ARC Geophysical Research

A Sandwich with Water: Bayesian & Frequentist Inference Under Model Misspecification

Jasper A. Vrugt*[1], Cees G.H. Diks[2], Ramon de Punder[2], and Peter Grünwald[3,4]

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

[2]Faculty Economics and Business, University of Amsterdam

[3]National Research Institute for Mathematics and Computer Science, Amsterdam

[4]Leiden University

Abstract

In this paper we review the foundation of frequentist inference, specifically maximum likelihood (ML) and M-estimation to point out a critical flaw of Bayesian methods for hydrologic model training and uncertainty quantification. Under model misspecification, the sensitivity $\widehat{\mathbf{A}}_n$ and variability $\widehat{\mathbf{B}}_n$ matrices of the ML model parameter values $\widehat{\boldsymbol{\theta}}_n$ provide conflicting information about the observed Fisher information $\widehat{\mathcal{I}}_n = n\,\widehat{\mathbf{A}}_n$ of the data $\omega_1,\ldots,\omega_n$ for $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)^\top$. As a result, the estimated ML parameter covariance matrix does not simplify to $\widehat{\mathcal{I}}_n^{-1}$ the inverse of the observed Fisher information as suggested by naive ML estimators and Bayesian methods but amounts instead to the sandwich matrix $\widehat{\mathcal{G}}_n^{-1} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1}$, where the observed Godambe information $\widehat{\mathcal{G}}_n$ is the fundamental currency of data informativeness under model misspecification. The *sandwich* matrix is a metaphor for a *meat* matrix $\widehat{\mathbf{B}}_n$ between two *bread* matrices $\widehat{\mathbf{A}}_n$ and yields asymptotically valid "robust standard errors" even when the likelihood function $L_n(\boldsymbol{\theta})$ (model) is incorrectly specified. The implications of the sandwich variance estimator are demonstrated in three case studies involving the modeling of soil water infiltration, watershed hydrologic fluxes, and the rainfall-discharge relationship. First and foremost, analytic and numerical results demonstrate that the sandwich variance estimator increases substantially hydrologic model parameter and predictive uncertainty. The sandwich estimator is invariant to likelihood stretching practiced by Generalized Likelihood Uncertainty Estimation as a remedy for over-conditioning and requires adjustments to the likelihood to yield asymptotically valid sandwich parameter estimates and inference via Monte Carlo simulation.

Keywords: Maximum Likelihood, Bayesian Inference, Model Misspecification, Naive Variance, Sandwich Variance, Sensitivity Matrix, Variability Matrix, Fisher Information, Godambe Information, Markov Chain Monte Carlo Simulation, Bootstrap Method, Soil Water Infiltration, Rainfall-Runoff Transformation, Hydrologic Modeling

*Corresponding Author
E-mail address: jasper@uci.edu
doi:10.5149/ARC-GR.1824

1 Introduction

Frequentist and Bayesian methods have found widespread application in hydrology and hydrometeorology for tasks such as model training (calibration) and forecasting of hydrologic states and fluxes. A default assumption in much of this work is that the likelihood function (i.e., the model) is correctly specified. Yet hydrologic models have well-documented limitations in resolving internal catchment processes and capturing hydrological change [148]. Consequently, the likelihood function is almost always misspecified, which undermines model training and statistical inference. This paper builds on the first author's work on hydrograph functionals and scoring rules [157] and examines the consequences of model misspecification for confidence regions and intervals derived from maximum likelihood (ML) and Bayesian methods, as well as frequentist approaches.

The mathematical-statistical theory underlying ML and Bayesian methods is elegant and compelling, yet their application requires assumptions that are often difficult to justify in practice. These assumptions have been vigorously debated in the hydrologic literature [2, 13, 64, 77, 105, 108, 125, 132, 141, 174], and spurred the development of alternative strategies for hydrologic model calibration and for quantifying parameter and predictive uncertainty [13, 15, 26, 55, 61, 62, 88, 91–93, 161, 163, 164]. The culprits are structural model errors which, when juxtaposed with measurement errors in driving exogenous variables such as precipitation, evaporation, and transpiration, produce nontraditional residual time series characterized by substantial variation in bias (nonstationarity), variance (heteroskedasticity), and correlation structure under different hydrologic conditions. Distribution-adaptive likelihood functions [133, 166] may provide the most appropriate specification of the likelihood function $L_n(\boldsymbol{\theta})$ for model outputs $y_1(\boldsymbol{\theta}),\ldots,y_n(\boldsymbol{\theta})$ and observations $\omega_1,\ldots,\omega_n$, but they are not immune to model misspecification or to the effects of measurement errors in the driving exogenous variables. As a result, all likelihood functions are misspecified to some degree, and asymptotic $100(1-\alpha)\%$ credible intervals will have less than nominal frequentist coverage probabilities. This poor calibration of the posterior credible regions is referred to as over-conditioning [14, 16] and is a result of the customary aleatoric treatment of residuals when in fact they are nonrandom (systematic) in nature.

Kleijn and van der Vaart [89] derived a Bernstein-von Mises theorem for Bayesian posteriors under misspecification and showed that even if parameters are correctly determined, misspecification can still cause a lack of calibration. Variational Bayesian posteriors suffer the same problem [19]. Accounting for misspecification should therefore be an important step in any modeling study, particularly when strong assumptions are made for computational tractability to support inferences about high-dimensional data and/or spatiotemporal extremes [153].

M-estimation (for "maximum-likelihood-type") is a generalization of ML and least-squares estimation to situations in which the data *have* outliers, extreme observations, and/or do *not* follow a normal distribution, making such estimators more robust to outliers and misspecification [78, 81]. The motivating ideas and principles of M-estimation originate from the pioneering work of John Wilder Tukey (1915-2000), one of the most influential mathematical and theoretical statisticians of the 20th century. His early articles on nonparametric methods, rank-based inference, and order statistics express concern about departures from normality even before the term "robustness" was first used by Box [21], let alone fully developed. Tukey [149] realized early on the practical difficulties of inference with small samples when the data do not follow a Gaussian distribution and Tukey [150] pointed out the excessive sensitivity of classical statistical methods of regression analysis to small departures from idealized hypotheses. In the landmark treatise "The future of data analysis" Tukey [151, p. 3] set out an agenda for data analysis in mathematical statistics with emphasis on "*. . . developing organized techniques of dealing with outliers, wild shots, blunders, or large deviations*" when the hypotheses on which these methods customarily build do not hold. Tukey's ideas were developed further by Huber [78, 79] and in the PhD thesis of Hampel [66], which ultimately led to the emergence of the new field of *robust statistics* or M-estimation. Huber [81, p. 1] defines a robust statistic as one that is "*. . . resistant to errors in the results, produced by deviations from assumptions*". Accordingly, robust estimation seeks adaptation of conventional statistical methods to make them less sensitive to spurious measurements and departures from idealized assumptions [36].

The main idea of M-estimation is to estimate $\boldsymbol{\theta}$ by minimizing a sample-based measure of discrepancy between data $\omega_1, \ldots, \omega_n$ and model $y_1(\boldsymbol{\theta}), \ldots, y_n(\boldsymbol{\theta})$ that achieves desirable bias and efficiency when the assumed data-generating process is correct, while avoiding pathological behavior when the data deviate moderately from that assumption. More broadly, although so-called robust standard errors are ubiquitous in statistics, biostatistics, and econometrics, their use in hydrology remains surprisingly limited: a Google Scholar search (July 5, 2025) yields approximately 5.38 million records containing "robust standard error" in the title or abstract, yet only a small number of hydrologic applications, including rainfall-runoff modeling [28], precipitation frequency analysis [131], and studies of rainfall and streamflow extremes [86, 115, 124, 137, 175]. Related ideas appear in the work of Bárdossy and Singh [7], who explored robust estimation through data-depth concepts [152]. In our own research, M-estimation did not initially arise from a deliberate effort to develop a robust inferential framework for hydrologic modeling under epistemic uncertainty, but instead emerged serendipitously from work on probabilistic model evaluation, where it appears as the asymptotic distribution of so-called hydrograph functionals, watershed signatures embodied in strictly proper scoring rules [157]. This perspective extends naturally to machine learning, where M-estimators provide practical and theoretically grounded alternatives to classical ML or fully Bayesian approaches in the presence of outliers, noisy observations, or structural misspecification.

Our emphasis here is not on promoting specific robust loss functions for model training but on the *asymptotic behavior* common to M-estimators under misspecification. This exposes a critical deficiency of ML and Bayesian workflows in the presence of structural model and input-data errors. Under misspecification, the asymptotic covariance of M-estimators including ML is not the inverse Fisher information (as under correct specification) but the *sandwich* covariance $\frac{1}{n}\,\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\,\mathbf{B}_n(\widehat{\boldsymbol{\theta}}_n)\,\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$, where the *sensitivity* matrix ($\mathbf{A}_n(\boldsymbol{\theta})$, "bread") captures local curvature of the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$ and the *variability* matrix ($\mathbf{B}_n(\boldsymbol{\theta})$, "meat") measures the variability of the *score* $\mathbf{g}_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}_n(\boldsymbol{\theta})$. These sandwich estimates are also called "robust standard errors" or "Eicker-Huber-White standard errors" [47, 79, 171] and yield consistent large-sample standard errors and intervals under misspecification. With independent and identically distributed (i.i.d.) residuals, $\mathbf{B}_n(\boldsymbol{\theta})$ is equivalent to the *score* variance. When serial dependence is present among the residuals, it must be replaced by the long-run variance $\boldsymbol{\beta}_n(\boldsymbol{\theta})$, obtained as the sum of *score* autocovariances derived from so-called HAC estimators [5, 6, 113, 114].

Building on classical foundations, this paper explores the asymptotic distribution of the ML or maximum *a posteriori* (MAP) parameter values under model misspecification. We revisit the theory of frequentist inference, specifically ML and M-estimation, to point to a critical flaw of Bayesian methods in the face of model misspecification, which leads to a naive characterization of hydrologic model parameter and/or predictive uncertainty. Model misspecification gives rise to the so-called sandwich variance estimator of Huber [79] and its generalization to dependent data in a time series context by Cameron and Trivedi [27]. The implications of the sandwich variance estimator for hydrologic model training and uncertainty quantification are illustrated through three case studies involving (i) soil water infiltration, (ii) watershed hydrologic fluxes, and (iii) the rainfall-discharge transformation. The first two case studies provide an analytic demonstration of the consequences of model misspecification on parameter uncertainty as described by ML and Bayesian methods. The third and last case study illustrates both the naive and sandwich variance estimators by applying them to measured streamflow data using a collection of different likelihood functions.

This paper is accompanied by DREAM-Suite[1], a Matlab-Python software package for Bayesian model training, evaluation, and diagnostics [156]. The package, available from the first author's GitHub account (`https://github.com/jaspervrugt`), includes the case studies presented herein and also provides an option for sandwich-adjusted Markov chain Monte Carlo simulation, as introduced by Vrugt and Diks [160] in a companion paper.

2 Reader's Guide and Six-Point Primer

Before we introduce our usage of notation in Section 3, we briefly explain how to navigate the paper and what the main ideas are.

2.1 Roadmap

This paper draws on M-estimation theory and the associated sandwich covariance to: (i) translate these results into a unified $\mathbf{A}_n$–$\mathbf{B}_n$ framework for hydrologic model training and uncertainty quantification; (ii) diagnose their implications for ML/Bayesian workflows under misspecification (over-conditioning); and (iii) present worked case studies. Sections 3–4 present the classical robust/M-estimation (sandwich) framework. The same $\mathbf{A}_n$–$\mathbf{B}_n$ machinery underlies heteroskedasticity-consistent [47, 79, 170] standard errors for regression models, the generalized method of moments [71] in econometrics, and generalized estimating equations [42, 100, 117] in biostatistics, enabling robust uncertainty estimation under misspecification. While this foundational theory can be demanding for non-specialists, we include it in the main text to make the statistical underpinning explicit. Readers primarily interested in applications may skip the derivations, consult the six-point primer below, and proceed directly to the case studies in Section 5. All key equations are referenced inline.

2.2 Six-Point Primer

We give a succinct primer on M-estimation, specifically, its historical lineage, terminology, and implications for Bayesian inference.

(i) Uncertainty is characterized by two $d \times d$ matrices: the sensitivity or "bread" matrix $\mathbf{A}_n(\boldsymbol{\theta})$, which captures the local curvature of the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$, and the variability or "meat" matrix $\mathbf{B}_n(\boldsymbol{\theta})$, which measures the variability of the *score* $\mathbf{g}_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}_n(\boldsymbol{\theta})$.

(ii) When the *scores* exhibit serial dependence, $\mathbf{B}_n(\boldsymbol{\theta})$ must be replaced by its long-run variance $\boldsymbol{\beta}_n(\boldsymbol{\theta})$, which we can estimate using a so-called HAC estimator. Under i.i.d. data, $\boldsymbol{\beta}_n(\boldsymbol{\theta}) = \mathbf{B}_n(\boldsymbol{\theta})$.

(iii) Under correct specification, the information identity holds, $\mathbb{E}[\mathbf{A}_n(\boldsymbol{\theta})] = \mathbb{E}[\mathbf{B}_n(\boldsymbol{\theta})]$, for every $\boldsymbol{\theta}$ if the expectation is taken under the model density. Consequently, the Fisher-based ("naive") covariance $\boldsymbol{\Sigma}_n^{\text{naive}} = \frac{1}{n}\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$ correctly characterizes the uncertainty of the ML/MAP estimates $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,d})^\top$.

[1] DREAM is an acronym for DiffeRential Evolution Adaptive Metropolis

(iv) Under misspecification, $\mathbb{E}[\mathbf{A}_n(\boldsymbol{\theta})] \neq \mathbb{E}[\mathbf{B}_n(\boldsymbol{\theta})]$, and the information identity no longer holds. In this case, the correct large-sample variance-covariance matrix of the ML/MAP estimates is the sandwich matrix $\boldsymbol{\Sigma}_n^{\text{sand}} = \frac{1}{n}\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\,\mathbf{B}_n(\widehat{\boldsymbol{\theta}}_n)\,\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$, which corresponds to the inverse of the Godambe information matrix.

(v) When sampling from a Bayesian posterior using Markov chain Monte Carlo (MCMC) under flat or weakly informative priors, the target density is proportional to $\exp\big(\mathcal{L}_n(\boldsymbol{\theta})\big)$. The local spread of this target is governed by the curvature $\frac{1}{n}\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$ at the mode. Consequently, unless adjustments are made for misspecification, posterior draws inherit the naive covariance $\boldsymbol{\Sigma}_n^{\text{naive}}$ rather than the sandwich covariance $\boldsymbol{\Sigma}_n^{\text{sand}}$.

(vi) To obtain MCMC-based uncertainty intervals consistent with $\boldsymbol{\Sigma}_n^{\text{sand}}$, one may (i) post-process posterior draws via the open-faced sandwich adjustment [134], (ii) sample from an adapted score-based target density [52], or (iii) sample under magnitude-, curvature-, or sandwich-adjusted likelihoods [119, 127, 160]. These approaches modify the statistical target rather than the MCMC sampling algorithm itself.

To make the ideas more concrete, Appendix B works through a simple illustrative example based on a normal data-generating process, $\Omega \sim \mathcal{N}(\mu, \sigma^2)$. Using a sample $\omega_1, \ldots, \omega_n$ drawn from this distribution, we derive analytic expressions for the ML estimators $(\widehat{\mu}_n, \widehat{\sigma}^2_n)$ as well as for the associated sensitivity matrix $\mathbf{A}_n(\boldsymbol{\theta})$ and variability matrix $\mathbf{B}_n(\boldsymbol{\theta})$. These derivations confirm the information identity under correct specification as stated in item (iii) of the six-point primer above. Figure 1 displays the log-likelihood surface of the normal distribution model together with the 95% confidence region around the ML estimate $(\widehat{\mu}_n, \widehat{\sigma}^2_n)$. In the correctly specified case, the naive and sandwich regions nearly coincide up to finite-sample

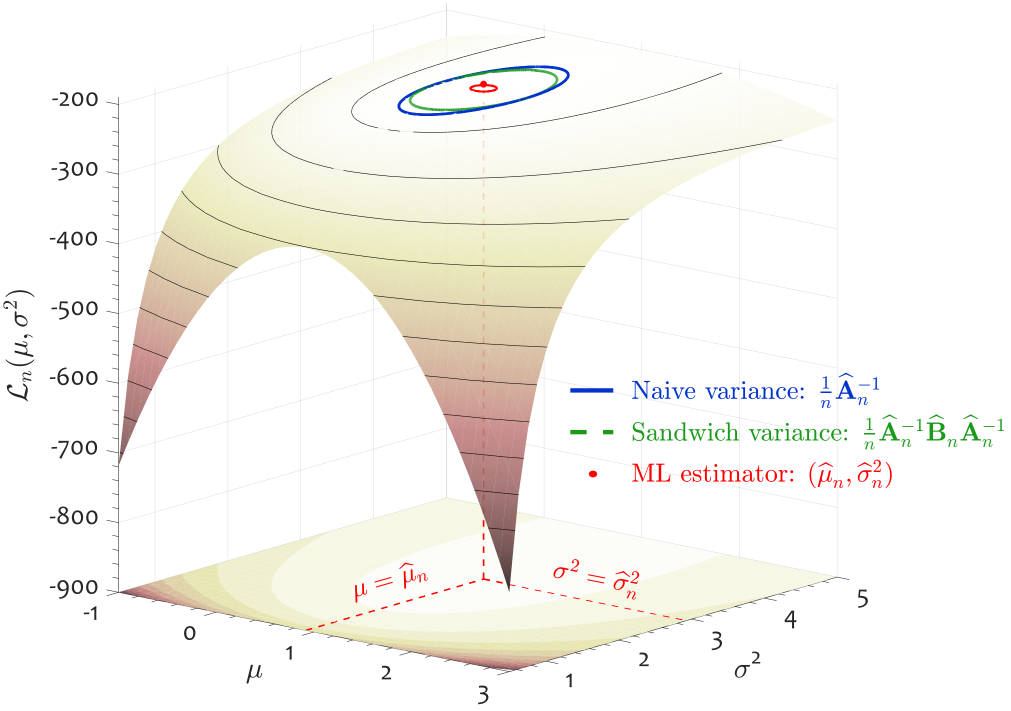

Figure 1: Surface and contours of the log-likelihood $\mathcal{L}_n(\mu, \sigma^2)$ for the normal model $\mathcal{N}(\mu, \sigma^2)$ in Appendix B over $\mu \in [-1, 3]$ and $\sigma^2 \in [0.5, 5]$, based on $n = 100$ data points $\omega_1, \ldots, \omega_{100}$ drawn from $\mathcal{N}(1, 3)$. The red dot marks the ML estimate $(\widehat{\mu}_n, \widehat{\sigma}^2_n)$ and blue and green ellipses show 95% confidence regions from the naive and sandwich estimators, respectively; $\widehat{\mathbf{A}}_n = \mathbf{A}_n(\widehat{\mu}_n, \widehat{\sigma}^2_n)$ and $\widehat{\mathbf{B}}_n = \mathbf{B}_n(\widehat{\mu}_n, \widehat{\sigma}^2_n)$.

differences. Under misspecification the ellipses separate in size and orientation, foreshadowing the results that follow.

3 Notation

Boldface uppercase letters denote matrices, $\mathbf{A}$, boldface lowercase letters signify vectors, $\mathbf{a}$, and italic lowercase letters are used for scalars, a. The superscripts "$\top$" and "-1" stand for matrix transpose and matrix inverse, respectively. By default, we assume column vectors and, thus, $\mathbf{a} = (a_1, \ldots, a_m)^\top$ is a $m \times 1$ vector. If $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is a vector of d random variables then we say its *expectation* is the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ and write $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$, thus combining d scalar equations into one vector equation. The *variance* of random vector $\mathbf{X}$ is the $d \times d$ matrix $\boldsymbol{\Sigma}$ whose (i,j)th element is

$$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)],$$

where $i, j \in (1, \ldots, d)$. In vector notation we write

$$\begin{aligned}\mathrm{Var}[\mathbf{X}] &= \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\big] \\ &= \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top,\end{aligned} \tag{1}$$

thus, combining d^2 scalar equations into one matrix equation. In the above formulation, $\mathbf{X} - \boldsymbol{\mu}$, equals a $d \times 1$ vector and the vector outer (cross) product of $\mathbf{X} - \boldsymbol{\mu}$ and $(\mathbf{X} - \boldsymbol{\mu})^\top$ returns a $d \times d$ matrix. The inner (dot) product of two d-vectors $\mathbf{a}$ and $\mathbf{b}$ is equal to $\mathbf{a}^\top\mathbf{b}$ and returns a scalar. For notational convenience, we write $\mathbf{Z}_n^z(\boldsymbol{\theta})$ instead of $\big(\mathbf{Z}(\boldsymbol{\theta})\big)^z$, where the superscript $z \in \{-1, \top\}$. This convention applies to any matrix $\mathbf{Z}$, such as $\mathbf{A}$, $\mathbf{B}$, $\mathbf{J}$, $\mathcal{I}$ and $\mathcal{G}$. In the same spirit, we write $\nabla_{\boldsymbol{\theta}}^\top \mathcal{L}_\omega(\boldsymbol{\theta})$ to denote the transpose of the gradient vector $\nabla_{\boldsymbol{\theta}} \mathcal{L}_\omega(\boldsymbol{\theta})$, so that outer products are written compactly.

