase the coefficient of variation of the BMA distribution forecasts. The BMA mixture distribution admits an analytic expression for its C_v (see Appendix J), otherwise time-averaged values of this metric are difficult to compute and interpret, particularly for skewed distribution forecasts with mean close to zero. As a result, the C_v correlates only mildly ($r = 0.589$) with the average spread W of the 95% BMA prediction intervals.
7. All distributions of Table 7 achieve an approximately adequate coverage at significance level $\alpha = 0.05$. This inspires confidence in the formulation of the BMA log-likelihood function and maximum likelihood BMA weights and shape parameters inferred by the DREAM algorithm. The coverage uses only the α-quantiles of a predictive distribution and, thus, does not judge the overall quality of a distribution forecast. Section 7 turns the coverage into a *proper* scoring rule.
8. The combined performance metrics single out the normal and/or generalized normal distribution. These two conditional PDFs minimize the coefficient of variation C_v and width W of the BMA distribution forecasts with a reliability R_1 and coverage C comparable to the other distributions.
9. Tabulated data highlight the limitations of the conjectured sharpness principle of Gneiting et al. (2007). The generalized normal PDF minimizes the spread of the 95% BMA prediction intervals, but this distribution does not receive most support from the scoring rules and log-likelihood. Sharpness is a property of the forecast distribution only, thus is an *improper* scoring rule.
10. The RMSE, NSE, and KGE appear almost unaffected by the choice of conditional PDF and forecast variance. The mean functional $T_{\text{mean}}(P)$ voluntarily relinquishes information about the underlying distribution of P for a performance assessment of the weighted-average forecast. Scoring functions such as the RMSE, NSE, and KGE are unresponsive therefore to the members' predictive distributions unless the choice of conditional PDF strongly controls the mean forecast. The BMA log-likelihood, on the other hand, is strictly local as well but does not suffer this same limitation as it evaluates the PDF of the BMA distribution forecasts at the materialized streamflows. As a result, $\ell\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}}|\boldsymbol{\omega}\right)$ can differentiate between the six conditional PDFs.
11. The logarithmic scoring rule is an affine transformation of the BMA log-likelihood and, thus, $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ and LS exhibit a perfect linear relationship (see Table 9). The BMA log-likelihood also displays a relatively strong relationship with the CRPS.

Strictly proper scoring rules equip hydrologists with an arsenal of robust measures for the quality-of-fit of distribution forecasts. The QS, LS, SS, CRPS, and ES favor the use of skewed component PDFs for the BMA mixture distribution of the measured hydrograph. This is a testament to the asymmetric (zero-bounded) distribution of discharge. The R_1, C_v, C, and W performance metrics, on the other hand, favor a symmetric conditional PDF for each model of the BMA ensemble. Summary metrics of the weighted-average BMA forecasts are not helpful for choosing an appropriate conditional PDF.

To provide insights into the temporal behavior of the scoring rules, please consider Figure 9 which presents time series plots of the quadratic (green), logarithmic (red), spherical (blue), continuous ranked probability (yellow), and energy (cyan) scoring rules. The scoring rules vary dynamically in time and display an intermittent pattern of

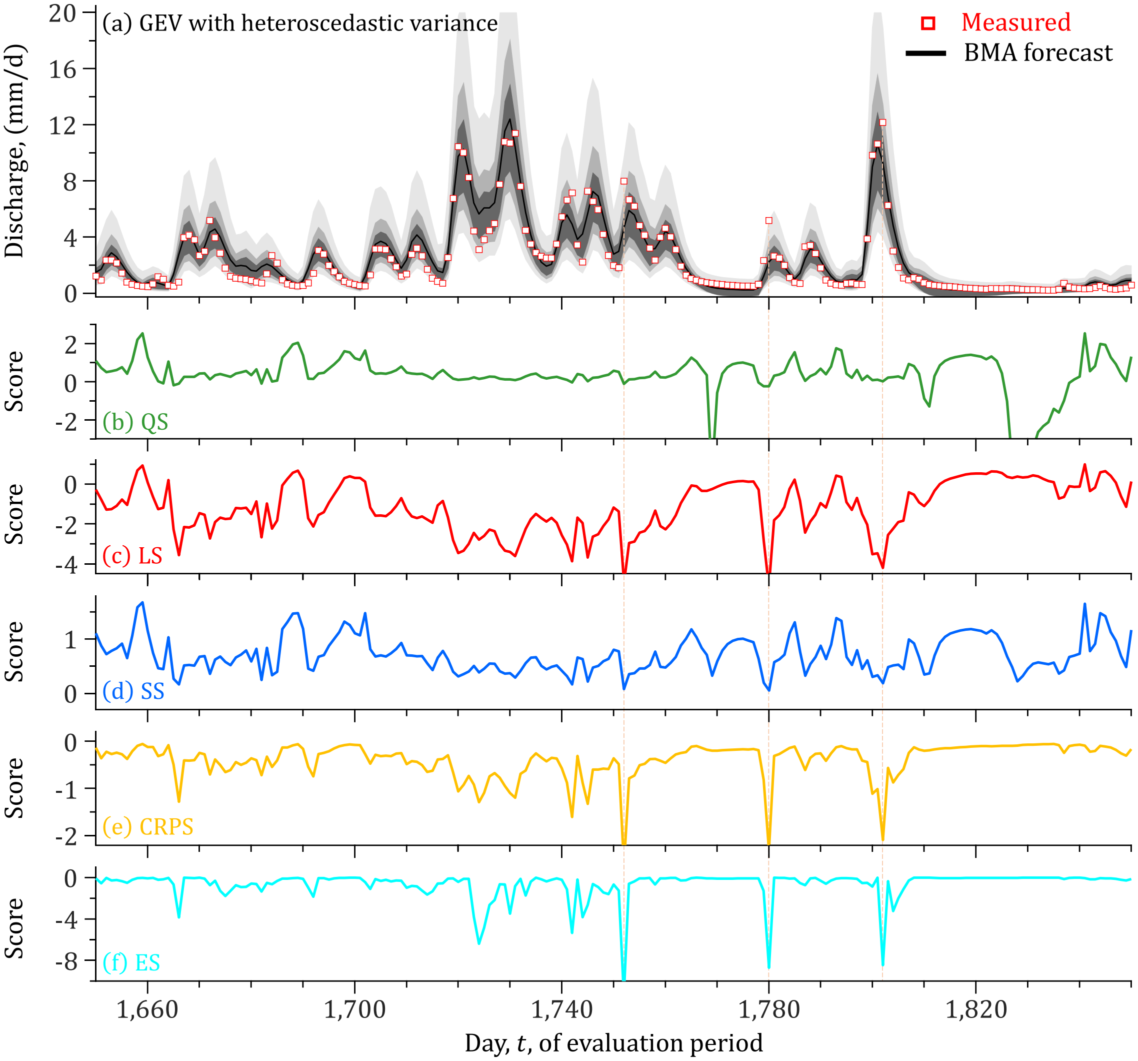

Figure 9. (a) 50%, 75%, and 95% BMA prediction intervals for the 180-day evaluation period using the GEV distribution with a nonconstant forecast variance and associated traces of the (b) quadratic, (c) logarithmic, (d) spherical, (e) continuous ranked probability, and (f) energy scoring rules.

smooth day-to-day variations followed by sudden small and occasionally larger fluctuations independent of flow level. The LS appears most responsive to the BMA forecast distribution and expresses the largest interdaily variability across the hydrograph and evaluation record. This is followed by the spherical and then the QS. Their two traces show many similarities confirming a strong affinity of their mathematical definitions. The CRPS and ES appear unresponsive at low and intermediate streamflows and react strongly when the discharge measurements materialize in the tails of the BMA distribution forecasts.

The scoring rules show important similarities and differences. The QS, LS, and SS attain their largest values when the discharge event materializes within the high probability density region (dark gray) of the BMA mixture distribution and this distribution forecast is compact so as to maximize $f_P(\omega)$. These two conditions are easiest to satisfy at the lowest flow levels, hence, the QS, LS, and SS attain their largest values in the nondriven slow part of the hydrograph. But the more compact or leptokurtic the BMA distribution forecast is, the larger its L_2-norm $\| f_P \|_2$ will be. Then, if as on days 1,827–1,830 the discharge event materializes outside the high density region of the BMA mixture distribution this results in very low values of the QS and to a lesser extent the SS. The CRPS and ES also value compactness but reward more closeness of the BMA distribution forecasts and the verifying discharge measurements. They attain values of near zero when the BMA mixture distribution centers on the materializing discharge event and this distribution forecast is as sharp as possible. Most scoring rules attain their lowest values when the discharge events materialize in the upper or lower tails of the BMA forecast distribution (see dotted vertical lines).

In principle, any of the *strictly proper* rules should suffice for evaluating the quality of the distribution forecasts of the BMA model. Yet, as model structural errors will obscure the uniqueness of the true forecast distribution there will be benefits to using multiple scoring rules simultaneously. This yields a more robust ranking of the conditional PDFs. The question which scoring rule(s) to use in practice depends on application specifics and goals. For the BMA model, the scoring rules will yield a similar characterization of predictive uncertainty and the LS will maximize general utility.

6.2.4. Case Study VI: The Flow Duration Curve

One obvious hydrologic application of the scoring rules is the 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

$$R_X(x) = \mathbb{P}(X > x) = \int_x^{\infty} f_X(t)\,\mathrm{d}t = 1 - \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = 1 - F_X(x), \tag{46}$$

and reconfirm that the FDC is the complement of the streamflow CDF, $F_X(x)$ (Vogel & Fennessey, 1994). Existing studies compare the slope (McMillan et al., 2017; Sawicz et al., 2011; Yadav et al., 2007), discharge values at given exceedance probabilities (Vogel & Fennessey, 1994), concavity index (Zhang et al., 2016) and ratio of high and low flow percentiles (Olden & Poff, 2003; Sadegh et al., 2015) of measured and simulated FDCs. These approaches elicit only partial information from the measured FDC, lack theoretical rigor and do not support formal Bayesian estimation.

Suppose $P_{\boldsymbol{\theta}} = (y_1, \ldots, y_n)$ is the simulated discharge time series of a hydrologic model indexed by the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\top}$. Now, if we enter the above relationship, $F_X(x) = 1 - R_X(x)$, in Equation 40

$$d_{\mathrm{CRPS}}(P,Q) = \int_0^{\infty} \big((1 - R_P(z)) - (1 - R_Q(z))\big)^2 \mathrm{d}z, \tag{47}$$

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

$$d_{\mathrm{FDC}}(P,Q) = \int_0^{\infty} \big(R_Q(z) - R_P(z)\big)^2\,\mathrm{d}z = d_{\mathrm{CRPS}}(P,Q). \tag{48}$$

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

$$d_{\mathrm{FDC}}(P,Q) = \mathbb{E}_{P,Q}[|\,y - \omega|] - \frac{1}{2}\big(\mathbb{E}_P[|\,y - y^*|] + \mathbb{E}_Q[|\omega - \omega^*|]\big), \tag{49}$$

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 49

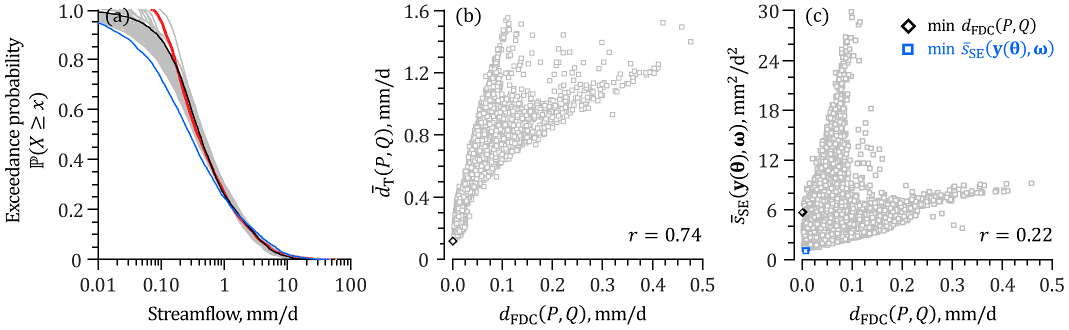

Figure 10. The flow duration curve (FDC) divergence score: (a) measured FDC (red dots) and SAC-SMA simulated reliability functions (gray lines) of the 50 discharge records with lowest values of $d_{\mathrm{FDC}}(P, Q)$ and (b, c) scatter diagrams of $d_{\mathrm{FDC}}(P, Q)$ and (b) mean taxicab distance $\bar{d}_{\mathrm{T}}(P,Q)$ and (c) mean squared residual $\bar{s}_{\mathrm{SE}}(\mathbf{y}(\boldsymbol{\theta}),\boldsymbol{\omega})$ of Equation 51 using all 25,000 simulated discharge records. The black and blue reliability functions correspond to the minima of the FDC divergence and mean squared residual, respectively.

$$d_{\mathrm{FDC}}(P,Q) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i-\omega_j| - \frac{1}{2}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{|y_i-y_j|+|\omega_i-\omega_j|\}, \tag{50}$$

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

We illustrate the usefulness of the FDC divergence of Equation 50 by application to the Sacramento Soil Moisture Accounting (SAC-SMA) model of Burnash et al. (1973). Appendix K presents our numerical implementation of the SAC-SMA model along with a description of its parameters. We simulate the rainfall-discharge relationship of the Leaf River watershed for the $n = 3{,}000$-day training record 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. A 1-year spin-up period eliminates the impact of state variable initialization.

We draw $m = 10{,}000$ vectors $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^m\}$ from the prior parameter ranges using Latin hypercube sampling. Then, for each parameter vector $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)^\top$ we simulate the daily discharge record, $P_{\boldsymbol{\theta}} = \left(y_1(\boldsymbol{\theta}),\ldots,y_n(\boldsymbol{\theta})\right)^\top$ and compute the FDC divergence $d_{\mathrm{FDC}}(P_{\boldsymbol{\theta}}, Q)$ and mean squared residual

$$\bar{s}_{\mathrm{SE}}(\mathbf{y}(\boldsymbol{\theta}),\boldsymbol{\omega}) = \frac{1}{n}\sum_{t=1}^{n}\left(\omega_t - y_t(\boldsymbol{\theta})\right)^2, \tag{51}$$

Figure 10 compares measured (red) and SAC-SMA simulated (gray lines) reliability functions of the 3,000-day training record using only the 50 best parameter vectors according to $d_{\mathrm{FDC}}(P, Q)$. The black and blue lines correspond to the minima of $d_{\mathrm{FDC}}(P, Q)$ and $\bar{s}_{\mathrm{SE}}(\mathbf{y}(\boldsymbol{\theta}),\boldsymbol{\omega})$, respectively. To benchmark the FDC divergence, we also present scatter plots of (b) $d_{\mathrm{FDC}}(P, Q)$ and the mean taxicab distance, $\bar{d}_{\mathrm{T}}(P,Q) = \frac{1}{n}\sum_{t=1}^{n}|\omega'_t - y'_t(\boldsymbol{\theta})|$, where $\omega'_1,\ldots,\omega'_n$ and $y'_1,\ldots,y'_n$ are ordered records of measured and simulated streamflows, respectively, and (c) $d_{\mathrm{FDC}}(P, Q)$ and $\bar{s}_{\mathrm{SE}}(\mathbf{y}(\boldsymbol{\theta}),\boldsymbol{\omega})$. Each square is a different parameter vector.

The FDCs of the 50 ensemble members with lowest values of $d_{\mathrm{FDC}}(P, Q)$ are in close agreement with the measured reliability function of the Leaf River. The relatively large discrepancies for the lowest flows are a result of the logarithmic streamflow scale. The FDC divergence correlates quite well ($r = 0.74$) with the taxicab distance $\bar{d}_{\mathrm{T}}(P,Q)$ of measured and simulated reliability functions. This confirms that lower values of $d_{\mathrm{FDC}}(P, Q)$ generally imply better agreement with the measured FDC. But the FDC divergence is not exactly equal to a Manhattan (=Euclidean) distance and, thus, we find minimum values of $d_{\mathrm{FDC}}(P, Q)$ over a range of taxicab distances. The FDC divergence correlates poorly with the mean squared residual of the simulated discharge records and attains its lowest value (black triangle) well removed from the minimum of $\bar{s}_{\mathrm{SE}}\,(\mathbf{y}(\boldsymbol{\theta}),\boldsymbol{\omega})$ (blue square). The optimal reliability functions of the FDC divergence (black line) and mean squared residual (blue line) differ substantially. This confirms once again that purely statistical metrics of the goodness of fit compromise the SAC-SMA model's ability to describe hydrologically relevant signatures of watershed behavior. This well-known trade-off between "statistical" and "hydrologic" model training testifies to the added value of hydrograph functionals for hydrologic model training. As a side note, we could apply differential weighting $\int_0^\infty w(z)\,\mathrm{d}z < \infty$ to the exceedance probabilities

$$d_{\mathrm{WFDC}}(P,Q) = \int_0^\infty w(z)\big(R_Q(z) - R_P(z)\big)^2\,\mathrm{d}z, \tag{52}$$

to emphasize particular flow levels of the FDC. When $P,Q \in \mathcal{P}_1$ the above expression is a valid generalization of the FDC divergence score. So-called localization and censored scoring rules help evaluate a model's ability in predicting accurately certain parts of a distribution, for example, the frequency of extreme events (de Punder et al., 2023; Diks et al., 2011).

6.2.5. Case Study VII: Hydrograph Recession Analysis

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)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -ay^b, \tag{53}$$

where a (d$^{-1/b}$) and b (−) are unknown recession constants that depend on watershed characteristics. BN77 estimate a and b using a $\log_b - \log_b$ graph of $-\mathrm{d}y/\mathrm{d}t$ versus y but this graphical interpretation of recession hydrographs is not without practical problems and has been subject to active debate in the hydrologic literature (Kirchner, 2009; Roques et al., 2017; Rupp & Selker, 2006; Tashie et al., 2020; Thomas et al., 2013). In fact, we do not need to know the values of a and b for a meaningful model evaluation 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. This necessitates use of a bivariate form of the scoring rules.

Suppose F_Q and F_P are bivariate CDFs of the *true* and *simulated* $\big(\log_{10}(y),\log_{10}(-\mathrm{d}y/\mathrm{d}t)\big)$ point clouds and $\boldsymbol{\omega} = (\omega_1,\omega_2)^\top$ is a sample drawn at random from Q. The multivariate CRPS is

$$S_{\mathrm{MCRPS}}(P,\boldsymbol{\omega}) = -\int_\Omega (F_P(\mathbf{u}) - \mathbb{1}\{\boldsymbol{\omega} \le \mathbf{u}\})^2\mathrm{d}\mathbf{u}, \tag{54}$$

with associated divergence function

$$d_{\mathrm{MCRPS}}(P,Q) = \int_\Omega \big(F_P(\mathbf{u}) - F_Q(\mathbf{u})\big)^2\mathrm{d}\mathbf{u}, \tag{55}$$

where $\mathbf{u} \in \Omega \subseteq \mathbb{R}^2$. By expanding the integrand of Equation 54 we yield a term $\int_\Omega 1\{\boldsymbol{\omega} \le \mathbf{u}\}^2\mathrm{d}\mathbf{u}$, which depends only on $\boldsymbol{\omega}$ and not F_P. Thus,

$$S^*_{\mathrm{MCRPS}}(P,\boldsymbol{\omega}) = -\int_\Omega F_P^2(\mathbf{u})\mathrm{d}\mathbf{u} + 2\int_\Omega F_P(\mathbf{u})\mathbb{1}\{\boldsymbol{\omega} \le \mathbf{u}\}\mathrm{d}\mathbf{u}, \tag{56}$$

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 $\mathbf{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_m\} \in \mathbb{R}^{2\times m}$ of $\big(\log_{10}(y),\log_{10}(-\mathrm{d}y/\mathrm{d}t)\big)$ data pairs, $\mathbf{y}_i = (y_{i1},y_{i2})^\top \in \mathbb{R}^{2\times 1}$, then the bivariate eCDF at point $\mathbf{u} = (u_1,u_2)^\top$ is equal to

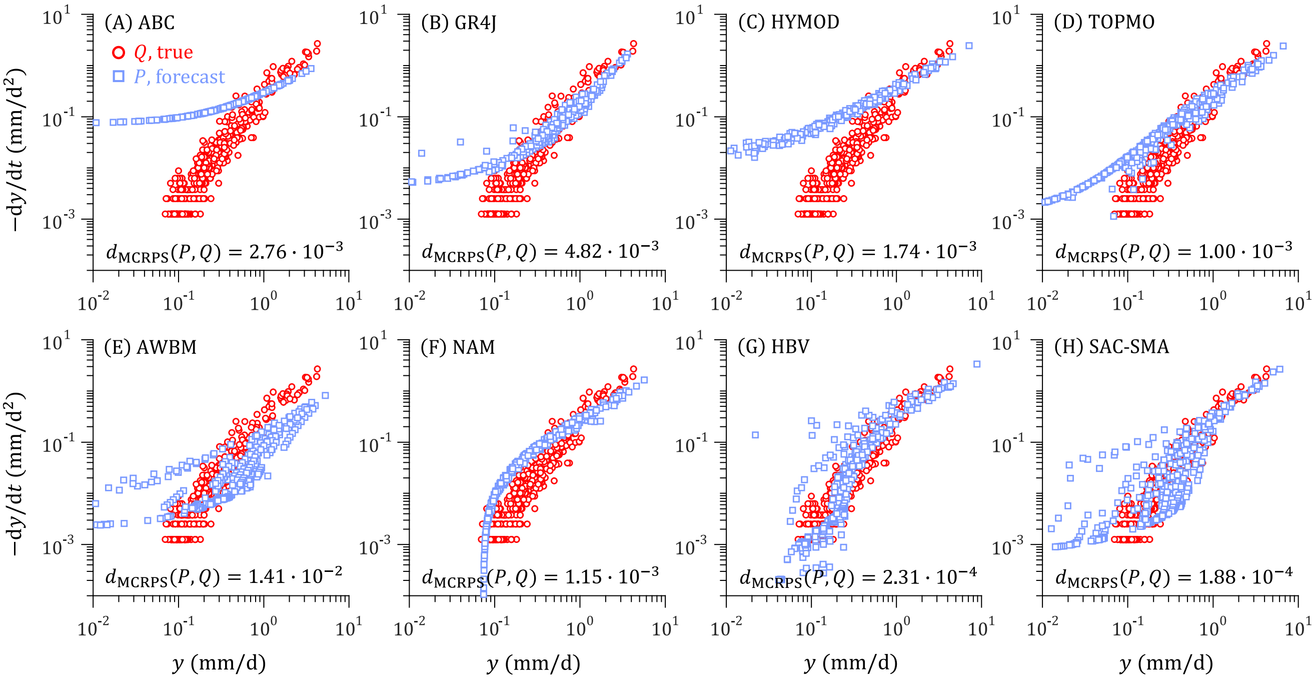

Figure 11. Scatter plots (blue squares) of the $\log_{10}(y) - \log_{10}(-\mathrm{d}y/\mathrm{d}t)$ relationship for the different models of the BMA discharge ensemble. The red dots correspond to the measured discharge record.

$$F_P(\mathbf{u}) = F_P(\mathbf{u}|\mathbf{Y}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{y_{i1} \le u_1, y_{i2} \le u_2\}. \tag{57}$$

Langrené and Warin (2021) present two algorithms for CPU-efficient estimation of multivariate eCDFs necessitating $\mathcal{O}(m)$ operations if samples are sorted. This formulation can be inserted in Equation 55 and yields the statistical distance $d_{\mathrm{MCRPS}}(P, Q)$ of the bivariate distributions P and Q. The mass-conservative numerical solver of the watershed models will help remedy artifacts of BN77 analysis due to a constant time step between successive discharge observations.