We follow the standard mathematical notation for functions. Thus, $f(x) : \mathbb{R} \to \mathbb{R}$ is a scalar-valued function of x, $\mathbf{f}(x) : \mathbb{R}^1 \to \mathbb{R}^n$ and $\mathbf{f}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^n$ are vector-valued functions of x and $\mathbf{x}$, respectively. We denote statistical distributions with calligraphic letters. For example, if random variable X follows a univariate normal distribution, we write $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, we write $X \sim \mathcal{U}(a, b)$ for a uniform distribution on $[a, b]$; $b > a$, $X \sim \mathcal{T}(\nu)$ for a Student's t distribution with $\nu > 0$ degrees of freedom, $X \sim \mathcal{F}(d_1, d_2)$ for a Fisher-Snedecor F distribution with $d_1 > 0$ and $d_2 > 0$ degrees of freedom, and $X \sim \mathcal{E}(\boldsymbol{\delta})$ for some arbitrary error distribution with scale, shape and possibly other parameters $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)^\top$. If $\mathbf{X}$ is multivariate normally distributed with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and $d \times d$ covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}]$, we write $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and use $\mathbf{X} \sim \mathcal{U}_d(\mathbf{a}, \mathbf{b})$ for the continuous d-variate uniform distribution on the closed-region $[\mathbf{a}, \mathbf{b}]$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{d \times 1}$ and $a_j < b_j$ for all $j = 1, \ldots, d$. We recall that if $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $a\mathbf{X} \sim \mathcal{N}_d(a\boldsymbol{\mu}, a^2\boldsymbol{\Sigma})$. It is customary to use a capital letter for a cumulative distribution function (cdf), as opposed to a lower case letter for its probability density function (pdf) or probability mass function. Specifically, we write $F(x; d_1, d_2)$ for the cdf of the F distribution $\mathcal{F}(d_1, d_2)$ and $T(x; \nu)$ for the cdf of the Student's t distribution $\mathcal{T}(\nu)$. Then, the Greek letter $\alpha \in (0, 1)$ denotes the significance level, and $\gamma = 1 - \alpha$ the confidence level.

4 Maximum Likelihood Estimation

4.1 Definitions and Terminology

Throughout, let Ω denote the continuous random variable of interest (not the sample space), with realizations ω_t. Similarly, let Y denote the model-simulated equivalent with realizations

y_t. For a record of length n, $\boldsymbol{\omega}_n = (\omega_1,\dots,\omega_n)^\top$ and $\mathbf{y}_n = (y_1,\dots,y_n)^\top$ are $n \times 1$ vectors of materialized and modeled outcomes, respectively. Suppose $\boldsymbol{\theta}_0$ are the *true* (but unknown) parameter values of the data-generating process $\mathfrak{S}$ and let $\mathcal{M}_t(\boldsymbol{\theta};\mathbf{I}_t)$ denote a deterministic forward model which links parameters $\boldsymbol{\theta} = (\theta_1,\dots,\theta_d)^\top$ and exogenous inputs $\mathbf{I}_t$ (e.g., time t, rainfall p_t) to model output $y_t = \mathcal{M}_t(\boldsymbol{\theta};\mathbf{I}_t)$. Residuals, $e_t = \omega_t - y_t$, are modeled as $e_t \sim \mathcal{E}(\boldsymbol{\delta})$ with nuisance parameters $\boldsymbol{\delta}$. Assuming conditional independence across t, the likelihood is $L_n(\boldsymbol{\theta} \mid \boldsymbol{\delta}) = \prod_{t=1}^n f(e_t;\boldsymbol{\delta})$, where $f(\cdot\,;\boldsymbol{\delta})$ is the pdf of the error distribution $\mathcal{E}$.

We summarize the key scalars, vectors, and matrices used in ML and Bayesian theory.

1. The likelihood is a scalar and written $L_\omega(\boldsymbol{\theta})$ for a single datum ω. For a data set $\omega_1,\dots,\omega_n$, we write $L_n(\boldsymbol{\theta})$. The symbol $\mathcal{L}_n(\boldsymbol{\theta})$ denotes the natural logarithm of $L_n(\boldsymbol{\theta})$.

2. The *score* $\mathbf{g}_\omega(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}_\omega(\boldsymbol{\theta})$ is the $d \times 1$ vector of first-order partial derivatives of the log-likelihood $\mathcal{L}_\omega(\boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\theta} = (\theta_1,\dots,\theta_d)^\top$.

3. The $n \times d$ Jacobian matrix $\mathbf{J}_n(\boldsymbol{\theta})$ stores as rows the *score functions* $\mathbf{g}_{\omega_t}(\boldsymbol{\theta})$ corresponding to each observation $\omega_1,\dots,\omega_n$.

4. The $d \times d$ Hessian matrix $\mathbf{H}_\omega(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}}\mathcal{L}_\omega(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{g}_\omega(\boldsymbol{\theta})$ contains the second-order partial derivatives of $\mathcal{L}_\omega(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The total Hessian is given by $\mathbf{H}_n(\boldsymbol{\theta}) = \sum_{t=1}^n \mathbf{H}_{\omega_t}(\boldsymbol{\theta})$, equivalently $\mathbf{H}_n(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}}\mathcal{L}_n(\boldsymbol{\theta})$.

5. The $d \times d$ sensitivity matrix is defined as $\mathbf{A}_n(\boldsymbol{\theta}) = -\frac{1}{n}\mathbf{H}_n(\boldsymbol{\theta})$ with probability limit $\mathbf{A}_0 = \operatorname{plim}\mathbf{A}_n(\boldsymbol{\theta}_0)$.

6. The $d \times d$ variability matrix is $\mathbf{B}_n(\boldsymbol{\theta}) = \frac{1}{n}\mathbf{J}_n^\top(\boldsymbol{\theta})\mathbf{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^n \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\omega_t}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^\top\mathcal{L}_{\omega_t}(\boldsymbol{\theta})$, and its probability limit in $\boldsymbol{\theta}_0$ is denoted by $\mathbf{B}_0 = \operatorname{plim}\mathbf{B}_n(\boldsymbol{\theta}_0)$.

7. The $d \times d$ Fisher information matrix $\mathcal{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}_n(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^\top\mathcal{L}_n(\boldsymbol{\theta})]$ is the expectation with respect to $\boldsymbol{\theta}$ (i.e., law of Ω under $\boldsymbol{\theta}$) of the outer product of the gradient of the log-likelihood at $\boldsymbol{\theta}$. This is also referred to as the expected Fisher information.

8. The matrix inverse of the expected Fisher information $\mathcal{I}_n^{-1}(\boldsymbol{\theta}_0)$ evaluated at $\boldsymbol{\theta}_0$ is a $d \times d$ covariance matrix. This naive variance is the asymptotic variance of the ML estimator under *correct specification*.

9. The $d \times d$ Godambe information matrix $\mathcal{G}_n(\boldsymbol{\theta}) = n\mathbf{A}_n(\boldsymbol{\theta})\mathbf{B}_n^{-1}(\boldsymbol{\theta})\mathbf{A}_n(\boldsymbol{\theta})$, with $\mathcal{G}_0 = \operatorname{plim}\frac{1}{n}\mathcal{G}_n(\boldsymbol{\theta}_0) = \mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{A}_0$.

10. The matrix inverse of Godambe information $\mathcal{G}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$ is a $d \times d$ covariance matrix. This robust or sandwich variance is the asymptotic variance of the ML estimator under *misspecification*.

The only matrix whose dimensions grows with sample size is the $n \times d$ Jacobian matrix $\mathbf{J}_n(\boldsymbol{\theta})$. All other matrices have fixed dimensions. The entries of the $d \times d$ matrices $\mathcal{I}_n$, $\mathbf{H}_n$, and $\mathcal{G}_n$ increase with sample size n. These "information" matrices grow linearly (on average) with n reflecting a steadily accumulating amount of information about the unknown parameters with more data. In contrast, $\mathbf{A}_n$ and $\mathbf{B}_n$ are sample averages of the sensitivity and variability matrices for n data points. Cameron and Trivedi [27] treat these two $d \times d$ matrices as estimators of $\mathbf{A}_0$ and $\mathbf{B}_0$, respectively, the probability limits under the *true* but unknown parameters $\boldsymbol{\theta}_0$, hence their subscript '0'.

4.2 Theory

Suppose for now that data $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ are independent and identically distributed realizations of some random variable Ω with underlying cdf Q. Let $q_0(\omega) = q_\Omega(\omega; \boldsymbol{\theta}_0)$ be the *true* pdf of Ω for the $d \times 1$ vector $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0d})^\top$ of parameter values of the data-generating process $\mathfrak{S}$. The *true* parameter values $\boldsymbol{\theta}_0$ are interior to the feasible parameter space $\boldsymbol{\Theta}$ which, in turn, is a compact subset of $\mathbb{R}^d$. The pdf $q_\Omega(\omega; \boldsymbol{\theta})$ describes the probability that we observe the outcome ω given the values of $\boldsymbol{\theta}$. The likelihood $L_\omega(\boldsymbol{\theta})$ that parameters $\boldsymbol{\theta}$ have generated observation ω is now equal to $q_\Omega(\omega; \boldsymbol{\theta})$. We write $L_\omega(\boldsymbol{\theta})$ as a shorthand notation for $L(\omega \mid \boldsymbol{\theta})$ and drop the explicit dependence of the likelihood on the nuisance variables. Then the joint likelihood $L_n(\boldsymbol{\theta})$ for the n-vector of observations $\omega_1, \ldots, \omega_n$ is equal to the product of $L_{\omega_1}(\boldsymbol{\theta}), \ldots, L_{\omega_n}(\boldsymbol{\theta})$ as follows

$$L_n(\boldsymbol{\theta}) = \prod_{t=1}^{n} q_\Omega(\omega_t; \boldsymbol{\theta}), \tag{2}$$

and the natural logarithm of the likelihoods becomes

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \bigl\{ \log\bigl(q_\Omega(\omega_t; \boldsymbol{\theta})\bigr) \bigr\}. \tag{3}$$

Subscript n in $L_n(\boldsymbol{\theta})$ and $\mathcal{L}_n(\boldsymbol{\theta})$ explicates the dependence of the likelihood and log-likelihood functions on all observations $\omega_1, \ldots, \omega_n$. The independence assumption that is evoked by Equations (2) and (3) may be wrong without affecting the consistency of ML estimators [41]. We make the default assumptions that (1) the first- and second-order partial derivatives of $\mathcal{L}_\omega(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist almost everywhere, (2) the support of Ω does not depend on $\boldsymbol{\theta}$ and (3) the integral of $\mathcal{L}_\omega(\boldsymbol{\theta})$ can be differentiated under the integral sign with respect to $\boldsymbol{\theta}$. Although often elided, these regularity conditions usually hold for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with use of a continuously differentiable likelihood function and fixed parameter ranges. Derivatives must be bounded to permit differentiation under the integral sign [29].

The gradient of the log-likelihood $\mathcal{L}_\omega(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is called the *score function*. We denote it by the bold lowercase letter $\mathbf{g}$ for gradient vector

$$\begin{aligned} \mathbf{g}_\omega(\boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} \mathcal{L}_\omega(\boldsymbol{\theta}) &\equiv \frac{\partial \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \begin{bmatrix} \dfrac{\partial \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_d} \end{bmatrix}. \end{aligned} \tag{4}$$

The *total score* $\mathbf{g}_n(\boldsymbol{\theta})$ for $\omega_1, \ldots, \omega_n$ is the sum of the individual *score* contributions

$$\mathbf{g}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \mathbf{g}_{\omega_t}(\boldsymbol{\theta}), \tag{5}$$

and is also conveniently written $\mathbf{g}_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$. The entries of this $d \times 1$ vector play a central role in (generalized) estimating equations [59, 60, 100] and are a key ingredient of extremum or M-estimators [80]. An estimating function is unbiased if its expectation vanishes for all parameter values, i.e., $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{g}_n(\boldsymbol{\theta})] = \mathbf{0}_d$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\mathbf{0}_d$ is the $d \times 1$ zero vector and the expectation is taken under the data-generating process indexed by $\boldsymbol{\theta}$ [34, 58, 104]. This property ensures consistency (under stipulated regularity conditions) and guarantees that ML estimators converge to the *true* parameter values $\boldsymbol{\theta}_0$ of $\mathfrak{S}$ as sample size n increases. Stacking the single-observation *score* vectors as rows yields a Jacobian matrix $\mathbf{J}_n(\boldsymbol{\theta}) \in \mathbb{R}^{n \times d}$ whose (i,j)th entry is $[\mathbf{J}_n(\boldsymbol{\theta})]_{ij} = \partial \mathcal{L}_{\omega_t}(\boldsymbol{\theta})/\partial \theta_j$. The *total score* is then $\mathbf{g}_n(\boldsymbol{\theta}) = \mathbf{J}_n^\top(\boldsymbol{\theta})\mathbf{1}_n$, and its jth component $g_{nj}(\boldsymbol{\theta})$ equals the sum of the jth column of $\mathbf{J}_n(\boldsymbol{\theta})$ or $g_{nj}(\boldsymbol{\theta}) = \sum_{t=1}^n [\mathbf{J}_n(\boldsymbol{\theta})]_{ij}$.

The second derivative of $\mathcal{L}_\omega(\boldsymbol{\theta})$ measures the curvature of the log-likelihood surface with respect to the parameters. This square $d \times d$ Hessian or sensitivity matrix $\mathbf{H}_\omega(\boldsymbol{\theta})$ has entries

$$\begin{aligned}\mathbf{H}_\omega(\boldsymbol{\theta}) \equiv \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_\omega(\boldsymbol{\theta}) &\equiv \frac{\partial}{\partial \boldsymbol{\theta}^\top}\left(\frac{\partial \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \\ &= \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_1^2} & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_d} \\ \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_1} & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_2} & \cdots & \dfrac{\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})}{\partial \theta_d^2} \end{bmatrix}.\end{aligned}$$

The order of differentiation does not matter; $\partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})/\partial \theta_i \partial \theta_j = \partial^2 \mathcal{L}_\omega(\boldsymbol{\theta})/\partial \theta_j \partial \theta_i$ for all $i \neq j \in (1,\dots,d)$ and, thus, the Hessian is a symmetric matrix. Through matrix addition we yield the total Hessian

$$\mathbf{H}_n(\boldsymbol{\theta}) = \sum_{t=1}^n \mathbf{H}_{\omega_t}(\boldsymbol{\theta}), \tag{6}$$

which is also written $\mathbf{H}_n(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$.

By definition, the ML parameter estimates $\widehat{\boldsymbol{\theta}}_n$ maximize the log-likelihood

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\; \mathcal{L}_n(\boldsymbol{\theta}),$$

with feasible set $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$. Under regularity, the log-likelihood is locally concave at $\widehat{\boldsymbol{\theta}}_n$, so the Hessian $\mathbf{H}_n(\widehat{\boldsymbol{\theta}}_n) = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)$ is negative (semi)definite and has non-positive diagonal entries. It is therefore convenient to work with the negative Hessian or Fisher information matrix

$$\mathcal{I}_n(\boldsymbol{\theta}) = -\,\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})\big],$$

as measure of the local curvature of the log-likelihood function. A small value of $[\mathcal{I}_n(\widehat{\boldsymbol{\theta}}_n)]_{jj}$ indicates that $\mathcal{L}_n(\boldsymbol{\theta})$ is relatively flat in θ_j near the optimum, whereas a large value indicates sharper curvature and, hence, greater local information about θ_j. Hence, Fisher information measures the average amount of information that an observable random variable Ω carries about one or more unknown parameters $\theta_1,\dots,\theta_d$ upon which the probabilities of $\omega_1,\dots,\omega_n$ depend. The variable $\mathcal{I}_n(\boldsymbol{\theta}_0)$ is also referred to as the *expected Fisher information* whereas its sample equivalent $-\mathbf{H}_n(\widehat{\boldsymbol{\theta}}_n)$ of Equation (6) is known as the *observed Fisher information.* From now on, we omit the subscript $\boldsymbol{\theta}$ in the vector differential operator ∇ as differentiation of the log-likelihood function is always with respect to the parameters.

Hessian matrices play a central role in Newton-type optimization as coefficient matrix of the quadratic term in the local second-order Taylor expansion of the objective function $\Phi(\cdot)$

$$\Phi(\boldsymbol{\theta} + \mathbf{h}) = \Phi(\boldsymbol{\theta}) + \nabla\Phi(\boldsymbol{\theta})\mathbf{h} + \tfrac{1}{2}\mathbf{h}^\top \nabla^2 \Phi(\boldsymbol{\theta})\mathbf{h} + \text{h.o.t.}, \tag{7}$$

where 'h.o.t.' stands for *higher order terms* in $\mathbf{h}$. Under standard regularity conditions, the ML estimator is consistent, thus, as $n \to \infty$ the ML parameter values $\widehat{\boldsymbol{\theta}}_n$ will converge to their *true* values $\boldsymbol{\theta}_0 = (\theta_{01}, \dots, \theta_{0d})^\top$ for data generated under $\mathfrak{S}$. Equivalently, if $\mathbf{h}_n = \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \in \mathbb{R}^d$, then $\mathbf{h}_n \xrightarrow{\text{p}} \mathbf{0}_d$. Following Equation (7) the *score function* $\Phi(\widehat{\boldsymbol{\theta}}_n) = \nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)$ can be expanded around $\boldsymbol{\theta}_0$ to yield

$$\nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n) = \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) + \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \text{h.o.t.}, \tag{8}$$

where $\mathbf{H}_n(\boldsymbol{\theta}_0) = \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)$ is the Hessian matrix where the higher-order terms formally are of order $o_{\text{p}}(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|)$.[2] For strictly concave log-likelihood functions a unique maximum $\widehat{\boldsymbol{\theta}}_n$ can be solved by finding the point $\boldsymbol{\theta}$ at which $\nabla \mathcal{L}_n(\boldsymbol{\theta}) = \mathbf{0}_d$. Thus, $\widehat{\boldsymbol{\theta}}_n$ will be a root of the *total score* $\mathbf{g}_n(\boldsymbol{\theta})$ defined in Equation (5).

Equation (8) reduces to

$$\mathbf{0}_d = \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) + \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \text{h.o.t.} \tag{9}$$

If we pre-multiply each term with $\big(\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)\big)^{-1}$ and rearrange the expression, we yield

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = -\big(\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)\big)^{-1} \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) + \text{h.o.t.}, \tag{10}$$

which is the starting point of the asymptotic analysis in the next subsection.

4.3 Asymptotic Behavior

The realized matrices $\mathbf{H}_n(\boldsymbol{\theta}_0)$, $\mathcal{I}_n(\boldsymbol{\theta}_0)$ and $\mathcal{G}_n(\boldsymbol{\theta}_0)$ are of order n as the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$ is a function of n data points, $\omega_1, \dots, \omega_n$. To understand the asymptotic behavior of the ML parameter estimates $\widehat{\boldsymbol{\theta}}_n$ we multiply both sides of Equation (10) with $\sqrt{n}$ as follows

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \sqrt{n}\big(-\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)\big)^{-1} \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) + \text{h.o.t.} \\ &= \big(-\tfrac{1}{n}\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)\big)^{-1} \tfrac{1}{\sqrt{n}} \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) + \text{h.o.t.}\end{aligned} \tag{11}$$

By the (weak) law of large numbers, the matrix inside brackets on the right-hand side

$$\mathbf{A}_n(\boldsymbol{\theta}_0) = -\tfrac{1}{n}\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0) \xrightarrow{\text{p}} \mathbf{A}_0,$$

converges in probability "$\xrightarrow{\text{p}}$" to $\mathbf{A}_0$, where the $d \times d$ matrix

$$\mathbf{A}_0 = \text{plim}\, \tfrac{1}{n}\mathbf{A}_n(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \tfrac{1}{n}\mathbb{E}_{\boldsymbol{\theta}_0}[-\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_0)],$$

is the expected Fisher information per observation. Under i.i.d. assumption, this simplifies to $\mathbf{A}_0 = \mathbb{E}_{\boldsymbol{\theta}_0}[-\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta}_0)]$. By the continuous mapping theorem, $\mathbf{A}_n(\boldsymbol{\theta}_0)^{-1} \xrightarrow{\text{p}} \mathbf{A}_0^{-1}$. According to the central limit theorem, the scaled *score* on the right-hand side of Equation (11)

$$\tfrac{1}{\sqrt{n}} \nabla \mathcal{L}_n(\boldsymbol{\theta}_0) \xrightarrow{\text{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{B}_0),$$

converges in distribution "$\xrightarrow{\text{d}}$" to a zero-mean d-variate normal random vector $\mathcal{N}_d(\mathbf{0}, \mathbf{B}_0)$ with covariance

$$\mathbf{B}_0 = \underset{n\to\infty}{\text{plim}}\, \tfrac{1}{n}\nabla \mathcal{L}_n(\boldsymbol{\theta}_0)\nabla^\top \mathcal{L}_n(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \tfrac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\boldsymbol{\theta}_0}[\nabla \mathcal{L}_{\Omega_t}(\boldsymbol{\theta}_0)\nabla^\top \mathcal{L}_{\Omega_t}(\boldsymbol{\theta}_0)].$$

[2] We use $a_n = o_{\text{p}}(b_n)$ to mean $a_n/b_n \xrightarrow{\text{p}} 0$, and $\|\cdot\|$ denotes the Euclidean norm.

Under the i.i.d. assumption, this reduces to $\mathbf{B}_0 = \mathbb{E}_{\boldsymbol{\theta}_0}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta}_0)\nabla^\top \mathcal{L}_\Omega(\boldsymbol{\theta}_0)]$, the variability matrix for a single observation. Finally, by Slutsky's theorem and the continuous mapping theorem, we arrive at the limiting distribution of the ML estimator

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}),$$

and the estimator $\widehat{\boldsymbol{\theta}}_n$ itself is said to be asymptotically normally distributed with asymptotic mean $\boldsymbol{\theta}_0$ and asymptotic covariance $\frac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$. The bread and meat matrices can be written using the *score function* of Equation (4) as follows

$$\mathbf{A}_n(\boldsymbol{\theta}) = -\tfrac{1}{n}\sum_{t=1}^{n} \nabla \mathbf{g}_{\omega_t}(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{B}_n(\boldsymbol{\theta}) = \tfrac{1}{n}\sum_{t=1}^{n} \mathbf{g}_{\omega_t}(\boldsymbol{\theta})\mathbf{g}_{\omega_t}^\top(\boldsymbol{\theta}).$$

In plain words, the $d \times d$ sensitivity matrix of a typical datum ω is equal to the expected value of the gradient vector of the negative log-likelihood function evaluated at the ML parameter values $\widehat{\boldsymbol{\theta}}_n$. The $d \times d$ variability matrix of a typical datum ω is equal to the expected value of the vector outer product of the gradient vector of the log-likelihood function evaluated at $\widehat{\boldsymbol{\theta}}_n$. The expectation in both cases is with respect to the random variable Ω.

The $d \times d$ matrix product $\frac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$ is traceable back at least to Eicker [47], Huber [79], and White [171] and yields asymptotically consistent covariance matrix estimates even when the fitted parametric model fails to hold, or is not even specified [87]. This estimator has become particularly popular in econometrics and is often referred to as *sandwich* (co)variance estimator in the context of generalized estimating equations [42, 87, 100, 101] such as is the *score* $\mathbf{g}_\omega(\boldsymbol{\theta})$ of Equation (4). The *sandwich* terminology is a metaphor for a *ham* or *meat* matrix $\mathbf{B}_n(\boldsymbol{\theta})$ between two *bread* matrices $\mathbf{A}_n(\boldsymbol{\theta})$. The square roots of the diagonal elements of $\frac{1}{n}\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\mathbf{B}_n(\widehat{\boldsymbol{\theta}}_n)\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$ are called "robust standard errors" or "Eicker-Huber-White standard errors" [47, 79, 171] and originate from M-estimation. The product of three matrices

$$\mathcal{G}_n(\boldsymbol{\theta}_0) = n\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{A}_0,$$

is referred to by various names including the Godambe [60] information matrix, sandwich information matrix and the robust information criterion [18, 44, 49, 109, 127, 134, 154]. Thus, under model misspecification, not the Fisher but the Godambe information is the fundamental currency of informativeness of the random variable Ω for the unknown parameters $\theta_1, \ldots, \theta_d$.

4.4 Correctly Specified Likelihood Function

First, recall that

$$\int q_\Omega(\omega; \boldsymbol{\theta})\,\mathrm{d}\omega = 1, \tag{12}$$

for any pdf $q_\Omega(\omega; \boldsymbol{\theta})$, where the integral is over the entire support of Ω. Standard regularity conditions justify differentiating under the integral sign. We obtain

$$\begin{aligned}
\mathbf{0}_d &= \frac{\partial}{\partial \boldsymbol{\theta}} \int q_\Omega(\omega \mid \boldsymbol{\theta})\,\mathrm{d}\omega \\
&= \int \nabla q_\Omega(\omega \mid \boldsymbol{\theta})\,\mathrm{d}\omega \\
&= \int \frac{\nabla q_\Omega(\omega \mid \boldsymbol{\theta})}{q_\Omega(\omega \mid \boldsymbol{\theta})}\, q_\Omega(\omega \mid \boldsymbol{\theta})\,\mathrm{d}\omega\ ^{3} \\
&= \int \nabla \log\bigl(q_\Omega(\omega \mid \boldsymbol{\theta})\bigr)\, q_\Omega(\omega \mid \boldsymbol{\theta})\,\mathrm{d}\omega \\
&= \mathbb{E}_{\boldsymbol{\theta}}\bigl[\nabla \log\bigl(q_\Omega(\Omega \mid \boldsymbol{\theta})\bigr)\bigr] = \mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})].
\end{aligned} \tag{13}$$

Equation (13) implies that

$$\int \nabla \mathcal{L}_\omega(\boldsymbol{\theta})\, q_\Omega(\omega \mid \boldsymbol{\theta})\, \mathrm{d}\omega = \mathbf{0}_d.$$

Differentiating both sides with respect to $\boldsymbol{\theta}$, and invoking standard regularity conditions that allow us to interchange differentiation and integration, we obtain

$$\mathbf{0}_{d\times d} = \frac{\partial}{\partial \boldsymbol{\theta}^\top} \int \nabla \mathcal{L}_\omega(\boldsymbol{\theta})\, q_\Omega(\omega \mid \boldsymbol{\theta})\, \mathrm{d}\omega. \tag{14}$$

We can now apply the product rule to the integrand

$$\nabla_{\boldsymbol{\theta}}\big(\nabla \mathcal{L}_\omega(\boldsymbol{\theta})\, q_\Omega(\omega \mid \boldsymbol{\theta})\big) = \underbrace{\nabla^2 \mathcal{L}_\omega(\boldsymbol{\theta})}_{d\times d}\, \underbrace{q_\Omega(\omega \mid \boldsymbol{\theta})}_{1\times 1} + \underbrace{\nabla \mathcal{L}_\omega(\boldsymbol{\theta})}_{d\times 1}\, \underbrace{\nabla^\top q_\Omega(\omega \mid \boldsymbol{\theta})}_{1\times d}.$$

Substituting this expression back into (14) gives

$$\mathbf{0}_{d\times d} = \int \nabla^2 \mathcal{L}_\omega(\boldsymbol{\theta})\, q_\Omega(\omega \mid \boldsymbol{\theta})\, \mathrm{d}\omega + \int \nabla \mathcal{L}_\omega(\boldsymbol{\theta})\, \nabla^\top q_\Omega(\omega \mid \boldsymbol{\theta})\, \mathrm{d}\omega.$$

The first integral is simply equal to the expected curvature of $\mathcal{L}_\omega(\boldsymbol{\theta})$ under the data-generating process. For the second integral, we can repeat the steps that led up to (13) as follows

$$\begin{aligned}
\mathbf{0}_{d\times d} &= \mathbb{E}_{\boldsymbol{\theta}}[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})] + \int \nabla \mathcal{L}_\omega(\boldsymbol{\theta}) \frac{\nabla^\top q_\Omega(\omega \mid \boldsymbol{\theta})}{q_\Omega(\omega \mid \boldsymbol{\theta})}\, q_\Omega(\omega \mid \boldsymbol{\theta}) \mathrm{d}\omega \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})] + \int \nabla \mathcal{L}_\omega(\boldsymbol{\theta}) \nabla^\top \mathcal{L}_\omega(\boldsymbol{\theta})\, q_\Omega(\omega \mid \boldsymbol{\theta}) \mathrm{d}\omega \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta}) \nabla^\top \mathcal{L}_\Omega(\boldsymbol{\theta})].
\end{aligned} \tag{15}$$

The second term can be rewritten using the variance rule of Equation (1) to yield

$$\mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta}) \nabla^\top \mathcal{L}_\Omega(\boldsymbol{\theta})] = \mathrm{Var}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{\theta}}[\nabla^\top \mathcal{L}_\Omega(\boldsymbol{\theta})].$$

Since, the expected *score* is zero or $\mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] = \mathbf{0}_d$, the second term on the right-hand side vanishes (a $d \times d$ null matrix), and we obtain the following equality

$$\mathrm{Var}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta}) \nabla^\top \mathcal{L}_\Omega(\boldsymbol{\theta})], \tag{16}$$

which states that the variance of the *score* is equal to the expected value of the outer product of the *score*. If we combine the above expression with Equation (15)

$$\mathrm{Var}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] = -\mathbb{E}_{\boldsymbol{\theta}}[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})]. \tag{17}$$

we end up with two expressions for the expected Fisher information $\mathcal{I}_\Omega(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})]$. Under stated regularity conditions, the Fisher information is defined as the variance of the *score* [51], and, when $\mathcal{L}_\omega(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$, it can equivalently be expressed as the expected curvature (second derivative) of $\mathcal{L}_\omega(\boldsymbol{\theta})$ [99]. Importantly, the expected Fisher information $\mathcal{I}_\Omega(\boldsymbol{\theta})$ does not depend on a particular datum ω since it is averaged across all possible outcomes of ω with respect to the distribution of Ω.

Equation (17) confirms the second Bartlett (or information) identity [8, 9]

$$\underbrace{\mathbb{E}_{\boldsymbol{\theta}}[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})]}_{= -\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_\Omega(\boldsymbol{\theta})]} + \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})]}_{= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{B}_\Omega(\boldsymbol{\theta})]} = 0. \tag{18}$$

[3] We use the following identity $\dfrac{\mathrm{d}}{\mathrm{d}x} \log(f(x)) = \dfrac{f'(x)}{f(x)}$.

Note the link to the variance identity of Equation (1). Thus, the variability matrix equals the variance of the *score*, $\mathbf{B}_\omega(\boldsymbol{\theta}) \equiv \mathrm{Var}_{\boldsymbol{\theta}}[\nabla\mathcal{L}_\omega(\boldsymbol{\theta})]$, and, since the expected *score* is zero according to Equation (13), $\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbf{B}_\Omega(\boldsymbol{\theta})\big] \equiv \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla\mathcal{L}_\Omega(\boldsymbol{\theta})\,\nabla^\top\mathcal{L}_\Omega(\boldsymbol{\theta})\big]$. If the second Bartlett identity holds, $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_\Omega(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{B}_\Omega(\boldsymbol{\theta})]$, and likelihood is *information-unbiased* in the sense of Lindsay [102].

The results for a single random variable Ω generalize to

$$\mathcal{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla\mathcal{L}_n(\boldsymbol{\theta})\nabla^\top\mathcal{L}_n(\boldsymbol{\theta})] = -\mathbb{E}_{\boldsymbol{\theta}}[\nabla^2\mathcal{L}_n(\boldsymbol{\theta})] = -\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{H}_n(\boldsymbol{\theta})],$$

and dividing both sides by n yields the information matrix in terms of bread and meat matrices

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{B}_n(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_n(\boldsymbol{\theta})], \tag{19}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, thus, also at $\boldsymbol{\theta}_0$. This information matrix equality is a finite sample result, meaning that it holds for finite n. Taking limits on both sides of Equation (19) gives the weaker result, $\mathbf{B}_0 = \mathbf{A}_0$, in terms of probability limits. Thus, under correct model specification, the sensitivity and variability matrices align perfectly and the asymptotic sandwich covariance simplifies to

$$\mathrm{asVar}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}}_n] = \tfrac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\,\mathbf{A}_0^{-1} = \tfrac{1}{n}\mathbf{B}_0^{-1}, \tag{20}$$

which is short-hand notation for $\operatorname{plim} n\,\mathrm{Var}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}}_n] = \frac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\,\mathbf{A}_0^{-1} = \frac{1}{n}\mathbf{B}_0^{-1}$. An equivalent expression in terms of $\mathbf{A}_n(\boldsymbol{\theta}_0)$ follows similarly

$$\mathrm{asVar}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}}_n] = \tfrac{1}{n}\mathbf{A}_0^{-1} = (-\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{H}_n(\boldsymbol{\theta}_0)])^{-1}, \tag{21}$$

and also follows from the information matrix equality and Equation (20).

The asymptotic covariance matrix, $\frac{1}{n}\mathbf{A}_0^{-1}$, is typically unknown but usually estimated by $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\mathbf{A}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$, which is consistent under correct specification. This provides a geometric description of the $100(1-\alpha)\%$ confidence regions of the ML/MAP parameter estimates $\widehat{\boldsymbol{\theta}}_n$ at significance level $\alpha \in (0,1)$

$$(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n) \leq d\,F^{-1}(1-\alpha; d, n-d), \tag{22}$$

where $F^{-1}(1-\alpha; d, n-d)$ is the upper $1-\alpha$ quantile (critical value) of the F distribution with d and $n-d$ degrees of freedom. Multivariate confidence regions are difficult to visualize, and pseudo-univariate $100(1-\alpha)\%$ confidence intervals may be documented instead using the diagonal entries of the $d \times d$ parameter covariance matrix

$$(\widehat{\boldsymbol{\theta}}_{n,\frac{1}{2}\alpha}, \widehat{\boldsymbol{\theta}}_{n,1-\frac{1}{2}\alpha}) = \widehat{\boldsymbol{\theta}}_n \pm T^{-1}(1-\tfrac{1}{2}\alpha; n-d)\sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_n)}, \tag{23}$$

where $T^{-1}(p_\alpha; \nu)$ is the quantile function of a standard Student's t distribution with $\nu = n-d$ degrees of freedom, evaluated at $p_\alpha = \frac{1}{2} \pm \big(\frac{1}{2} - \frac{1}{2}\alpha\big)$.

Equation (20) is arguably the most important equation in theoretical statistics. Thus, if the likelihood function is correctly specified then the ML estimate $\widehat{\boldsymbol{\theta}}_n$ obtained from $\omega_1, \ldots, \omega_n$ is asymptotically normal under weak regularity conditions [153, 171] with $d \times d$ covariance matrix $\boldsymbol{\Sigma}_n$ equal to the inverse of the *expected Fisher information* matrix, $\mathcal{I}_n(\boldsymbol{\theta}_0)$. If the data $\omega_1, \ldots, \omega_n$ are i.i.d., the second Bartlett identity holds for every observation and we can use the Fisher information of a single datum to yield the Fisher information for sample size $n > 1$

$$\begin{aligned}
\mathcal{I}_n(\boldsymbol{\theta}_0) &= -\mathbb{E}_{\boldsymbol{\theta}_0}\left[\nabla^2 \mathcal{L}_n(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}\right] \\
&= -\mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\sum_{t=1}^{n}\left\{\log\bigl(q_\Omega(\omega_t \mid \boldsymbol{\theta})\bigr)\right\}\Big|_{\boldsymbol{\theta}_0}\right] \\
&= -\sum_{t=1}^{n}\left\{\mathbb{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\log\bigl(q_\Omega(\Omega_t \mid \boldsymbol{\theta})\bigr)\Big|_{\boldsymbol{\theta}_0}\right]\right\} \\
&= -\sum_{t=1}^{n}\left\{\mathbb{E}_{\boldsymbol{\theta}_0}\left[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}\right]\right\} \\
&\equiv n\,\mathcal{I}_\Omega(\boldsymbol{\theta}_0).
\end{aligned}$$

This fundamental identity follows from $\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{A}_n(\boldsymbol{\theta}_0)] = \mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{A}_\Omega(\boldsymbol{\theta}_0)]$, simplifying Equation (21). Thus, if the training data are i.i.d., then theory dictates that each observation has the same expected information, and the total information contained in $\omega_1, \ldots, \omega_n$ is n times that contributed by a single datum ω. This proposition can have far-reaching consequences for a misspecified likelihood and this will be addressed in Section 4.5.

At this point, we do not know the ML parameter values $\widehat{\boldsymbol{\theta}}_n$, yet we do have a large-sample approximation to their distribution. Two remarks are in order. First, the limiting distribution

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \mathcal{N}_d\bigl(\mathbf{0}, \mathcal{I}_\Omega^{-1}(\boldsymbol{\theta}_0)\bigr), \tag{24}$$

has only been proven under certain regularity conditions. Albeit quite general, those conditions do not always hold in practice, for example, when the first- and second-order derivatives of $\mathcal{L}(\boldsymbol{\theta})$ violate continuity in the neighborhood of $\widehat{\boldsymbol{\theta}}_n$ due to the use of an improper numerical solver. It only asserts that for some sufficiently large n, perhaps much larger than the actual n of our data, the asymptotic approximation on the right-hand side of Equation (24) should be good. Second, and perhaps more important, we made the convenient but unrealistic assumption that the model (aka likelihood) is *exactly* correct and, thus, that the *true* distribution of the data has density $q_\Omega(\omega; \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0$. This assumption is questionable for hydrologic models. The implications of this are discussed next.

4.5 Misspecified Likelihood Function

The translation of a watershed $\mathfrak{S}$ to a hydrologic model that is suitable for computation (so-called computational model) inevitably requires making strong and simplifying assumptions about (among others) heterogeneity (surface and subsurface), material properties (soil and/or bedrock), streambed morphology and roughness, governing physical and hydrologic laws, and the dimensionality of the parameter and state space. The sheer complexity of watersheds renders impossible a perfect description of the distribution Q of variable Ω of interest, say discharge emanating from the catchment outlet. In plain words, all watershed models will be wrong to some extent, and, consequently, the likelihood will be misspecified too.

Suppose again that the ω's are i.i.d. random variables which have *true* but unknown pdf $q_\Omega(\omega; \cdot)$. However, we fit the incorrect family of densities given by $f(\omega; \boldsymbol{\theta}, \cdot)$ to the data using ML estimation. Figure 2 illustrates this situation wherein the data-generating process $\mathfrak{S}$ is a standard normal distribution $\Omega \sim \mathcal{N}(0, 1)$ and, thus, data $\omega_1, \ldots, \omega_{100}$ are drawn from a Gaussian distribution with zero-mean and unit variance, $\boldsymbol{\theta}_0 = (0, 1)^\top$. The assumed model $f(\omega; \boldsymbol{\theta})$ is a uniform distribution $\mathcal{U}(a, b)$ on the interval from $a = -4$ to $b = 4$, thus $\boldsymbol{\theta} = (a, b)^\top$.

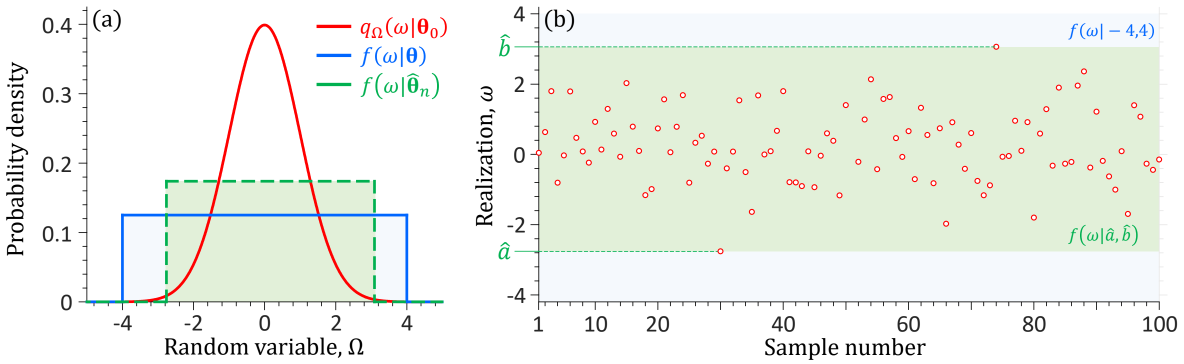

Figure 2: Illustration of model misspecification (a) PDFs of the standard normal data-generating process, $q_\Omega(\omega; 0, 1) = (2\pi)^{-1/2}\exp(-\omega^2/2)$ (red), the misspecified uniform model with prior $(a, b) = (-4, 4)$, $f(\omega; a, b) = 1/(b-a)$ (blue), and the ML fit $\widehat{\boldsymbol{\theta}}_n = (\widehat{a}, \widehat{b})^\top$ (dashed green), (b) trace of $\omega_1, \dots, \omega_{100}$ sampled from $\mathcal{N}(0, 1)$ (red dots), with feasible ranges under the prior uniform model (blue band) and under the ML estimates (green band).