We illustrate the application of Equations 55 and 57 to hydrograph recession analysis using the BMA model ensemble. Figure 11 compares measured (red circles) and simulated (blue squares) $\log_{10}(y)$ and $\log_{10}(-\mathrm{d}y/\mathrm{d}t)$ point clouds of the $K = 8$ watershed models. Each graph also lists the corresponding value of the divergence score $d_{\mathrm{MCRPS}}(P, Q)$. BN77 scatterplot analysis demonstrates that the recession curves of (a) NAM, (b) GR4J, (c) HYMOD, and (d) TOPMO compare poorly to the measured y and $-\mathrm{d}y/\mathrm{d}t$ point cloud. This may be a side-effect of linear bias-correction, nevertheless is sufficient grounds for removal of these four models from the BMA ensemble. The recession curves of the HBV and SAC-SMA models display the best match with the measured point cloud. The MCRPS divergence score, $d_{\mathrm{MCRPS}}(P, Q)$, confirms our visual assessment of the point clouds and attains its lowest values for the SAC-SMA model with HBV as runner-up. The MCRPS divergence score can serve as a loss function for model training to promote hydrologic characterization of recession periods.

6.3. Quantile and Interval Scoring Rules

We may summarize a distribution forecast of a continuous variable using predictive quantiles. Suppose a forecaster quotes quantiles $\mathbf{r} = (r_1, \ldots, r_k)^\top$ and x materializes, then the reward equals $S(\mathbf{r}; x)$ and the expected score $\mathcal{S}(\mathbf{r}; P)$ under probability measure $P \in \mathcal{P}$ becomes (Gneiting & Raftery, 2007)

$$\mathcal{S}(\mathbf{r}; P) = \int S(\mathbf{r}; x)\mathrm{d}P(x). \tag{58}$$

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

$$\mathcal{S}(q_1,\ldots,q_k;P) \geq \mathcal{S}(r_1,\ldots,r_k;P) \tag{59}$$

for all real numbers $r_1, \ldots, r_k$ and $P \in \mathcal{P}$. Gneiting and Raftery (2007) present a general form of a scoring rule for quantiles among which $S^{\tau}(r;x) = (\mathbb{1}\{x \leq r\} - \tau)(x - r)$ of Equation 33 discussed in the context of the CRPS.

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

$$S^{\alpha}_{\mathrm{IS}}(P,\omega) = (l-u) - \frac{2}{\alpha}(l-\omega)\mathbb{1}\{\omega \leq l\} - \frac{2}{\alpha}(\omega-u)\mathbb{1}\{\omega \geq u\}, \tag{60}$$

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 $S^{\alpha}_{\mathrm{IS}}(P,\omega)$ is positively oriented and incurs a penalty, the size of which depends on significance level α, if ω is outside the $[u, l]$ prediction interval. The first term, $l - u$, of Equation 60 confirms that $S^{\alpha}_{\mathrm{IS}}(P,\omega)$ rewards narrow prediction intervals. The interval score is *proper* and not *strictly proper* as shown in Appendix L. To guarantee an accurate description of the *true* forecast distribution one would need to compute $S^{\alpha}_{\mathrm{IS}}(\mathbf{P},\omega)$ for many different α-values. But this is a daunting task (see Christoffersen, 1998) 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

$$\overline{S}^{\alpha}_{\mathrm{IS}}(\mathbf{P},\boldsymbol{\omega}) = \frac{1}{n}\sum_{t=1}^{n} S^{\alpha}_{\mathrm{IS}}(P_t,\omega_t). \tag{61}$$

6.4. Multivariate Forecasts

Up until now, we have considered forecast distributions of only a single variable of interest, say, discharge, and verifying data measured at different times. The overall skill score, $\overline{S}(\mathbf{P},\boldsymbol{\omega})$, is then a time-averaged mean score as in Equation 45. We can expand this approach to multi-variable forecasts by treating the distribution of each variable, say discharge, soil moisture content, groundwater table and aspects of stream water chemistry, separately. But for such multi-variable forecasts, $P \in \mathcal{P} \in \mathbb{R}^{\zeta}$, we can also resort to multivariate scoring rules such as the energy score of Gneiting and Raftery (2007)

$$S_{\mathrm{ES}}(P,\boldsymbol{\omega}) = \frac{1}{2}\mathbb{E}_P\left[\|\mathbf{y}-\mathbf{y}^*\|_2^{\eta}\right] - \mathbb{E}_P\left[\|\mathbf{y}-\boldsymbol{\omega}\|_2^{\eta}\right], \tag{62}$$

where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^{\zeta}$, $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_{\zeta})^{\top}$, $\eta \in (0, 2)$, and $\mathbf{y} = (y_1,\ldots,y_{\zeta})^{\top}$ and $\mathbf{y}^* = (y_1^*,\ldots,y_{\zeta}^*)^{\top}$ are independent copies of the ζ-variate distribution, $P \in \mathcal{P}_{\eta}$. For $\zeta = 1$, the above expression reduces to $S_{\mathrm{ES}}(P, \omega)$ in Equation 41. The multivariate form of the energy score is a *strictly proper* score (Székely, 2003) and for $\eta \rightarrow 2$ reduces to

$$S_{\mathrm{SE}}(P,\boldsymbol{\omega}) = -\left\|\boldsymbol{\mu}_P - \boldsymbol{\omega}\right\|_2^2, \tag{63}$$

where $\boldsymbol{\mu}_P = (\mu_{P,1},\ldots,\mu_{P,\zeta})^{\top}$ is the mean of the distribution forecast. In analogy to the numerical form of the CRPS in Equation 37, 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)

$$S_{\mathrm{ES}}(P,\boldsymbol{\omega}) = \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^{\eta} - \frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{y}_i - \boldsymbol{\omega}\right\|_2^{\eta}. \tag{64}$$

Dawid (1998) and Dawid and Sebastiani (1999) studied scoring rules that depend only on the mean, $\boldsymbol{\mu}_P \in \mathbb{R}^{\zeta \times 1}$, and covariance matrix, $\boldsymbol{\Sigma}_P \in \mathbb{R}^{\zeta \times \zeta}$ of forecast distribution P. Their divergence function, $d_{\text{DDS}}(P, Q)$, in Equation 3 is linked to the *proper* scoring rule (Dawid & Sebastiani, 1999)

$$S_{\text{DSS}}(P,\boldsymbol{\omega}) = -\log_e(|\boldsymbol{\Sigma}_P|) - (\boldsymbol{\omega} - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_P), \tag{65}$$

with entropy function $H(P) = -\log_e(|\boldsymbol{\Sigma}_P|) - \zeta$. The DSS is equal to the unnormalized log-likelihood of a multivariate normal density. For a univariate forecast, $\zeta = 1$, the DSS simplifies to $S_{\text{DSS}}(P,\omega) = -\log_e(\sigma_P^2) - (\omega - \mu_P)^2 / \sigma_P^2$, where, again, μ_P and σ_P^2 are the mean and variance of the forecast distribution.

A multivariate forecast does not necessarily imply use of different variables but can equal a single variable that is forecasted at many different sites. Suppose $i, j \in (1, \ldots, \zeta)$ are linear indexes of points in a two-dimensional grid of ζ sites. Scheuerer and Hamill (2015) investigate the accuracy of the forecast distribution P in describing the spatial structure of gridded measurements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\zeta)^\top$ of wind speed forecasts. They introduced the so-called variogram score of order $\varsigma > 0$

$$S_{\text{VS}}^{\varsigma}(P,\boldsymbol{\omega}) = -\sum_{i=1}^{\zeta} \sum_{j=1}^{\zeta} w_{ij} \left(|\omega_i - \omega_j|^{\varsigma} - \mathbb{E}_P\left[|y_i - y_j|^{\varsigma} \right] \right)^2, \tag{66}$$

where $w_{ij} \geq 0$ is a nonnegative weight attached to the (i, j)th pair of sites and y_i and y_j are the ith and the jth elements (sites) of a random vector that is distributed according to P. The weights can be used to emphasize or downplay specific aspects of the distribution forecast. The variogram scoring rule is analogous to the squared error $S_{\text{SE}}(P, \boldsymbol{\omega})$ of Equation 63 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 P is given in the form of a m-member ensemble $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ the second term of $S_{\text{VS}}^{\varsigma}(P,\boldsymbol{\omega})$ can be approximated by $\mathbb{E}_P\left[|y_i - y_j|^{\varsigma} \right] = \frac{1}{m}\sum_{k=1}^{m} |y_{ik} - y_{jk}|^{\varsigma}$. The variogram score is more sensitive than the other scoring rules to multivariate site dependencies. But as it uses site differences, it is insensitive to location, thus, is a *proper* and not *strictly proper* scoring rule. Distribution forecasts that differ only in their mean will yield the same value of $S_{\text{VS}}^{\varsigma}(P,\boldsymbol{\omega})$.

We could ignore time as governing variable of the rainfall-discharge transformation and treat a modeled streamflow time series as a multivariate forecast $\mathbf{y} = (y_1, \ldots, y_\zeta)^\top$ with ζ the length n of the simulated record. Then, Equations 64–66 will yield $S_{\text{ES}}(P, \boldsymbol{\omega})$, $S_{\text{DSS}}(P, \boldsymbol{\omega})$ and $S_{\text{VS}}^{\varsigma}(P,\boldsymbol{\omega})$, respectively.

Atmospheric scientists have put forth the use of skill scores (Briggs & Ruppert, 2005; Murphy, 1973a). Skill scores are unitless and express the model's performance relative to a hypothetical ideal and reference (benchmark) model. Skill scores may help communicate model performance (Knoben, Freer, & Woods, 2019) but are improper even if the underlying scoring rule S is proper.

7. Decomposition of Scoring Rules for Categorical Forecasts

Strictly proper scoring rules condense the accuracy of a distribution forecast P to a scalar with attractive statistical properties. This compression simplifies forecast verification, model evaluation and likelihood function selection (e.g., Vrugt et al., 2022), but makes it difficult to detect which attributes of P are deficient and in need of improvement. Scoring rule decomposition yields attributes related to the overall consistency, accuracy and precision of distribution forecasts. Such deconstruction is well-known in the context of scoring functions such as the mean squared error (Gupta et al., 2009; Hodson et al., 2021). Kull and Flach (2015) presents a decomposition of the logarithmic and Brier scoring rules into an epistemic and aleatoric loss term. Decomposition of the expected loss into a calibration and refinement loss has stimulated the development of calibration methods (Bella et al., 2013). Refinement loss consists of an uncertainty and a resolution term (DeGroot & Fienberg, 1983; Murphy, 1973b). 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 $p \in [0, 1]$. Once we observe $\omega \in \Omega$, we assign a score $S(p,\omega)\mathbf{:}\ \mathcal{P}_2 \times \Omega \rightarrow \mathbb{R}$ to the prediction. Thus, ω is either 0 (*no rain*) or 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

$$\mathbb{E}[S(p,\omega)] = \mathbb{E}[\mathbb{E}[S(p,\omega)|\mathcal{D}]]. \tag{67}$$

For the Brier or QS, $S_{\mathrm{QS}}(p,\omega) = -\sum_{k=1}^{2}(\delta_{\omega k} - p_k)^2$, where $\delta_{\omega k} = 1$ if $\omega = k$ and $\delta_{\omega k} = 0$ otherwise, the conditional expectation can be decomposed to (see Appendix M1)

$$\mathbb{E}\big[S_{\mathrm{QS}}(p,\omega)\big] = -\mathrm{Var}[\omega] + \mathrm{Var}[\mathbb{E}[\omega|\mathcal{D}]] - \mathbb{E}\big[(p(\mathcal{D}) - \mathbb{E}[\omega|\mathcal{D}])^2\big], \tag{68}$$

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, $S(\mathbf{p},\omega)\mathbf{:}\ \mathcal{P}_m \times \Omega \rightarrow \mathbb{R}$ of a categorical forecast of $m \geq 2$ events to yield

$$\mathbb{E}[S(\mathbf{p},\omega)] = \underbrace{H(\overline{\mathbf{p}})}_{1} + \underbrace{\mathbb{E}[d(\overline{\mathbf{p}},\boldsymbol{\pi})]}_{2} - \underbrace{\mathbb{E}[d(\mathbf{p},\boldsymbol{\pi})]}_{3}, \tag{69}$$

where $\overline{\mathbf{p}} = (\overline{p}_1,\ldots,\overline{p}_m)^\top$ is the unconditional probability of ω (called climatology) 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 $\mathcal{P}_m$. The three terms are nonnegative for *strictly proper* scoring rules and referred to as (a) uncertainty (of ω), (b) resolution (or sharpness) and (c) reliability (Bröcker, 2009). Entropy and resolution have a positive effect on $\mathbb{E}[S(\mathbf{p},\omega)]$, whereas reliability decreases the expected score. Note that the minus sign of $-\mathrm{Var}[\omega]$ in Equation 68 has vanished from the uncertainty term as Bröcker (2009) uses $-H(\overline{\mathbf{p}})$ for the entropy function. Furthermore, our use of positively oriented scoring rules reverses the sign of the resolution and reliability terms. Appendix M2 explains in detail the three terms of Equation 69.

Equation 69 is a generalization of the well known decomposition of the Brier score of Murphy (1973b)

$$\text{1. uncertainty} = \overline{p}(1-\overline{p}) \qquad \text{2. resolution} = \mathbb{E}\big[(\overline{p} - \pi_1)^2\big] \qquad \text{3. reliability} = \mathbb{E}\big[(p - \pi_1)^2\big], \tag{70}$$

where $\pi_1 = \pi$ is the conditional probability of *rain* given p, $\pi_1 = \mathbb{P}(x = 1|\,p)$. Weijs, van Nooijen, et al. (2010) present a similar decomposition of the relative entropy, the divergence of the logarithmic scoring rule.

7.2. Case Study VII: Discharge Forecast Ensemble

We illustrate the analytic decomposition of Equation 69 by application to the multi-model ensemble of discharge forecasts displayed in Figure 7. Appendix M3 defines the unconditional $\overline{\mathbf{p}}$ and conditional $\boldsymbol{\pi}$ probabilities of the $K = 8$ watershed models. Table 10 lists the expected value of the QS, entropy, resolution and reliability of the BMA mixture distribution for the PDFs of Table 7 using a constant group and sole forecast variance, respectively. In each case, the forecast probabilities $\mathbf{p} = (p_1,\ldots,p_K)^\top$ are set equal to the maximum likelihood weights $\beta_1, \ldots, \beta_K$ of the BMA mixture distribution. Tabulated data confirm the decomposition of Equation 69. The QS (top row) is indeed equal to the sum of the entropy $H(\overline{\mathbf{p}})$ and 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 $\overline{\mathbf{p}}$ and conditional $\boldsymbol{\pi}$ event frequencies, hence, are invariant to the constituent PDFs of the BMA mixture density. With exception of the generalized normal PDF, the use of a model-dependent nonconstant forecast variance increases forecast reliability. The forecast probabilities derived from the truncated normal PDF maximize the QS. This finding contradicts results from distribution-based model evaluation in Table 8 which ranks the truncated normal PDF last for

Table 10
Time-Averaged Values of the Strictly Proper Quadratic Scoring Rule $\mathbb{E}[S_{QS}(\mathbf{p},\omega)]$ and Entropy $H(\bar{\mathbf{p}})$, Resolution $\mathbb{E}[d(\bar{\mathbf{p}},\boldsymbol{\pi})]$ and Reliability $\mathbb{E}[d(\mathbf{p},\boldsymbol{\pi})]$ of BMA Density Forecasts of Equation 43 Using the Normal, Lognormal, Generalized Normal, Truncated Normal, Gamma, and GEV Distributions With a Nonconstant Group and Single Variance

	Normal[a]		Lognormal		Gen. Normal		Trunc. Normal		Gamma		GEV	
	Group	Single	Group	Single	Group	Single	Group	Single	Group	Single	Group	Single
$\mathbb{E}[S_{\text{QS}}(\mathbf{p},\omega)]$	−0.013	0.002	−0.026	−0.015	−0.002	−0.011	0.022	0.035	−0.012	−0.009	−0.074	−0.069
$H(\bar{\mathbf{p}})$	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132
$\mathbb{E}[d(\bar{\mathbf{p}},\boldsymbol{\pi})]$	0.155	0.155	0.155	0.155	0.155	0.155	0.155	0.155	0.155	0.155	0.155	0.155
$\mathbb{E}[d(\mathbf{p},\boldsymbol{\pi})]$	0.300	0.285	0.313	0.302	0.289	0.298	0.265	0.253	0.299	0.296	0.361	0.356
Sum	−0.013	0.002	−0.026	−0.015	−0.002	−0.011	0.022	0.035	−0.012	−0.009	−0.074	−0.069

Note. The bottom row completes the decomposition of Equation 69. [a]We fix $\tau = 2$ in the PDF of the generalized normal distribution (Table 7).

the QS. This testifies to this paper's premise that only distribution-based evaluation provides an accurate assessment of model adequacy.

8. Outlook

Scoring rules guarantee a more robust and complete evaluation of hydrologic models but may satisfy other purposes as well. We describe a few avenues for future work.

8.1. 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 parameters will convey which assumption about the mean of the distribution is most supported by the data (Luke et al., 2017). Scoring rules come in handy for hypothesis testing of stationary and nonstationary flood frequency models. Suppose we use the Pearson type III distribution $P = \mathcal{P}_{\text{III}}(\mu,\sigma^2,\rho)$ for log-transformed annual maxima discharges. If we reparameterize the location μ, shape ρ and scale σ^2 of $\mathcal{P}_{\text{III}}(\mu,\sigma^2,\rho)$ 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 & Wallis, 1997; Tegos et al., 2022)

$$f_P(x,\xi,a,b) = \frac{|x-\xi|^{a-1}}{b^a\Gamma(a)}\exp\left(-b^{-1}|x-\xi|\right), \tag{71}$$

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 LS, $S_{\text{LS}}(P,\omega) = \log_b\left(f_P(\omega)\right)$, of distribution forecast $P = \mathcal{P}_{\text{III}}(\xi,a,b)$ for the materialized outcome ω_t is now equal to

$$S_{\text{LS}}(\mathcal{P}_{\text{III}}(\xi,a,b),\omega_t) = (a-1)\log_e(|\omega_t - \xi|) - a\log_e(b) - \log_e(\Gamma(a)) - b^{-1}|\omega_t - \xi|, \tag{72}$$

with Euler's number as logical choice for the score base. The model that maximizes the LS is preferred by the data. The QS, PSS, and SS are readily computed as well but a closed-form expression for the CRPS of a PIII distribution forecast is more involved (see Appendix I2)

$$\begin{aligned} S_{\text{CRPS}}(\mathcal{P}_{\text{III}}(\xi,a,b),\omega) = {} & 2\times\frac{4^{-a}b}{B(a,a)} - ab + |\omega-\xi| + 2abF_{\mathcal{G}}(|\omega-\xi|,a+1,b) \\ & -2|\omega-\xi|F_{\mathcal{G}}(|\omega-\xi|,a,b), \end{aligned} \tag{73}$$

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 quotient in the above expression is equal to $\mathbb{E}_P[|y - y^*|]$ in Equation 36 and can be rewritten using the concentration index G of Gini (1909) (McDonald & Jensen, 1979; Scheuerer & Möller, 2015)

$$\frac{1}{2}\mathbb{E}_P[|y-y^*|] = 2\times\frac{4^{-a}b}{B(a,a)} = abG = ab\frac{\Gamma\left(a+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma(a+1)}. \tag{74}$$

Then, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ and $\frac{ab}{\pi}B\left(a+\frac{1}{2},\frac{1}{2}\right)$ is a numerically more stable expression of the first term of Equation 73. For completeness, Appendix I3 also derives an analytic expression for the CRPS of a generalized extreme value distribution forecast $P = \mathcal{GEV}(\mu,\sigma^2,\xi)$ and verifying observation $\omega \in \Omega$.

8.2. Bayesian Model Selection

Let us look in more detail at the normalization constant $p(\mathcal{D}|\mathcal{H})$ of Bayes theorem

$$p(\mathcal{D}|\mathcal{H}) = \int_{\boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})\,\mathrm{d}\boldsymbol{\theta}. \tag{75}$$

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. For two competing model hypothesis, $\mathcal{H}_0$ and $\mathcal{H}_1$, with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in a d-dimensional Euclidean space, we wish to determine which hypothesis is most supported by a sample $\mathcal{D} = (\omega_1,\ldots,\omega_n)^\top$ of outcomes of $\boldsymbol{\Omega}$. The Bayes factor $B_{1,0}$ for $\mathcal{H}_1$ against $\mathcal{H}_0$

$$B_{1,0} = \frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_0)}, \tag{76}$$

summarizes the evidence provided by the data $\mathcal{D}$ in favor of hypothesis $\mathcal{H}_1$ as opposed to the null hypothesis $\mathcal{H}_0$ (Jeffreys, 1939; Kass & Raftery, 1995). Good (1952) established a simple relationship between the LS and the logarithmic value of the Bayes factor

$$\log_{\mathfrak{b}}(B_{1,0}) = \log_{\mathfrak{b}}\left(\frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_0)}\right) = S_{\mathrm{LS}}(\mathcal{H}_1,\mathcal{D}) - S_{\mathrm{LS}}(\mathcal{H}_0,\mathcal{D}), \tag{77}$$

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 Figure 2a. This confirms that P_3 is 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. For this same reason, the time-averaged values $\overline{S}_{\mathrm{LS}}(P,\boldsymbol{\omega})$ of the LS in Table 8 convey that there is overwhelming evidence for the lognormal conditional PDF of the BMA model.

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

$$p(\omega_t|\mathcal{D}^{t-1},\mathcal{H}) = \int_{\boldsymbol{\Theta}} p(\omega_t|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{D}^{t-1},\mathcal{H})\mathrm{d}\boldsymbol{\theta}. \tag{78}$$

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 LS for ω_t is now equal to

$$S_{\mathrm{LS}}(\mathcal{H},\omega_t) = \log_{\mathfrak{b}}\left(p(\omega_t|\mathcal{D}^{t-1},\mathcal{H})\right), \tag{79}$$

and the total score in Equation 77 becomes

$$S_{\mathrm{LS}}(\mathcal{H},\mathcal{D}) = \sum_{t=1}^{n}\log_{\mathfrak{b}}\left(p(\omega_t|\mathcal{D}^{t-1},\mathcal{H})\right). \tag{80}$$

and is asymptotically equivalent to Bayes information criterion (Dawid, 1984). Thus, the LS does not only help evaluate distribution forecasts but has broader application to model selection.

8.3. Sensitivity Analysis

Suppose $\boldsymbol{\chi} = (\chi_1,\ldots,\chi_m)^\top$ are the m variables of the data generating process that determines river discharge y via a mapping or aggregation function $g : \mathbb{R}^m \rightarrow \mathbb{R}$, such that $y = g(\boldsymbol{\chi})$. Not all governing factors $\boldsymbol{\chi}$ of y are known and/or observable and the available information is summarized into a vector $\mathbf{X} = (x_1,\ldots,x_r)^\top$ of r variables, where $r \ll m$. This may include rainfall and temperature data, land-surface characteristics, soil properties, and parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)^\top$ of model hypothesis $\mathcal{H}$. Sensitivity analysis is an essential step in model development and ascertains the relative importance of input factors $\mathbf{X}$ in determining model output, $\mathbf{y} = (y_1,\ldots,y_n)^\top$ (Saltelli et al., 2008). But this analysis does not relate to materialized events $\omega \in \Omega$. A question that has received much less attention is how sensitive is y with respect to $\mathbf{X}$ or more specifically, what is the gain in predictive accuracy for y when knowing $\mathbf{X}$? Truthful prediction amounts to specifying the correct conditional distribution of ω given $\mathbf{X}$ or a functional thereof. If this functional T is elicitable and $s(y, \omega)$ is a strictly consistent scoring function, then the reduction in predictive uncertainty, $\mathbb{E}_{\omega \sim Q}[s(T(P), \omega)] - \mathbb{E}_{\omega \sim Q}[s(T(P|\mathbf{X}), \omega)]$, equals the information value of $\mathbf{X}$ for y (Borgonovo et al., 2021). Fissler and Pesenti (2023) extend this notion to a score-based sensitivity measure of ω to information $\mathbf{X}$

$$\xi_s(\omega, \mathbf{X}) = \frac{\mathbb{E}_{\omega \sim Q}[s(T(P), \omega)] - \mathbb{E}_{\omega \sim Q}[s(T(P|\mathbf{X}), \omega)]}{\mathbb{E}_{\omega \sim Q}[s(T(P), \omega)]}. \tag{81}$$

This unitless sensitivity measure varies between 0 and 1 and quantifies the relative improvement in predictive accuracy when input factors $\mathbf{X}$ are optimally used. The well-known Sobol (1993) indices correspond to T_{mean} and a squared loss strictly consistent scoring function (Borgonovo et al., 2021).