The *true* but unknown data-generating density $q_\Omega(\omega; \cdot)$ that produced $\omega_1, \dots, \omega_n$ is not of the form $f(\omega; \boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In particular, q_Ω is not parameterized by $\boldsymbol{\theta}$. Therefore, identities such as $\int q_\Omega(\omega; \boldsymbol{\theta})\,\mathrm{d}\omega = 1$ in Equation (12) are inapplicable under misspecification. Instead, the correct normalizations are $\int q_\Omega(\omega)\,\mathrm{d}\omega = 1$ and $\int f(\omega; \boldsymbol{\theta})\,\mathrm{d}\omega = 1$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This is referred to as the M-open case in [11] and the consequences are profound. First, the values of the ML parameters $\widehat{\boldsymbol{\theta}}_n$ for $\omega_1, \dots, \omega_n$ will diverge from their *true* values $\boldsymbol{\theta}_0$ of the data-generating process $\mathfrak{S}$. There is nothing we can do about this, as the model parameters $\boldsymbol{\theta}$ lie in a different space than their counterparts of the data-generating process, $\mathfrak{S}$. A second and from the perspective of this article more important concern is, that the resulting asymptotic $100\gamma\%$ confidence regions of the model parameters and simulated output will differ substantially from well-calibrated $100\gamma\%$ confidence regions [103, 159, 164]. This over-conditioning issue led Keith Beven and his coworkers to advocate for Generalized Likelihood Uncertainty Estimation (GLUE) [13, 15]. GLUE assigns weights to parameter vectors based on pseudo-likelihoods, thereby constructing predictive distributions without requiring a fully-fledged probabilistic model. At the time, statisticians were only beginning to recognize the potential of Monte Carlo methods and in particular the then novel MCMC techniques for Bayesian inference [56, 138]. However, as we will demonstrate, the foundation of the GLUE methodology rests on a fundamental misconception. We return to this point later, but first explore in more detail the consequences of model misspecification on ML parameter estimates.

4.5.1 The M-Open Case: Parameter Interpretability, Meaning and Underpinning

A misspecified model (likelihood) undermines parameter meaning and interpretability. In the toy example of Fig. 2, the hypothesized uniform distribution $\mathcal{U}(a, b)$ is parameterized by support bounds (a, b), whereas the *true* normal distribution $\mathcal{N}(\mu, \sigma^2)$ uses location and variance (μ, σ^2). These parameter spaces are not comparable $(a, b) \in \mathbb{R}^2$ versus $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Consequently, maximizing a uniform likelihood on data from $\mathcal{N}(0, 1)$ does not recover the *true* parameters $\boldsymbol{\theta}_0 = (0, 1)^\top$ of the data-generating process. The ML estimates of $\widehat{a} = -2.786$ and $\widehat{b} = 3.03$ merely "shrink-wrap" the data $\omega_1, \dots, \omega_{100}$ and have no direct interpretation as (μ, σ^2). The hypothesized uniform distribution does not tolerate $\omega_t < a$ or $\omega_t > b$ otherwise $f(\omega_t; a, b) = 0$ and finite multiplication of the densities $L_n(\boldsymbol{\theta}) = \prod_{t=1}^{n} f(\omega_t; \boldsymbol{\theta})$ of $\omega_1, \dots, \omega_n$ produces a zero likelihood and log-likelihood of minus infinity.

To understand what the ML estimator is actually estimating in the case of a misspecified model, we use the law of large numbers to obtain the probability limit of the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$

$$\begin{aligned}\tfrac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}) = \tfrac{1}{n}\sum_{t=1}^{n}\bigl\{\log\bigl(f(\omega_t;\boldsymbol{\theta})\bigr)\bigr\} &\xrightarrow{\text{p}} \mathbb{E}_q\bigl[\log\bigl(f(\Omega;\boldsymbol{\theta})\bigr)\bigr] \\ &= \int q_\Omega(\omega \mid \cdot)\log\bigl(f(\omega;\boldsymbol{\theta})\bigr)\mathrm{d}\omega \end{aligned} \qquad (25)$$

uniformly in $\boldsymbol{\theta}$. When confronted with model misspecification, we should not write $\mathbb{E}_{\boldsymbol{\theta}}$ and $\text{Var}_{\boldsymbol{\theta}}$ as we did (among others) in Equations (13), (17) and (18). Instead, we write $\mathbb{E}_q$ and Var_q. Suppose that the expectation $\mathbb{E}_q[\mathcal{L}_n(\boldsymbol{\theta})]$ of the log-likelihood function with respect to the *true* distribution Q of Ω reaches its maximum value at $\boldsymbol{\theta}_* \in \boldsymbol{\Theta}$. Then it is clear that $\widehat{\boldsymbol{\theta}}_n = \arg\max\, \mathcal{L}_n(\boldsymbol{\theta})$ is an estimator of

$$\boldsymbol{\theta}_* = \underset{\boldsymbol{\theta}\,\in\,\boldsymbol{\Theta}}{\arg\max} \int q_\Omega(\omega;\cdot)\log\bigl(f(\omega;\boldsymbol{\theta})\bigr)\mathrm{d}\omega, \qquad (26)$$

where $\boldsymbol{\theta}_* = (\theta_{1,*},\ldots,\theta_{d,*})^\top$ specifies the *best* distribution out of all distributions in the misspecified parametric family. In our illustrative example of Fig. 2, the *best* distribution $\mathcal{U}(-2.786, 3.03)$ maximizes the integrand of Equation (26) over the $d = 2$-dimensional hypercube $\boldsymbol{\Theta} \in [-4, 4]^2$. The pdf of this ML solution (dashed green line) deviates substantially from the standard normal pdf of the data-generating process. So, then what is *best* about the ML parameter values $\widehat{\boldsymbol{\theta}}_n$ under misspecification? In Equations (25) and (26) we have established that the parameters $\boldsymbol{\theta}_*$ are best according to the criterion $\int q_\Omega(\omega;\cdot)\log\bigl(f(\omega;\boldsymbol{\theta})\bigr)\mathrm{d}\omega$. To determine how well this describes the unknown distribution $q_\Omega(\omega;\cdot)$ of the data-generating process, we compare it to the integrand of the log-likelihood $\int q_\Omega(\omega;\cdot)\log\bigl(q_\Omega(\omega;\cdot)\bigr)\mathrm{d}\omega$ under the correct distribution $q_\Omega(\omega;\cdot)$. In other words, the closer the difference

$$\int q_\Omega(\omega;\cdot)\log\bigl(q_\Omega(\omega;\cdot)\bigr)\mathrm{d}\omega - \int q_\Omega(\omega;\cdot)\log\bigl(f(\omega;\boldsymbol{\theta})\bigr)\mathrm{d}\omega, \qquad (27)$$

is to zero, the better the parameters $\boldsymbol{\theta}$ describe the distribution $q_\Omega(\omega;\cdot)$. The above expression is strictly positive and simplifies to

$$\int q_\Omega(\omega;\cdot)\log\left(\frac{q_\Omega(\omega;\cdot)}{f(\omega;\boldsymbol{\theta})}\right)\mathrm{d}\omega = d_{\text{KL}}\bigl(q_\Omega(\omega;\cdot), f(\omega;\boldsymbol{\theta})\bigr), \qquad (28)$$

the Kullback and Leibler [95] divergence between the unknown *true* $q_\Omega(\omega;\cdot)$ and hypothesized $f(\omega;\boldsymbol{\theta})$ pdfs of the data-generating process. Using Jensen's inequality [84] it is not difficult to demonstrate that the right integral of Equation (27) is nonnegative and zero if and only if $f(\omega;\boldsymbol{\theta}) = q_\Omega(\omega;\cdot)$. Thus, the closer $f(\omega;\boldsymbol{\theta})$ is to $q_\Omega(\omega;\cdot)$, the smaller $d_{\text{KL}}\bigl(q_\Omega(\omega;\cdot), f(\omega;\boldsymbol{\theta})\bigr)$ and the better the model is. Thus, we can write for the *pseudo-true* parameter values

$$\boldsymbol{\theta}_* = \underset{\boldsymbol{\theta}\,\in\,\boldsymbol{\Theta}\,\subseteq\,\mathbb{R}^d}{\arg\min}\; d_{\text{KL}}\bigl(q_\Omega(\omega;\cdot), f(\omega;\boldsymbol{\theta})\bigr).$$

The KL-divergence or relative entropy is also known as the I-divergence [33] and coincides with the divergence induced by the logarithmic score [157]. With the growing interest in information theory, we point out that the first integral of Equation (27) equals negative Shannon entropy, $-\mathbb{H}(q)$, of the *true* distribution [135, 136] and the second integral is known as the cross-entropy $\mathbb{H}(q, f)$ between the *true* and assumed distributions. Since the data-generating process is fixed, $\mathbb{H}(q)$ is constant and, thus, ML estimation is asymptotically equivalent to minimizing the cross-entropy $\mathbb{H}(q, f)$ between the *true* and assumed distributions.

4.5.2 The M-Open Case: Parameter Uncertainty

If our model is misspecified, which will almost always be the case in hydrology, how should we quantify parameter uncertainty? We must look in more detail to the limiting distribution of the parameters at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$. If differentiation under the integral sign is possible, then the gradient of the log-likelihood function at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$ should be zero on average, that is,

$$\mathbb{E}_q[\nabla \mathcal{L}_{\Omega_t}(\boldsymbol{\theta}_*)] = \mathbf{0}_d.$$

It is difficult to satisfy the above condition for every time t when the likelihood function is misspecified. We can relax this requirement and use a weaker condition instead. To guarantee a consistent estimator for $\boldsymbol{\theta}_*$ we must ascertain that when n grows large the average of $\partial \mathcal{L}_\omega(\boldsymbol{\theta}_*)/\partial \boldsymbol{\theta}$ goes to zero or

$$\lim_{n \to \infty} \tfrac{1}{n} \sum_{t=1}^{n} \{\mathbb{E}_q[\nabla \mathcal{L}_{\Omega_t}(\boldsymbol{\theta}_*)]\} = \mathbf{0}_d.$$

A more fundamental problem, however, is that the second Bartlett identity (18) will not hold under misspecification

$$\mathrm{Var}_q[\nabla \mathcal{L}_\Omega(\boldsymbol{\theta})] \neq -\mathbb{E}_q[\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta})],$$

for some generic observation ω and the associated likelihood is *information-biased* in the terminology of Lindsay [102]. As a result, Equation (19) can no longer be used to simplify the sandwich estimator [79]. Recalling the derivation of the asymptotic distribution of the ML estimator, results remain to hold upon replacing the *true* parameter values $\boldsymbol{\theta}_0$ by the *pseudo-true* parameter $\boldsymbol{\theta}_*$, at which the *score* is zero. The bread and meat matrices become

$$\begin{aligned} \mathbf{A}_n(\boldsymbol{\theta}_*) &= \tfrac{1}{n} \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_*) \\ \mathbf{B}_n(\boldsymbol{\theta}_*) &= \tfrac{1}{n} \nabla \mathcal{L}_n(\boldsymbol{\theta}_*) \nabla^\top \mathcal{L}_n(\boldsymbol{\theta}_*), \end{aligned} \tag{29}$$

satisfying $\mathbb{E}_q[\mathbf{A}_n(\boldsymbol{\theta}_*)] = \mathbb{E}_q[\mathbf{A}_\Omega(\boldsymbol{\theta}_*)]$ and $\mathbb{E}_q[\mathbf{B}_n(\boldsymbol{\theta}_*)] = \mathbb{E}_q[\mathbf{B}_\Omega(\boldsymbol{\theta}_*)]$ under i.i.d. data. The asymptotic variance now is

$$\mathrm{asVar}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}}_n] = \tfrac{1}{n} \mathbf{A}_*^{-1} \mathbf{B}_* \, \mathbf{A}_*^{-1}, \tag{30}$$

which is now *not equal to* either $\mathbf{A}_*^{-1}$ or $\mathbf{B}_*^{-1}$. Performing the Taylor expansion given by Equation (9) in $\boldsymbol{\theta}_*$ instead of $\boldsymbol{\theta}_0$ now yields the analogue of Equation (11) under misspecification

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \xrightarrow{\mathrm{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{A}_*^{-1} \mathbf{B}_* \mathbf{A}_*^{-1}),$$

where $\mathbf{A}_* = \mathrm{plim}\, \tfrac{1}{n} \mathbb{E}_q[\mathbf{A}_n(\boldsymbol{\theta}_*)]$ and $\mathbf{B}_* = \mathrm{plim}\, \tfrac{1}{n} \mathbb{E}_q[\mathbf{B}_n(\boldsymbol{\theta}_*)]$, for the $d \times d$ parameter covariance matrix. The sandwich matrix $\mathbf{A}_*^{-1} \mathbf{B}_* \mathbf{A}_*^{-1}$ is the relevant asymptotic covariance whenever the likelihood is misspecified. Although $\mathrm{asVar}_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}}_n] = \tfrac{1}{n} \mathbf{A}_*^{-1} \mathbf{B}_* \, \mathbf{A}_*^{-1}$ is typically unknown, it can be consistently estimated by its sample analogue $\boldsymbol{\Sigma}_n^{\mathrm{sand}} = \tfrac{1}{n} \mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1} \, \mathbf{B}_n(\widehat{\boldsymbol{\theta}}_n) \, \mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$ introduced in item (iv) of the six-point primer in Section 2.

In summary, the naive variance $\boldsymbol{\Sigma}_n^{\mathrm{naive}} = \tfrac{1}{n} \mathbf{A}_0^{-1}$ is appropriate when the likelihood is correctly specified. In this case, the residuals behave exactly as expected and the second Bartlett (or information) identity (18) holds. When the likelihood is misspecified, $\mathbb{E}_q[\mathbf{A}_n(\boldsymbol{\theta})] \neq \mathbb{E}_q[\mathbf{B}_n(\boldsymbol{\theta})]$, and this information mismatch requires the use of the sandwich variance $\boldsymbol{\Sigma}_n^{\mathrm{sand}} = \tfrac{1}{n} \mathbf{A}_*^{-1} \mathbf{B}_* \mathbf{A}_*^{-1}$ for a robust characterization of the confidence intervals of the ML/MAP parameter estimates $\widehat{\boldsymbol{\theta}}_n$.

4.6 Caveats

The presented theory, albeit elegant, is still incomplete for hydrologic application. We must remedy at least one practical problem and that is how do we determine the sensitivity $\mathbf{A}_n(\boldsymbol{\theta}_*)$ and variability $\mathbf{B}_n(\boldsymbol{\theta}_*)$ matrices if we do not know the *pseudo-true* parameter values $\boldsymbol{\theta}_*$ of the data-generating process? Furthermore, what do we do if observations $\omega \in \Omega$ violate the independence assumption? Next, we address both questions.

4.6.1 Observed Fisher and Godambe Information

The information, sensitivity, and variability matrices require knowledge of the *true* parameter values $\boldsymbol{\theta}_0$ of the data-generating process $\mathfrak{S}$ (or $\boldsymbol{\theta}_*$ under misspecification). Since these values are rarely known, the matrices $\mathbf{A}_\omega(\boldsymbol{\theta}_*)$, $\mathbf{A}_n(\boldsymbol{\theta}_*)$, $\mathbf{B}_\omega(\boldsymbol{\theta}_*)$, $\mathbf{B}_n(\boldsymbol{\theta}_*)$, $\mathcal{I}_n(\boldsymbol{\theta}_*)$, and $\mathcal{G}_n(\boldsymbol{\theta}_*)$ cannot be computed directly. Instead, we work with their empirical, sample-based counterparts evaluated at the ML or MAP solution $\widehat{\boldsymbol{\theta}}_n$, denoted by $\widehat{\mathbf{A}}_\omega$, $\widehat{\mathbf{A}}_n$, $\widehat{\mathbf{B}}_\omega$, $\widehat{\mathbf{B}}_n$, $\widehat{\mathcal{I}}_n$, and $\widehat{\mathcal{G}}_n$, respectively. The hat notation $\widehat{\cdot}$ thus signifies an empirical or sample-based estimate. If $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator, then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_*$ as $n \to \infty$, and the empirical estimates of the sensitivity, variability, Fisher, and Godambe information matrices converge in probability to their asymptotic population counterparts [171]. Under correct specification, these limits are attained at the true parameter vector $\boldsymbol{\theta}_0$. Under misspecification, they are attained at the *pseudo-true* parameter vector $\boldsymbol{\theta}_* \neq \boldsymbol{\theta}_0$ instead.

Accordingly, the asymptotic covariance of the ML/MAP estimator can be expressed in terms of its empirical analogues as

$$\widehat{\boldsymbol{\Sigma}}_n = \begin{cases} \widehat{\boldsymbol{\Sigma}}_n^{\text{naive}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{I}}_n^{-1} & \text{naive variance,} \quad (31\text{a}) \\ \widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{G}}_n^{-1} & \text{sandwich variance.} \quad (31\text{b}) \end{cases}$$

where $\widehat{\mathcal{I}}_n$ and $\widehat{\mathcal{G}}_n$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$ denote the *observed* Fisher and Godambe information, respectively, and must be distinguished from their *expected* counterparts, $\mathcal{I}_n$ and $\mathcal{G}_n$, which are defined at the true (or *pseudo-true*) parameter values. By definition, both $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ are positive semi-definite; throughout this paper, they are assumed to be positive definite. When Fisher information is estimated as $\widehat{\mathbf{A}}_n = -\frac{1}{n}\mathbf{H}_n(\widehat{\boldsymbol{\theta}}_n)$, the sandwich variance estimator can be as $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = n\,\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{\mathbf{B}}_n\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$.

4.6.2 Dependent Realizations

Up to now we have made the convenient assumption that the random variables $\Omega_1, \ldots, \Omega_n$ are independent. This may be reasonable for memoryless (zero-state) responses, but cannot be justified for watersheds whose behavior is controlled by state variables such as groundwater water storage. This memory induces serial dependence between successive Ω's, say, soil moisture content or river discharge unless sampling is coarse relative to spatiotemporal system dynamics. We therefore model the conditional mean with a deterministic forward model $y_t = \mathcal{M}_t(\boldsymbol{\theta}; \mathbf{I}_t)$, and define a distribution $e_t \sim \mathcal{E}(\boldsymbol{\delta})$ for the residuals $e_t(\boldsymbol{\theta}) = \omega_t - y_t(\boldsymbol{\theta})$. If the residuals are i.i.d., then $L_n(\boldsymbol{\theta} \mid \boldsymbol{\delta}) = \prod_{t=1}^n f(e_t; \boldsymbol{\delta})$. Otherwise, if the sequence $\mathbf{e}_n = \{e_t(\boldsymbol{\theta})\}_{t=1}^n$ displays autocorrelation and/or heteroskedasticity, we replace the product of marginals with a joint density $f_n(\mathbf{e}_n; \boldsymbol{\delta})$ [1, 20, 45, 65, 166] taking into account the autocorrelation and/or using a so-called HAC estimator for matrix $\widehat{\mathbf{B}}_n$ [5, 71, 112–114, 171] as we show next in estimating-equations form. Dependence in $\Omega_1, \ldots, \Omega_n$ is therefore not problematic per se as long as $\mathbf{A}_n(\boldsymbol{\theta}_*) \xrightarrow{\text{p}} \mathbf{A}_*$ and $\mathbf{B}_n(\boldsymbol{\theta}_*) \xrightarrow{\text{p}} \mathbf{B}_*$ as $n \to \infty$ (see Equation (29)).

The Huber-sandwich estimator extends to dependent realizations $\omega_1, \ldots, \omega_n$ via the generalized method of moments [71, 172] and generalized estimating equations [42, 100, 117]

$$\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \tfrac{1}{n}(\widehat{\mathbf{A}}_n^\top \mathbf{W}_d \widehat{\mathbf{A}}_n)^{-1} \widehat{\mathbf{A}}_n^\top \mathbf{W}_d \, \widehat{\boldsymbol{\beta}}_n \mathbf{W}_d \widehat{\mathbf{A}}_n (\widehat{\mathbf{A}}_n^\top \mathbf{W}_d \widehat{\mathbf{A}}_n)^{-1}, \tag{32}$$

where $\widehat{\boldsymbol{\beta}}_n$ is a HAC estimator (see item (ii) of the six-point primer in Section 2) of the long-run covariance $\widehat{\mathbf{B}}_n$ of the individual *score* contributions $\{\widehat{\mathbf{g}}_{\omega_t}\}_{t=1}^n$ applicable when the sequence exhibits serial dependence and/or heteroskedasticity, and $\mathbf{W}_d$ is a symmetric positive-definite $d \times d$ weighting matrix for the d estimating equations. The *efficient* choice is $\mathbf{W}_d = \widehat{\boldsymbol{\beta}}_n^{-1}$, yet, in our just-identified setting (number of estimating equations s equals number of estimable parameters d) any invertible $\mathbf{W}_d$ gives the same asymptotic covariance $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$ as is shown later.

Newey and West [113] introduced a so-called heteroskedasticity- and autocorrelation-consistent (HAC) estimator $\widehat{\boldsymbol{\beta}}_n$ for the variability matrix

$$\widehat{\boldsymbol{\beta}}_n = \widehat{\mathfrak{B}}_0 + \sum_{\ell=1}^{b_n} w(\ell, b_n)(\widehat{\mathfrak{B}}_\ell + \widehat{\mathfrak{B}}_\ell^\top), \tag{33}$$

where

$$\widehat{\mathfrak{B}}_\ell = \tfrac{1}{n} \sum_{t=\ell+1}^{n} \widehat{\mathbf{g}}_{\omega_t} \widehat{\mathbf{g}}_{\omega_{t-\ell}}^\top,$$

is the sample *autocovariance* matrix of $\widehat{\mathbf{g}}_{\omega_t}$ and $\widehat{\mathbf{g}}_{\omega_{t-\ell}}$ at lag $\ell \geq 0$, $b_n \in \mathbb{N}_+$ denotes the HAC bandwidth (truncation lag) beyond which autocovariances are treated as negligible, and $w(\ell, b_n)$ is a taper or weight function that smooths the sample autocovariance function [3]. For $\ell = 0$, we yield $\widehat{\mathfrak{B}}_0 = \frac{1}{n}\sum_{t=1}^n \widehat{\mathbf{g}}_{\omega_t} \widehat{\mathbf{g}}_{\omega_t}^\top$, the empirical second moment of the *scores*. According to Equation (16), this matrix equals the sample variance of the ML/MAP *scores* provided the first-order condition or estimating equation (A.1) holds at $\widehat{\boldsymbol{\theta}}_n$, i.e., $\widehat{\mathbf{g}}_n = \frac{1}{n}\sum_{t=1}^n \widehat{\mathbf{g}}_{\omega_t} = \mathbf{0}_d$, so the average *score* at $\widehat{\boldsymbol{\theta}}_n$ is zero. Under correct specification and negligible autocorrelation of the *score* sequence $\{\widehat{\mathbf{g}}_{\omega_t}\}$, the lagged autocovariance matrices $\widehat{\mathfrak{B}}_\ell$ vanish for $\ell \geq 1$, leaving only the contemporaneous covariance $\widehat{\mathfrak{B}}_0$. In this case, $\widehat{\boldsymbol{\beta}}_n = \widehat{\mathbf{B}}_n = \widehat{\mathfrak{B}}_0$.

Following Definition 6 on p. 7, the HAC estimator can be written in quadratic form

$$\widehat{\boldsymbol{\beta}}_n = \tfrac{1}{n} \widehat{\mathbf{J}}_n^\top \mathbf{W}_{b_n} \widehat{\mathbf{J}}_n \qquad [\mathbf{W}_{b_n}]_{ij} = w\big(|i-j|, b_n\big)$$

where $\mathbf{W}_{b_n}$ is a symmetric $n \times n$ lag-window matrix. We use the Bartlett (triangular) window,

$$w(\ell, b_n) = \begin{cases} 1 - \dfrac{|\ell|}{1 + b_n} & \text{if } |\ell| \leq b_n, \\ 0 & \text{otherwise,} \end{cases}$$

and note that other common choices include the Parzen [118], Tukey-Hanning (tapered-cosine) [5, 72], and quadratic spectral [5] windows. Under standard mixing conditions with $b_n \to \infty$ and $b_n/n \to 0$, all such kernels yield consistent long-run variance estimates. Finite-sample bias-variance tradeoffs differ, with bandwidth b_n as the primary tuning parameter. In practice, a data-driven bandwidth is advisable and its value can be guided by inspecting the sample autocorrelation function of the columns of $\widehat{\mathbf{J}}_n$, the $n \times d$ Jacobian matrix (stacked *scores*) of the ML/MAP estimator. More formal rules of thumb include $b_n = \lfloor c\, n^{1/4} \rfloor$ for small $c > 0$ [5] and the Newey and West [114] bandwidth $b_n = \lfloor 4\,(n/100)^{2/9} \rfloor$, where $\lfloor \cdot \rfloor$ rounds down to the nearest integer. Once a sensible bandwidth b_n is chosen, the HAC meat matrix $\widehat{\boldsymbol{\beta}}_n$ is typically far less sensitive to the specific (reasonable) kernel choice.

Equation (32) permits time-series correlation in the MAP/ML *score* vectors $\widehat{\mathbf{g}}_\omega$ in addition to contemporaneous correlation [27, 38, 65]. For just-identified models the choice of weighting matrix $\mathbf{W}_d$ is inconsequential. Indeed, when $s = d$ then $\widehat{\mathbf{A}}_n$, $\mathbf{W}_d$, and $\widehat{\boldsymbol{\beta}}_n$ are square invertible matrices and $(\widehat{\mathbf{A}}_n^\top \mathbf{W}_d \widehat{\mathbf{A}}_n)^{-1} = \widehat{\mathbf{A}}_n^{-1}\mathbf{W}_d^{-1}(\widehat{\mathbf{A}}_n^\top)^{-1}$. Then Equation (32) equals

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} &= \tfrac{1}{n}\widehat{\mathbf{A}}_n^{-1}\mathbf{W}_d^{-1}(\widehat{\mathbf{A}}_n^\top)^{-1}\widehat{\mathbf{A}}_n^\top \mathbf{W}_d\,\widehat{\boldsymbol{\beta}}_n \mathbf{W}_d \widehat{\mathbf{A}}_n \widehat{\mathbf{A}}_n^{-1}\mathbf{W}_d^{-1}(\widehat{\mathbf{A}}_n^\top)^{-1}\\ &= \tfrac{1}{n}\widehat{\mathbf{A}}_n^{-1}\mathbf{W}_d^{-1}\mathbf{I}_d\mathbf{W}_d\,\widehat{\boldsymbol{\beta}}_n \mathbf{W}_d \mathbf{I}_d \mathbf{W}_d^{-1}(\widehat{\mathbf{A}}_n^\top)^{-1}.\end{aligned}$$

As the sensitivity matrix is symmetric, $\widehat{\mathbf{A}}_n^{-1} = (\widehat{\mathbf{A}}_n^\top)^{-1}$ and $\mathbf{I}_d = \mathbf{W}_d^{-1}\mathbf{I}_d\mathbf{W}_d$, we end up with

$$\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1},$$

the now familiar sandwich variance estimator of Equation (31b), except with $\widehat{\mathbf{B}}_n$ replaced by $\widehat{\boldsymbol{\beta}}_n$. Thus, nothing fundamentally changes when the realizations $\omega_1, \ldots, \omega_n$ of the random variable Ω are dependent. The only requirement is to use a consistent estimator $\widehat{\boldsymbol{\beta}}_n$ of the variability matrix in order to safeguard the estimated variance of the *score* against autocorrelation. In practice, a consistent estimator of the variability matrix should always be employed. Note that if the ML *scores* are uncorrelated, then $\widehat{\boldsymbol{\beta}}_n = \widehat{\mathbf{B}}_n$.

4.6.3 Variance and Bias

As watersheds do not admit a *perfect* characterization of the data-generating process $\mathfrak{S}$, the likelihood function will almost surely be misspecified and the sandwich variance estimator $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}$ should be the default approach for quantifying model parameter and predictive uncertainty in hydrologic modeling. The "Huber sandwich estimator" will yield a robust description of the parameter credible regions in the face of epistemic and forcing data errors. Though, when misspecification is severe enough, the parameters being estimated by ML or Bayesian methods are likely to be meaningless except perhaps as descriptive statistics [53]. Then, considerable optimism is warranted as most of the estimated quantities will be subject to bias. This bias may be of greater interest than the variance of model parameters and simulated (predicted) quantities, particularly when the sample size n is large. Even then, the "Huber sandwich estimator" ensures a more defensible and robust inference of hydrologic model parameters. Moreover, the degree of misalignment of the bread and meat matrices can be measured with a divergence score. This *strictly proper* measure of the degree of model misspecification serves different purposes including model ranking, selection and refinement.

5 Case Studies

We demonstrate the naive and sandwich covariance estimators by application to three case studies involving soil water infiltration, watershed hydrologic modeling and rainfall-discharge simulation. In the first two studies, the mathematical-physical description of the data-generating process lends support to symbolic differentiation enabling an analytic demonstration of how model misspecification affects parameter uncertainty in ML and Bayesian methods. The third study applies the naive and sandwich variance estimators to measured streamflows using numerical methods and a suite of likelihood functions. This study confirms that conventional curvature-based approaches mischaracterize parameter and predictive uncertainty under misspecification.

5.1 Case Study I: Soil Water Infiltration

In his classic essay on the mathematical-physical description of infiltration, Philip [121] derived quasi-analytic solutions to the general flow equation for cumulative infiltration $I(t)$ [L] into homogeneous soils at uniform initial moisture content. For vertical infiltration, the solution to Richards [128]' equation[4] admits a series expansion, typically truncated after $d \geq 3$ terms

$$I(t) = a_1 t^{1/2} + a_2 t + a_3 t^{3/2} + \ldots + a_d t^{d/2} = \sum_{j=1}^{d} a_j t^{j/2}, \tag{34}$$

where t (T) is time and a_j are soil-dependent coefficients (free parameters) with dimensions $[a_j] = \mathrm{L\,T}^{-j/2}$. Thus, d is the number of retained terms and equals the number of free coefficients $a_1, \ldots, a_d$. Philip [121] showed that a_1 is synonymous to the sorptivity, S ($\mathrm{L\,T}^{-1/2}$), a measure of the soil's capacity to take up and release liquids by capillarity, and a_2 is equal to a unitless multiple, c, of the saturated hydraulic conductivity, K_s ($\mathrm{L\,T}^{-1}$), which measures a soil's ability to transmit water under the influence of gravity. Note that the sorptivity S is not an invariant soil property but a function of the soil's initial and final moisture contents.

The most popular (and robust) variant of Equation (34) retains only the first two terms

$$I(t) = S t^{1/2} + cK_\mathrm{s} t, \tag{35}$$

with coefficients $a_1 = S$ and $a_2 = cK_\mathrm{s}$ that are strictly positive. This two-term form is easy to use in practice and has a rigorous mathematical-physical foundation [82, 121], but its validity is limited in time [123, 167]. This issue is often overlooked in the practical application of Equation (35) and will inevitably corrupt the estimates of S and the product, cK_s, in curve fitting. Values of the multiplicative coefficient c for different soil types are presented in Table 6 of Jaiswal et al. [83]. We conveniently assume that $c = 1$ and restrict our attention to estimation of S and K_s from cumulative infiltration measurements $\omega_1, \ldots, \omega_n$.

To estimate the ML values of S and K_s, we write Philip's two-term expression as an inner product of the design vector $\mathbf{d}(t) = (t^{1/2}, t)^\top$ and vector of parameters $\boldsymbol{\theta} = (a_1 = S, a_2 = K_\mathrm{s})^\top$

$$\omega_t = \mathbf{d}(t_i)^\top \boldsymbol{\theta} + e_i, \tag{36}$$

where the residuals $e_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ for all $i = 1, \ldots, n$ are assumed to be independent and zero-mean normally distributed with variance σ_ϵ^2. For simplicity, σ_ϵ^2 is assumed known. This assumption is made without loss of generality, as the error variance can be estimated jointly with the model parameters if desired. Thus, residuals are expected to behave as measurement errors of $\omega_1, \ldots, \omega_n$. Structural errors arising from an imperfect model are assumed to be small (inconsequential) and absorbed into the residuals. We can generalize the above form to n different measurement times $t_1, \ldots, t_n$

$$\boldsymbol{\omega}_n = \mathbf{D}_n \boldsymbol{\theta} + \mathbf{e}_n,$$

where the $n \times d$ design matrix $\mathbf{D}_n$ consists of the stacked vectors $\mathbf{d}(t_1)^\top, \ldots, \mathbf{d}(t_n)^\top$ and $\mathbf{e}_n = (e_1, \ldots, e_n)^\top$ is the $n \times 1$ residual vector. Under the stated residual assumptions, the ML estimates $\widehat{\boldsymbol{\theta}}_n$ of the Philip coefficients are a solution to the minimization problem [25]

$$\widehat{\boldsymbol{\theta}}_n = \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{bmatrix} = \underset{\boldsymbol{\theta} \in \mathbb{R}_+^2}{\arg\min} \; \frac{1}{2} \mathrm{WSSR}_n(\boldsymbol{\theta}),$$

where

$$\begin{aligned} \mathrm{WSSR}_n(\boldsymbol{\theta}) &= (\boldsymbol{\omega}_n - \mathbf{D}_n \boldsymbol{\theta})^\top \boldsymbol{\Sigma}_\epsilon^{-1} (\boldsymbol{\omega}_n - \mathbf{D}_n \boldsymbol{\theta}) \\ &= \mathbf{e}_n(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{e}_n(\boldsymbol{\theta}), \end{aligned}$$

[4] The general flow equation is commonly credited to Lorenzo A. Richards (1904–1993) in the vadose zone literature, but as Knight and Raats [90] note, it was introduced earlier by Lewis F. Richardson (1881–1953) in the context of heat and mass transfer in the atmosphere [130].

is the weighted sum of squared residuals, $\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2\,\mathbf{I}_n$ denotes the $n \times n$ covariance matrix of the measurement errors, $\mathbb{R}_+^2$ is the feasible parameter space and the $n \times n$ identity matrix $\mathbf{I}_n$ has ones on the main diagonal and zeros elsewhere. Under Gaussian errors with covariance $\boldsymbol{\Sigma}_\epsilon$, the generalized least squares or ML estimator is $\widehat{\boldsymbol{\theta}}_n = (\mathbf{D}_n^\top \boldsymbol{\Sigma}_\epsilon^{-1}\,\mathbf{D}_n)^{-1}\,\mathbf{D}_n^\top \boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\omega}_n$. If component-wise nonnegativity is required and some entries of $\widehat{\boldsymbol{\theta}}_n$ are negative, the constrained ML estimator should be obtained by imposing zero lower bounds via bounded-variable (nonnegative) least squares [97, 140].

The ML solution maximizes the normal likelihood

$$\begin{aligned} L_n^{\mathrm{n}}(\boldsymbol{\theta}) &= f_{\mathcal{N}_n}(\mathbf{y}_n; \boldsymbol{\omega}_n, \boldsymbol{\Sigma}_\epsilon) \\ &= (2\pi)^{-n/2}|\boldsymbol{\Sigma}_\epsilon|^{-1/2}\exp\bigl(-\tfrac{1}{2}(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})^\top \boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr), \end{aligned} \tag{37}$$

where $\mathbf{y}_n = \mathbf{D}_n\boldsymbol{\theta}$ denotes the simulated cumulative infiltration curve, $|\cdot|$ is the determinant, and the argument of the exponential equals $-\mathrm{WSSR}_n(\boldsymbol{\theta})/2$. This expression can be generalized to a composite or power likelihood $L_n^{\mathrm{p}}(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta})^\lambda$, in log-likelihood form

$$\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta}) = \lambda\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta}) = -\tfrac{1}{2}n\lambda\log(2\pi) - \tfrac{1}{2}\lambda\log(|\boldsymbol{\Sigma}_\epsilon|) - \tfrac{1}{2}\lambda\mathrm{WSSR}_n(\boldsymbol{\theta}), \tag{38}$$

where the positive, dimensionless, scalar $\lambda > 0$ controls the peakedness of the likelihood and is often referred to as a learning rate. Larger values of λ increase the curvature $\nabla^2\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta})$ at $\widehat{\boldsymbol{\theta}}_n$ and make the data $\omega_1, \ldots, \omega_n$ more informative about the parameters $\boldsymbol{\theta}$. Thus, $\lambda > 0$ scales the curvature $\mathcal{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[-\nabla_{\boldsymbol{\theta}}^2\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta})]$ of the power log-likelihood associated with the data. In theory, $\lambda = 1$ recovers the standard likelihood, but choosing $\lambda < 1$ can be advantageous to temper learning in the presence of model structural error.

The $d \times 1$ *score* of the $\underline{\text{n}}$ormal $\underline{\text{p}}$ower log-likelihood $\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta})$ is equal to

$$\begin{aligned} \mathbf{g}_n^{\mathrm{np}}(\boldsymbol{\theta}) \equiv \nabla\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta}) &= \lambda\nabla\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta}) \\ &= -\tfrac{1}{2}\lambda\nabla\bigl((\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})^\top\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr) \\ &= -\tfrac{1}{2}\lambda\,\sigma_\epsilon^{-2}\nabla\bigl((\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})^\top\mathbf{I}_n(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr) = \lambda\,\sigma_\epsilon^{-2}\,\mathbf{D}_n^\top(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta}), \end{aligned}$$

and we yield for its $d \times d$ Hessian matrix

$$\begin{aligned} \mathbf{H}_n^{\mathrm{np}}(\boldsymbol{\theta}) \equiv \nabla^2\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta}) &= \nabla\bigl(\lambda\nabla\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta})\bigr) \\ &= \nabla\bigl(\lambda\,\sigma_\epsilon^{-2}\,\mathbf{D}_n^\top(\boldsymbol{\omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr) = -\lambda\,\sigma_\epsilon^{-2}(\mathbf{D}_n^\top\,\mathbf{D}_n) = -n\mathbf{A}_n^{\mathrm{np}}. \end{aligned}$$

Thus, given σ_ϵ^2 the Hessian matrix $\mathbf{H}_n^{\mathrm{np}} = -n\mathbf{A}_n^{\mathrm{np}}$ does not depend on the ML estimates $\widehat{\boldsymbol{\theta}}_n$, or on $\boldsymbol{\theta}$ more generally, but is constant across the feasible parameter space $\boldsymbol{\Theta} = \mathbb{R}_+^2$. This property holds for any vector-valued regression function $\mathbf{y}_n = \mathbf{f}(\boldsymbol{\theta}, \cdot)$ whose output $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$ is a linear multiple of the parameters $\boldsymbol{\theta}$.

The $d \times d$ variability matrix $\mathbf{B}_n^{\mathrm{np}}(\boldsymbol{\theta})$ of the normal power log-likelihood $\mathcal{L}_n^{\mathrm{np}}(\boldsymbol{\theta})$ is now equal to

$$\begin{aligned} \mathbf{B}_n^{\mathrm{np}}(\boldsymbol{\theta}) &= \tfrac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\bigl[\mathbf{g}_n^{\mathrm{np}}(\boldsymbol{\theta})\mathbf{g}_n^{\mathrm{np}}(\boldsymbol{\theta})^\top\bigr] \\ &= \tfrac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\bigl[\bigl(\lambda\,\sigma_\epsilon^{-2}\,\mathbf{D}_n^\top(\boldsymbol{\Omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr)\bigl(\lambda\,\sigma_\epsilon^{-2}\,\mathbf{D}_n^\top(\boldsymbol{\Omega}_n - \mathbf{D}_n\boldsymbol{\theta})\bigr)^\top\bigr] \\ &= \tfrac{1}{n}\lambda^2\,\sigma_\epsilon^{-4}\,\mathbf{D}_n^\top\,\mathbb{E}_{\boldsymbol{\theta}}\bigl[(\boldsymbol{\Omega}_n - \mathbf{D}_n\boldsymbol{\theta})(\boldsymbol{\Omega}_n - \mathbf{D}_n\boldsymbol{\theta})^\top\bigr]\,\mathbf{D}_n. \end{aligned} \tag{39}$$

Inside the expectation $\boldsymbol{\omega}_n$ is replaced by the random vector $\boldsymbol{\Omega}_n = (\Omega_1, \ldots, \Omega_n)^\top$, since $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ is taken with respect to the law of $\boldsymbol{\Omega}_n$. Let $\boldsymbol{\epsilon}_n = \boldsymbol{\Omega}_n - \mathbf{D}_n\boldsymbol{\theta}$ denote the population residual vector. Under the assumption of independent measurement errors in the data-generating process $\epsilon_1, \ldots, \epsilon_n$ are independent with $\mathbb{E}[\epsilon_i = 0]$ and $\mathbb{E}[\epsilon_i^2] = \sigma_\epsilon^2$, it follows that $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ for all $i \neq j$ and, hence, $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top] = \sigma_\epsilon^2 \mathbf{I}_n$. Then $\mathbf{B}_n^{\text{np}}(\boldsymbol{\theta})$ in Equation (39) simplifies to

$$\mathbf{B}_n^{\text{np}} = \tfrac{1}{n}\lambda^2\,\sigma_\epsilon^{-4}\,\mathbf{D}_n^\top\,\sigma_\epsilon^2\,\mathbf{I}_n\,\mathbf{D}_n = \tfrac{1}{n}\lambda^2\sigma_\epsilon^{-2}(\mathbf{D}_n^\top\,\mathbf{D}_n),$$

and the $d \times d$ sensitivity and variability matrices amount to

$$\begin{aligned} \mathbf{A}_n^{\text{np}} &= \tfrac{1}{n}\lambda\,\sigma_\epsilon^{-2}(\mathbf{D}_n^\top\,\mathbf{D}_n) && \text{sensitivity matrix} \\ \mathbf{B}_n^{\text{np}} &= \tfrac{1}{n}\lambda^2\sigma_\epsilon^{-2}(\mathbf{D}_n^\top\,\mathbf{D}_n) && \text{variability matrix.} \end{aligned}$$

The mathematical expressions for $\mathbf{A}_n^{\text{np}}$ and $\mathbf{B}_n^{\text{np}}$ are equivalent when the learning rate $\lambda = 1$, otherwise $\mathbf{A}_n^{\text{np}} \neq \mathbf{B}_n^{\text{np}}$, the second Bartlett identity (18) fails and the normal power likelihood $L_n^{\text{np}}(\boldsymbol{\theta})$ is *information-biased.* The estimated naive and sandwich variances of the ML parameter estimates for $L_n^{\text{np}}(\boldsymbol{\theta}) = L_n^{\text{n}}(\boldsymbol{\theta})^\lambda$ become

$$\widehat{\boldsymbol{\Sigma}}_n^{\text{np}} = \begin{cases} \widehat{\boldsymbol{\Sigma}}_n^{\text{naive}} = \tfrac{1}{n}(\mathbf{A}_n^{\text{np}})^{-1} = \lambda^{-1}\sigma_\epsilon^2(\mathbf{D}_n^\top\,\mathbf{D}_n)^{-1} & \text{naive variance} \quad (40\text{a}) \\ \widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \tfrac{1}{n}(\mathbf{A}_n^{\text{np}})^{-1}\mathbf{B}_n^{\text{np}}(\mathbf{A}_n^{\text{np}})^{-1} = \sigma_\epsilon^2(\mathbf{D}_n^\top\,\mathbf{D}_n)^{-1} & \text{sandwich variance.} \quad (40\text{b}) \end{cases}$$

The above results are both remarkable and troubling. The expression for the naive variance $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}$ confirms the widely held belief that parameter uncertainty can be tuned by applying an arbitrary power $\lambda > 0$ to the likelihood function. This is the basis of the GLUE method of Beven and Binley [15]. Values $0 < \lambda < 1$ expand parameter confidence regions, while $\lambda > 1$ shrinks the "posterior" distribution of $\widehat{\boldsymbol{\theta}}_n$, thereby reducing parameter uncertainty. This elastic stretching of the likelihood may appear pragmatic for addressing over-conditioning, but it lacks theoretical support, as shown by the closed-form sandwich variance $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$. In the product $(\mathbf{A}_n^{\text{np}})^{-1}\mathbf{B}_n^{\text{np}}(\mathbf{A}_n^{\text{np}})^{-1}$, the arbitrary power λ cancels out, implying that λ has no effect on parameter (or predictive) uncertainty under misspecification. It is therefore a misconception that parameter uncertainty is determined solely by the curvature of the negative log-likelihood as is measured by the Hessian $\mathbf{H}_n(\boldsymbol{\theta}) = -\nabla^2\mathcal{L}_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$, or equivalently the negative ML bread matrix $-n\,\widehat{\mathbf{A}}_n$. Under misspecification, the variability of the *score* $\nabla\mathcal{L}_n(\boldsymbol{\theta})$ also governs the width of confidence intervals for $\widehat{\boldsymbol{\theta}}_n$. We explore the implications of these findings in the case study below.