8.4. Localized Scoring Rules: Extreme Events

One may only be interested in certain aspects of a forecast distribution, for example, the probability of extreme events in its lower and/or upper tail. The FDC divergence $d_{\text{WFDC}}(P, Q)$ in Equation 52 allows differential weighting of the flow levels that make up the FDC. If all weights are positive, then this localization should not sacrifice strict propriety of the FDC 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 $S^{\varsigma}_{\text{VS}}(P, \boldsymbol{\omega})$ favors an accurate probabilistic description of the powered differences between nearby sites over such differences of distant sites. But as all sites with $|i-j| \geq 3$ receive a zero weight, such formulation of the variogram score is locally proper at best (de Punder et al., 2023; Diks et al., 2011). The weight function also allows users to incorporate soft information in model evaluation.

8.5. Synthesis With Model Diagnostics

There is an urgent need for *proper* scoring rules of hydrologic functionals in support of diagnostic model evaluation. We proposed steps in this direction for hydrograph recession and flow duration curves using the bivariate form of the CRPS and FDC divergence score, respectively, but are in need of a much larger family of (strictly) consistent scoring functions and *strictly proper* scoring rules for hydrograph functionals. This paper focused attention on scoring rules but a consistent scoring function is a special case of a *proper* scoring rule that depends on the predictive distribution via a target functional only, such as the mean, median or a quantile (Gneiting & Katzfuss, 2014).

Let us restrict attention to numeral descriptors of the stream hydrograph such as the baseflow index, runoff ratio and flashiness index (Baker et al., 2004). Time-averaged values of these hydrograph functionals are commonly used for model evaluation, but this mapping of the hydrograph to a handful of points implies a significant loss of information about the watershed's response to rainfall. This loss does not have to be as colossal if we work instead with distribution functions of hydrograph functionals. The moving-block bootstrap of Kunsch (1989) will yield frequency distributions of hydrograph functionals 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 controls the smoothness of the signatures' distribution functions. Distribution-based model diagnostics is more robust and complete as it (a) compares measured and simulated signature distributions and not just their mean values, (b) acknowledges temporal variability of hydrograph functionals and (c) accounts implicitly for signature uncertainty. We can compute the CRPS divergence for each hydrograph functional separately or quantify at once the divergence of measured and simulated signature distributions using $d_{\text{MCRPS}}(P, Q)$ in Equation 55. Confidence intervals can be derived from the Dvoretzky-Kiefer-Wolfowitz-Massart inequality (Dvoretzky et al., 1956) and its extension to multivariate distributions by Naaman (2021).

8.6. Standardization of Model Evaluation Metrics

The past decades have witnessed an unbridled growth in the number of performance measures used to evaluate hydrologic models. This proliferation is in large part a result of the lack of conforming theory and principles for metric development. Widely used scoring function in hydrology such as the NSE quantify the model's ability in describing all of the measured hydrograph. This is a desirable quality for any hydrologic model but “describe all” is not an elicitable quantity nor is the common verbiage “as closely and consistently as possible.” The exception is the ideal situation where the NSE takes on a unit value and the model matches exactly any data functional whether it is statistical or hydrograph-based. In all other cases, the directive “describe all” is ambiguous and cannot be used to determine a model's success in learning watershed behavior from the verifying data. The KGE on the other hand is explicit in what it expects the model to do. The model is directed to match the mean and variance of the discharge data while simultaneously maximizing the correlation between measured and simulated streamflow records. In this context, the KGE is a large improvement over the NSE. Yet, as the KGE is a summary measure of three functionals it is not a consistent scoring function. Scoring functions should be (strictly) consistent for hydrograph functionals, in the sense that they optimize the expected score when following the directive. Scoring rules should at least be *proper* to warrant an accurate characterization of the distribution of hydrograph functionals.

Principles of elicitability and propriety from scoring functions and scoring rules in forecast verification extend to model simulation and set much higher standards for metric development of hydrograph functionals, thereby promoting metric standardization and reproducibility and reducing a model's susceptibility to misinformation and unfinished learning. There are ample opportunities to expand the use of scoring rules to hydrologic functionals derived from high-dimensional data of ground-based sensor networks and Earth-observing satellites, possibly with extensions to the spectral domain.

9. Conclusions

Scoring functions such as the NSE and KGE have found widespread application and use to quantify the agreement between a point forecast (simulation) and materialized outcome. This point-valued mapping necessarily implies a loss of information about model performance. This paper was concerned with the basic question of how we should evaluate simulation distributions of observed quantities. A simulation distribution summarizes the diversity of model responses (behaviors) across the model input space and coalesces information about model functioning, behavior, robustness, sensitivity, and uncertainty that is not available in single-valued model output. But such distributions demand a fundamentally different approach to model evaluation and diagnostics. We discussed past developments that led to the current state-of-the-art of distribution-based evaluation in hydrology and brought scoring rules to the attention of hydrologists. 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 and density forecasts, discussed the importance of scoring function elicitability and scoring rule propriety, presented diagnostically appealing *strictly proper* divergence scores for flow duration and recession curves and addressed the decomposition of scoring rules into a sharpness, reliability and entropy term.

We first summarize the main conclusions of the theoretical treatise on scoring rules. These conclusions may be known to statisticians but are of importance to model evaluation in general.

1. Any generalized entropy function, $H(P)$, has a corresponding expression for its scoring rule, $S(P, \omega)$, expected score function, $\mathcal{S}(P, Q)$, and score divergence, $d(P, Q)$.
2. If $H(P)$ is cup-shaped then the score divergence $d(P, Q)$ is a Bregman distance, strictly positive and zero only when the distribution forecast P is equal to the true but unknown distribution Q. This guarantees *strict propriety* of the scoring rule $S(P, \omega)$ and score divergence $d(P, Q)$.

3. The LS, $S_{\mathrm{LS}}(P, \omega)$, is directly related to the log-likelihood function $\ell(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\omega})$ and Bayes factor B and has negative Shannon entropy $-\mathbb{H}(P)$ as its generalized entropy function $H(P)$ and the well-known KL-divergence (or relative entropy) as its score divergence $d(P, Q)$.

The power and usefulness of distribution-based (probabilistic) model evaluation by means of *strictly proper* scoring rules was demonstrated using simple illustrative examples, 24-hr forecasts of daily rainfall and discharge distributions simulated with conceptual watershed models using Bayesian model averaging and random sampling. Our most important practical findings are as follows.

1. Diagnostically appealing distribution verification metrics such as the coefficient of variation, reliability, width and coverage do not provide a complete evaluation of distribution forecasts.
2. Scoring functions such as the RMSE, NSE, and KGE which quantify model performance using single-valued output are insensitive to the underlying distribution of this output. The likelihood is also strictly local but preserves information about the distribution of simulated quantities.
3. The *strictly proper* quadratic, logarithmic, spherical, continuous ranked probability, and energy scoring rules enable a complete evaluation of distribution forecasts and offer robust metrics for probabilistic model evaluation. This warrants a honest and fair assessment of model adequacy. Given one strictly proper scoring rule, one can construct others by affine transformation (shifting and scaling). The CRPS is a generalization of the mean absolute error to distribution forecasts.
4. The directive to describe a measured discharge record "as closely and consistently as possible" is ambiguous and cannot be used to determine a model's success in learning watershed behavior.

In analogy to statistical functionals such as the mean and variance, we coined the term hydrologic functional for a scalar-valued mapping of the catchment's response to rainfall in ways that correspond to major behavioral functions of watershed behavior. The hydrograph functionals are (a) rooted in statistical and information theory, (b) have a strong and compelling diagnostic power and (c) remove susceptibility of model evaluation to misinformation and incomplete learning. Frequency distributions of hydrograph functionals derived from a moving-block bootstrap method admit the application of *strictly proper* scoring rules to model diagnostics. In this context, we introduced diagnostically appealing *strictly proper* divergence scores for flow duration and recession curves.

1. The FDC divergence score $d_{\mathrm{FDC}}(P, Q)$ measures in a single real number the distance between a measured R_Q and simulated R_P FDC. 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 absolute error of two point measures of the survival function.
2. The bivariate form of the CRPS offers a *strictly proper* scoring rule for hydrograph recession analysis. The MCRPS divergence $d_{\mathrm{MCRPS}}(P, Q)$ measures in a single numerical value the distance between measured and simulated bivariate distributions of the time rate of change in discharge $-\mathrm{d}y/\mathrm{d}t$ and discharge y itself. This avoids many of the problems reported with analysis of the $-\mathrm{d}y/\mathrm{d}t$ and y point clouds and simplifies model diagnostics and selection.

Furthermore, we also presented a closed-form expression of the CRPS for flood frequency analysis with the Pearson type III distribution and discussed differential weighting (censoring) as a means for characterizing better the distribution of extreme events. In general, watershed model diagnostics would benefit from decision-theoretically principled hydrograph functionals.

Finally, elicitability and propriety offer two useful working paradigms for the development and application of scoring functions and rules for hydrograph functionals. These principles set universal standards for metric development thereby promoting metric standardization, reproducibility and comparative analysis across models and data sets. Furthermore, information-theoretic principled metrics reduce a model's susceptibility to misinformation and unfinished learning.

Then, a final remark. The past decades have witnessed important developments in statistics and mathematics (inverse methods!) to help bridge the gap between hydrologic theory and data. You would expect this work to fit into hydrometrics by analogy with biometrics, bibliometrics and econometrics. But the application of statistical and mathematical methods to hydrologic data does not fall under the umbrella of hydrometry as presently defined by the International Organization for Standardization as "*…the science of monitoring water in natural water* *resources*". To rectify this inconsistency, we should think of broadening the scope of hydrometry. This would create the new job title of hydrometrician.

Appendix A: On Gibbs' Inequality

Gibbs' inequality

$$\mathbb{H}(Q,P) \geq \mathbb{H}(Q) \tag{A1}$$

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 A1 may now be written in discretized form

$$\mathbb{H}(\mathbf{q}, \mathbf{p}) \geq \mathbb{H}(\mathbf{q}) \tag{A2}$$

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

$$\sum_{x \in \Omega} q(x)\log_b\left(\frac{1}{p(x)}\right) \geq \sum_{x \in \Omega} q(x)\log_b\left(\frac{1}{q(x)}\right), \tag{A3}$$

which is equal to

$$\sum_{x \in \Omega} q(x)\log_b(q(x)) - \sum_{x \in \Omega} q(x)\log_b(p(x)) \geq 0, \tag{A4}$$

and simplifies further to

$$\begin{aligned} &\sum_{x \in \Omega} q(x)\big(\log_b(q(x)) - \log_b(p(x))\big) \geq 0 \\ &\sum_{x \in \Omega} q(x)\log_b\left(\frac{q(x)}{p(x)}\right) \geq 0 \\ &\Rightarrow d_{\mathrm{KL}}(Q, P) \geq 0. \end{aligned} \tag{A5}$$

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

$$\log_b(t(x)) \leq t(x) - 1 \tag{A6}$$

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

$$-\log_b\left(\frac{q(x)}{p(x)}\right) \leq \frac{p(x)}{q(x)} - 1, \tag{A7}$$

and, thus, we yield

$$\log_b\left(\frac{q(x)}{p(x)}\right) \geq 1 - \frac{p(x)}{q(x)}. \tag{A8}$$

We can multiply both sides of Equation A8 with $q(x)$

$$\begin{aligned}\sum_{x\in\Omega} q(x)\log_b\left(\frac{q(x)}{p(x)}\right) &\geq \sum_{x\in\Omega} q(x)\left(1-\frac{p(x)}{q(x)}\right)\\ &\geq \sum_{x\in\Omega} q(x) - \sum_{x\in\Omega} p(x),\end{aligned} \tag{A9}$$

to arrive at the following inequality

$$d_{\mathrm{KL}}(Q,P) \geq 1 - \sum_{x\in\Omega} p(x). \tag{A10}$$

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

$$1 - \sum_{x\in\Omega} p(x) \geq 0 \quad \Longleftrightarrow \quad \sum_{x\in\Omega} p(x) \leq 1. \tag{A11}$$

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, (a) all non-zero $q(x)$ values will sum to one and so do their $p(x)$ values, and (b) 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: Analytic Expressions of the Relative Entropy

The relative entropy $d_{\mathrm{KL}}(P,Q)$ is a measure of the statistical distance between a probability distribution P and reference probability distribution Q. In this Appendix, we derive closed-form expressions of $d_{\mathrm{KL}}(P,Q)$ and the reverse KL-divergence $d_{\mathrm{KL}}(Q,P)$ for some well-known *forecast* P and *true* Q distributions in $\mathbb{R}^{\zeta}$. We consider univariate ($\zeta = 1$) and multivariate ($\zeta > 1$) distribution forecasts and confirm that the relative entropy does not satisfy the triangle inequality.

B1. Univariate Distribution Forecast

B1.1. Uniform Forecast and Normal True Distribution

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

$$d_{\mathrm{KL}}(Q,P) = \mathbb{H}(Q,P) - \mathbb{H}(Q), \tag{B1}$$

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

$$\begin{aligned}\mathbb{H}(Q,P) &= -\int_{a_P}^{b_P} Q(x)\log_b(P(x))\,\mathrm{d}x\\ &= -\int_{a_P}^{b_P} Q(x)\log_e\left(\frac{1}{b_P - a_P}\right)\mathrm{d}x\\ &= \log_e(b_P - a_P)\int_{a_P}^{b_P} Q(x)\mathrm{d}x\\ &= \log_e(b_P - a_P),\end{aligned} \tag{B2}$$

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

$$\begin{aligned}
\mathbb{H}(Q) &= -\int_{a_P}^{b_P} Q(x)\log_{\mathfrak{b}}(Q(x))\,\mathrm{d}x\\
&= -\int_{a_P}^{b_P} Q(x)\log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\left(x-\mu_Q\right)^2}{\sigma_Q^2}\right)\right]\mathrm{d}x\\
&= -\int_{a_P}^{b_P} Q(x)\left(\log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\right]-\frac{1}{2}\frac{\left(x-\mu_Q\right)^2}{\sigma_Q^2}\right)\mathrm{d}x\\
&= \log_e\left(2\pi\sigma_Q^2\right)^{1/2}\int_{a_P}^{b_P} Q(x)\mathrm{d}x+\frac{1}{2\sigma_Q^2}\int_{a_P}^{b_P}\left(x-\mu_Q\right)^2 Q(x)\mathrm{d}x\\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2)\times 1+\frac{1}{2\sigma_Q^2}\times\sigma_Q^2\\
&= \frac{1}{2}\log_e(2e\pi\sigma_Q^2).
\end{aligned}\tag{B3}$$

The relative entropy in units of nats is now equal to

$$d_{\mathrm{KL}}(Q,P)=\mathbb{H}(Q,P)-\mathbb{H}(Q)=\log_e(b_P-a_P)-\frac{1}{2}\log_e(2e\pi\sigma_Q^2),\tag{B4}$$

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 $d_{\mathrm{KL}}(Q,P)=\log_e(\nu)-\frac{1}{2}\log_e(2e\pi)$.

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

$$\begin{aligned}
\mathbb{H}(P,Q) &= -\int_{a_P}^{b_P} P(x)\log_{\mathfrak{b}}(Q(x))\,\mathrm{d}x\\
&= -\int_{a_P}^{b_P} P(x)\log_{\mathfrak{b}}\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\left(x-\mu_Q\right)^2}{\sigma_Q^2}\right)\right]\mathrm{d}x\\
&= -\int_{a_P}^{b_P} P(x)\left(\log_e\left[\frac{1}{\sigma_Q\sqrt{2\pi}}\right]-\frac{1}{2}\frac{\left(x-\mu_Q\right)^2}{\sigma_Q^2}\right)\mathrm{d}x\\
&= \log_e\left(2\pi\sigma_Q^2\right)^{1/2}\int_{a_P}^{b_P} P(x)\mathrm{d}x+\frac{1}{2\sigma_Q^2}\int_{a_P}^{b_P}\left(x-\mu_Q\right)^2 P(x)\mathrm{d}x\\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2)+\frac{1}{2(b_P-a_P)\sigma_Q^2}\left|-\frac{1}{3}\left(\mu_Q-x\right)^3\right|_{a_P}^{b_P}\\
&= \frac{1}{2}\log_e(2\pi\sigma_Q^2)+\frac{\left(\mu_Q-a_P\right)^3-\left(\mu_Q-b_P\right)^3}{6(b_P-a_P)\sigma_Q^2}.
\end{aligned}\tag{B5}$$

The entropy of P is equal to Equation B2 to yield

$$\begin{aligned}\mathbb{H}(P) &= -\int_{a_P}^{b_P} P(x)\log_{\mathrm{b}}(P(x))\,\mathrm{d}x \\ &= -\int_{a_P}^{b_P} P(x)\log_e\left(\frac{1}{b_P - a_P}\right)\mathrm{d}x \\ &= \log_e(b_P - a_P)\int_{a_P}^{b_P} P(x)\mathrm{d}x \\ &= \log_e(b_P - a_P),\end{aligned} \tag{B6}$$

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

$$\begin{aligned}d_{\mathrm{KL}}(P,Q) &= \mathbb{H}(P,Q) - \mathbb{H}(P) \\ &= \frac{1}{2}\log_e(2\pi\sigma_Q^2) + \frac{(\mu_Q - a_P)^3 - (\mu_Q - b_P)^3}{6(b_P - a_P)\sigma_Q^2} - \log_e(b_P - a_P).\end{aligned} \tag{B7}$$

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

B1.2. Triangular Forecast and Uniform True Distribution

Suppose that the *true* probability distribution Q of the quantity of interest x equals the uniform distribution, $Q = \mathcal{U}(a, b)$ on the closed interval $\Omega = [a, b]$ with PDF, $f_{\mathcal{U}}(x, a, b) = 1/(b - a)$ (see Figure B1). The distribution *forecast* P of x is a symmetric triangular distribution, $P = \mathcal{T}(a,b)$, with midpoint $c = (a + b)/2$ and PDF

$$f_{\mathcal{T}}(x, a, b) = \frac{2(b - a) - 2|a + b - 2x|}{(b - a)^2}, \tag{B8}$$

where $|\cdot|$ denotes the absolute value. If we enter the analytic expressions of the PDFs of P and Q into the integral of Equation 1 we yield

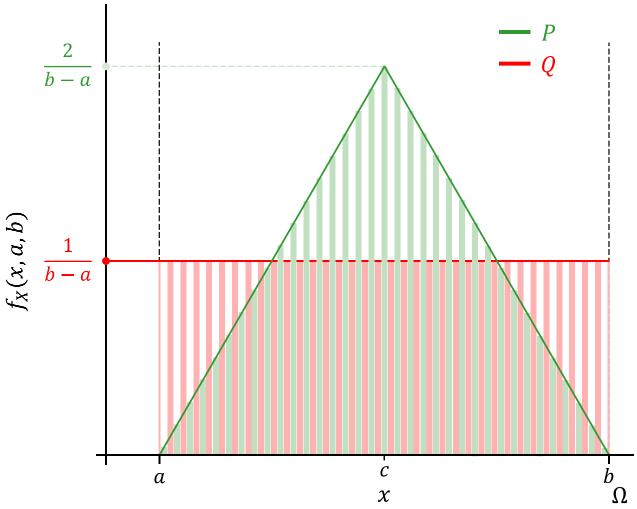

Figure B1. PDFs of the uniform true distribution $Q = \mathcal{U}(a, b)$ and symmetric triangular forecast distribution $P = \mathcal{T}(a, b)$ of random variable x on the closed sample space $\Omega = [a, b]$.

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \int_a^b \frac{1}{b-a}\log_{\mathfrak{b}}\left(\frac{(b-a)^2}{(b-a)(2(b-a)-2|a+b-2x|)}\right)\mathrm{d}x \\ &= \frac{1}{b-a}\int_a^c \log_{\mathfrak{b}}\left(\frac{b-a}{2(b-a)-2(a+b-2x)}\right)\mathrm{d}x + \frac{1}{b-a}\int_c^b \log_{\mathfrak{b}}\left(\frac{b-a}{2(b-a)-2(2x-a-b)}\right)\mathrm{d}x \\ &= \frac{1}{b-a}\int_a^c \log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{x-a}\right)\mathrm{d}x + \frac{1}{b-a}\int_c^b \log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{b-x}\right)\mathrm{d}x \\ &= \frac{1}{b-a}\left|(x-a)\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{a-b}{a-x}\right)+1\right]\right|_a^c + \frac{1}{b-a}\left|(x-b)\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{b-x}\right)+1\right]\right|_c^b. \end{aligned} \tag{B9}$$

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

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \left(\frac{\frac{1}{2}(b-a)}{b-a}\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{a-b}{\frac{1}{2}(a-b)}\right)+1\right]-0\right)+\left(0-\frac{\frac{1}{2}(a-b)}{b-a}\left[\log_{\mathfrak{b}}\left(\frac{1}{4}\frac{b-a}{\frac{1}{2}(b-a)}\right)+1\right]\right) \\ &= \frac{1}{2}\left[\log_{\mathfrak{b}}\left(\frac{1}{2}\right)+1\right]+\frac{1}{2}\left[\log_{\mathfrak{b}}\left(\frac{1}{2}\right)+1\right] \\ &= 1-\log_{\mathfrak{b}}(2). \end{aligned} \tag{B10}$$

Note that $d_{\mathrm{KL}}(Q,P)$ does not depend on the width of the sample space 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 and compute the relative entropy from Q to P we yield

$$\begin{aligned} d_{\mathrm{KL}}(P,Q) &= \int_a^b \frac{2(b-a)-2|a+b-2x|}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{(b-a)(2(b-a)-2|a+b-2x|)}{(b-a)^2}\right)\mathrm{d}x \\ &= \int_a^c \frac{4(x-a)}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{4(x-a)}{b-a}\right)\mathrm{d}x+\int_c^b \frac{4(b-x)}{(b-a)^2}\log_{\mathfrak{b}}\left(\frac{4(b-x)}{b-a}\right)\mathrm{d}x \\ &= \left|\frac{(4a-4x)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{4(a-x)}{a-b}\right)-\frac{1}{2}\right]\right|_a^{\frac{a+b}{2}}+\left|-\frac{(4b-4x)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{4(x-b)}{a-b}\right)-\frac{1}{2}\right]\right|_{\frac{a+b}{2}}^b \\ &= \left(\frac{4(a-b)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{2(a-b)}{a-b}\right)-\frac{1}{2}\right]-0\right)+\left(0+\frac{4(b-a)^2}{8(a-b)^2}\left[\log_{\mathfrak{b}}\left(\frac{2(a-b)}{a-b}\right)-\frac{1}{2}\right]-0\right) \\ &= \frac{1}{2}\left[\log_{\mathfrak{b}}(2)-\frac{1}{2}\right]+\frac{1}{2}\left[\log_{\mathfrak{b}}(2)-\frac{1}{2}\right] \\ &= \log_{\mathfrak{b}}(2)-\frac{1}{2}. \end{aligned} \tag{B11}$$

Thus, $d_{\mathrm{KL}}(P,Q)=\log_{\mathfrak{b}}(2)-\frac{1}{2}$ or $d_{\mathrm{KL}}(P,Q)=0.2787$ bits.

B1.3. Univariate Normal Forecast and True Distribution

In the special case of two univariate normal distributions, $Q\sim\mathcal{N}(\mu_Q,\sigma_Q^2)$ and $P=\mathcal{N}(\mu_P,\sigma_P^2)$, the relative entropy $d_{\mathrm{KL}}(Q,P)$ is equal to

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \frac{1}{2}\left[\log_e\left(\frac{\sigma_P^2}{\sigma_Q^2}\right) - 1 + \frac{\sigma_Q^2}{\sigma_P^2} + \frac{\left(\mu_Q - \mu_P\right)^2}{\sigma_P^2}\right] \\
&= \log_e\left(\frac{\sigma_P}{\sigma_Q}\right) + \frac{\sigma_Q^2 + \left(\mu_Q - \mu_P\right)^2 - \sigma_P^2}{2\sigma_P^2}.
\end{aligned}
\tag{B12}
$$

This expression follows directly from the multivariate case of the relative entropy discussed in the next section. Specifically, Equation B25 reduces to the above expression for a univariate normal *true* and *forecast* distribution.

B2. Multivariate Distribution Forecast

B2.1. Normal Forecast and True Distribution

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

$$
f_{\mathcal{N}}(\mathbf{x},\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\zeta/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),
\tag{B13}
$$

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 $\mathbb{R}^\zeta$. Indeed, we can write

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \mathbb{E}_Q\left[\log_e\left(\frac{Q(x)}{P(x)}\right)\right] \\
&= \mathbb{E}_Q\left[\log_e\left(f_{\mathcal{N}}(\mathbf{x},\boldsymbol{\mu}_Q,\boldsymbol{\Sigma}_Q)\right) - \log_e\left(f_{\mathcal{N}}(\mathbf{x},\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)\right)\right].
\end{aligned}
\tag{B14}
$$

To cancel the exponential function in Equation B13, 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 B14 the normal density of Equation B13 we yield

$$
\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \mathbb{E}_Q\Bigg[\log_e\left(\frac{1}{(2\pi)^{\zeta/2}\left|\boldsymbol{\Sigma}_Q\right|^{1/2}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_Q)^\top\boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q) \\
&\qquad - \left[\log_e\left(\frac{1}{(2\pi)^{\zeta/2}|\boldsymbol{\Sigma}_P|^{1/2}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\right]\Bigg] \\
&= \mathbb{E}_Q\Bigg[-\frac{\zeta}{2}\log_e(2\pi) - \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_Q)^\top\boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q) + \frac{\zeta}{2}\log_e(2\pi) \\
&\qquad + \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_P|) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\Bigg] \\
&= \frac{1}{2}\mathbb{E}_Q\Bigg[\log_e(|\boldsymbol{\Sigma}_P|) - \log_e(|\boldsymbol{\Sigma}_Q|) + (\mathbf{x}-\boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P) \\
&\qquad - (\mathbf{x}-\boldsymbol{\mu}_Q)^\top\boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\Bigg]
\end{aligned}
\tag{B15}
$$

The expected value of a constant, $\mathbb{E}[c]$, is equal to c itself and, thus, Equation B15 becomes

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \frac{1}{2}\log_e\left(\frac{|\boldsymbol{\Sigma}_P|}{|\boldsymbol{\Sigma}_Q|}\right) + \frac{1}{2}\mathbb{E}_Q\left[(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P) - \left(\mathbf{x}-\boldsymbol{\mu}_Q\right)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\right] \\ &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) + \frac{1}{2}\mathbb{E}_Q\left[(\mathbf{x}-\boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)\right] \\ &\quad - \frac{1}{2}\mathbb{E}_Q\left[\left(\mathbf{x}-\boldsymbol{\mu}_Q\right)^\top \boldsymbol{\Sigma}_Q^{-1}(\mathbf{x}-\boldsymbol{\mu}_Q)\right]. \end{aligned} \tag{B16}$$

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

$$\begin{aligned} \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] &= \mathbb{E}\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right)\right] \\ &= \mathrm{tr}\left(\mathbb{E}[\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\right) \\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\right), \end{aligned} \tag{B17}$$

where the trace function

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{\zeta} a_{ii} = a_{11} + a_{22} + \cdots + a_{\zeta\zeta}, \tag{B18}$$

computes the sum of the elements on the main diagonal of the $\zeta \times \zeta$ matrix, $\boldsymbol{\Sigma}^{-1}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]$. We can use Equation B17 to write Equation B16 as follows

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q\left[(\mathbf{x}-\boldsymbol{\mu}_P)(\mathbf{x}-\boldsymbol{\mu}_P)^\top\right]\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_Q^{-1}\mathbb{E}_Q\left[(\mathbf{x}-\boldsymbol{\mu}_Q)\left(\mathbf{x}-\boldsymbol{\mu}_Q\right)^\top\right]\right), \end{aligned} \tag{B19}$$

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

$$\begin{aligned} d_{\mathrm{KL}}(Q,P) &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_Q) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q[(\mathbf{x}-\boldsymbol{\mu}_P)(\mathbf{x}-\boldsymbol{\mu}_P)^\top]\right) \\ &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \frac{1}{2}\zeta + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q[\mathbf{x}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}_P^\top - \boldsymbol{\mu}_P\mathbf{x}^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top]\right) \\ &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \frac{1}{2}\zeta + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_Q[\mathbf{x}\mathbf{x}^\top - 2\mathbf{x}\boldsymbol{\mu}_P^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top]\right), \end{aligned} \tag{B20}$$

where the sum of the diagonal elements of the $\zeta \times \zeta$ identity matrix, $\mathbf{I}_\zeta = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}$, equals the dimension, $\zeta \in \mathbb{N}_+$, of the multivariate normal distribution and the $\zeta \times \zeta$ 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 $Q \sim \mathcal{N}_\zeta(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$, and compute numerically the expected (mean) value of the $\zeta \times \zeta$ 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 B20. 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 B20 as follows

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{x}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}\boldsymbol{\mu}^\top] - \mathbb{E}[\boldsymbol{\mu}\mathbf{x}^\top] + \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^\top] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbb{E}[\mathbf{x}^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&\Rightarrow \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top.
\end{aligned} \tag{B21}$$

Thus, Equation B20 is equal to

$$\begin{aligned}
d_{\mathrm{KL}}(Q,P) &= \frac{1}{2}\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \frac{1}{2}\zeta + \frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\Sigma}_Q + \boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top - 2\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top + \boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top)\Big) \\
&= \frac{1}{2}\Big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \zeta + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q + \boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top - 2\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top + \boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top)\Big] \\
&= \frac{1}{2}\Big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \zeta + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_Q^\top) \\
&\quad -2\mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q\boldsymbol{\mu}_P^\top) + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\boldsymbol{\mu}_P^\top)\Big].
\end{aligned} \tag{B22}$$

As corollary of Equation B17, we yield

$$\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\big) = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}), \tag{B23}$$

and, consequently, Equation B22 may be written as follows

$$\begin{aligned}
d_{\mathrm{KL}}(Q,P) = \frac{1}{2}\Big[&\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \zeta + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + \boldsymbol{\mu}_Q^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_Q - 2\boldsymbol{\mu}_Q^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P \\
&+ \boldsymbol{\mu}_P^\top\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\mu}_P\Big].
\end{aligned} \tag{B24}$$

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

$$d_{\mathrm{KL}}(Q,P) = \frac{1}{2}\Big[\log_e(|\boldsymbol{\Sigma}_Q^{-1}\boldsymbol{\Sigma}_P|) - \zeta + \mathrm{tr}(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + (\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\top\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)\Big]. \tag{B25}$$

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

The use of the natural logarithm in our derivation of Equation B25 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.

B3. Triangle Inequality

Suppose that we have three distributions, $Q = \mathcal{N}(\mu_Q,\sigma_Q^2)$, $P = \mathcal{U}(a_P,b_P)$ and $R = \mathcal{N}(\mu_R,\sigma_R^2)$. Equation B4 and B12 will help demonstrate that the relative entropy does not honor the triangle inequality, $d_{\mathrm{KL}}(Q,\ P) \le d_{\mathrm{KL}}(Q,\ R) + d_{\mathrm{KL}}(R,\ P)$. Indeed, we yield

$$\begin{aligned}
d_{\mathrm{KL}}\Big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{U}(a_P,d_P)\Big) &\leq d_{\mathrm{KL}}\Big(\mathcal{N}(\mu_Q,\sigma_Q^2),\mathcal{N}(\mu_R,\sigma_R^2)\Big) + d_{\mathrm{KL}}\big(\mathcal{N}(\mu_R,\sigma_R^2),\mathcal{U}(a_P,d_P)\big)\\
\log_e(b_P - a_P) - \frac{1}{2}\log_e(2e\pi\sigma_Q^2) &\leq \frac{1}{2}\log_e\left(\frac{\sigma_R^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + \Big(\mu_Q - \mu_R\Big)^2 - \sigma_R^2}{2\sigma_R^2} + \log_e(b_P - a_P) - \frac{1}{2}\log_e(2e\pi\sigma_R^2)\\
-\frac{1}{2}\log_e(\sigma_Q^2) &\leq \frac{1}{2}\log_e\left(\frac{\sigma_R^2}{\sigma_Q^2}\right) + \frac{\sigma_Q^2 + \Big(\mu_Q - \mu_R\Big)^2 - \sigma_R^2}{2\sigma_R^2} - \frac{1}{2}\log_e(\sigma_R^2)\\
\Rightarrow \sigma_Q^2 + \Big(\mu_Q - \mu_R\Big)^2 - \sigma_R^2 &\geq 0.
\end{aligned} \tag{B26}$$

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.

Appendix C: Numerical Computation of the KL-Divergence

To illustrate the computation and interpretation of the KL-divergence, please consider the PMFs of Q and P depicted in Figure C1. The *true distribution* Q of the quantity of interest x 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

$$\begin{aligned}
d_{\mathrm{KL}}(\mathbf{q},\mathbf{p}) &= \sum_{x=0}^{4} f_{\mathcal{B}}\left(x,4,\frac{1}{2}\right)\log_{\mathfrak{b}}\left(\frac{f_{\mathcal{B}}\left(x,4,\frac{1}{2}\right)}{f_{\mathcal{U}_{\mathrm{d}}}(x,5)}\right)\\
&= \frac{1}{16}\log_{\mathfrak{b}}\left(\frac{\frac{1}{16}}{\frac{1}{5}}\right) + \frac{1}{4}\log_{\mathfrak{b}}\left(\frac{\frac{1}{4}}{\frac{1}{5}}\right) + \frac{6}{16}\log_{\mathfrak{b}}\left(\frac{\frac{6}{16}}{\frac{1}{5}}\right) + \frac{1}{4}\log_{\mathfrak{b}}\left(\frac{\frac{1}{4}}{\frac{1}{5}}\right) + \frac{1}{16}\log_{\mathfrak{b}}\left(\frac{\frac{1}{16}}{\frac{1}{5}}\right)\\
&= \frac{1}{8}\log_{\mathfrak{b}}\left(\frac{5}{16}\right) + \frac{1}{2}\log_{\mathfrak{b}}\left(\frac{5}{4}\right) + \frac{3}{8}\log_{\mathfrak{b}}\left(\frac{30}{16}\right),
\end{aligned} \tag{C1}$$

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 $x_k \in \Omega$, the summand, $q_k\log_{\mathfrak{b}}(q_k)$, 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

$$\begin{aligned}
d_{\mathrm{KL}}(\mathbf{p},\mathbf{q}) &= \sum_{x=0}^{4} f_{\mathcal{U}_{\mathrm{d}}}(x,5)\log_{\mathfrak{b}}\left(\frac{f_{\mathcal{U}_{\mathrm{d}}}(x,5)}{f_{\mathcal{B}}\left(x,4,\frac{1}{2}\right)}\right)\\
&= \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{16}}\right) + \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{4}}\right) + \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{6}{16}}\right) + \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{4}}\right) + \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{\frac{1}{5}}{\frac{1}{16}}\right)\\
&= \frac{2}{5}\log_{\mathfrak{b}}\left(\frac{16}{5}\right) + \frac{2}{5}\log_{\mathfrak{b}}\left(\frac{4}{5}\right) + \frac{1}{5}\log_{\mathfrak{b}}\left(\frac{16}{30}\right),
\end{aligned} \tag{C2}$$

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.

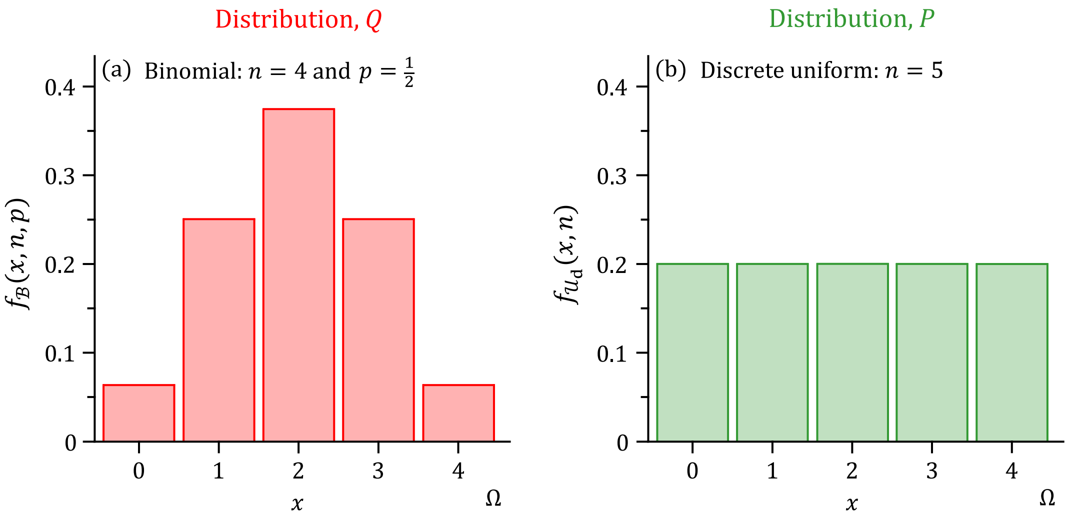

Figure C1. Illustration of the computation of the relative entropy $d_{\text{KL}}(Q, P)$ on a sample space $\Omega = \{0, 1, 2, 3, 4\}$: (a) *true distribution*, $Q = \mathcal{B}(n,p)$, and (b) *forecast distribution*, $P = \mathcal{U}_{\text{d}}(n)$. According to data, the variable of interest x follows a binomial distribution with $n = 4$, $p = \frac{1}{2}$ and probability mass function $f_{\mathcal{B}}(x,n,p) = c(n,x)p^x(1-p)^{n-x}$ where $c(a, b) = a!/b!(a-b)!$ denotes the binomial coefficient and ! is the factorial function. Theory predicts a discrete uniform distribution for x with equal density $f_{\mathcal{U}_{\text{d}}}(x,n) = 1/n$ for all $n = 5$ values.

Appendix D: Strictly Proper Categorical Scoring Rules

In this Appendix we present analytic derivations of the quadratic, logarithmic, and (pseudo)spherical scores for a categorical forecast.

D1. Quadratic Score

For a categorical forecast of m events with true probabilities, $p_1, \ldots, p_m$ issued on the probability simplex $P \in \mathcal{P}_m$ the entropy function of the QS 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). If we enter the gradient of the QS, $\nabla H(\mathbf{p}) = (2p_1,\ldots,2p_m)^\top$, into Equation 18 we yield the divergence function of the QS

$$\begin{aligned} d_{\text{QS}}(\mathbf{p},\mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q} \rangle \\ &= \sum_{k=1}^{m} q_k^2 - \sum_{k=1}^{m} p_k^2 + \langle (2p_1,\ldots,2p_m), (p_1 - q_1,\ldots,p_m - q_m) \rangle \\ &= \sum_{k=1}^{m} q_k^2 - \sum_{k=1}^{m} p_k^2 + 2\langle \mathbf{p},\mathbf{p} \rangle - 2\langle \mathbf{p},\mathbf{q} \rangle \\ &= \sum_{k=1}^{m} q_k^2 + \sum_{k=1}^{m} p_k^2 - 2\sum_{k=1}^{m} p_k q_k \\ &= \sum_{k=1}^{m} (p_k - q_k)^2 \end{aligned} \tag{D1}$$

According to Equation 20 the scoring rule of the QS is now equal to

$$\begin{aligned} S_{\mathrm{QS}}(\mathbf{p},j) &= H(\mathbf{p}) - \langle \nabla H(\mathbf{p}),\mathbf{p}\rangle + \nabla H_j(\mathbf{p}) \\ &= \sum_{k=1}^{m} p_k^2 - 2\langle \mathbf{p},\mathbf{p}\rangle + 2p_j \\ &= 2p_j - \sum_{k=1}^{m} p_k^2 \end{aligned} \tag{D2}$$

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

$$S_{\mathrm{BS}}(\mathbf{p},j) = -\sum_{k=1}^{m}\left(\delta_{jk} - p_k\right)^2 = 2p_j - \sum_{k=1}^{m} p_k^2 - 1, \tag{D3}$$

where the Kronecker symbol $\delta_{jk} = 1$ when the jth event materializes ($j = k$) and $\delta_{jk} = 0$ otherwise.

If the true probabilities $\mathbf{q} = \left(q_1,\ldots,q_m\right)^{\top}$ are known then Equation 21 will yield the expected score of the quadratic rule

$$\begin{aligned} \mathcal{S}_{\mathrm{QS}}(\mathbf{p},\,\mathbf{q}) = \sum_{j=1}^{m} q_j S_{\mathrm{QS}}(\mathbf{p},\,j) &= \sum_{j=1}^{m} q_j\left(2p_j - \sum_{k=1}^{m} p_j^2\right) \\ &= 2\sum_{j=1}^{m} p_j q_j - \sum_{j=1}^{m} q_j \sum_{k=1}^{m} p_k^2 \\ &= 2\sum_{j=1}^{m} p_j q_j - \sum_{k=1}^{m} p_k^2 \end{aligned} \tag{D4}$$

and we can confirm the divergence function of the QS in Table 2 and Equation D1

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

D2. Logarithmic Score

The LS 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 & Rainer, 2001) and ignorance score (Roulston & Smith, 2002). The entropy function of the LS $H(\mathbf{p}) = \sum_{k=1}^{m} p_k \log_{\mathfrak{b}}(p_k)$ is equal to negative Shannon entropy $-\mathbb{H}(P)$ with gradient function $\nabla H(\mathbf{p}) = \left(\log_{\mathfrak{b}}(p_1) + 1,\ldots,\log_{\mathfrak{b}}(p_m) + 1\right)^{\top}$. The divergence function $d_{\mathrm{LS}}(\mathbf{p},\,\mathbf{q})$ of the LS may now be derived from Equation 18 to yield

$$\begin{aligned}
d_{\mathrm{LS}}(\mathbf{p},\mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \langle \nabla H(\mathbf{p}), \mathbf{p}-\mathbf{q}\rangle \\
&= \sum_{k=1}^{m} q_k \log_b(q_k) - \sum_{k=1}^{m} p_k \log_b(p_k) + \big\langle \big(\log_b(p_1)+1,\cdots,\log_b(p_m)+1\big),(p_1-q_1,\cdots,p_m-q_m)\big\rangle \\
&= \sum_{k=1}^{m} q_k \log_b(q_k) - \sum_{k=1}^{m} p_k \log_b(p_k) + \langle \log_b(\mathbf{p}),\mathbf{p}\rangle + \langle 1_m,\mathbf{p}\rangle - \langle \log_b(\mathbf{p}),\mathbf{q}\rangle - \langle 1_m,\mathbf{q}\rangle \\
&= \sum_{k=1}^{m} q_k \log_b(q_k) - \sum_{k=1}^{m} q_k \log_b(p_k) + \sum_{k=1}^{m} p_k - \sum_{k=1}^{m} q_k \\
&= \sum_{k=1}^{m} q_k \big(\log_b(q_k) - \log_b(p_k)\big) \\
&= \sum_{k=1}^{m} q_k \log_b\left(\frac{q_k}{p_k}\right).
\end{aligned} \tag{D5}$$

According to Equation 20, the scoring rule of the LS is now equal to

$$\begin{aligned}
S_{\mathrm{LS}}(\mathbf{p},j) &= H(\mathbf{p}) - \langle \nabla H(\mathbf{p}),\mathbf{p}\rangle + \nabla H_j(\mathbf{p}) \\
&= \sum_{k=1}^{m} p_k \log_b(p_k) \; - \; \big\langle \big(\log_b(p_1)+1,\ldots,\log_b(p_1)+1\big),(p_1,\ldots,p_m)\big\rangle + \log_b(p_j) + 1 \\
&= \sum_{k=1}^{m} p_k \log_b(p_k) - \sum_{k=1}^{m} p_k \log_b(p_k) - \sum_{k=1}^{m} p_k + \log_b(p_j) + 1 \\
&= \log_b(p_j).
\end{aligned} \tag{D6}$$

Thus the LS has negative Shannon entropy as its generalized entropy function and the reverse KL-divergence, $d_{\mathrm{LS}}(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{m} q_k \log_b(q_k/p_k)$, as its associated score divergence. Roulston and Smith (2002) provide an information-theoretic perspective on the LS and advocate using the so-called ignorance score, $S_{\mathrm{IS}}(\mathbf{p},j) = -\log_b(p_j)$.