We create a synthetic data set $\boldsymbol{\omega}_n = (\omega_1, \ldots, \omega_{20})^\top$ of cumulative infiltration values by evaluating Philip's two-term expression in Equation (36) for $n = 20$ logarithmically equally spaced infiltration times between $t = 0.05$ and $t = 10$ hr using $S = 0.5$ cm/h$^{1/2}$ and $K_{\text{s}} = 1.0$ cm/h, $\sigma_\epsilon^2 = 0.01$ cm^2 and $\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2\,\mathbf{I}_n$. Next, we compute the ML values of the soil sorptivity and saturated soil hydraulic conductivity

$$\widehat{\boldsymbol{\theta}}_n = \begin{bmatrix} \widehat{S} \\ \widehat{K}_{\text{s}} \end{bmatrix} = (\mathbf{D}_n^\top\boldsymbol{\Sigma}_\epsilon^{-1}\,\mathbf{D}_n)^{-1}\,\mathbf{D}_n^\top\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\omega}_n, \tag{41}$$

and calculate the naive $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}$ and sandwich $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$ variance matrices of Equation (40) for a learning rate $\lambda = 0.1$. In M-estimation, we replace $\boldsymbol{\Sigma}_\epsilon^{-1}$ by a diagonal weight matrix based on an influence function. Figure 3 presents the confidence regions of $\widehat{S}$ and $\widehat{K}_{\text{s}}$ obtained with (a) the naive and (b) the sandwich estimator, for significance levels $\alpha = 0.01$ (dark shading), $\alpha = 0.05$ (medium shading), $\alpha = 0.10$ (light-medium shading), and $\alpha = 0.50$ (light shading). The 100γ% confidence region of the sorptivity S and saturated soil hydraulic conductivity K_{s} is an ellipsoid that centers on the ML solution (red cross). The length and direction of the two principal axes of the ellipse are given by the eigenvalues and eigenvectors, respectively, of the parameter covariance matrix, $\widehat{\boldsymbol{\Sigma}}_n$. As a result, the 99% confidence region of $\widehat{S}$ and $\widehat{K}_{\text{s}}$ is simply

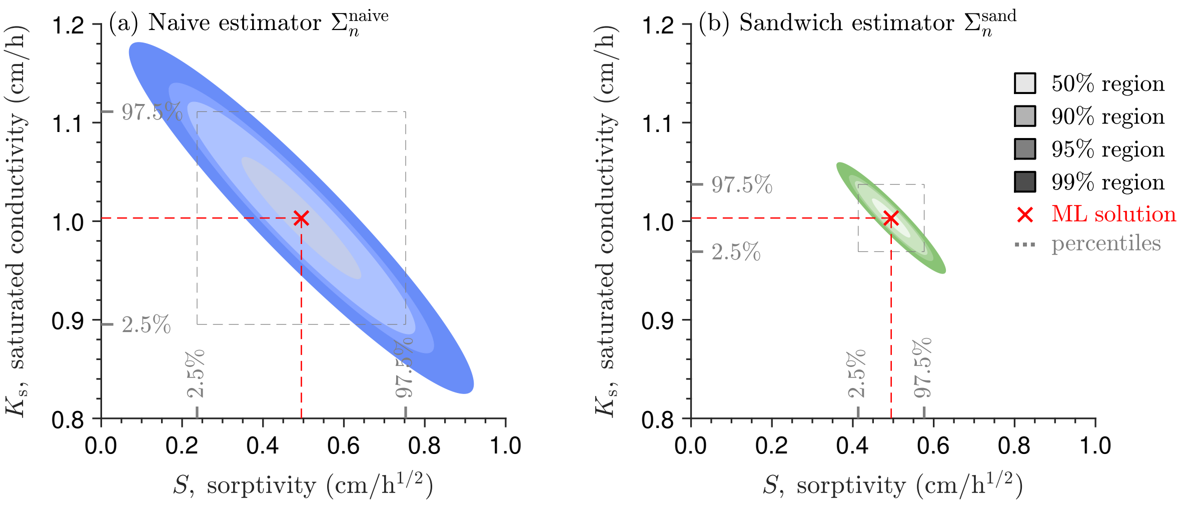

Figure 3: Maximum likelihood estimates (red cross), confidence intervals (dashed gray lines, $\alpha = 0.05$), and confidence regions (shaded areas) of the soil sorptivity S (cm/h$^{1/2}$) and saturated hydraulic conductivity K_{s} (cm/h) of Philip's two-term infiltration equation, $I(t) = St^{1/2} + K_{\text{s}}t$ for the (a) naive (blue) and (b) sandwich (green) estimator of the ML parameter covariance matrix $\widehat{\mathbf{\Sigma}}_n$. The bivariate $100\gamma\%$ confidence regions of $\widehat{S}$ and $\widehat{K}_{\text{s}}$ are shaded according to significance level: $\gamma = 0.50$ (light shading), $\gamma = 0.90$ (light-medium shading), $\gamma = 0.95$ (medium shading), and $\gamma = 0.99$ (dark shading).

a multiple of their 95% region. The ellipses enclose values of the sorptivity and saturated soil hydraulic conductivity which are statistically acceptable at a given confidence level $\gamma = 1 - \alpha$. The two graphs confirm our earlier finding. The confidence regions and intervals of the naive variance are much larger than their counterparts of the sandwich variance. This is the immediate effect of the learning rate $\lambda = 0.1$ and $\widehat{\mathbf{\Sigma}}_n^{\text{naive}} = \lambda^{-1}\widehat{\mathbf{\Sigma}}_n^{\text{sand}}$. The results merely illustrate that the sandwich estimator is much more confident about the optimal values of the model parameters. As is common, the univariate 95% confidence intervals (blue dotted lines) of $\widehat{S}$ and $\widehat{K}_{\text{s}}$ underestimate their bivariate estimates. The hypercube defined by (23) can be very different from the exact confidence regions of the parameters. This discrepancy is well known and univariate confidence intervals are usually presented only [43].

Next, we turn parameter uncertainty into $100\gamma\%$ confidence intervals $[l, u]$ of Philip's infiltration function. Confidence intervals of ML simulated infiltration, $\widehat{I}(t) = \mathbf{d}(t)^\top \widehat{\boldsymbol{\uptheta}}_n$

$$\left[\widehat{I}(t)_{\frac{1}{2}\alpha}, \widehat{I}(t)_{1-\frac{1}{2}\alpha}\right] = \widehat{I}(t) \pm T^{-1}(p_\alpha; n-d)\sqrt{\mathbf{d}(t)^\top\, \widehat{\mathbf{\Sigma}}_n^{\text{np}}\, \mathbf{d}(t)} \tag{42}$$

are presented in Figure 4 for the (a) naive and (b) sandwich variance estimators using $\alpha = 0.05$. Note that the quantity under the square root in Equation (42) equals the variance of $\widehat{I}(t)$ for any fixed $t > 0$. As expected, Philip's two-term expression accurately describes the measured infiltration curve. The naive confidence intervals encapsulate all but two of the cumulative infiltration measurements. In contrast, the sandwich confidence intervals are much smaller than their naive counterparts and are barely visible at early times without proper magnification. This discrepancy is caused by the learning rate $\lambda = 0.1$ and grows with time t as the design vector $\mathbf{d}(t)$ increases in magnitude.

Substituting the naive and sandwich covariance expressions of Equations (40a) and (40b) into formula (42) yields $\text{Var}^{\text{naive}}\big(\widehat{I}(t)\big) = \lambda^{-1/2}\,\text{Var}^{\text{sand}}\big(\widehat{I}(t)\big)$, where $\lambda^{-1/2} = 3.1623$. Under homoskedastic residuals $\mathbf{\Sigma}_\epsilon = \sigma_\epsilon^2\,\mathbf{I}_n$, the sandwich prediction variance reduces to

$$\text{Var}^{\text{sand}}\big(\widehat{I}(t)\big) = \sigma_\epsilon^2\,\mathbf{d}(t)^\top\big(\mathbf{D}_n^\top\,\mathbf{D}_n\big)^{-1}\mathbf{d}(t).$$

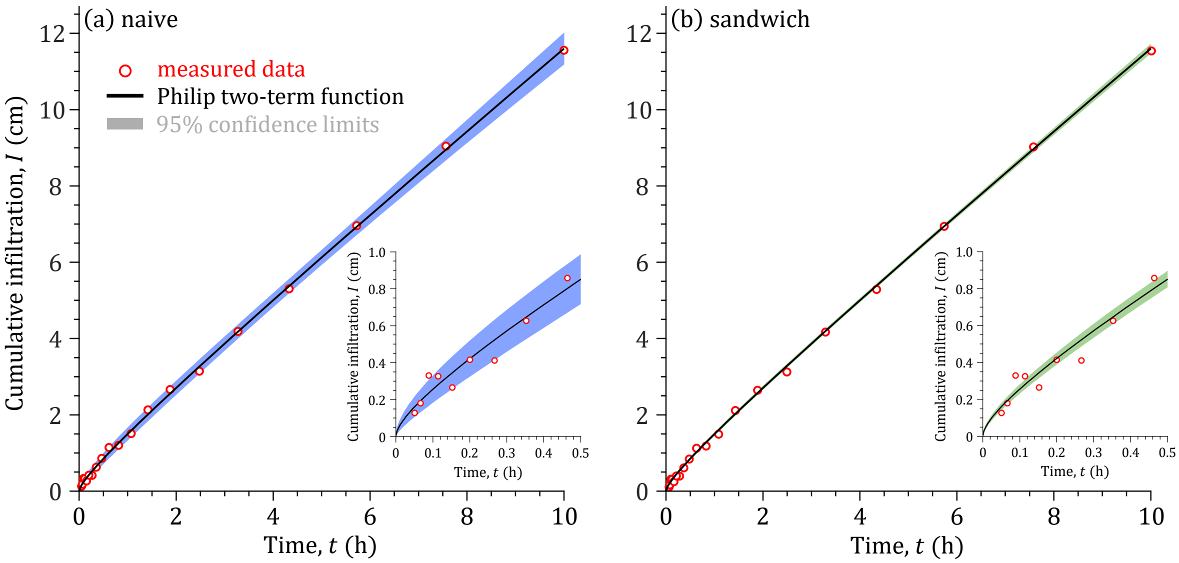

Figure 4: Comparison of observed cumulative infiltration (red dots) and ML-simulated curves (black line) for Philip's two-term infiltration equation. Shaded regions show pointwise 95% confidence bands from (a) the naive (blue) and (b) the sandwich (green) variance estimators. Insets zoom the first 30 minutes.

Thus, the naive confidence intervals half-widths are inflated by the factor $\lambda^{-1/2}$ relative to the sandwich ones, consistent with the plotted intervals.

What is the effect of the learning rate λ on the confidence regions of S and K_{s} when these regions are estimated by numerical methods? Figure 5 presents confidence regions of S and K_{s} derived from (a) MCMC sampling with the DREAM$_{(\mathrm{ZS})}$ algorithm and (b) the bootstrap method. The MCMC-sampled confidence intervals and regions for S and K_{s} are almost an exact copy of those obtained for the naive variance estimator in Fig. 3a. The only noticeable difference is in the 99% confidence region of $\widehat{S}$ and $\widehat{K}_{\mathrm{s}}$, where the outer edges do not extend as far into the plotted domain as the dark gray ellipsoid of the naive variance. This outer perimeter of the high probability density (HPD) region is difficult to delineate exactly by sampling. Much more important, however, is the demonstration that the limiting distribution of the sampled Markov chains is commensurate with the inverse of a single slice of bread, $\widehat{\mathbf{A}}_n^{\mathrm{np}}$. This justifies the terminology of a naive (quasi-)posterior distribution for S and K_{s} and highlights a fundamental limitation of unadjusted posterior sampling under model misspecification. This limitation is echoed in item (v) of the six-point primer in Section 2. When a regular prior is combined with a misspecified likelihood, the resulting posterior distribution, whether explored by MCMC or any other sampling scheme, has local curvature governed by the inverse Hessian $\mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$ and therefore inherits the naive covariance $\boldsymbol{\Sigma}_n^{\mathrm{naive}}$ rather than the robust sandwich covariance $\boldsymbol{\Sigma}_n^{\mathrm{sand}}$. This limitation concerns the target distribution itself, not the MCMC algorithm used to explore it, and motivates the need for sandwich-adjusted or score-based posterior constructions under misspecification.

The bootstrap results in Fig. 5b confirm that the (misspecification-robust) sandwich covariance $\boldsymbol{\Sigma}_n^{\mathrm{sand}} = \sigma_\epsilon^2(\mathbf{D}_n^\top \mathbf{D}_n)^{-1}$ provides the consistent large-sample covariance of the ML estimates. This is expected. For linear regression functions such as Philip's two-term model, the bootstrap variance of the parameters

$$\boldsymbol{\Sigma}_n^{\mathrm{boot}} = \sigma_\epsilon^2(\mathbf{D}_n^\top \mathbf{D}_n)^{-1}\left(\sum_{t=1}^{n} \mathbf{d}(i)\,\mathbf{d}(i)^\top\right)(\mathbf{D}_n^\top \mathbf{D}_n)^{-1}, \tag{43}$$

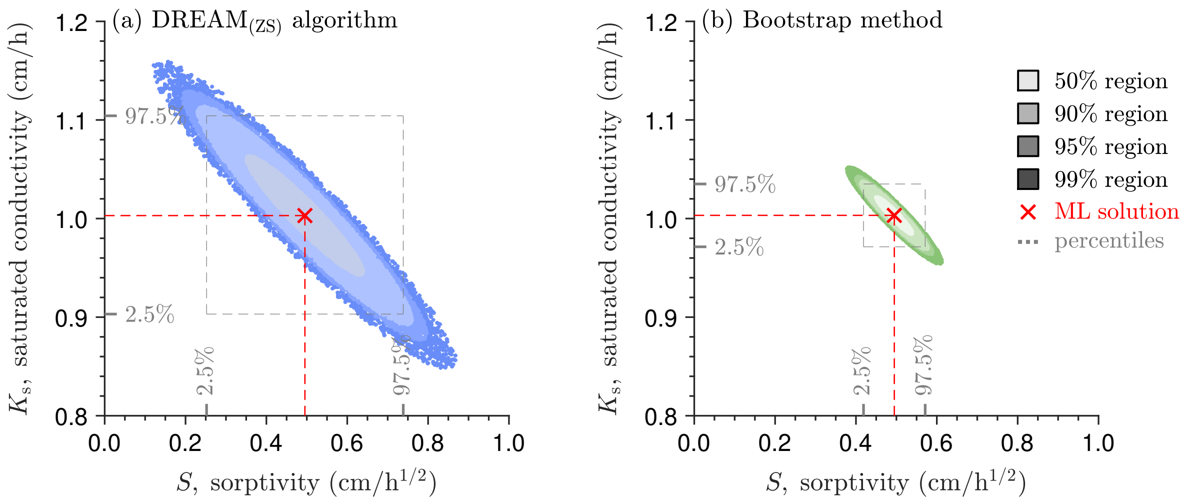

Figure 5: ML solution (red cross), 95% marginal confidence intervals (dashed gray; $\alpha = 0.05$), and joint confidence regions (shaded) for soil sorptivity S and saturated hydraulic conductivity K_s in Philip's two-term infiltration model obtained via Monte Carlo simulation: (a) multi-chain MCMC targeting the normal-power log-likelihood $\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ in Equation (38) with $\lambda = 0.1$ and box constraints on $\boldsymbol{\theta} = (S, K_\text{s})^\top$ and; (b) parametric bootstrap using repeated application of Equation (41) to $m = 10^4$ sampled infiltration curves $\widetilde{\boldsymbol{\omega}}_n^i = \boldsymbol{\omega}_n + \boldsymbol{\epsilon}_n$ with $\boldsymbol{\epsilon}_n \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ and $i = 1, \ldots, m$. Shading shows nested regions (50%, 90%, 95%, 99%; light to dark). Legend and axes as in Fig. 3.

does not depend on the learning rate λ. Since the sum in parentheses equals $\mathbf{D}_n^\top \mathbf{D}_n$, it follows that $\boldsymbol{\Sigma}_n^{\text{boot}} = \sigma_\epsilon^2\,(\mathbf{D}_n^\top \mathbf{D}_n)^{-1}$ which is identical to the sandwich covariance in Equation (40b). Consequently, the bootstrap confidence intervals and regions for $\widehat{S}$ and $\widehat{K}_\text{s}$ in Fig. 5b coincide with their sandwich counterparts in Fig. 3b.

Apart from minor differences in the principal axes of the (S, K_s) ellipses, this example shows that the sandwich variance is not an elusive theoretical construct but an empirically tangible quantity. It also underscores item (vi) of the six-point primer in Section 2. Under likelihood misspecification, Bayesian posterior uncertainty whether explored by MCMC or any other sampling scheme must be adjusted either through post-processing or by modifying the target distribution so that credible regions are asymptotically valid and reflect $\boldsymbol{\Sigma}_n^{\text{sand}}$ rather than $\boldsymbol{\Sigma}_n^{\text{naive}}$.

5.2 Case Study II: Watershed Hydrology

Our second case study analyzes relationships among precipitation, evapotranspiration, groundwater storage and streamflow using the ABC model of Fiering [50]. This model was primarily developed for didactic purposes, and its simplicity provides an excellent opportunity to demonstrate the application of both the naive and sandwich variance estimators to watershed modeling through analytic differentiation. Our use of lowercase letters for water fluxes, p rainfall, i infiltration and e evaporation (all with units L/T) conflicts with capital letters used in surface hydrology but follows the convention defined in Section 3. The symbol e for evaporation should not be confused with the same symbol used for the residuals.

The ABC model describes in a highly simplified manner the main hydrologic fluxes in a watershed using mass balance equations and a groundwater reservoir (see Figure 6). The groundwater reservoir describes subsurface storage as a result of infiltration (percolation) and baseflow. The ABC model is named after its three dimensionless parameters, a, b, and c,

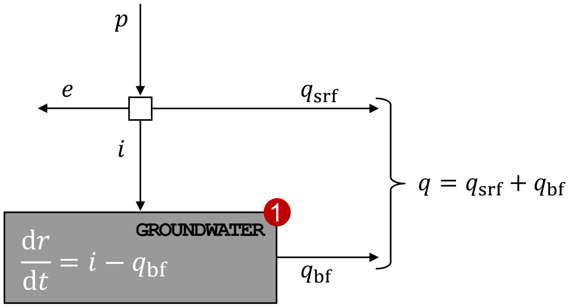
p
e
q_{srf}
i
$q = q_{\mathrm{srf}} + q_{\mathrm{bf}}$
1
GROUNDWATER
$\frac{\mathrm{d}r}{\mathrm{d}t} = i - q_{\mathrm{bf}}$
q_{bf}

Figure 6: Schematic illustration of the ABC model of Fiering [50]. The gray box, labeled in red, is a fictitious subsurface control volume. Its state variable, r, is the water storage in the groundwater reservoir and exerts control on the rainfall-runoff transformation. Arrows portray the fluxes into and out of this compartment, including precipitation, p, precipitation converted into infiltration, i, evaporation, e, and surface runoff, q_{srf}, and groundwater flow or baseflow, q_{bf}. These fluxes are computed as follows, $i_t = a\,p_t$, $e_t = b\,p_t$, $q_{\mathrm{srf}t} = (1-a-b)p_t$, $q_{\mathrm{bf}t} = c\,r_{t-1}$ and $q_t = q_{\mathrm{srf}t} + q_{\mathrm{bf}t}$, where a, b and c are unitless model parameters. Model equations are solved using a mass-conservative second-order integration method. Adaptive time stepping guarantees a robust and accurate numerical solution.

that are presumed to have some physical significance (see Table 1). Since the parameters represent fractions they have upper and lower limits $0 \le a, b, c \le 1$, and since infiltration and evapotranspiration combined cannot exceed total precipitation, $0 \le a + b \le 1$ [155].