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

$$\mathcal{S}_{\mathrm{LS}}(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{m} q_j S_{\mathrm{LS}}(\mathbf{p},j) = \sum_{j=1}^{m} q_j \log_b(p_j). \tag{D7}$$

Next, we verify the divergence score of the logarithmic rule in Equation D5 and Table 2

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

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

D3. 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 $\mathbf{p} = (p_1,\ldots,p_m)^{\top}$

$$\|\mathbf{p}\|_\eta^1 = \left(p_1^\eta + \cdots + p_m^\eta\right)^{1/\eta}, \tag{D8}$$

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

$$\nabla H(\mathbf{p}) = \left(\frac{p_1^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}}, \frac{p_2^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}}, \ldots, \frac{p_m^{\eta-1}}{\|\mathbf{p}\|_\eta^{\eta-1}}\right)^\top, \tag{D9}$$

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 can be written as a scalar-vector product, $\nabla H(\mathbf{p}) = \|\mathbf{p}\|_\eta^{1-\eta}\mathbf{p}^{\eta-1}$. According to Equation 18 the score divergence $d_{\mathrm{PSS}}(\mathbf{p}, \mathbf{q})$ of the pseudospherical score becomes

$$\begin{aligned}
d_{\mathrm{PSS}}(\mathbf{p},\mathbf{q}) &= H(\mathbf{q}) - H(\mathbf{p}) + \langle \nabla H(\mathbf{p}), \mathbf{p} - \mathbf{q}\rangle \\
&= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{p} - \mathbf{q}\rangle \\
&= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{p}\rangle - \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle \\
&= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^1 + \|\mathbf{p}\|_\eta^{1-\eta}\|\mathbf{p}\|_\eta^{\eta} - \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle \\
&= \|\mathbf{q}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle,
\end{aligned} \tag{D10}$$

for $\mathbf{p}$, $\mathbf{q} \in \mathcal{P}_m$ and $\eta > 1$. Next, we yield the scoring rule $S_{\mathrm{PSS}}(\mathbf{p}, j)$ of the pseudospherical score

$$\begin{aligned}
S_{\mathrm{PSS}}(\mathbf{p}, j) &= H(\mathbf{p}) - \langle \nabla H(\mathbf{p}), \mathbf{p}\rangle + \nabla H_j(\mathbf{p}) \\
&= \|\mathbf{p}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{p}\rangle + \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\
&= \|\mathbf{p}\|_\eta^1 - \|\mathbf{p}\|_\eta^{1-\eta}\|\mathbf{p}\|_\eta^{\eta} + \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\
&= \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1}.
\end{aligned} \tag{D11}$$

With true probabilities, $\mathbf{q} = \left(q_1, \ldots, q_m\right)^\top$, we yield the expected pseudospherical score $\mathcal{S}_{\mathrm{PSS}}(\mathbf{p}, \mathbf{q})$

$$\begin{aligned}
\mathcal{S}_{\mathrm{PSS}}(\mathbf{p},\mathbf{q}) &= \sum_{j=1}^m q_j S_{\mathrm{PSS}}(\mathbf{p}, j) \\
&= \sum_{j=1}^m q_j \|\mathbf{p}\|_\eta^{1-\eta} p_j^{\eta-1} \\
&= \|\mathbf{p}\|_\eta^{1-\eta}\langle \mathbf{p}^{\eta-1}, \mathbf{q}\rangle,
\end{aligned} \tag{D12}$$

and we can confirm the divergence function of the pseudospherical score in Table 2 and Equation D10

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

For $\eta = 2$ the pseudospherical rule in Equation D12 reduces to the well known spherical scoring rule

$$\mathcal{S}_{\mathrm{SS}}(\mathbf{p},\mathbf{q}) = \|\mathbf{p}\|_2^{-1}\langle \mathbf{p}, \mathbf{q}\rangle, \tag{D13}$$

with associated divergence function

$$d_{\mathrm{SS}}(\mathbf{p},\mathbf{q}) = \|\mathbf{q}\|_2^1 - \|\mathbf{p}\|_2^{-1}\langle \mathbf{p},\mathbf{q}\rangle. \tag{D14}$$

Appendix E: Rainfall Data

Table E1 is taken from Hughes and Topp (2015) and summarizes a data set of $n = 346$ forecasts of 24-hr 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.

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. The raw precipitation data can be found at https://www.cawcr.gov.au/projects/

Table E1
Rainfall Data From Table 1 of Hughes and Topp (2015) for the City of Tampere, Finland

a: Original data

k	p_k	n_k	o_k	$\overline{o}_k$	n_k/n
1	0.05	46	1	0.0217	0.1329
2	0.1	55	1	0.0182	0.1590
3	0.2	59	5	0.0847	0.1705
4	0.3	41	5	0.1220	0.1185
5	0.4	19	4	0.2105	0.0549
6	0.5	22	8	0.3636	0.0636
7	0.6	22	6	0.2727	0.0636
8	0.7	34	16	0.4706	0.0983
9	0.8	24	16	0.6667	0.0694
10	0.9	11	8	0.7273	0.0318
11	0.95	13	11	0.8462	0.0376
Σ		346	81		1.0000

b: Adapted data

k	p_k	q_k
1	0.1727	0.1359
2	0.1636	0.1364
3	0.1455	0.1272
4	0.1273	0.1220
5	0.1091	0.1097
6	0.0909	0.0884
7	0.0727	0.1011
8	0.0545	0.0736
9	0.0364	0.0463
10	0.0182	0.0379
11	0.0091	0.0214
Σ	1.0000	1.0000

Table E2
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 E1

	Quadratic score			Logarithmic score, $\mathfrak{b} = 2$			Spherical score		
Fcst	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$ Equation D4	$d(\mathbf{p}, \mathbf{q})$ Equation D1	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$ Equation D7	$d(\mathbf{p}, \mathbf{q})$ Equation D5	$H(\mathbf{p})$	$\mathcal{S}(\mathbf{p},\mathbf{q})$ Equation D13	$d(\mathbf{p}, \mathbf{q})$ Equation D14
p	0.124	0.103	0.0043	−3.156	−3.351	0.045	0.352	0.323	0.0052
q	0.108	0.108	0.0000	−3.305	−3.305	0.0000	0.328	0.328	0.0000

verification/POP3/POP_3cat_2003.txt and was used by Hughes and Topp (2015) to provide a diagrammatic interpretation of the Brier scoring rule and associated score divergences.

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. Table E2 reports the entropy, expected score and score divergence of the quadratic, logarithmic and spherical scoring rules for the normalized probabilities of *true* and *forecasted* rainfall. The bottom row presents the function values when the forecaster quotes the *true* rainfall probabilities, $\mathbf{p} = \mathbf{q}$.

This case study is intentionally presented to assist in testing, benchmarking and evaluating numerical implementations of the scoring rules in coding languages other than MATLAB.

Appendix F: The Lebesgue Measure

Scoring rules for density forecasts are defined up to a so-called Lebesgue measure μ. Henri Lebesgue describes this measure in his PhD dissertation together with the Lebesgue integral (Lebesgue, 1902). To explain the Lebesgue measure, please consider Figure F1 which displays an example Lebesgue density of the standard normal distribution. The Riemann integral partitions the domain of a function into a collection of small intervals and bars are constructed to meet the height of the graph. Then in $\mathbb{R}^2$ the resulting rectangles make up the area under the graph. The Lebesgue integral also uses rectangles, but these rectangles are formed by partitioning the function's range (also called codomain) into different intervals. For each horizontal slice, a rectangle is drawn with height of the corresponding function value and width equal to the length of all intervals on the real line $\mathbb{R}$ (e.g., sample space) where the function reaches approximately this height. This horizontal slicing of the codomain leads to much more complicated sets of x values, certainly for multivariate densities. Thus, the Lebesgue definition

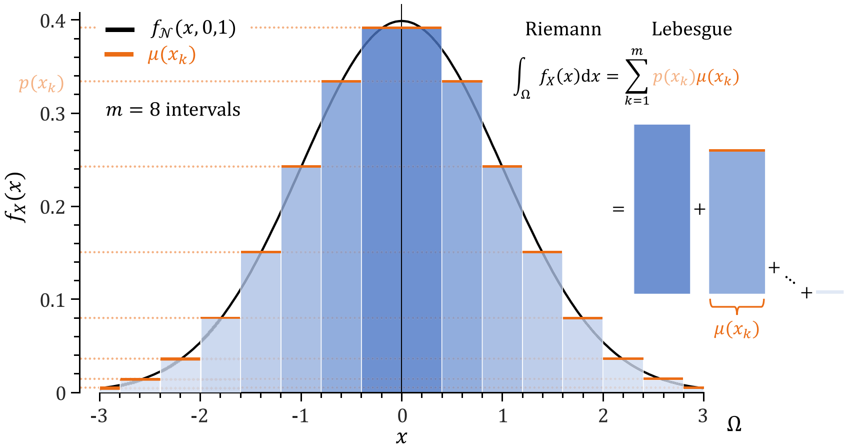

Figure F1. Illustration of the standard normal Lebesgue density on sample space $\Omega = [-3, 3]$ with range (codomain) of $\mathcal{N}(0,1)$ partitioned into $m = 8$ small intervals. We use color coding to discern the intervals of x. The Lebesgue measure $\mu(x_k)$ is equal to the length of each color coded interval containing x_k. The Lebesgue density $f_X(x)$ is constant in each interval $x_1, \ldots, x_m$ of x values. The sum of the areas of the rectangles is equal to the Lebesgue integral.

extends integral calculation to a much broader class of functions. Now, the Lebesgue measure μ is equal to the width of each slice, which, in turn, is the sum of the widths of all rectangles with the same height. For univariate densities, we can divide each interval of x 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).

Appendix G: Impropriety of Linear Scoring Rule

The linear score

$$S_{\mathrm{LinS}}(P,\,\omega) = f_P(\omega), \tag{G1}$$

may seem appealing but is improper as we will demonstrate next.

Suppose $f_P(\epsilon) = \frac{1}{2\epsilon}$ and $f_Q(\epsilon) = \frac{1}{\sqrt{2\pi}}\exp(-\epsilon^2/2)$ are Lebesgue densities of the uniform *forecast distribution* P and standard normal *true distribution* Q on the closed interval $[-\epsilon,\,\epsilon]$. According to Equation 12, the expected score of probabilistic forecasts P and Q under true distribution Q becomes

$$\begin{array}{rclrcl}
\mathcal{S}_{\mathrm{LinS}}(P,Q) & = & \displaystyle\int_{-\epsilon}^{\epsilon} S_{\mathrm{LinS}}(P,\omega)\mathrm{d}Q(\omega) & \qquad \mathcal{S}_{\mathrm{LinS}}(Q,Q) & = & \displaystyle\int_{-\infty}^{\infty} S_{\mathrm{LinS}}(Q,\omega)\mathrm{d}Q(\omega) \\
& = & \displaystyle\int_{-\epsilon}^{\epsilon} \frac{1}{2\epsilon}\frac{1}{\sqrt{2\pi}}\exp(-\omega^2/2)\mathrm{d}\omega & & = & \displaystyle\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}}\exp(-\omega^2/2)\right)^2\mathrm{d}\omega \\
& = & \displaystyle\frac{1}{2\epsilon}\frac{1}{\sqrt{2\pi}}\int_{-\epsilon}^{\epsilon}\exp(-\omega^2/2)\mathrm{d}\omega & & = & \displaystyle\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-\omega^2)\mathrm{d}\omega \\
& = & \displaystyle\frac{1}{2\epsilon}\frac{1}{\sqrt{2\pi}}\left|\sqrt{2\pi}\,\mathrm{erf}\left(\omega/\sqrt{2}\right)\right|_0^{\epsilon} & & = & \displaystyle\frac{1}{2\pi}\left|\sqrt{\pi}\,\mathrm{erf}(\omega)\right|_0^{\infty} \\
& = & \displaystyle\frac{1}{2\epsilon}\mathrm{erf}\left(\epsilon/\sqrt{2}\right) & & = & \displaystyle\frac{1}{2\sqrt{\pi}},
\end{array} \tag{G2}$$

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

$$\begin{aligned}
d_{\mathrm{LinS}}(P,Q) &= \mathcal{S}_{\mathrm{LinS}}(Q,Q) - \mathcal{S}_{\mathrm{LinS}}(P,Q) \\
&= \frac{1}{2\sqrt{\pi}} - \frac{1}{2\epsilon}\mathrm{erf}(\epsilon/\sqrt{2}).
\end{aligned} \tag{G3}$$

Figure G1 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, $\epsilon = 1.5634$, of Equation G3. Thus, the score divergence of $S_{\mathrm{LinS}}(P,\,\omega) = f_P(\omega)$ does not have a proper zero point. At this point distribution forecast $P = \mathcal{U}[-1.56, 1.56]$ is certainly not equal to the standard normal *true distribution* Q. Thus, $S_{\mathrm{LinS}}(P,\,\omega)$ is an improper scoring rule.

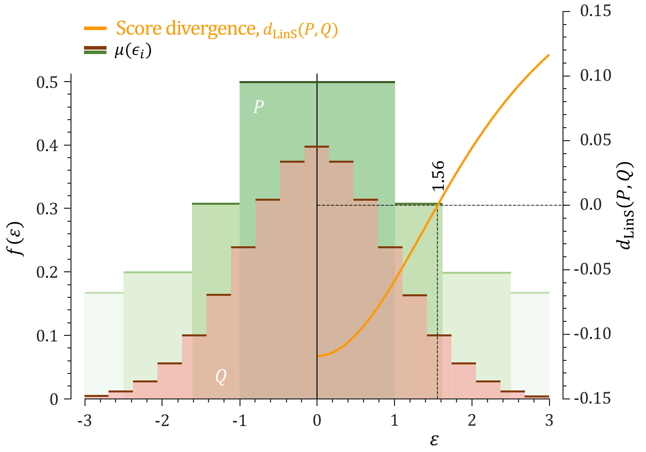

Figure G1. Score divergence $d_{\mathrm{LinS}}(P, Q)$ of the linear scoring rule $S_{\mathrm{LinS}}(P, \omega) = f_P(\omega)$ for a uniform probabilistic forecast P under a standard Gaussian true distribution Q and symmetric interval $[-\epsilon, \epsilon]$ with $\epsilon \in (0, 3]$. The horizontal green and blue lines correspond to the Lebesgue measure $\mu(\epsilon_i)$ or the length of the interval containing event ϵ_i.

Appendix H: 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 (Laio & Tamea, 2007)

$$S_{\mathrm{CRPS}}(P, \omega) = \int_0^1 S_{\mathrm{QNT}}^{\tau}(P, \omega)\,\mathrm{d}\tau, \tag{H1}$$

and, thus, we yield

$$S_{\mathrm{CRPS}}(P,\omega) = -2\int_0^1 (1\{\omega < y_\tau\} - \tau)(y_\tau - \omega)\mathrm{d}\tau. \tag{H2}$$

If we work out the product under the integral sign we yield

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

The $\tau \in [0, 1]$-quantile forecast of P is equal to $y_\tau = F_P^{-1}(\tau)$ and, thus, the above expression equals

$$\begin{aligned}
S_{\mathrm{CRPS}}(P, \omega) &= 2\int_0^1 \tau\big(F_P^{-1}(\tau) - \omega\big)\,\mathrm{d}\tau - 2\int_0^1 \mathbb{1}\{F_P(\omega) < \tau\}\big(F_P^{-1}(\tau) - \omega\big)\,\mathrm{d}\tau \\
&= 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\omega\int_0^1 \tau\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 \big(F_P^{-1}(\tau) - \omega\big)\,\mathrm{d}\tau \\
&= 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\omega\int_0^1 \tau\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau + 2\omega\int_{F_P(\omega)}^1 \mathrm{d}\tau \\
&= 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\omega\left[\frac{1}{2}\tau^2\right]_0^1 - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau + 2\omega[\tau]_{F_P(\omega)}^1 \\
&= 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - \omega - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau + 2\omega - 2\omega F_P(\omega) \\
&= 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau + \omega - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\omega F_P(\omega),
\end{aligned} \tag{H3}$$

and finally, we yield

$$S_{\mathrm{CRPS}}(P,\,\omega)\;=\;\omega(1-2F_P(\omega))+2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau-2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau. \tag{H4}$$

This concludes the derivation.

Appendix I: Analytic Expressions for Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) is given by

$$S_{\mathrm{CRPS}}(P,\,\omega)=-\int_{-\infty}^{\infty}(F_P(z)-\mathbb{1}\{z\geq\omega\})^2\mathrm{d}z, \tag{I1}$$

where $F_P(z)$ denotes the CDF of P and the indicator function, $\mathbb{1}\{a\}$, returns 1 if a is true and zero otherwise. In the next sections, we derive closed-form expressions of the CRPS for a univariate normal, $P=\mathcal{N}(\mu,\sigma^2)$, Pearson type III, $P=\mathcal{P}_{\mathrm{III}}(\mu,\sigma^2,\rho)$, and generalized extreme value, $P=\mathcal{GEV}(\mu,\sigma^2,\xi)$, distribution forecast and verifying observation $\omega\in\Omega$. Some of the derivations in this Appendix were completed before turning my attention to the quantile form of the CRPS in Equation H4. The use of this quantile form may have simplified analytic derivation of some of the CRPS expressions presented below.

I1. Continuous Ranked Probability Score for $\mathcal{N}(\mu,\sigma^2)$

If we make the convenient assumption that the probability measure is univariate normal, $P=\mathcal{N}(\mu,\sigma^2)$, then the CDF of P has a closed-form expression

$$F_P(x,\mu,\sigma^2)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \tag{I2}$$

where $\operatorname{erf}(x)$ is the error function for element x

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x\exp\left(-t^2\right)\mathrm{d}t, \tag{I3}$$

and the CRPS becomes

$$S_{\mathrm{CRPS}}\left(\mathcal{N}(\mu,\sigma^2),\omega\right)=-\int_{-\infty}^{\omega}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]\right)^2\mathrm{d}z-\int_{\omega}^{\infty}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]-1\right)^2\mathrm{d}z. \tag{I4}$$

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

$$
\begin{aligned}
-\int_{-\infty}^{\omega}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right]\right)^{2}\mathrm{d}z = -\Bigg|\frac{z}{4}+\frac{(\mu-z)}{2}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right) - \frac{(\mu-z)}{4}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)^{2} \\
+\frac{\sigma}{2\sqrt{\pi}}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right)+\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right) \\
-\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right)\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)\Bigg|_{-\infty}^{\omega} \\
= \Bigg|-\frac{z}{4}-\frac{(\mu-z)}{2}f(z,\mu,\sigma)+\frac{(\mu-z)}{4}f(z,\mu,\sigma)^{2}-\frac{\sigma}{2\sqrt{\pi}}g(z,\mu,\sigma) \\
-\frac{\sigma}{\sqrt{2\pi}}h(z,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\mu,\sigma)h(z,\mu,\sigma)\Bigg|_{-\infty}^{\omega},
\end{aligned}
\tag{I5}
$$

where $f(z,\mu,\sigma) = \operatorname{erf}(\sqrt{2}(\mu-z)/(2\sigma))$, $g(z,\mu,\sigma) = \operatorname{erf}((\mu-z)/\sigma)$ and $h(z,\mu,\sigma) = \exp(-(\mu-z)^{2}/(2\sigma^{2}))$. We follow a similar recipe for the right integral of Equation I4 to yield

$$
\begin{aligned}
-\int_{\omega}^{\infty}\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right]-1\right)^{2}\mathrm{d}z = -\Bigg|\frac{z}{4}-\frac{(\mu-z)}{2}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right) - \frac{(\mu-z)}{4}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)^{2} \\
+\frac{\sigma}{2\sqrt{\pi}}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right)-\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right) \\
-\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right)\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)\Bigg|_{\omega}^{\infty} \\
= \Bigg|-\frac{z}{4}+\frac{(\mu-z)}{2}f(z,\mu,\sigma)+\frac{(\mu-z)}{4}f(z,\mu,\sigma)^{2}-\frac{\sigma}{2\sqrt{\pi}}g(z,\mu,\sigma) \\
+\frac{\sigma}{\sqrt{2\pi}}h(z,\mu,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\mu,\sigma)h(z,\mu,\sigma)\Bigg|_{\omega}^{\infty}.
\end{aligned}
\tag{I6}
$$

Before admitting the integral bounds, we first perform limit analysis of the constituent functions

$$
\begin{aligned}
\lim_{z\to-\infty} f(z,\mu,\sigma) &= \lim_{z\to-\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)=1 \\
\lim_{z\to\infty} f(z,\mu,\sigma) &= \lim_{z\to\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma\sqrt{2}}\right)=-1 \\
\lim_{z\to-\infty} g(z,\mu,\sigma) &= \lim_{z\to-\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right)=1 \\
\lim_{z\to\infty} g(z,\mu,\sigma) &= \lim_{z\to\infty}\operatorname{erf}\left(\frac{\mu-z}{\sigma}\right)=-1 \\
\lim_{z\to-\infty} h(z,\mu,\sigma) &= \lim_{z\to-\infty}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right)=0 \\
\lim_{z\to\infty} h(z,\mu,\sigma) &= \lim_{z\to\infty}\exp\left(-\frac{(\mu-z)^{2}}{2\sigma^{2}}\right)=0.
\end{aligned}
\tag{I7}
$$

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 = {} & \left|-\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)\right.\\
& \left.-\frac{\sigma}{\sqrt{2\pi}}h(z,\,\mu,\,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\,\mu,\,\sigma)h(z,\,\mu,\,\sigma)\right|_{-\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.\\
& \left.-\frac{\sigma}{\sqrt{2\pi}}h(\omega,\,\mu,\,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(\omega,\,\mu,\,\sigma)h(\omega,\,\mu,\,\sigma)\right)\\
& -\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.\\
& \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.\\
& \left.-\frac{\sigma}{\sqrt{2\pi}}h(\omega,\,\mu,\,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(\omega,\,\mu,\,\sigma)h(\omega,\,\mu,\,\sigma)\right)\\
& -\lim_{z\to-\infty}\left(-\frac{z}{4}-\frac{(\mu-z)}{2}\times 1+\frac{(\mu-z)}{4}\times 1^{2}-\frac{\sigma}{2\sqrt{\pi}}\times 1\right.\\
& \left.-\frac{\sigma}{\sqrt{2\pi}}\times 0+\frac{\sigma}{\sqrt{2\pi}}\times 0\times 1\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.\\
& \left.-\frac{\sigma}{\sqrt{2\pi}}h(\omega,\,\mu,\,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(\omega,\,\mu,\,\sigma)h(\omega,\,\mu,\,\sigma)\right)\\
& -\lim_{z\to-\infty}\left(-\frac{z}{4}+\frac{z}{2}-\frac{\mu}{2}-\frac{z}{4}+\frac{\mu}{4}-\frac{\sigma}{2\sqrt{\pi}}\right)\\
= {} & -\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)+\frac{\mu}{4}+\frac{\sigma}{2\sqrt{\pi}},
\end{aligned}
\tag{I8}
$$

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 = {} & \left|-\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)\right. \\
& \left.+\frac{\sigma}{\sqrt{2\pi}}h(z,\,\mu,\,\sigma)+\frac{\sigma}{\sqrt{2\pi}}f(z,\,\mu,\,\sigma)h(z,\,\mu,\,\sigma)\right|_{\omega}^{\infty} \\
= {} & \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. \\
& \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}\right. \\
& \left.-\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)\right) \\
= {} & \lim_{z\to\infty}\left(-\frac{z}{4}+\frac{(\mu-z)}{2}\times(-1)\;+\;\frac{(\mu-z)}{4}\times(-1)^{2}-\frac{\sigma}{2\sqrt{\pi}}\times(-1)\right. \\
& \left.+\frac{\sigma}{\sqrt{2\pi}}\times 0+\frac{\sigma}{\sqrt{2\pi}}\times 0\times(-1)\right)-\left(-\frac{\omega}{4}+\frac{(\mu-\omega)}{2}f(\omega,\,\mu,\,\sigma)\right. \\
& +\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) \\
& \left.+\frac{\sigma}{\sqrt{2\pi}}f(\omega,\,\mu,\,\sigma)h(\omega,\,\mu,\,\sigma)\right) \\
= {} & \lim_{z\to\infty}\left(-\frac{z}{4}+\frac{z}{2}-\frac{\mu}{2}-\frac{z}{4}+\frac{\mu}{4}+\frac{\sigma}{2\sqrt{\pi}}\right) \\
& -\left(-\frac{\omega}{4}+\frac{(\mu-\omega)}{2}f(\omega,\,\mu,\,\sigma)\;+\;\frac{(\mu-\omega)}{4}f(\omega,\,\mu,\,\sigma)^{2}\right. \\
& \left.-\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)\right) \\
= {} & -\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).
\end{aligned}
\tag{I9}
$$

We can now add up Equations I8 and I9 to yield the CRPS of $P = \mathcal{N}(\mu,\sigma^{2})$ in Equation I4

$$\begin{aligned} S_{\mathrm{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) &= \Bigg(-\frac{\omega}{4} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) + \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 - \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) \\ &\quad - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma) + \frac{\mu}{4} + \frac{\sigma}{2\sqrt{\pi}}\Bigg) + \\ &\quad \Bigg(-\frac{\mu}{4} + \frac{\sigma}{2\sqrt{\pi}} + \frac{\omega}{4} - \frac{(\mu-\omega)}{2} f(\omega,\mu,\sigma) - \frac{(\mu-\omega)}{4} f(\omega,\mu,\sigma)^2 \\ &\quad + \frac{\sigma}{2\sqrt{\pi}} g(\omega,\mu,\sigma) - \frac{\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) - \frac{\sigma}{\sqrt{2\pi}} f(\omega,\mu,\sigma) h(\omega,\mu,\sigma)\Bigg) \\ &= -\frac{2(\mu-\omega)}{2} f(\omega,\mu,\sigma) - \frac{2\sigma}{\sqrt{2\pi}} h(\omega,\mu,\sigma) + \frac{2\sigma}{2\sqrt{\pi}} \\ &= -(\mu-\omega)\mathrm{erf}\left(\frac{\mu-\omega}{\sigma\sqrt{2}}\right) - \frac{\sigma\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{(\mu-\omega)^2}{2\sigma^2}\right) + \frac{\sigma}{\sqrt{\pi}}. \end{aligned} \tag{I10}$$

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 I2 as follows

$$\begin{aligned} S_{\mathrm{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) &= -(\omega-\mu)\mathrm{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right) - 2\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mu-\omega)^2}{2\sigma^2}\right) + \frac{\sigma}{\sqrt{\pi}} \\ &= -(\omega-\mu)\left[1 + \mathrm{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right] + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}} \\ &= -2(\omega-\mu)\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\omega-\mu}{\sigma\sqrt{2}}\right)\right] + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}} \\ &= -2(\omega-\mu) F_{\mathcal{N}}(\omega,\mu,\sigma^2) + (\omega-\mu) - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) + \frac{\sigma}{\sqrt{\pi}}, \end{aligned} \tag{I11}$$

which may be rearranged and simplified to

$$S_{\mathrm{CRPS}}\big(\mathcal{N}(\mu,\sigma^2),\omega\big) = \frac{\sigma}{\sqrt{\pi}} - 2\sigma^2 f_{\mathcal{N}}(\omega,\mu,\sigma^2) - (\omega-\mu)\big(2F_{\mathcal{N}}(\omega,\mu,\sigma^2) - 1\big). \tag{I12}$$

Note that the quantile form of the CRPS in Equation 34 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 $\omega \in \Omega$.

I2. Continuous Ranked Probability Score for $\mathcal{P}_{\mathrm{III}}(\mu,\sigma^2,\rho)$

The CRPS is equal to the integral of the quantile scores

$$S_{\mathrm{CRPS}}(P,\,\omega) = -\int_{-\infty}^{\infty} S^{\tau}(F_P(z) - \mathbb{1}\{z \geq \omega\})^2 \mathrm{d}z, \tag{I13}$$

where $F_P(z)$ denotes the CDF 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 CDF of $P = \mathcal{P}_{\mathrm{III}}(\xi,\,a,\,b)$ simplifies to (Tegos et al., 2022)

$$F_P(x,\ \xi,\ a,\ b)\ =\ \begin{cases} \dfrac{1}{\Gamma(a)}\gamma\big(a,b^{-1}(x-\xi)\big) & \text{if } \rho>0 \\ \dfrac{1}{\Gamma(a)}\Gamma\big(a,b^{-1}(\xi-x)\big) & \text{if } \rho<0 \end{cases} \tag{I14}$$

where $x,\ \xi,\ a,\ b \in \mathbb{R},\ a > 0,\ b > 0$ and

$$\Gamma(a,\ q) = \int_q^\infty t^{a-1} \exp(-t)\, \mathrm{d}t \tag{I15}$$

and

$$\gamma(a,\ q) = \int_0^q t^{a-1} \exp(-t)\, \mathrm{d}t \tag{I16}$$

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 $\omega \in \Omega$.

I2.1. Analytic Expression for $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)_{\rho>0}$: 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 I13 is now equal to

$$S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho>0} = -\int_0^{z_\omega} \left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)\right]^2 \mathrm{d}z - \int_{z_\omega}^{\infty} \left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z) - 1\right]^2 \mathrm{d}z, \tag{I17}$$

where $z_\omega = \omega - \xi$. We first derive an expression for the left integral using integration by parts

$$\int u\mathrm{d}v = uv - \int v\mathrm{d}u, \tag{I18}$$

with $u = \gamma^2(a,\ b^{-1}z)$ and $\mathrm{d}v = 1\mathrm{d}z$. Then

$$\mathrm{d}u = \frac{2}{z}\exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^a\gamma\left(a,b^{-1}z\right) \tag{I19}$$

and $v = z$ to yield

$$\begin{aligned} -\frac{1}{\Gamma^2(a)}\int \gamma^2\left(a,b^{-1}z\right)\mathrm{d}z &= -\frac{1}{\Gamma^2(a)}\left(z\gamma^2\left(a,b^{-1}z\right) - \int z\frac{2}{z}\exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^a\gamma\left(a,b^{-1}z\right)\right) \\ &= -\frac{1}{\Gamma^2(a)}\left(z\gamma^2\left(a,b^{-1}z\right) - 2\int \exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^a\gamma\left(a,b^{-1}z\right)\right). \end{aligned} \tag{I20}$$

The online calculator of Wolfram|Alpha (Wolfram Research, 2024) returns a closed-form expression for the indefinite integral

$$\begin{aligned}\int \exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^{a}\gamma\left(a,b^{-1}z\right)\mathrm{d}z &= \int \exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^{a}\left(\Gamma(a)-\Gamma\left(a,b^{-1}z\right)\right)\mathrm{d}z\\ &= b\exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^{a}\Gamma\left(a,b^{-1}z\right)+\frac{1}{2}ab\Gamma^{2}\left(a,b^{-1}z\right)\\ &\quad - 4^{-a}b\Gamma\left(2a,2b^{-1}z\right)-b\Gamma(a)\Gamma\left(a+1,b^{-1}z\right)+C.\end{aligned} \tag{I21}$$

If we substitute Equation I21 into Equation I20

$$\begin{aligned}-\frac{1}{\Gamma^{2}(a)}\int \gamma^{2}(a,z)\mathrm{d}z = &-\frac{1}{\Gamma^{2}(a)}\left(z\gamma^{2}\left(a,b^{-1}z\right)-2\left(b\exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^{a}\Gamma\left(a,b^{-1}z\right)+\frac{1}{2}ab\Gamma^{2}\left(a,b^{-1}z\right)\right.\right.\\ &\left.\left.-4^{-a}b\Gamma\left(2a,2b^{-1}z\right)-b\Gamma(a)\Gamma\left(a+1,b^{-1}z\right)\right)\right)+C,\end{aligned} \tag{I22}$$

and admit the integral limits

$$\begin{aligned}-\frac{1}{\Gamma^{2}(a)}\int_{0}^{z_{\omega}}\gamma^{2}\left(a,b^{-1}z\right)\mathrm{d}z = &-\frac{1}{\Gamma^{2}(a)}\left|z\gamma^{2}\left(a,b^{-1}z\right)-2\left(b\exp\left(-b^{-1}z\right)\left(b^{-1}z\right)^{a}\Gamma\left(a,b^{-1}z\right)\right.\right.\\ &\left.\left.+\frac{1}{2}ab\Gamma^{2}\left(a,b^{-1}z\right)-4^{-a}b\Gamma\left(2a,2b^{-1}z\right)-b\Gamma(a)\Gamma\left(a+1,b^{-1}z\right)\right)\right|_{0}^{z_{\omega}},\end{aligned} \tag{I23}$$

then we yield the following expression for the left integral of Equation I17

$$\begin{aligned}-\frac{1}{\Gamma^{2}(a)}\int_{0}^{z_{\omega}}\gamma^{2}\left(a,b^{-1}z\right)\mathrm{d}z = &-\frac{1}{\Gamma^{2}(a)}\left(z_{\omega}\gamma^{2}\left(a,b^{-1}z_{\omega}\right)-2\left(b\exp\left(-b^{-1}z_{\omega}\right)\left(b^{-1}z_{\omega}\right)^{a}\Gamma\left(a,b^{-1}z_{\omega}\right)\right.\right.\\ &\left.\left.+\frac{1}{2}ab\Gamma^{2}\left(a,b^{-1}z_{\omega}\right)-4^{-a}b\Gamma\left(2a,2b^{-1}z_{\omega}\right)-b\Gamma(a)\Gamma\left(a+1,b^{-1}z_{\omega}\right)\right)\right)\\ &+\frac{1}{\Gamma^{2}(a)}\left(-2\left(\frac{1}{2}ab\Gamma^{2}(a)-4^{-a}b\Gamma(2a)-b\Gamma(a)\Gamma(a+1)\right)\right)\\ = &-\frac{1}{\Gamma^{2}(a)}\left(z_{\omega}\gamma^{2}\left(a,b^{-1}z_{\omega}\right)-2\left(b\exp\left(-b^{-1}z_{\omega}\right)\left(b^{-1}z_{\omega}\right)a\Gamma\left(a,b^{-1}z_{\omega}\right)\right.\right.\\ &\left.\left.+\frac{1}{2}ab\Gamma^{2}\left(a,b^{-1}z_{\omega}\right)-4^{-a}b\Gamma\left(2a,2b^{-1}z_{\omega}\right)-b\Gamma(a)\Gamma\left(a+1,b^{-1}z_{\omega}\right)\right)\right)\\ &+\frac{2}{\Gamma^{2}(a)}\left(4^{-a}b\Gamma(2a)+b\Gamma(a)\Gamma(a+1)\right)-ab.\end{aligned} \tag{I24}$$

Next, we proceed with the right integral of Equation I17

$$\begin{aligned}-\int_{z_{\omega}}^{\infty}\left[\frac{1}{\Gamma(a)}\gamma\left(a,b^{-1}z\right)-1\right]^{2}\mathrm{d}z &= -\int_{z_{\omega}}^{\infty}\left(\frac{1}{\Gamma^{2}(a)}\gamma^{2}\left(a,b^{-1}z\right)-\frac{2}{\Gamma(a)}\gamma\left(a,b^{-1}z\right)+1\right)\mathrm{d}z\\ &= -\frac{1}{\Gamma^{2}(a)}\int_{z_{\omega}}^{\infty}\gamma^{2}\left(a,b^{-1}z\right)\mathrm{d}z+\frac{2}{\Gamma(a)}\int_{z_{\omega}}^{\infty}\gamma\left(a,b^{-1}z\right)\mathrm{d}z-\int_{z_{\omega}}^{\infty}\mathrm{d}z.\end{aligned} \tag{I25}$$

The first of three integrals is equal to Equation I22, and the other two integrals yield

$$\begin{aligned}\frac{2}{\Gamma(a)}\int_{z_{\omega}}^{\infty}\gamma\left(a,b^{-1}z\right)\mathrm{d}z-\int_{z_{\omega}}^{\infty}\mathrm{d}z &= \left|\frac{2}{\Gamma(a)}\left(z\Gamma(a)-z\Gamma\left(a,b^{-1}z\right)+b\Gamma\left(a+1,b^{-1}z\right)\right)-z\right|_{z_{\omega}}^{\infty}\\ &= \left|z-\frac{2}{\Gamma(a)}\left(z\Gamma\left(a,b^{-1}z\right)-b\Gamma\left(a+1,b^{-1}z\right)\right)\right|_{z_{\omega}}^{\infty}.\end{aligned} \tag{I26}$$

Thus, Equation I25 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)}\bigg| 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)\bigg|_{z_\omega}^{\infty} \\
&+ \bigg| z - \frac{2}{\Gamma(a)}\big(z\Gamma(a,b^{-1}z) - b\Gamma(a+1,b^{-1}z)\big)\bigg|_{z_\omega}^{\infty} \\
= &-\frac{1}{\Gamma^2(a)}\bigg| 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) - z\Gamma^2(a) \\
&+ 2\Gamma(a)\big(z\Gamma(a,b^{-1}z) - b\Gamma(a+1,b^{-1}z)\big)\bigg|_{z_\omega}^{\infty} \\
= &\lim_{z\to\infty} -\frac{1}{\Gamma^2(a)}\Big(\cancelto{\Gamma^2(a)}{z\gamma^2(a,b^{-1}z)} - 2\big(b\exp(-b^{-1}\cancelto{0}{z)(b^{-1}z)^a\Gamma(a,b^{-1}z)} \\
&+ \cancelto{0}{\tfrac{1}{2}ab\Gamma^2(a,b^{-1}z)} - 4^{-a}b\cancelto{0}{\Gamma(2a,2b^{-1}z)} - b\Gamma(a)\cancelto{0}{\Gamma(a+1,b^{-1}z)}\big) - z\Gamma^2(a) \\
&+ 2\Gamma(a)\big(\cancelto{0}{z\Gamma(a,b^{-1}z)} - \cancelto{0}{b\Gamma(a+1,b^{-1}z)}\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) + \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{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).
\end{aligned}
\tag{I27}
$$

The sum of Equations I24 and I27 equals $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\, a,\, b),\, \omega)_{\rho>0}$, whence we can write

$$
\begin{aligned}
S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho>0} = & -\frac{1}{\Gamma^2(a)}\big(z_\omega\gamma^2(a,b^{-1}z_\omega) - 2\big(b\exp(-b^{-1}z_\omega)\big(b^{-1}z_\omega\big)^a\Gamma(a,b^{-1}z_\omega) \\
& + \frac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega) - 4^{-a}b\Gamma(2a,2b^{-1}z_\omega) - b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\big)\big)\big) \\
& + \frac{2}{\Gamma^2(a)}(4^{-a}b\Gamma(2a) + b\Gamma(a)\Gamma(a+1)) - ab \\
& + \frac{1}{\Gamma^2(a)}\big(z_\omega\gamma^2(a,b^{-1}z_\omega) - 2\big(b\exp(-b^{-1}z_\omega)\big(b^{-1}z_\omega\big)^a\Gamma(a,b^{-1}z_\omega) \\
& + \frac{1}{2}ab\Gamma^2(a,b^{-1}z_\omega) - 4^{-a}b\Gamma(2a,2b^{-1}z_\omega) - b\Gamma(a)\Gamma(a+1,b^{-1}z_\omega)\big)\big) \\
& -z_\omega\Gamma^2(a) + 2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega) - b\Gamma(a+1,b^{-1}z_\omega)\big)\big)\big) \\
& = \frac{2}{\Gamma^2(a)}(4^{-a}b\Gamma(2a) + b\Gamma(a)\Gamma(a+1)) - ab - \frac{1}{\Gamma^2(a)}\big(z_\omega\Gamma^2(a) \\
& -2\Gamma(a)\big(z_\omega\Gamma(a,b^{-1}z_\omega) - b\Gamma(a+1,b^{-1}z_\omega)\big)\big)\big) \\
& = 2\,\frac{4^{-a}b\Gamma(2a)}{\Gamma^2(a)} + 2b\,\frac{\Gamma(a+1)}{\Gamma(a)} - ab - z_\omega + 2\,\frac{z_\omega\Gamma(a,b^{-1}z_\omega)}{\Gamma(a)} \\
& -\ 2b\,\frac{\Gamma(a+1,b^{-1}z_\omega)}{\Gamma(a)} \\
& = 2\,\frac{4^{-a}b}{B(a,a)} + ab - z_\omega + 2z_\omega(1 - F_{\mathcal{G}}(z_\omega,a,b)) \\
& -\ 2ab(1 - F_{\mathcal{G}}(z_\omega,a+1,b)),
\end{aligned}
\tag{I28}
$$

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

$$
F_{\mathcal{G}}(z,a,b) = \frac{1}{\Gamma(a)}\gamma(a,b^{-1}z), \tag{I29}
$$

with $z > 0$, unitless shape parameter $a > 0$ and scale parameter $b > 0$. We can rearrange Equation I28 to yield a final expression for the CRPS of a PIII distribution forecast $P = \mathcal{P}_{\mathrm{III}}(\xi, a, b)$ with positive skewness, $\rho > 0$ and $z_\omega = \omega - \xi$

$$
S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho>0} = 2\,\frac{4^{-a}b}{B(a,a)} + ab(2F_{\mathcal{G}}(z_\omega,a+1,b) - 1) + z_\omega(1 - 2F_{\mathcal{G}}(z_\omega,a,b)). \tag{I30}
$$

I2.2. Analytic Expression for $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi, a, b), \omega)_{\rho<0}$: 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 I13 becomes

$$
S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho<0} = -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\Gamma(a,b^{-1}z)\right]^2\mathrm{d}z - \int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\Gamma(a,b^{-1}z) - 1\right]^2\mathrm{d}z. \tag{I31}
$$

We take advantage of the following identity

$$
\gamma(a,b^{-1}z) + \Gamma(a,b^{-1}z) = \Gamma(a), \tag{I32}
$$

to write Equation I31 in another form

$$\begin{aligned} S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho<0} &= -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\big(\Gamma(a)-\gamma(a,b^{-1}z)\big)\right]^2\mathrm{d}z \\ &\quad -\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\big(\Gamma(a)-\gamma(a,b^{-1}z)\big)-1\right]^2\mathrm{d}z \\ &= -\int_0^{z_\omega}\left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)-1\right]^2\mathrm{d}z-\int_{z_\omega}^{\infty}\left[\frac{1}{\Gamma(a)}\gamma(a,b^{-1}z)\right]^2\mathrm{d}z, \end{aligned} \tag{I33}$$

which is identical to Equation I17 but with left and right integrals swapped. This confirms that

$$S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)_{\rho<0} = S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)_{\rho>0}, \tag{I34}$$

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 $z_\omega = \xi - \omega$

$$S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,a,b),\omega)_{\rho<0} = 2\frac{4^{-a}b}{B(a,a)} + ab(2F_{\mathcal{G}}(z_\omega,a+1,b)-1) + z_\omega(1-2F_{\mathcal{G}}(z_\omega,a,b)), \tag{I35}$$

I2.3. Analytic Expression for $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)$: Positive/Negative Skewness

We can combine the mathematical expressions of $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)_{\rho>0}$ and $S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\ \omega)_{\rho<0}$ into one general Equation for the CRPS of $P = \mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b)$ and verifying observation $\omega \in \Omega$

$$\begin{aligned} S_{\mathrm{CRPS}}(\mathcal{P}_{\mathrm{III}}(\xi,\ a,\ b),\omega) = {} & 2\frac{4^{-a}b}{B(a,\ a)} - ab + |\omega-\xi| + 2abF_{\mathcal{G}}(|\omega-\xi|,\ a+1,\ b) \\ & -2|\omega-\xi|F_{\mathcal{G}}(|\omega-\xi|,\ a,\ b), \end{aligned} \tag{I36}$$

where $z_\omega = |\omega - \xi$l. The first term in the above expression is equal to the first term, $\frac{1}{2}\mathbb{E}_P[|y-y^*|]$, of Equation 36 and the sum of all the remaining terms of Equation I36 equals $-\mathbb{E}_P[|y-\omega|]$.

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 I36 simplifies to

$$S_{\mathrm{CRPS}}(\mathcal{G}(a,\ b),\ \omega) = 2\frac{4^{-a}b}{B(a,\ a)} - ab + |\omega| + 2abF_{\mathcal{G}}(|\omega|,\ a+1,b) - 2|\omega|F_{\mathcal{G}}(|\omega|,\ a,\ b). \tag{I37}$$

This expression for the CRPS of $P = \mathcal{G}(a,\ b)$ matches the numerical estimates of the CRPS shown in Figure 6 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 I12. This concludes the derivation of the CRPS for a Pearson type III distribution forecast $P = \mathcal{P}_{\mathrm{III}}(\mu,\sigma^2,\rho)$ and verifying observation $\omega \in \Omega$.

I3. Continuous Ranked Probability Score for $\mathcal{GEV}(\mu,\sigma^2,\xi)$

We revisit the quantile form of the CRPS

$$S_{\mathrm{CRPS}}(P,\ \omega) = \omega(1-2F_P(\omega)) + 2\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau - 2\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\,\mathrm{d}\tau. \tag{I38}$$

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 $\xi \in \mathbb{R}$. Appendix A of Friederichs and Thorarinsdottir (2012) derives a closed-from expression for $S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,\xi),\omega\big)$. We present our own analytic derivation for the CRPS of $P = \mathcal{GEV}(\mu,\sigma^2,\xi)$ and verifying observation $\omega \in \Omega$.

The CDF of the GEV distribution equals

$$F_{\mathcal{GEV}}(x,\mu,\sigma^2,\xi) \;=\; \begin{cases} \exp\left[-\exp\left(-\dfrac{\xi}{\sigma}(x-\mu)\right)\right] & \text{if } \xi = 0 \\ \exp\left[-\left(1+\dfrac{\xi}{\sigma}(x-\mu)\right)^{-1/\xi}\right] & \text{if } \xi < 0 \text{ and } x < -\dfrac{1}{\xi} \\ 1 & \text{if } \xi < 0 \text{ and } x \geq -\dfrac{1}{\xi} \\ 0 & \text{if } \xi > 0 \text{ and } x \leq -\dfrac{1}{\xi} \\ \exp\left[-\left(1+\dfrac{\xi}{\sigma}(x-\mu)\right)^{-1/\xi}\right] & \text{if } \xi > 0 \text{ and } x > -\dfrac{1}{\xi}, \end{cases} \tag{I39}$$

and its quantile function has the following explicit expression

$$F^{-1}_{\mathcal{GEV}}(x,\mu,\sigma^2,\xi) \;=\; \begin{cases} \mu - \sigma\log_e\left(-\log_e(\tau)\right) & \text{if } \xi = 0 \text{ and } \tau \in (0,1) \\ \mu + \dfrac{\sigma}{\xi}\left(\left(-\log_e(\tau)\right)^{-\xi} - 1\right) & \text{if } \xi \neq 0, \end{cases} \tag{I40}$$

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. We must separately consider the cases $\xi < 1$ and $\xi = 0$.