Table 1: ABC Model Parameters and Their Symbols, Units and Lower and Upper Bounds.

Parameter	Description	Units	Min.	Max.
a	Fraction of precipitation that becomes surface runoff	–	0	1
b	Fraction of infiltrated water	–	0	1
c	Fraction of groundwater storage that becomes baseflow	–	0	1

We write the ABC-modeled time series of discharge $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$ as a vector-valued regression function $\mathbf{y}_n = \boldsymbol{\mathcal{M}}_{\mathrm{abc}}(\boldsymbol{\theta}; \mathbf{p}_n)$, where $\boldsymbol{\theta} = (a, b, c)^\top$ and $\mathbf{p}_n = (p_1, \ldots, p_n)^\top$ is the rainfall record. The normal power likelihood of the ABC parameters is

$$\begin{aligned} L_n^{\mathrm{np}}(\boldsymbol{\theta}) &= f_{\mathcal{N}_n}^{\lambda}(\mathbf{y}_n; \boldsymbol{\omega}_n, \boldsymbol{\Sigma}_\epsilon) \\ &= (2\pi)^{-\frac{1}{2}n\lambda} |\boldsymbol{\Sigma}_\epsilon|^{-\frac{1}{2}\lambda} \exp\Big(-\tfrac{1}{2}\big(\boldsymbol{\omega}_n - \boldsymbol{\mathcal{M}}_{\mathrm{abc}}(\boldsymbol{\theta}; \mathbf{p}_n)\big)^\top \boldsymbol{\Sigma}_\epsilon^{-1} \big(\boldsymbol{\omega}_n - \boldsymbol{\mathcal{M}}_{\mathrm{abc}}(\boldsymbol{\theta}; \mathbf{p}_n)\big)\Big)^{\lambda}. \end{aligned} \quad (44)$$

Analytic differentiation of the per-observation log-likelihood $\mathcal{L}_{\omega_t}^{\mathrm{np}}(\boldsymbol{\theta})$ is straightforward, but treating ABC-simulated streamflow y_t as a function only of p_t, r_{t-1} and a, b and c yields inaccurate gradients $\partial \mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)/\partial a$ and $\partial \mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)/\partial c$, because the dependence of r_{t-1} on ABC parameters a and c extends further back in time. We therefore perform *backward substitution* of the groundwater-storage recursion $r_{t-1} = f(r_{t-2} \mid \cdot)$ over $\Delta t > 0$ days so that r_{t-1} is expressed in terms of $r_{t-\Delta t-1}$, past and present rainfall, $p_{t-\Delta t}, p_{t-\Delta t+1}, \ldots, p_t$, and ABC parameters, properly capturing the long-memory effects of a and c on y_t. Details of this procedure are provided in Appendix C, which derives closed-form expressions for $\mathbf{A}_{\omega}^{\mathrm{np}}(a,b,c)$ and $\mathbf{B}_{\omega}^{\mathrm{np}}(a,b,c)$ using $\Delta t = 5$ days. These expressions are accurate for large c (e.g., $c > 0.5$) when groundwater residence times are relatively short (a few days). Otherwise, Δt should be increased accordingly. We use Matlab's Symbolic Math Toolbox [145] and Python's SymPy [144] to automate symbolic differentiation and to generate the ABC sensitivity and variability matrices for any $\Delta t \in \mathbb{N}_+$.

We simulate a six-year daily discharge time series with using $a = 0.2$, $b = 0.5$, $c = 0.6$, starting from a near-empty initial groundwater storage $r_0 = 10^{-5}$. We discard the first year as spin-up to remove dependence on the initial state, leaving a five-year record $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$ of $n = 1827$ daily values. We then add heteroskedastic measurement errors, $\boldsymbol{\epsilon}_n \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$, to obtain the observed series $\boldsymbol{\omega}_n = \mathbf{y}_n + \boldsymbol{\epsilon}_n$. The $n \times n$ measurement error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ is diagonal with entries $\sigma^2_{\epsilon,tt} = (s_0 + s_1\, y_t)^2$ using $s_0 = 10^{-2}$ mm/d and $s_1 = 0.1$ (unitless). We do not display ABC measured discharges $\omega_1, \ldots, \omega_n$, as our primary interest is the sensitivity and variability matrices evaluated at the ML estimates $\widehat{\boldsymbol{\theta}}_n = (0.2, 0.5, 0.6)^\top$. Table 2 reports analytic estimates of $\widehat{\mathbf{A}}_n^{\mathrm{np}}$ and $\widehat{\mathbf{B}}_n^{\mathrm{np}}$ for the normal power likelihood with $\lambda = 1$. In this correctly specified example with serially independent *scores*, the HAC variability matrix $\widehat{\boldsymbol{\beta}}_n^{\mathrm{np}}$ collapses to the contemporaneous covariance, $\widehat{\boldsymbol{\beta}}_n^{\mathrm{np}} = \widehat{\mathbf{B}}_n^{\mathrm{np}}$. By the second Bartlett identity, $\widehat{\mathbf{B}}_n^{\mathrm{np}} = \widehat{\mathbf{A}}_n^{\mathrm{np}}$, hence $\widehat{\mathbf{B}}_n^{\mathrm{np}}(\widehat{\mathbf{A}}_n^{\mathrm{np}})^{-1} \longrightarrow \mathbf{I}_d$ and the sandwich variance reduces to the naive variance $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}(\widehat{\mathbf{A}}_n^{\mathrm{np}})^{-1}$ which will provide an accurate description of ABC parameter uncertainty.

Table 2: Ideal case: Analytic estimates for the ABC model's sensitivity $\widehat{\mathbf{A}}_n^{\mathrm{np}}$ and variability $\widehat{\mathbf{B}}_n^{\mathrm{np}}$ matrices, evaluated at the ML estimate $\widehat{\boldsymbol{\theta}}_n$ under the normal power likelihood $L_n^{\mathrm{np}}(\boldsymbol{\theta})$ of Equation (44) with unit learning rate, $\lambda = 1$.

Sensitivity (bread) matrix	Variability (meat) matrix
$\widehat{\mathbf{A}}_n^{\mathrm{np}} = 10^2 \left[\begin{array}{ccc|c} a & b & c & \\ 5.505 & 1.040 & -1.683 & a \\ 1.040 & 2.316 & 0.043 & b \\ -1.683 & 0.043 & 2.038 & c \end{array}\right.$	$\widehat{\mathbf{B}}_n^{\mathrm{np}} = 10^2 \left[\begin{array}{ccc|c} a & b & c & \\ 5.340 & 0.876 & -1.683 & a \\ 0.876 & 2.141 & 0.033 & b \\ -1.683 & 0.033 & 1.953 & c \end{array}\right.$

Next, we induce ABC model inadequacy by perturbing each entry of the original hyetograph with an i.i.d. multiplicative error of ±25% applied to the "measured" rainfall. These uncorrelated, zero-mean perturbations are admittedly stylized, but they preserve nonnegativity (and zeros) and suffice to corrupt the state of the ABC groundwater reservoir and thereby bias the simulated water fluxes. More realistic rainfall error structures (e.g., temporal correlation, intensity-dependent bias and storm-intermittency errors) would only exacerbate these effects by driving state trajectories farther from the truth, shifting the *pseudo-true* parameters and inflating *score* variability. Despite the rainfall errors, the ABC parameter estimates and the sensitivity (curvature) matrix change little. The ML estimate remains close to the data-generating values, $\widehat{\boldsymbol{\theta}}_n = (0.200, 0.506, 0.597)^\top$, and the entries of $\widehat{\mathbf{A}}_n^{\mathrm{np}}$ in Table 3 differ only marginally from those reported previously for the ideal case (Table 2). The rainfall errors increase both the variability of, and the serial correlation among, the successive *score* vectors

Table 3: Semi-ideal case: ABC model's sensitivity and variability matrices of the ML estimate $\widehat{\boldsymbol{\theta}}_n$ under the normal power likelihood $L_n^{\mathrm{np}}(\boldsymbol{\theta})$ of Equation (44) with $\lambda = 1$. We list the consistent HAC estimator ($\widehat{\boldsymbol{\beta}}_n^{\mathrm{np}}$; top right) and, for comparison, the inconsistent i.i.d. estimator ($\widehat{\mathbf{B}}_n^{\mathrm{np}}$; bottom right) of the 3×3 variability matrix.

$\widehat{\mathbf{A}}_n^{\mathrm{np}} = 10^2 \left[\begin{array}{ccc|c} a & b & c & \\ 5.697 & 1.080 & -1.715 & a \\ 1.080 & 2.362 & 0.045 & b \\ -1.715 & 0.045 & 2.082 & c \end{array}\right.$	$\widehat{\boldsymbol{\beta}}_n^{\mathrm{np}} = 10^3 \left[\begin{array}{ccc|c} a & b & c & \\ 1.739 & 0.004 & -0.833 & a \\ 0.004 & 0.479 & 0.073 & b \\ -0.833 & 0.073 & 0.620 & c \end{array}\right.$
	$\widehat{\mathbf{B}}_n^{\mathrm{np}} = 10^2 \left[\begin{array}{ccc|c} 9.742 & 2.980 & -2.320 & a \\ 2.980 & 5.311 & 0.041 & b \\ -2.320 & 0.041 & 3.006 & c \end{array}\right.$

$\mathbf{g}_{\omega_1}(\widehat{\boldsymbol{\theta}}_n), \ldots, \mathbf{g}_{\omega_n}(\widehat{\boldsymbol{\theta}}_n)$. Consequently, we must replace the i.i.d.-based variability estimator $\widehat{\mathbf{B}}_n^{\text{np}}$ with the HAC estimator $\widehat{\boldsymbol{\beta}}_n^{\text{np}}$ of Equation (33).

While the rainfall errors do not alter much the local curvature of the log-likelihood as is evidenced by $\widehat{\mathbf{A}}_n^{\text{np}}$, they do affect the variability matrix $\widehat{\boldsymbol{\beta}}_n^{\text{np}}$. The two matrices, $\widehat{\mathbf{A}}_n^{\text{np}}$ and $\widehat{\boldsymbol{\beta}}_n^{\text{np}}$, now provide conflicting assessments of data informativeness under $\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$. This *information bias* signals misspecification and demands the sandwich variance $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\,(\widehat{\mathbf{A}}_n^{\text{np}})^{-1}\,\widehat{\boldsymbol{\beta}}_n^{\text{np}}\,(\widehat{\mathbf{A}}_n^{\text{np}})^{-1}$ for a robust description of ABC parameter uncertainty. The misalignment between matrices $\widehat{\mathbf{A}}_n^{\text{np}}$ and $\widehat{\boldsymbol{\beta}}_n^{\text{np}}$ is expected to increase with (i) complexity of the rainfall-error characteristics and (ii) residence time of the groundwater reservoir (i.e., for smaller c). To keep this synthetic experiment tractable, we limit the hydraulic memory by setting $c = 0.6$, so that a relatively large fraction of groundwater storage becomes baseflow. This shortens dependence in the *scores* and simplifies the analytic expressions for $\mathbf{A}_n^{\text{np}}(a,b,c)$ and $\mathbf{B}_n^{\text{np}}(a,b,c)$ in Appendix C.

Albeit small, unbiased and stylized, the precipitation errors have an undeniable impact on the uncertainty of the ABC model parameters (see Figure 7). The normal marginal distributions of the sandwich estimator (green line) display a noticeably larger spread than their counterparts from the naive estimator (blue line). The differences between the two estimators are relatively small in this semi-ideal case with synthetic discharge data, a perfect model and small hyetograph errors and are expected to substantially grow with the use of measured discharge data. The histograms of the MCMC-sampled posterior realizations (in blue) are in excellent agreement with the naive pdfs of the ABC model parameters. This reiterates our earlier finding that MCMC methods provide only a naive description of parameter uncertainty. This naive description is robust only if the ABC model is correctly specified. Note that the MCMC sampled posterior distribution of the ABC parameters does not necessarily have to match with the d-variate normal distribution $\mathcal{N}_d(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ of the naive estimator. The first-order approximation of $\widehat{\boldsymbol{\Sigma}}_n$ can be deficient if model nonlinearity is strong and $\widehat{\mathbf{A}}_n$ is a poor description for the region of parameter uncertainty [94, 158].

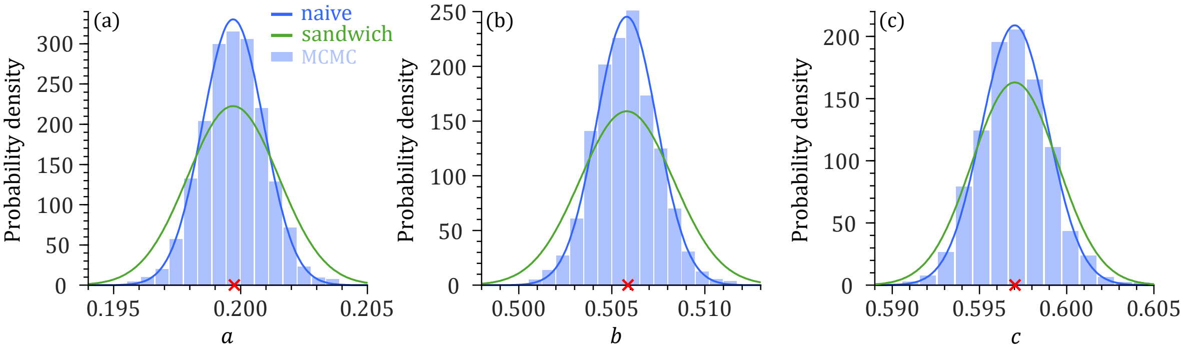

Figure 7: Naive (blue) and sandwich (green) estimates of the marginal densities for the ABC parameters (a) a, (b) b, and (c) c under the normal power likelihood $\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ of Equation (44) with learning rate $\lambda = 1$. The red cross marks the ML estimate. Light-blue histograms show the corresponding marginal parameter distributions obtained with the DREAM$_{\text{(ZS)}}$ algorithm.

Figure 8 compares bivariate 95% confidence regions of (a) (a,b), (b) (a,c) and (c) (b,c) for the normal power log-likelihood function $\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ with $\lambda = 1$. We separately also display the 95% confidence regions (gray dots) of the ABC model parameters derived from MCMC simulation using the DREAM$_{\text{(ZS)}}$ algorithm with a uniform prior on $\boldsymbol{\theta} = (a,b,c)^\top$. The principal axes of the sandwich ellipsoids (green region) align with those of the naive ellipsoids (blue region) but are longer. Consequently, the naive confidence regions are interior to the sandwich ellipsoids. The bivariate scatterplots confirm that MCMC methods such as the DREAM$_{\text{(ZS)}}$ algorithm provide a naive description of ABC parameter uncertainty. This *quasi-*

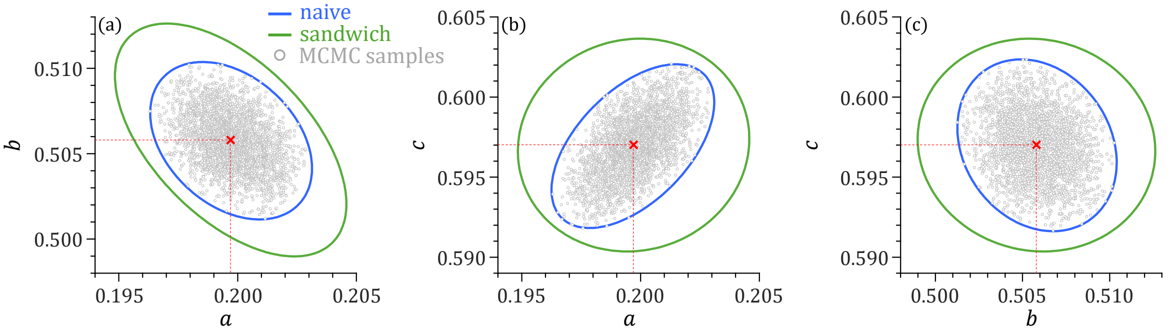

Figure 8: 95% confidence regions of the (a) (a, b), (b) (a, c) and (c) (b, c) parameter pairs of the ABC model for the naive (blue) and sandwich (green) variance estimators using the normal power likelihood $\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ of Equation (44) with a unit learning rate. The red cross marks the ML solution. The gray dots correspond to the "best" 95% of the posterior realizations sampled by the DREAM$_{(\text{ZS})}$ algorithm according to the normal power log-likelihood function.

posterior distribution is valid only if the ABC model is correctly specified. Note that the DREAM$_{(\text{ZS})}$-sampled chains do not provide a uniform coverage of the confidence regions of the ABC parameters. Sample concentration decreases away from the ML solution proportional to the density of the *quasi-posterior* distribution, $p_n(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})L_n^{\text{np}}(\boldsymbol{\theta}) \sim \mathcal{N}_3(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\Sigma}}_n^{\text{naive}})$, of the ABC parameters where $p(\boldsymbol{\theta})$ signifies the prior density.

The expression for the naive variance suggests one can artificially inflate uncertainty by raising the likelihood function $L_n(\boldsymbol{\theta})$ to an arbitrary power $0 < \lambda < 1$, yielding the adjusted covariance estimate $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\lambda^{-1}\widehat{\mathbf{A}}_n^{-1}$. Learning rates $0 < \lambda < 1$ expand confidence regions of the estimated parameters, while values $\lambda > 1$ contract the "posterior" distribution of $\widehat{\boldsymbol{\theta}}_*$, thereby reducing parameter uncertainty. This so-called *likelihood stretching* or tempering is widely practiced in the GLUE method of Beven and Binley [15] to counteract over-conditioning, model inadequacy, and flawed statistical assumptions. For example, it is well known that treating model structural and input data errors as *aleatory* in nature leads to an overestimation of the information content in the data $\omega_1, \ldots, \omega_n$. In the words of Watson and Holmes [169, p. 476], "*there is less information in the experimental data than is conditioned on.*" As a result, tempering (annealing) the observed information $n\widehat{\mathbf{A}}_n$ with a multiplicative constant $0 < \lambda < 1$ may seem defensible when facing misinformation or disinformation. However, this reasoning rests on a false premise. Under model misspecification, the correct measure of data informativeness is *not* the observed information $n\,\widehat{\mathbf{A}}_n$ (i.e., the negative Hessian), but the Godambe information $n\,\widehat{\mathbf{A}}_n\widehat{\mathbf{B}}_n^{-1}\widehat{\mathbf{A}}_n$ which, by standard sandwich theory, does not depend on λ. This invariance was demonstrated in the first case study by application of the bootstrap method to a non-unit learning rate $\lambda = 0.1$ and is confirmed by the analytic expressions of $\mathbf{A}_n^{\text{np}}(a, b, c)$ and $\mathbf{B}_n^{\text{np}}(a, b, c)$ in Appendix C. The sensitivity matrix $\mathbf{A}_n^{\text{np}}(a, b, c)$ of the normal power log-likelihood depends linearly on λ and, as a result, the naive variance is inversely proportional to the learning rate. The variability matrix $\mathbf{B}_n^{\text{np}}(a, b, c)$ of $\mathcal{L}_n^{\text{np}}(a, b, c)$ scales linearly with the square of the learning rate. As a result, λ cancels in the bread-meat-bread matrix product of Equation (40b) and the sandwich variance is invariant to the learning rate. Thus, tempering $L_n(\boldsymbol{\theta})$ does not alter the asymptotically valid uncertainty for $\boldsymbol{\theta}$. Moreover, tempered likelihood weights are not likelihoods in the statistical sense, so the resulting inferences lack a rigorous probabilistic interpretation [17, 96, 165].

Emphasizing Godambe rather than Fisher information helps produce uncertainty estimates that better reflect model and data limitations, aligning with the desideratum expressed by Beven and Binley [15, p. 285] of producing "*an estimate of uncertainty that is consistent* *with the limitations of the model(s) and data used.*" Sandwich variance estimation already goes a long way toward meeting this goal, offering a statistically coherent approach to uncertainty quantification even under misspecification. In this sense, GLUE can be viewed as defining estimating equations that resemble, but do not coincide with, M-estimators. The value of the sandwich approach is that it provides asymptotically correct variance estimates for such GLUE-type estimators without requiring them to arise from a likelihood or loss function, thereby offering a path to place GLUE-style inference on firmer statistical ground without altering its foundational philosophy.