I3.1. Analytic Expression for $S_{\text{CRPS}}\left(\mathcal{GEV}(\mu,\sigma^2,\xi),\omega\right)_{\xi<1,\,\xi\neq 0}$

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 I38 can be expressed analytically using integration by parts

$$\begin{aligned} \int \tau F_P^{-1}(\tau)\,\mathrm{d}\tau = {} & \frac{\tau}{\xi}\left(\mu\xi\tau + \sigma\Gamma\left(1-\xi,\, -\log_e(\tau)\right) - \sigma\tau\right) + C \\ & - \int \frac{1}{\xi}\left(\mu\xi\tau + \sigma\Gamma\left(1-\xi,\, -\log_e(\tau)\right) - \sigma\tau\right)\mathrm{d}\tau \\ = {} & \frac{\tau}{\xi}\left(\mu\xi\tau + \sigma\Gamma\left(1-\xi,\, -\log_e(\tau)\right) - \sigma\tau\right) \\ & - \frac{1}{2\xi}\left[\tau\left(\mu\tau\xi + 2\sigma\Gamma\left(1-\xi,\, -\log_e(\tau)\right) - \sigma\tau\right) - 2^{\xi}\sigma\Gamma\left(1-\xi,\, -2\log_e(\tau)\right)\right] + C, \end{aligned} \tag{I41}$$

where

$$\Gamma(a,q) = \int_q^{\infty} t^{a-1}\exp(-t)\,\mathrm{d}t \tag{I42}$$

denotes the upper incomplete gamma function. If we now admit the lower and upper limits of the quantiles, we yield

$$\begin{aligned}\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \Bigg[\frac{\tau}{\xi}\big(\mu\xi\tau + \sigma\Gamma\big(1-\xi,\,-\log_e(\tau)\big) - \sigma\tau\big)\\ &\quad -\frac{1}{2\xi}\big[\tau\big(\mu\tau\xi + 2\sigma\Gamma\big(1-\xi,\,-\log_e(\tau)\big) - \sigma\tau\big) - 2^{\xi}\sigma\Gamma\big(1-\xi,\,-2\log_e(\tau)\big)\big] + C\Bigg]_0^1\\ &= \left(\mu + \frac{\sigma}{\xi}\Gamma(1-\xi) - \frac{\sigma}{\xi} - \frac{\mu}{2} - \frac{\sigma}{\xi}\Gamma(1-\xi) + \frac{\sigma}{2\xi} + \frac{2^{\xi}\sigma\Gamma(1-\xi)}{2\xi}\right) - (0)\\ &= \frac{\mu}{2} + \frac{\sigma}{2\xi}\big(2^{\xi}\Gamma(1-\xi) - 1\big).\end{aligned} \tag{I43}$$

The second or right integral of the CRPS in Equation I38 results in

$$\begin{aligned}\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\mathrm{d}\tau &= \left[\frac{1}{\xi}\big(\mu\xi\tau + \sigma\Gamma\big(1-\xi,\,-\log_e(\tau)\big) - \sigma\tau\big) + C\right]_{F_P(\omega)}^1\\ &= \left(\mu + \frac{\sigma}{\xi}\Gamma(1-\xi) - \frac{\sigma}{\xi}\right) - \Bigg(\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\\ &\quad + \frac{\sigma}{\xi}\Gamma\big(1-\xi,\,-\log_e\left(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\right)\big) - \frac{\sigma}{\xi}F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\Bigg)\\ &= \mu - \mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi) + \frac{\sigma}{\xi}\big[\Gamma(1-\xi) + F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\\ &\quad -\ \Gamma\big(1-\xi,\,-\log_e\left(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\right)\big) - 1\big]\end{aligned} \tag{I44}$$

Next, we can insert the analytic expressions of the two integrals into Equation I38

$$\begin{aligned}S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,\xi),\omega\big)_{\xi<1,\,\xi\neq0} =&\ \omega\big(1 - 2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\big) + \mu + \frac{\sigma}{\xi}\big(2^{\xi}\Gamma(1-\xi) - 1\big)\\ &-\ 2\mu + 2\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi) - \frac{2\sigma}{\xi}\big[\Gamma(1-\xi) + F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\\ &-\Gamma\big(1-\xi,\,-\log_e\left(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\right)\big) - 1\big].\end{aligned}$$

The above expression may be rearranged and rewritten to yield

$$\begin{aligned}S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,\xi),\omega\big)_{\xi<1,\,\xi\neq0} &= (\mu-\omega)\big(2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi) - 1\big) + \frac{\sigma}{\xi}\big(1 - (2-2^{\xi})\Gamma(1-\xi)\big)\\ &\quad + \frac{2\sigma}{\xi}\big[\Gamma\big(1-\xi,\,-\log_e\left(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\right)\big)\\ &\quad -F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,\xi)\big].\end{aligned} \tag{I45}$$

I3.2. Analytic Expression for $S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,0),\omega\big)$

Next, we consider $\xi = 0$. We use integration by parts

$$\int u\mathrm{d}v = uv - \int v\mathrm{d}u, \tag{I46}$$

and yield the following expression for the first of two integrals of Equation I38

$$\begin{aligned}\int \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \frac{1}{2}\mu\tau^2 - \frac{1}{2}\sigma\tau^2\log_e\big(-\log_e(\tau)\big) + C - \int -\frac{\sigma\tau^2}{2\tau\log_e(\tau)}\,\mathrm{d}\tau\\ &= \frac{1}{2}\mu\tau^2 - \frac{1}{2}\sigma\tau^2\log_e\big(-\log_e(\tau)\big) + C + \frac{1}{2}\sigma\int\frac{\tau}{\log_e(\tau)}\,\mathrm{d}\tau\\ &= \frac{1}{2}\mu\tau^2 - \frac{1}{2}\sigma\tau^2\log_e\big(-\log_e(\tau)\big) + \frac{1}{2}\sigma\mathrm{Ei}\big(2\log_e(\tau)\big) + C,\end{aligned} \tag{I47}$$

where Ei(x) is the exponential integral function. If we admit the integral limits, we yield

$$\begin{aligned}\int_0^1 \tau F_P^{-1}(\tau)\,\mathrm{d}\tau &= \left[\frac{1}{2}\mu\tau^2 - \frac{1}{2}\sigma\tau^2\log_e\big(-\log_e(\tau)\big) + \frac{1}{2}\sigma\mathrm{Ei}\big(2\log_e(\tau)\big) + C\right]_0^1\\ &= \frac{1}{2}\mu - \frac{1}{2}\sigma\lim_{\tau\to 1^-}\Big(\tau^2\log_e\big(-\log_e(\tau)\big) - \mathrm{Ei}\big(2\log_e(\tau)\big)\Big)\\ &= \frac{1}{2}\mu + \frac{1}{2}\sigma\big(\gamma_{\mathrm{c}} + \log_e(2)\big),\end{aligned} \tag{I48}$$

where $\gamma_{\mathrm{c}} = 0.57721566\ldots$ is the Euler-Mascheroni constant. The right integral of Equation I38 becomes

$$\begin{aligned}\int_{F_P(\omega)}^1 F_P^{-1}(\tau)\mathrm{d}\tau &= \big[\mu\tau + \sigma\mathrm{Li}(\tau) - \sigma\tau\log_e\big(-\log_e(\tau)\big)\big]_{F_P(\omega)}^1\\ &= \Big(\mu + \lim_{\tau\to 1^-}\big(\sigma\mathrm{Li}(\tau) - \sigma\tau\log_e\big(-\log_e(\tau)\big)\big)\Big)\\ &\quad -\ \Big(\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) + \sigma\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\\ &\quad - \sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big)\Big)\\ &= \mu + \gamma_{\mathrm{c}}\sigma - \mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) - \sigma\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\\ &\quad +\ \sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big).\end{aligned} \tag{I49}$$

where Li(x) signifies the logarithmic integral function. If we substitute Equations I48 and I49 into Equation I38, we yield the following expression for the CRPS of $P = \mathcal{GEV}(\mu,\sigma^2,0)$

$$\begin{aligned}S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,0),\omega\big) &= \omega\big(1 - 2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) + \mu + \sigma\big(\gamma_{\mathrm{c}} + \log_e(2)\big) - 2\mu\\ &\quad - 2\gamma_{\mathrm{c}}\sigma + 2\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) + 2\sigma\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\\ &\quad - 2\sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big)\\ &= \omega\big(1 - 2F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big) - \mu - \gamma_{\mathrm{c}}\sigma + \sigma\log_e(2)\\ &\quad + 2\mu F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0) + 2\sigma\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\\ &\quad - 2\sigma F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big)\\ &= \omega - \mu - \gamma_{\mathrm{c}}\sigma + \sigma\log_e(2) + 2\sigma\mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\\ &\quad - 2\big[\omega - \mu + \sigma\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big)\big]F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0).\end{aligned} \tag{I50}$$

Per the quantile function in Equation I40, we find that

$$-\mu + \sigma\log_e\big(-\log_e\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big)\big) = -\omega, \tag{I51}$$

and, thus, the CRPS of $P = \mathcal{GEV}(\mu,\sigma^2,0)$ simplifies to

$$S_{\mathrm{CRPS}}\big(\mathcal{GEV}(\mu,\sigma^2,0),\omega\big) \quad = \omega - \mu - \gamma_{\mathrm{c}}\sigma + \sigma \log_e(2) + 2\sigma \mathrm{Li}\big(F_{\mathcal{GEV}}(\omega,\mu,\sigma^2,0)\big). \tag{I52}$$

This concludes the derivation of the CRPS for a generalized extreme value distribution forecast $P = \mathcal{GEV}(\mu,\sigma^2,\xi)$ and verifying observation $\omega \in \Omega$.

Appendix J: BMA Model Training and Evaluation

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 7 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 43 is equal to

$$\sigma_{P_t}^2 = \sum_{k=1}^{K} \beta_k(\sigma_k^2 + y_{kt}^2) - \mu_{P_t}^2 \tag{J1}$$

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 each conditional PDF of Table 7 has an analytic or closed-form solution for its variance, σ_k^2; $k = (1, \dots, K)$, the coefficient of variation of the BMA distribution forecast is exactly defined at each time t; $C_{\mathrm{v},t} = \sigma_{P_t}/\mu_{P_t}$. Lower and upper endpoints of the $\gamma = 100(1 - \alpha)\%$ prediction interval of the BMA mixture density can be derived from the CDF

$$F_{P_t}(y|\boldsymbol{\beta},\boldsymbol{\psi}) = \sum_{k=1}^{K} \beta_k F_k(y|\, y_{kt},\psi_k), \tag{J2}$$

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 J2 at each verifying observation and sort the resulting values in ascending order, then the reliability R_1 of the BMA forecast distribution is easily computed using Equation 11. 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 7 along with performance metrics, scoring rules and discharge prediction intervals of the BMA forecast density.

We use different metrics to evaluate the performance of the BMA model. This includes the log-likelihood

$$\ell(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\omega}) = \sum_{t=1}^{n}\left\{\log_e\left(\sum_{k=1}^{K}\beta_k f_k(\omega_t|\, y_{kt},\psi_k)\right)\right\}, \tag{J3}$$

of the maximum likelihood weights $\hat{\boldsymbol{\beta}}$ and shape parameters $\hat{\boldsymbol{\psi}}$ and the RMSE

$$\bar{s}_{\mathrm{RMSE}}(\boldsymbol{\mu}_P,\boldsymbol{\omega}) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\omega_t - \mu_{P_t})^2}, \tag{J4}$$

Nash and Sutcliffe (1970) efficiency

$$\bar{s}_{\mathrm{NSE}}(\boldsymbol{\mu}_P,\boldsymbol{\omega}) = 1 - \frac{\sum_{t=1}^{n}(\omega_t - \mu_{P_t})^2}{\sum_{t=1}^{n}(\omega_t - m_\omega)^2}, \tag{J5}$$

and KG efficiency (Gupta et al., 2009)

$$\bar{s}_{\mathrm{KGE}}(\boldsymbol{\mu}_P,\boldsymbol{\omega}) = 1 - \sqrt{(a-1)^2 + (b-1)^2 + (r-1)^2}, \tag{J6}$$

of the weighted-average BMA forecast $\boldsymbol{\mu}_P = \left(\mu_{P_1},\ldots,\mu_{P_n}\right)^\top$ of Equation 44 using the maximum likelihood weights, 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, and r is the sample Pearson correlation coefficient of the n-data pairs, (ω_t,μ_{P_t}); $t = (1,\ldots,n)$.

Appendix K: Description of SAC-SMA Model

The Sacramento Soil Moisture Accounting (SAC-SMA) model of Burnash et al. (1973) is used by the National Weather Service River Forecast System for flood forecasting throughout the United States. The model converts areal average precipitation into streamflow. Our model implementation in MATLAB and C++ follows Clark et al. (2008) and is presented in Figure K1. A mass-conservative second-order integration method with adaptive time stepping solves the state variables, u_t, u_f, l_t, l_fp, l_fs, r_1, r_2, r_3, and fluxes, q_xx, of the control volumes using daily time series of areal average rainfall $\left(p_1,\ldots,p_n\right)^\top$ and potential evapotranspiration $(e_{\mathrm{p}1},\ldots,e_{\mathrm{p}n})^\top$ and values of the model parameters listed in Table K1. A 1-year spin-up period eliminates the impact of state variable initialization.

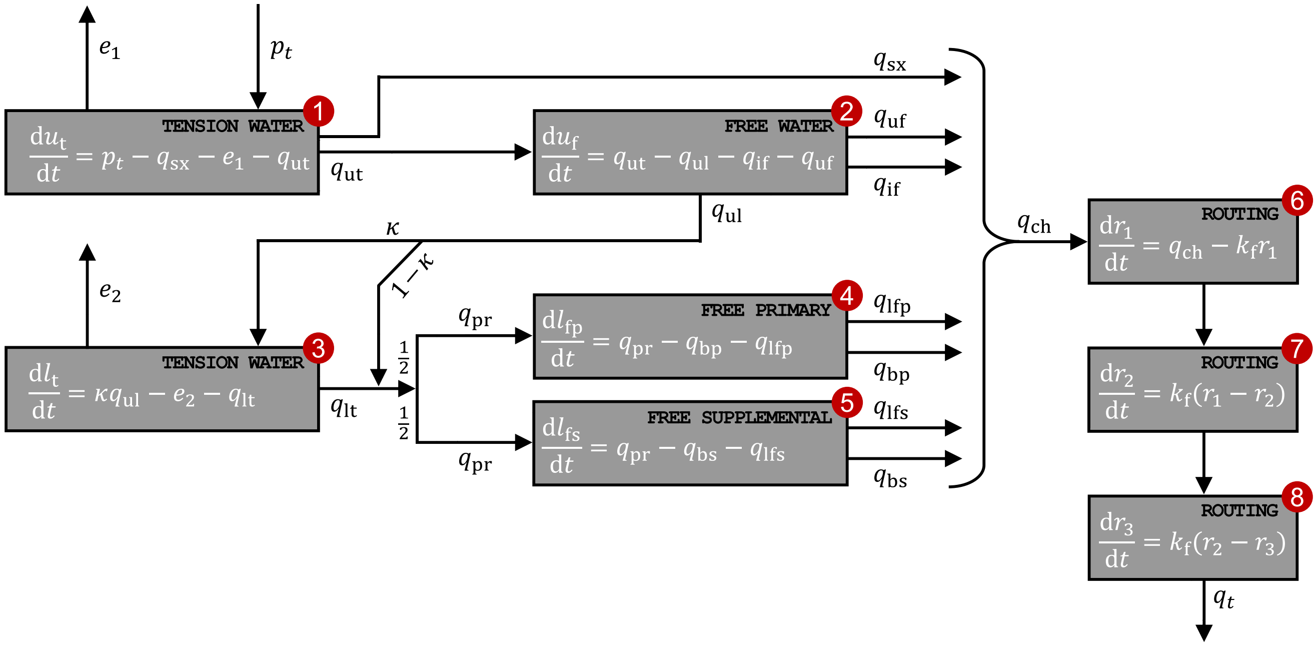
e_1
p_t
q_sx
TENSION WATER
1
$\frac{du_\mathrm{t}}{dt} = p_t - q_\mathrm{sx} - e_1 - q_\mathrm{ut}$
q_ut
FREE WATER
2
$\frac{du_\mathrm{f}}{dt} = q_\mathrm{ut} - q_\mathrm{ul} - q_\mathrm{if} - q_\mathrm{uf}$
q_uf
q_if
q_ul
κ
$1-\kappa$
e_2
q_ch
ROUTING
6
$\frac{dr_1}{dt} = q_\mathrm{ch} - k_\mathrm{f}r_1$
q_pr
FREE PRIMARY
4
$\frac{dl_\mathrm{fp}}{dt} = q_\mathrm{pr} - q_\mathrm{bp} - q_\mathrm{lfp}$
q_lfp
q_bp
TENSION WATER
3
$\frac{dl_\mathrm{t}}{dt} = \kappa q_\mathrm{ul} - e_2 - q_\mathrm{lt}$
q_lt
$\frac{1}{2}$
$\frac{1}{2}$
ROUTING
7
$\frac{dr_2}{dt} = k_\mathrm{f}(r_1 - r_2)$
FREE SUPPLEMENTAL
5
$\frac{dl_\mathrm{fs}}{dt} = q_\mathrm{pr} - q_\mathrm{bs} - q_\mathrm{lfs}$
q_pr
q_lfs
q_bs
ROUTING
8
$\frac{dr_3}{dt} = k_\mathrm{f}(r_2 - r_3)$
q_t

Figure K1. Schematic illustration of the SAC-SMA model after Burnash et al. (1973) and Clark et al. (2008). Gray boxes labeled in red correspond to fictitious control volumes which govern the rainfall-runoff transformation. The model has eight state variables, including the free water storages of the upper soil layer, u_f, and primary, l_fp, and secondary, l_fs, base flow reservoirs, the tension water storages of the upper, u_t, and lower, l_t, soil layers and the water levels, r_1, r_2, and r_3, of the three routing reservoirs. Arrows portray fluxes in and out of the compartments, including precipitation, p_t, evaporation from the upper soil layer, e_1, overflow from tension storage in upper soil layer, q_ut, surface runoff, q_sx, overflow from free storage in upper soil layer, q_uf, interflow, q_if, percolation from upper to lower layer, q_ul, evaporation from lower soil layer, e_2, overflow from tension storage in lower soil layer, q_lt, flow into primary and supplemental storage, q_pr, overflow from primary q_lfp and secondary q_lfs base flow storage in the lower soil layer and base flow from primary q_bp and secondary q_bs reservoirs. These fluxes are computed as follows, $e_1 = e_\mathrm{p}(u_\mathrm{t}/u_\mathrm{t,max})$, $q_\mathrm{sx} = a_\mathrm{c,max}(u_\mathrm{t}/u_\mathrm{t,max})p_t$, $q_\mathrm{ut} = (p_t - q_\mathrm{sx})f(u_\mathrm{t},\ u_\mathrm{t,max})$, $q_{ul} = q_0 d_\mathrm{lz}(u_\mathrm{f}/u_\mathrm{f,max})$, $q_\mathrm{uf} = q_\mathrm{ut}f(u_\mathrm{f},\ u_\mathrm{f,max})$, $q_\mathrm{if} = k_\mathrm{i}(u_\mathrm{f}/u_\mathrm{f,max})$, $e_2 = (e_\mathrm{p} - e_1)(l_\mathrm{t}/l_\mathrm{t,max})$, $q_\mathrm{lt} = \kappa q_\mathrm{ul}f(l_\mathrm{t},\ l_\mathrm{t,max})$, $q_\mathrm{pr} = \frac{1}{2}(1-\kappa)q_\mathrm{ul} + \frac{1}{2}q_\mathrm{lt}$, $q_\mathrm{lfp} = q_\mathrm{pr}f(l_\mathrm{fp},\ l_\mathrm{fp,max})$, $q_\mathrm{lfs} = q_\mathrm{pr}f(l_\mathrm{fs},\ l_\mathrm{fs,max})$, $q_\mathrm{bp} = \nu_\mathrm{p}l_\mathrm{fp}$, $q_\mathrm{bs} = \nu_\mathrm{s}l_\mathrm{fs}$, where e_p is the potential evapotranspiration, $q_0 = \nu_\mathrm{p}l_\mathrm{fp,max} + \nu_\mathrm{s}l_\mathrm{fs,max}$, $d_\mathrm{lz} = 1 + \alpha\left[\left(l_\mathrm{t} + l_\mathrm{fp} + l_\mathrm{fs}\right)/\left(l_\mathrm{t,max} + l_\mathrm{fp,max} + l_\mathrm{fs,max}\right)\right]^\psi$, the smoothing function $f(x_1,x_2) = \{1 + \exp[(x_2 - \epsilon\rho x_2 - x_1)/(\rho x_2)]\}^{-1}$ with $\epsilon = 5$ and $\rho = 10^{-2}$, and $u_\mathrm{f,max}$, $u_\mathrm{t,max}$, $l_\mathrm{fp,max}$, $l_\mathrm{fs,max}$, $l_\mathrm{t,max}$, α, ψ, k_i, κ, ν_p, ν_s, and $a_\mathrm{c,max}$ are unknown parameters. Channel inflow $q_\mathrm{ch} = q_\mathrm{sx} + q_\mathrm{uf} + q_\mathrm{if} + q_\mathrm{lfp} + q_\mathrm{bp} + q_\mathrm{lfs} + q_\mathrm{bs}$ is routed through three linear reservoirs with common recession constant k_f and yields the streamflow $q_t = k_\mathrm{f}r_3$ at the watershed outlet.

Table K1
SAC-SMA Model Parameters and Their Symbols, Units, Lower, and Upper Bounds

Symbol	Description	Units	Min.	Max.
$u_{\mathrm{t,max}}$	Upper zone tension water maximum storage	mm	50	500
$u_{\mathrm{f,max}}$	Upper zone free water maximum storage	mm	10	500
$l_{\mathrm{t,max}}$	Lower zone tension water maximum storage	mm	10	500
$l_{\mathrm{fp,max}}$	Lower zone free water primary maximum storage	mm	10	1,000
$l_{\mathrm{fs,max}}$	Lower zone free water supplemental maximum storage	mm	10	1,000
α	Percolation multiplier for the lower layer	–	1	250
ψ	Percolation exponent for the lower layer	–	1	5
k_{i}	Upper zone free water lateral depletion rate (interflow rate)	mm d^{-1}	10^{-2}	100
κ	Fraction of percolation to tension storage in the lower layer	–	0.05	0.95
ν_{p}	Base flow depletion rate for primary reservoir	d^{-1}	10^{-3}	0.25
ν_{s}	Base flow depletion rate for secondary reservoir	d^{-1}	10^{-3}	0.25
$a_{\mathrm{c,max}}$	Maximum fraction of saturated area	–	0.05	0.95
k_{f}	Recession constant of routing reservoirs	d^{-1}	10^{-1}	5

Appendix L: The Interval Score

We perform a simple numerical experiment to demonstrate that $S^{\alpha}_{\mathrm{IS}}(P,\omega)$ 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 60 then $l = -1.96$ and $u = 1.96$ should maximize the expected value of $S^{\alpha}_{\mathrm{IS}}(P,\omega)$. We verify this assertion in Figure L1 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 $S^{\alpha}_{\mathrm{IS}}(P,\omega)$ on the choice 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.

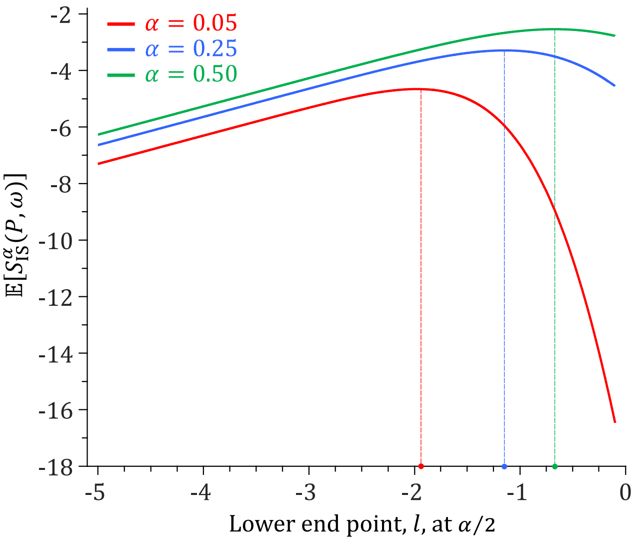

Figure L1. Traces of the expected value of the interval score $S^{\alpha}_{\mathrm{IS}}(P,\omega)$ as function of the lower endpoint l of the $100(1-\alpha)$% prediction interval using $\alpha = 0.05$ (red), $\alpha = 0.25$ (blue) and $\alpha = 0.50$ (green). The colored dots are a projection of the maximum interval score on the x-axis.

Appendix M: Decomposition of Scoring Rules

M1. Decomposition of Conditional Expectation of Quadratic Score

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 $p \in [0, 1]$. Once we observe $\omega \in \Omega$, we can assign a score $S(p,\omega)\mathbf{:}\ \mathcal{P}_2 \times \Omega \to \mathbb{R}$ to the prediction. Thus, ω takes on values of 0 (*no rain*) and 1 (*rain*).

Suppose we use the Brier or QS, $S_{\mathrm{QS}}(p,\omega) = -\sum_{k=1}^{2}(\delta_{\omega k} - p_k)^2$, where $\delta_{\omega k} = 1$ if $\omega = k$ and $\delta_{\omega k} = 0$ otherwise. Then we can decompose the conditional expectation $\mathbb{E}[S_{\mathrm{QS}}(p,\omega)\big|\mathcal{D}]$ of the QS as follows

$$\begin{aligned}\mathbb{E}[S_{\mathrm{QS}}(p,\omega)|\mathcal{D}] &= \mathbb{E}\big[-(\omega - p(\mathcal{D}))^2\big|\mathcal{D}\big]\\ &= -\mathbb{E}\big[(\omega - \mathbb{E}[\omega|\mathcal{D}] + \mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D}))^2\big|\mathcal{D}\big]\\ &= -\mathrm{Var}[\omega|\mathcal{D}] - (\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D}))^2,\end{aligned} \tag{M1}$$

and insert Equation M1 into Equation 67 to yield

$$\begin{aligned}\mathbb{E}[S_{\mathrm{QS}}(p,\omega)] &= \mathbb{E}\big[-\mathrm{Var}[\omega|\mathcal{D}] - (\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D}))^2\big]\\ &= -\mathbb{E}[\mathrm{Var}[\omega|\mathcal{D}]] - \mathbb{E}\big[(\mathbb{E}[\omega|\mathcal{D}] - p(\mathcal{D}))^2\big]\\ &= -\mathrm{Var}[\omega] + \mathrm{Var}[\mathbb{E}[\omega|\mathcal{D}]] - \mathbb{E}\big[(p(\mathcal{D}) - \mathbb{E}[\omega|\mathcal{D}])^2\big],\end{aligned} \tag{M2}$$

where $\mathbb{E}[\omega|\mathcal{D}]$ is simply equal to the conditional probability of *rain* and $p(\mathcal{D})$ equals the unconditional *rain* probability.

Table M1
Unconditional, $\overline{\mathbf{p}}$*, and Conditional,* $\boldsymbol{\pi}$*, Probabilities of the Watershed Models Estimated From the 3,000-Day Training Data Record:* $\pi_{jk} = \mathbb{P}(\omega = y_j\big|\,y_k)$ *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

Note. The values on the main diagonal are highlighted in bold.

M2. Explanation of Terms of General Decomposition *Strictly Proper* Scoring Rule

The first term of the decomposition of Bröcker (2009) in Equation 69 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 **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.

M3. Case Study VII: Discharge Forecast Ensemble

We apply the analytic decomposition of Equation 69 to the multi-model ensemble of discharge forecasts displayed in Figure 7. 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. As index $\hbar_t \in (1, \ldots, K)$ of this "best" forecast at time t we take the index of the minimum entry of the K-vector of absolute residuals $(|\omega_t - y_{1t}|, \ldots, |\omega_t - y_{Kt}|)$ of forecasted discharges $y_{1t}, \ldots, y_{Kt}$ by the different models

$$\hbar_t = \underset{\hbar \in (1,\ldots,K)}{\arg\min} |\omega_t - y_{\hbar t}|, \tag{M3}$$

Now the n-vector $(\hbar_1, \ldots, \hbar_n)$ stores the indices of the best models in the ensemble for our training record of $n = 3{,}000$ days, we yield the event frequencies $\overline{\mathbf{p}} = \left(\overline{p}_1, \ldots, \overline{p}_m\right)^\top$

$$\overline{p}_k = \mathbb{P}(\omega = y_k) = \frac{1}{n}\sum_{t=1}^{n} 1\{\hbar_t = k\}, \tag{M4}$$

and the conditional forecast probabilities

$$\begin{aligned}\pi_{jk} = \mathbb{P}(\omega = y_j | \, y_k) &= \frac{\sum_{t=2}^{n} 1\{\hbar_t = j | \hbar_{t-1} = k\}}{\sum_{t=1}^{n} 1\{\hbar_t = k\}} \\ &= \frac{\sum_{t=2}^{n} 1\{\hbar_t = j | \hbar_{t-1} = k\}}{n\overline{p}_k},\end{aligned} \tag{M5}$$

where $j, k = (1, \ldots, K)$. Table M1 reports the unconditional and conditional probabilities of the watershed models. 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.

Data Availability Statement

The case studies of this paper are part of the MATLAB toolbox `Scoring_Rules`. This toolbox can be obtained from GitHub at https://github.com/jaspervrugt/Scoring_rules. 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. The CAMELS data set is described in Newman et al. (2015) and can be downloaded from https://doi.org/10.5065/D6MW2F4D.

Acknowledgments

I would like to thank the Associate Editor, Juliane Mai, Tim Hodson, Alberto Montanari, and three anonymous referees for their time and effort spent on reviewing this manuscript. Their constructive comments and feedback have led to a significantly improved manuscript.

References

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

Alkema, L., Raftery, A. E., & Clark, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Annals of Applied Statistics*, *1*(1), 229–248. https://doi.org/10.1214/07-AOAS111

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

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, *33*(3), 1148–1159. https://doi.org/10.1214/aoms/1177704477

Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (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. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, *88*(1), 190–206. https://doi.org/10.1016/S0047-259X(03)00079-4

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

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. https://doi.org/10.1098/rstl.1763.0053

Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2013). On the effect of calibration in classifier combination. *Applied Intelligence*, *38*(4), 566–585. https://doi.org/10.1007/s10489-012-0388-2

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, *7*(3), 686–690. https://doi.org/10.1214/aos/1176344689

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

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

Beven, K., & Freer, J. (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. https://doi.org/10.1016/S0022-1694(01)00421-8

Borgonovo, E., Hazen, G. B., Jose, V. R. R., & Plischke, E. (2021). Probabilistic sensitivity measures as information value. *European Journal of Operational Research*, *289*(2), 595–610. https://doi.org/10.1016/j.ejor.2020.07.010

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

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

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

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. https://doi.org/10.1016/0041-5553(67)90040-7

Brehmer, J. R., & Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, *13*(2). https://doi.org/10.1214/19-ejs1622

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Briggs, W., & Ruppert, D. (2005). Assessing the skill of yes/no predictions. *Biometrics*, *61*(3), 799–807. https://doi.org/10.1111/j.1541-0420.2005.00347.x

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

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, *60*(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

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

Buja, A., Stuetzle, W., & Shen, Y. (2005). *Loss functions for binary class probability estimation and classification: Structure and applications, technical report*. Statistics Department, The Wharton School, University of Pennsylvania.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2 ed., p. 488). Springer. https://doi.org/10.1007/b97636

Casella, G., & Berger, R. L. (2002). *Statistical inference, duxbury advanced series* (2 ed., p. 660). Duxbury.

Cervera, J. L., & Muñoz, J. (1996). Proper scoring rules for fractiles. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5* (pp. 513–519). Oxford University Press.

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

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

Cheng, H.-Y., Wu, Y. C., Lin, M. H., Liu, Y. L., Tsai, Y. Y., Wu, J. H., et al. (2020). Applying machine learning models with an ensemble approach for accurate real-time influenza forecasting in Taiwan: Development and validation study. *Journal of Medical Internet Research*, *22*(8), e15394. https://doi.org/10.2196/15394

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

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862. https://doi.org/10.2307/2527341

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12), 604. https://doi.org/10.1029/2007WR006735

Clausen, B., & Biggs, B. (2000). Flow variables for ecological studies in temperate streams: Groupings based on covariance. *Journal of Hydrology*, *237*(3), 184–197. https://doi.org/10.1016/S0022-1694(00)00306-1

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory, telecommunications and signal processing* (2 ed., p. 784). John Wiley & Sons, Inc.

Cox, D. R., & Lewis, P. A. W. (1966). Statistical analysis of series of events. In *Methuen's monographs on applied probability and statistics (MMAPS)* (p. 285). Methuen.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, *1928*(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862

Csiszar, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*(1), 146–158. https://doi.org/10.1214/aop/1176996454

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*, *147*(2), 278–292. https://doi.org/10.2307/2981683

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

Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, *72*(2), 169–183. https://doi.org/10.1007/s40300-014-0039-y

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

de Finetti, B. (2017). *Theory of probability: A critical introductory treatment*. Wiley Series in Probability and Statistics, John Wiley & Sons Ltd.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(1/2), 12–22. https://doi.org/10.2307/2987588

de Punder, R., Diks, C., Laeven, R., & van Dijk, D. (2023). *Localising strictly proper scoring rules, technical report*. Erasmus University Rotterdam, Tinbergen Institute.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883. https://doi.org/10.2307/2527342

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

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

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

Dunsmore, I. R. (1968). A Bayesian approach to calibration. *Journal of the Royal Statistical Society: Series B*, *30*(2), 396–405. https://doi.org/10.1111/j.2517-6161.1968.tb00740.x

Dvoretzky, A., Kiefer, J., & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, *27*(3), 642–669. https://doi.org/10.1214/aoms/1177728174

Eckhardt, K. (2005). How to construct recursive digital filters for baseflow separation. *Hydrological Processes*, *19*(2), 507–515. https://doi.org/10.1002/hyp.5675

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), 505–562. https://doi.org/10.1111/rssb.12154

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

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

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*(11), W11510. https://doi.org/10.1029/2010WR010174

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

Fissler, T., Hlavinová, J., & Rudloff, B. (2021). Elicitability and identifiability of set-valued measures of systemic risk. *Finance and Stochastics*, *25*(1), 133–165. https://doi.org/10.1007/s00780-020-00446-z

Fissler, T., & Pesenti, S. M. (2023). Sensitivity measures based on scoring functions. *European Journal of Operational Research*, *307*(3), 1408–1423. https://doi.org/10.1016/j.ejor.2022.10.002

Freer, J., Beven, K., & Ambroise, B. (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. https://doi.org/10.1029/95WR03723

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

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

Gao, Y., Sahin, A., & Vrugt, J. A. (2023). Probabilistic sensitivity analysis with dependent variables: Covariance-based decomposition of hydrologic models. *Water Resources Research*, *59*(4), e2022WR032. https://doi.org/10.1029/2022WR032834

Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling: An application to the U.K. economy. *Journal of the American Statistical Association*, *98*(464), 829–838. https://doi.org/10.1198/016214503000000765

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

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

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

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

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2005). *Probabilistic forecasts, calibration, and sharpness, technical report*. Department of Statistics, University of Washington.

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

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2005). *Strictly proper scoring rules, prediction, and estimation, technicla report 463R*. Department of Statistics, University of Washington.

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

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110

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

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B*, *14*(1), 107–114. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x

Good, I. J. (1971). Discussion of "Measuring information and uncertainty" by R. J. Buehler. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 337–339). Holt, Rinehardt and Winston.

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

Grayson, R., & Blöschl, G. (2001). *Spatial patterns in catchment hydrology: Observations and modelling* (p. 416). Cambridge University Press.

Grimit, E. P., Gneiting, T., Berrocal, V. J., & Johnson, N. A. (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. https://doi.org/10.1256/qj.05.235

Groen, J. J. J., Paap, R., & Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, *31*(1), 29–44. https://doi.org/10.1080/07350015.2012.727718

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

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

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

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

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

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

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

Heinrich, C. (2014). The mode functional is not elicitable. *Biometrika*, *101*(1), 245–251. https://doi.org/10.1093/biomet/ast048

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

Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12), e2021MS002681. https://doi.org/10.1029/2021MS002681

Holton, G. A. (2013). Value-at-risk: Theory and practice (2nd ed.). Retrieved from https://www.value-at-risk.net

Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, *306*(5696), 640–643. https://doi.org/10.1126/science.1104635

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on L-moments* (p. 224). Cambridge University Press. https://doi.org/10.1017/cbo9780511529443

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

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., & Ames, D. P. (2019). Introductory overview: Error metrics for hydrologic modelling—A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, *119*, 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001

Jaynes, E. T. (1963). *Information theory and statistical mechanics* (pp. 181–218). W. A. Benjamin, Inc.

Jeffreys, H. (1939). The theory of probability. In *Oxford classic texts in the physical sciences* (3 ed., p. 470). Oxford University Press.

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.

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

Jolliffe, I. T., & Stephenson, D. B. (2011). *Forecast verification: A practitioner's guide in atmospheric science* (2 ed., Vol. 296). Wiley Blackwell.

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

Jordan, T., Chen, Y. T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., et al. (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation. *Annals of Geophysics*, *54*, 315–391. https://doi.org/10.4401/ag-5350

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, *55*(4), 582–590. https://doi.org/10.1287/mnsc.1080.0955

Jowett, I. G., & Duncan, M. J. (1990). Flow variability in New Zealand rivers and its relationship to in stream habitat and biota. *New Zealand Journal of Marine & Freshwater Research*, *24*(3), 305–317. https://doi.org/10.1080/00288330.1990.9516427

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

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*(3), W03407. https://doi.org/10.1029/2005WR004368

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3), W03408. https://doi.org/10.1029/2005WR004376

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2), W02429. https://doi.org/10.1029/2008WR006912

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (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. https://doi.org/10.5194/gmd-12-2463-2019

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

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

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

Koutsoyiannis, D., & Montanari, A. (2022). Bluecat: A local uncertainty estimator for deterministic simulations and predictions. *Water Resources Research*, *58*(1), e2021WR031215. https://doi.org/10.1029/2021WR031215

Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*, 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, *249*(1–4), 2–9. https://doi.org/10.1016/s0022-1694(01)00420-6

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

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

Kull, M., & Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, & A. Jorge (Eds.), *Machine learning and knowledge discovery in databases* (pp. 68–85). Springer International Publishing.

Kullback, S. (1959). *Information theory and statistics*. John Wiley & Sons.

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

Kunsch, H. R. (1989). The Jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265

Lai, T. L., Gross, S. T., & Shen, D. B. (2011). Evaluating probability forecasts. *Annals of Statistics*, *39*(5), 2356–2382. https://doi.org/10.1214/11-aos902

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

Lambert, N. S., Pennock, D. M., & Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on electronic commerce, EC '08* (pp. 129–138). Association for Computing Machinery. https://doi.org/10.1145/1386790.1386813

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

Langrené, N., & Warin, X. (2021). Fast multivariate empirical cumulative distribution function with connection to kernel density estimation. *Computational Statistics & Data Analysis*, *162*, 107267. https://doi.org/10.1016/j.csda.2021.107267

Lebesgue, H. L. (1902). Intégrale, longueur, aire. *Annali di Matematica Pura ed Applicata*, *7*(1), 231–359. https://doi.org/10.1007/bf02420592

Liese, F., & Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*(10), 4394–4412. https://doi.org/10.1109/TIT.2006.881731

Liu, Y., Chen, W., Arendt, P., & Huang, H.-Z. (2011). Toward a better understanding of model validation metrics. *Journal of Mechanical Design*, *133*(7), 071005. https://doi.org/10.1115/1.4004223

Lu, D., Ye, M., & Neuman, S. P. (2011). Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. *Mathematical Geosciences*, *43*(8), 971–993. https://doi.org/10.1007/s11004-011-9359-0

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

Matheron, G. (1984). *The selectivity of the distributions and the "second principle of geostatistics"* (pp. 421–433). Springer. https://doi.org/10.1007/978-94-009-3699-7_24

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

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

McDonald, J. B., & Jensen, B. C. (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. https://doi.org/10.1080/01621459.1979.10481042

McInerney, D., Kavetski, D., Thyer, M., Lerat, J., & Kuczera, G. (2019). Benefits of explicit treatment of zero flows in probabilistic hydrological modeling of ephemeral catchments. *Water Resources Research*, *55*(12), 11035–11060. https://doi.org/10.1029/2018WR024148

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

McMillan, H., Westerberg, I., & Branger, F. (2017). Five guidelines for selecting hydrological signatures. *Hydrological Processes*, *31*(26), 4757–4761. https://doi.org/10.1002/hyp.11300

Meng, X., Taylor, J. W., Ben Taieb, S., & Li, S. (2022). *Scores for multivariate distributions and level sets, technical report*. University of Sussex, Falmer. https://doi.org/10.48550/arXiv.2002.09578

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

Murphy, A. H. (1973a). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology and Climatology*, *12*(1), 215–223. https://doi.org/10.1175/1520-0450(1973)012¡0215:HASSFP¿2.0.CO;2

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

Murphy, A. H., & Katz, R. W. (1985). *Probability, statistics, and decision making in the atmospheric sciences* (1 ed., Vol. 560). Westview Press.

Naaman, M. (2021). On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, *173*, 109088. https://doi.org/10.1016/j.spl.2021.109088

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

Neuman, S. P. (2003). Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, *17*(5), 291–305. https://doi.org/10.1007/s00477-003-0151-7

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., 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. https://doi.org/10.5194/hess-19-209-2015

Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research*, *48*(12), W12434. https://doi.org/10.1029/2011WR011128

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data*, *7*(1), 20. https://doi.org/10.1186/s40537-020-00299-5

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications*, *19*(2), 101–121. https://doi.org/10.1002/rra.700

Osband, K. H. (1985). *Providing incentives for better cost forecasting, Ph.D. thesis*. University of California Berkeley.

Pachepsky, Y. A., Martinez, G., Pan, F., Wagener, T., & Nicholson, T. (2016). Evaluating hydrological model performance using information theory-based metrics. In *Hydrology and Earth system sciences discussions, 2016* (pp. 1–24). https://doi.org/10.5194/hess-2016-46

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. https://doi.org/10.1256/0035900021643593

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

Pool, S., Vis, M. J. P., Knight, R. R., & Seibert, J. (2017). Streamflow characteristics from modeled runoff time series—Importance of calibration criteria selection. *Hydrology and Earth System Sciences*, *21*(11), 5443–5457. https://doi.org/10.5194/hess-21-5443-2017

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

Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays, chapter 7* (pp. 156–198). McMaster University Archive for the History of Economic Thought.

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

Reich, N. G., Lessler, J., Sakrejda, K., Lauer, S. A., Iamsirithaworn, S., & Cummings, D. A. T. (2016). Case study in evaluating time series prediction models using the relative mean absolute error. *The American Statistician*, *70*(3), 285–292. https://doi.org/10.1080/00031305.2016.1148631

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5), W05521. https://doi.org/10.1029/2009WR008328

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, *47*(11), W11516. https://doi.org/10.1029/2011WR010643

Resin, J. (2023). From classification accuracy to proper scoring rules: Elicitability of probabilistic top list predictions. *Journal of Machine Learning Research*, *24*, 1–21.

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

Roby, T. B. (1964). *Belief states: A preliminary empirical study, tech. Rep. ESD-TDR-64-238*. Tufts University.

Roccioletti, S. (2015). *Elicitability* (pp. 27–41). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-11908-9_3

Rockafellar, R. T. (1970). *Convex analysis, Princeton mathematical series* (p. 472). Princeton University Press.

Roques, C., Rupp, D. E., & Selker, J. S. (2017). Improved streamflow recession parameter estimation with attention to calculation of $-\mathrm{d}Q/\mathrm{d}t$. *Advances in Water Resources*, *108*, 29–43. https://doi.org/10.1016/j.advwatres.2017.07.013

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, *23*(3), 470–472. https://doi.org/10.1214/aoms/1177729394

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

Rupp, D. E., & Selker, J. S. (2006). Information, artifacts, and noise in $\mathrm{d}q/\mathrm{d}t - q$ recession analysis. *Advances in Water Resources*, *29*(2), 154–160. https://doi.org/10.1016/j.advwatres.2005.03.019

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

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

Sadegh, M., Vrugt, J. A., Xu, C., & Volpi, E. (2015). The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(\mathrm{ABC})}$. *Water Resources Research*, *51*(11), 9207–9231. https://doi.org/10.1002/2014WR016805

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global sensitivity analysis: The primer*. John Wiley & Sons.

Sankarasubramanian, A., Vogel, R. M., & Limbrunner, J. F. (2001). Climate elasticity of streamflow in the United States. *Water Resources Research*, *37*(6), 1771–1781. https://doi.org/10.1029/2000WR900330

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

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*(9), 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Scharnagl, B., Iden, S. C., Durner, W., Vereeken, H., & Herbst, M. (2015). Inverse modelling of in situ soil water dynamics: Accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals. In *Hydrology and Earth system sciences discussions* (Vol. 12, pp. 2155–2199).

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

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

Scheuerer, M., & Möller, D. (2015). Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Annals of Applied Statistics*, *9*(3), 1328–1349. https://doi.org/10.1214/15-aoas843

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

Schoups, G., & Vrugt, J. A. (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), W10531. https://doi.org/10.1029/2009WR008933

Schoups, G., Vrugt, J. A., Fenicia, F., & van de Giesen, N. C. (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), W10531. https://doi.org/10.1029/2009WR008648

Schwemmle, R., Demand, D., & Weiler, M. (2021). Technical note: Diagnostic efficiency—Specific evaluation of model performance. *Hydrology and Earth System Sciences*, *25*(4), 2187–2198. https://doi.org/10.5194/hess-25-2187-2021

Searcy, J. K. (1959). *Flow-duration curves, technical report 1542*. United States Geological Survey. https://doi.org/10.3133/wsp1542A

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

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

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

Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, *31*(2), 125–145. https://doi.org/10.1007/bf02289503

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

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling in Civil Engineering*, *1*(4), 407–414.

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

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

Spear, R. C., & Hornberger, G. (1980). Eutrophication in peel inlet-II. identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, *14*(1), 43–49. https://doi.org/10.1016/0043-1354(80)90040-8

Staël von Holstein, C.-A. S. (1970). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, *9*(3), 360–364. https://doi.org/10.1175/1520-0450(1970)009<0360:afosps>2.0.co;2

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

Storch, H. V., & Zwiers, F. W. (1999). *Statistical analysis in climate research* (1 ed., p. 484). Cambridge University Press. https://doi.org/10.1017/CBO9780511612336

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

Tashie, A., Pavelsky, T., & Band, L. E. (2020). An empirical reevaluation of streamflow recession analysis at the continental scale. *Water Resources Research*, *56*(1), e2019WR025448. https://doi.org/10.1029/2019WR025448

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

Tegos, S. A., Karagiannidis, G. K., Diamantoulakis, P. D., & Chatzidiamantis, N. D. (2022). New results for Pearson type iii family of distributions and application in wireless power transfer. *IEEE Internet of Things Journal*, *9*(23), 24038–24050. https://doi.org/10.1109/JIOT.2022.3189220

Thielen, J., Schaake, J., Hartman, R., & Buizza, R. (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*(2), 29–35. https://doi.org/10.1002/asl.168

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

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

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (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), W00B14. https://doi.org/10.1029/2008WR006825

Tracton, M. S., & Kalnay, E. (1993). Operational ensemble prediction at the national meteorological center: Practical aspects. *Weather and Forecasting*, *8*(3), 379–398. https://doi.org/10.1175/1520-0434(1993)008¡0379:OEPATN¿2.0.CO;2

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

Tyralis, H., & Papacharalampous, G. (2021). Quantile-based hydrological modelling. *Water*, *13*(23), 3420. https://doi.org/10.3390/w13233420

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.

Villez, K. (2017). *Analytical expressions to compute the continuous ranked probability score (CRPS), technical report 4*. Eawag, Aquatic Research, Swiss Federal Institute of Aquatic Science and Technology.

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

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

Von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer.

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. https://doi.org/10.1016/j.envsoft.2015.08.013

Vrugt, J. A. (2018). *Modelavg: A MATLAB toolbox for postprocessing of model ensembles, technical report*. University of California.

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

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

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., & Robinson, B. A. (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters*, *33*(19), L19817. https://doi.org/10.1029/2006GL027126

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

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., & Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, *39*(8), 1214. https://doi.org/10.1029/2002WR001746

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

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

Vrugt, J. A., Yumi de Oliveria, D., Schoups, G., & Diks, C. G. H. (2022). On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. *Journal of Hydrology*, *615*, 128542. https://doi.org/10.1016/j.jhydrol.2022.128542

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

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

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

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

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

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, *64*(327), 1073–1078. https://doi.org/10.2307/2283486

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., et al. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1–60. https://doi.org/10.1007/BF02562681

Wolfram Research, I. (2024). Mathematica online.

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Ye, M., Meyer, P. D., & Neuman, S. P. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, *44*(3), W03428. https://doi.org/10.1029/2008WR006803

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9), W09417. https://doi.org/10.1029/2007WR006716

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., & She, D. (2016). Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology*, *539*, 525–538. https://doi.org/10.1016/j.jhydrol.2016.05.053

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