5.3 Case Study III: The Rainfall-Discharge Transformation

The first two studies were purposely simple in support of an analytic demonstration of the naive and sandwich variance estimators. This theoretical treatise has surfaced several misconceptions about how we should quantify model parameter and predictive uncertainty in the face of epistemic and rainfall data errors. Specifically, the sandwich estimator invalidates the use of likelihood stretching as practiced by the GLUE method of Beven and Binley [15] to enlarge the parameter confidence intervals and regions under model misspecification. For all but the ideal cases, the arbitrary power λ of the likelihood $L_n(\boldsymbol{\theta})$ has no bearing on model parameter and predictive uncertainty. Furthermore, sandwich theory calls into question the validity of the posterior parameter distributions documented in the hydrologic literature, specifically the first author's past publications. Certainly, we should not expect the sensitivity $\widehat{\mathbf{A}}_n(\boldsymbol{\theta})$ and variability $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})$ matrices of the ML or MAP parameters of a conceptual watershed model to yield about equal estimates of the *observed* Fisher information. It is now time to put the sandwich theory into practice and explore in more detail the practical consequences of model misspecification on model parameter and predictive uncertainty.

Motivated by the well-known non-Gaussian behavior of streamflow residuals (e.g., skewed and heavy-tailed), we revisit our recently published work on distribution-adaptive likelihoods in Vrugt et al. [166] which develops and applies the Generalized Likelihood plus (GL$^+$) and Universal Likelihood (UL) functions, alongside the Student-t likelihood (SL) of Scharnagl et al. [132]. The GL$^+$, SL and UL do not assume a specific residual distribution; instead, inference is performed over the model parameters and the family of density functions, as defined by one or more nuisance variables. These so-called distribution-adaptive likelihood functions guarantee the most adequate form of the distribution $f(\omega \mid \boldsymbol{\theta})$ of residuals $e_1(\boldsymbol{\theta}),\dots,e_n(\boldsymbol{\theta})$ of observations $\boldsymbol{\omega}_n = (\omega_1,\dots,\omega_n)^\top$ and vector-valued model output $\mathbf{y}_n = \boldsymbol{\mathcal{M}}(\boldsymbol{\theta};\cdot)$. This begs the question of whether the GL$^+$, SL and UL are information-unbiased and offer protection against misspecification? We briefly review the different likelihood functions.

The GL$^+$ function

$$L_n^{\mathrm{g}+}(\boldsymbol{\theta}, s_0, \beta, \xi, \phi_1, \phi_2) = \prod_{t=1}^{n} f_{\mathrm{SEP}}(\underline{\varepsilon}_t; 0, 1, \beta, \xi),$$

admits standardized partial residuals $\underline{\varepsilon}_t(\boldsymbol{\theta}) = \varepsilon_t(\boldsymbol{\theta})/\sigma_\varepsilon$ of an AR(2)-process of the studentized discharge residuals $\underline{e}_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta})/s_{e_t}$

$$\underline{e}_t(\boldsymbol{\theta}) = \phi_1 \underline{e}_{t-1}(\boldsymbol{\theta}) + \phi_2 \underline{e}_{t-2}(\boldsymbol{\theta}) + \varepsilon_t,$$

to the standardized SEP density [133]

$$f_{\mathrm{SEP}}(\underline{\varepsilon}; 0, 1, \beta, \xi) = \frac{2\sigma_\xi \omega_\beta}{\xi + \xi^{-1}} \exp\left(-c_\beta \left| \frac{\mu_\xi + \sigma_\xi \underline{\varepsilon}}{\xi^{\mathrm{sign}(\mu_\xi + \sigma_\xi \underline{\varepsilon})}} \right|^{2/(1+\beta)} \right), \tag{45}$$

where $|\cdot|$ denotes the absolute value or modulus operator, $\mathrm{sign}(a) = |a|/a$, is the signum function, the scalars c_β, ω_β, μ_ξ and σ_ξ, are a function of the kurtosis, $\beta \in (-1, 1]$, and skewness, $\xi > 0$ and $\sigma_\varepsilon^2 = 1 + \phi_1^2 - \phi_2^2 - 2\phi_1^2/(1-\phi_2)$ is the theoretical variance of the partial discharge residuals, $\varepsilon_1(\boldsymbol{\theta}),\ldots,\varepsilon_n(\boldsymbol{\theta})$. The log-likelihood $\mathcal{L}_n^{\mathrm{g}+}(\boldsymbol{\theta}, s_0, \beta, \xi, \phi_1, \phi_2)$ becomes

$$\begin{aligned}\mathcal{L}_n^{\mathrm{g}+}(\boldsymbol{\theta}, s_0, \beta, \xi, \phi_1, \phi_2) \simeq n\log(2\sigma_\xi\omega_\beta) - n\log(\xi+\xi^{-1}) - \sum_{t=1}^{n}\bigl\{\log\bigl(|s_0 + s_1 y_t(\boldsymbol{\theta})|\bigr)\bigr\}\\ - \tfrac{1}{2}n\log(\sigma_\varepsilon^2) - c_\beta\sum_{t=1}^{n}\left|\frac{\mu_\xi + \sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta})}{\xi^{\mathrm{sign}\left(\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta})\right)}}\right|^{2/(1+\beta)},\end{aligned}$$

where analytic expressions for c_β, ω_β, μ_ξ and σ_ξ^2 are found in Schoups and Vrugt [133].

The SL function

$$L_n^{\mathrm{s}}(\boldsymbol{\theta}, s_0, \nu, \xi, \phi_1, \phi_2) = \prod_{t=1}^{n} f_{\mathrm{SST}}(\underline{\varepsilon}_t; 0, 1, \nu, \xi),$$

admits standardized partial discharge residuals $\underline{\boldsymbol{\varepsilon}}(\boldsymbol{\theta}) = (\underline{\varepsilon}_1(\boldsymbol{\theta}),\ldots,\underline{\varepsilon}_n(\boldsymbol{\theta}))^\top$ to the standardized skewed Student's t (SST) density of Scharnagl et al. [132]

$$f_{\mathrm{SST}}(\underline{\varepsilon}; 0, 1, \nu, \xi) = \frac{2\sigma_\xi}{(\xi+\xi^{-1})}\frac{\Gamma\bigl((\nu+1)/2\bigr)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\left[1 + \frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi\underline{\varepsilon}}{\xi^{\mathrm{sign}(\mu_\xi+\sigma_\xi\underline{\varepsilon})}}\right)^2\right]^{-(\nu+1)/2}, \quad (46)$$

where $\nu > 2$ is the degrees of freedom, $\xi > 0$ is the skewness and $\mu_\xi \in \mathbb{R}$ and $\sigma_\xi^2 > 0$ are shift and scale constants which standardize the SST density for $\xi \in \mathbb{R}_+$. The log-likelihood $\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, s_0, \nu, \xi, \phi_1, \phi_2)$ equals

$$\begin{aligned}\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta} \mid s_0, \nu, \xi, \phi_1, \phi_2) \simeq -\tfrac{1}{2}n\log(\sigma_\varepsilon^2) - \sum_{t=1}^{n}\bigl\{\log\bigl(|s_0 + s_1 y_t(\boldsymbol{\theta})|\bigr)\bigr\} + n\log(2) + n\log(\sigma_\xi)\\ + n\log\Bigl(\Gamma\bigl(\tfrac{1}{2}(\nu+1)\bigr)\Bigr) - n\log(\xi+\xi^{-1}) - n\log\bigl(\Gamma(\nu/2)\bigr) - \tfrac{1}{2}n\log(\pi)\\ - \tfrac{1}{2}n\log(\nu-2) - \frac{\nu+1}{2}\sum_{t=1}^{n}\left\{\log\left[1 + \frac{1}{\nu-2}\left(\frac{\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta})}{\xi^{\mathrm{sign}\left(\mu_\xi+\sigma_\xi\underline{\varepsilon}_t(\boldsymbol{\theta})\right)}}\right)^2\right]\right\},\end{aligned}$$

where analytic expressions for μ_ξ and σ_ξ^2 appear in Scharnagl et al. [132].

The UL represents a further generalization of the SL

$$L_n^{\mathrm{u}}(\boldsymbol{\theta}, s_0, p, q, \eta, \phi_1, \phi_2) = \prod_{t=1}^{n} f_{\mathrm{SGT}}(\underline{\varepsilon}_t; 0, 1, p, q, \eta),$$

and admits standardized partial discharge residuals to the standardized skewed generalized Student's t-(SGT)-distribution of Theodossiou [146]

$$f_{\mathrm{SGT}}(\underline{\varepsilon}; 0, 1, p, q, \eta) = \frac{p}{2\kappa_{pq\eta}B\bigl(1/p, q/p\bigr)}\left(1 + \left|\frac{\underline{\varepsilon}+\mu_{pq\eta}}{\kappa_{pq\eta}\bigl(1+\eta\,\mathrm{sign}(\underline{\varepsilon}+\mu_{pq\eta})\bigr)}\right|^p\right)^{-(q+1)/p}, \quad (47)$$

where $p > 0$ and $q > 0$ control the kurtosis of the distribution, $\eta \in (-1, 1)$ is the skewness, $\mu_{pq\eta}$ and $\kappa_{pq\eta}$, are shift and scale constants, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Euler integral of the first kind and $\Gamma(\cdot)$ denotes the gamma function. The scalars $\mu_{pq\eta}$ and $\kappa_{pq\eta}$ negate changes in the mean and variance of the SGT distribution imposed by variables, p, q and η. Analytic expressions are found in Theodossiou [146]. The universal log-likelihood then becomes

$$\begin{aligned}\mathcal{L}^{\mathrm{u}}_n(\boldsymbol{\theta} \mid s_0, p, q, \eta, \phi_1, \phi_2) \simeq & -\tfrac{1}{2} n \log(\sigma^2_{\varepsilon}) - \sum_{t=1}^{n} \bigl\{ \log\bigl(|s_0 + s_1 y_t(\boldsymbol{\theta})|\bigr) \bigr\} + n \log(p) \\ & - n\log(2) - n\log(\kappa_{pq\eta}) - n \log\bigl(B(1/p, q/p)\bigr) \\ & - \frac{q+1}{p} \sum_{t=1}^{n} \left\{ \log\left(1 + \left| \frac{\underline{\varepsilon}_t(\boldsymbol{\theta}) + \mu_{pq\eta}}{\kappa_{pq\eta}\bigl[1 + \eta\,\mathrm{sign}\bigl(\underline{\varepsilon}_t(\boldsymbol{\theta}) + \mu_{pq\eta}\bigr)\bigr]} \right|^{p} \right) \right\}.\end{aligned}$$

The heteroskedastic nature of discharge measurement errors motivates a flow-dependent error scale. Let $e_t(\boldsymbol{\theta}) = \omega_t - y_t(\boldsymbol{\theta})$ denote the discharge residual at time t. We model the measurement-error standard deviation of ω_t as a linear function of simulated discharge, $s_{e_t} = s_0 + s_1\, y_t(\boldsymbol{\theta})$, under model parameters $\boldsymbol{\theta}$, where the intercept s_0 (mm/d) and unitless slope $s_1 > 0$ may be estimated jointly with the other nuisance variables. Though, studentized residuals, $\underline{e}_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta})/s_{e_t}$ should have a unit variance otherwise the partial residuals $\varepsilon_t(\boldsymbol{\theta})$ (after AR filtering with coefficients ϕ_1, ϕ_2) will not attain the theoretic variance $\sigma^2_{\varepsilon} = 1 + \phi^2_1 - \phi^2_2 - 2\phi^2_1/(1-\phi_2)$, and the standardized partial residuals $\underline{\varepsilon}_1(\boldsymbol{\theta}), \ldots, \underline{\varepsilon}_n(\boldsymbol{\theta})$ will be improperly scaled (e.g. Hernández-López and Francés [74]) for the standardized SEP, SST and SGT densities. In practice, we found s_0 to be near zero and fix $s_0 = 10^{-3}$ mm/d for numerical stability. We then treat s_1 as an auxiliary ("phantom") parameter and estimate it by Newton's method so that $\mathrm{Var}[\underline{\mathbf{e}}(\boldsymbol{\theta})] = 1$. Further details are provided in Vrugt et al. [166].

To provide insights into the functional shape of the standardized SEP, SST and SGT densities of Equations (45), (46) and (47), please consider Figure 9 which presents graphs of (a) $f_{\mathrm{SEP}}(\underline{\varepsilon}; 0, 1, \beta, \xi)$, (b) $f_{\mathrm{SST}}(\underline{\varepsilon}; 0, 1, \nu, \xi)$ and (c) $f_{\mathrm{SGT}}(\underline{\varepsilon}; 0, 1, p, q, \eta)$ for $\underline{\varepsilon} \in [-3\frac{1}{2}, 3\frac{1}{2}]$ using different values of the kurtosis, β, ν and p and q and (b) skewness, ξ and η. As is evident from the three graphs, the SEP, SST and SGT densities are very flexible and their nuisance variables describe a large family of continuous probability distributions. This includes the generalized Student's t [70, 106], error [147], and exponential power [22, 143] distributions, as well as the skewed and symmetric Laplace, Cauchy, Student's t, normal, and uniform distributions (see, e.g., [85]). In short, the SEP and SST densities are symmetric for $\xi = 1$, positively skewed for $\xi > 1$ and negatively skewed for $\xi \in (0, 1)$. The SGT has negative skew for $\eta \in (-1, 0)$, positive skew for $\eta \in (0, 1)$ and is symmetric for $\eta = 0$. The kurtosis β, ν and p or q allow the SEP, SST and SGT densities to change from a uniform distribution, ($\beta \to -1$, $\nu \to 2$ and $p, q \to \infty$) to a double-exponential or Laplace distribution ($\beta = 1$ and

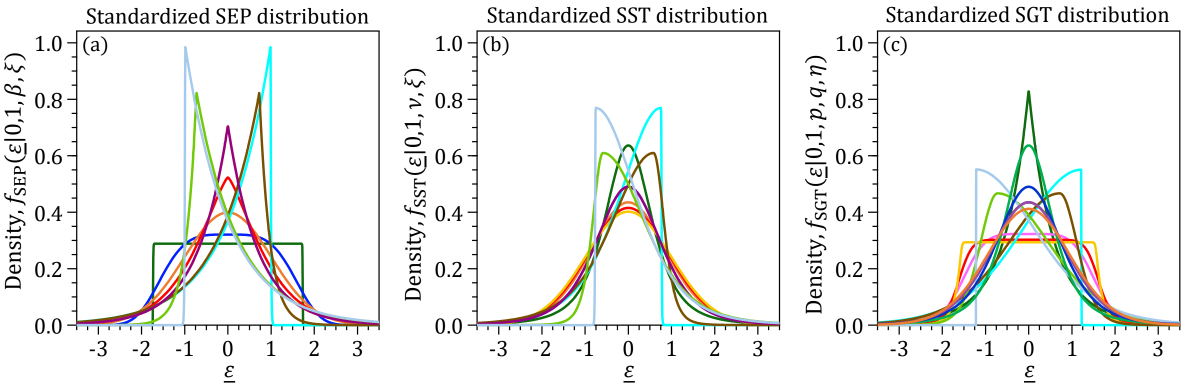

Figure 9: Probability density functions of the standardized (a) skewed exponential power (b) skewed Student's t- and (c) skewed generalized t-distributions of Equations (45), (46) and (47), respectively, for different values of the kurtosis (β, ν or p and q) and skewness (ξ or η).

$p = 1, q \to \infty$) and anything in between. Based on the insights gained from the M-estimators in Table A.1, leptokurtic or heavy-tailed distributions (Student's t, Laplace, Cauchy) can mitigate the undue influence of outliers on parameter estimation.

Vrugt et al. [166] evaluated the performance of the GL^{+}, SL and UL functions by application to the HYdrologic MODel of Boyle [23] using a five year discharge record (10/1/1999 - 9/30/2004) of the Leaf River near Collins, MS, USA. Appendix D provides a concise description of HYMOD, including all governing equations. In short, HYMOD represents the rainfall-discharge transformation through five conceptual control volumes that simulate processes such as evaporation, percolation, river inflow, and baseflow. The $d = 5$ HYMOD parameters, $r_{\mathrm{u,max}}$, a, b, k_{s} and k_{f}, in Table D.1 are subject to inference using the daily record of $n = 1,827$ measured streamflows. To facilitate comparison of the sensitivity and variability matrices, we transform the HYMOD parameters onto the unit hypercube $[0,1]^{d}$ and estimate their normalized counterparts, $\overline{r}_{\mathrm{u,max}}$, $\overline{a}$, $\overline{b}$, $\overline{k}_{\mathrm{s}}$, and $\overline{k}_{\mathrm{f}}$. Before executing HYMOD, these normalized parameters are mapped back to the physical ranges specified in Table D.1, yielding values of $r_{\mathrm{u,max}}$, a, b, k_{s}, and k_{f}.

We revisit the formulations of the UL, GL^{+} and SL functions documented in Table 8 of Vrugt et al. [166], and pay specific attention to the likelihoods illustrated in their Figures 9-11. Specifically, we focus on likelihood 2: $\mathcal{L}_{n}^{\mathrm{u}}(\overline{\boldsymbol{\theta}}, s_0, \eta, \phi_1 \mid p = 2, q = \infty)$, likelihood 12: $\mathcal{L}_{n}^{\mathrm{g+}}(\boldsymbol{\theta}, s_0, \xi, \phi_1 \mid \beta = 0)$ and likelihood 20: $\mathcal{L}_{n}^{\mathrm{s}}(\boldsymbol{\theta}, s_0, \xi, \phi_1 \mid \nu = n - d)$ but with the intercept of the discharge measurement error model fixed at $s_0 = 10^{-3}$. Note that our notation does not make explicit the estimation of normalized parameter values and nuisance variables by the log-likelihood functions. The MAP solution of the HYMOD parameters, $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^{\top}$, together with the nuisance variables, $\boldsymbol{\delta} = (\eta, \phi_1)^{\top}$ or $\boldsymbol{\delta} = (\xi, \phi_1)^{\top}$, is set equal to the posterior sample with the highest posterior density among the joint Markov chains generated by the DREAM$_{\mathrm{(ZS)}}$ algorithm. Next, the MAP sensitivity $\widehat{\mathbf{A}}_{n}$ and variability $\widehat{\boldsymbol{\beta}}_{n}$ matrices are determined by numerical differentiation using the DERIVESTsuite toolbox of D'Errico [40] in Matlab [145]. High-fidelity first- and second-order partial derivatives of the log-likelihood are estimated on the fly from a sequence of logarithmically spaced points away from $(\widehat{\boldsymbol{\theta}}_{n}, \widehat{\boldsymbol{\delta}}_{n})$ using a semi-adaptive central difference schemes of varying orders, combined with a generalized Richardson [129] extrapolation approach.

Table 4 documents the bread and meat matrices of the MAP solution under the UL. The bread and meat matrices of the HYMOD parameters and nuisance variables of the universal log-likelihood function provide conflicting information about the Fisher information of the streamflow record. The entries of the $\widehat{\mathbf{A}}_{n}^{\mathrm{u}}$ and $\widehat{\boldsymbol{\beta}}_{n}^{\mathrm{u}}$ are substantially different with some elements differing up to orders of magnitude. This failure of the second Bartlett identity is a testament to model misspecification. Although the SGT family can flexibly approximate a wide range of continuous distributions and thereby fit the (partial) discharge residuals well, this does not guarantee information-unbiasedness of the UL. In particular, under $\mathcal{L}_{n}^{\mathrm{u}}(\boldsymbol{\theta}, \boldsymbol{\delta} \mid \cdot)$ the sensitivity and variability matrices need not coincide. More broadly, distribution-adaptive likelihoods are not exempt from information bias (i.e., $\mathbf{A}_{n} \neq \mathbf{B}_{n}$), even when residual autocorrelation is explicitly modeled, albeit perhaps rudimentary, via an AR(1) or AR(2) process. Further evidence for this claim is provided in Appendix E, which tabulates the sensitivity matrix $\widehat{\mathbf{A}}_{n}$ and variability matrix $\widehat{\boldsymbol{\beta}}_{n}$ of the GL^{+}, SL, and Normal Likelihood (NL), estimated from the partial discharge residuals.

Next, we use the tabulated bread and meat matrices of the GL^{+}, SL and UL functions and compute the naive and sandwich variance matrices following Equation (31). Table 5 reports only the 5×5 covariance block corresponding to the HYMOD parameters, whereas the full 7×7 covariance matrices (including the nuisance variables) are not shown. For completeness, we also list $\widehat{\boldsymbol{\Sigma}}_{n}^{\mathrm{naive}}$ and $\widehat{\boldsymbol{\Sigma}}_{n}^{\mathrm{sand}}$ for the NL function $\mathcal{L}_{n}^{\mathrm{n}}(\boldsymbol{\theta}, \phi_1 \mid s_0 = 10^{-3})$ with an AR(1)-process of the studentized discharge residuals.

Table 4: Sensitivity $\widehat{\mathbf{A}}^{\mathrm{u}}_n$ and HAC variability $\widehat{\boldsymbol{\beta}}^{\mathrm{u}}_n$ matrices evaluated at the MAP solution for HYMOD parameters $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^\top$ and nuisance variables η and ϕ_1 under the universal likelihood $\mathcal{L}^{\mathrm{u}}_n(\boldsymbol{\theta}, \eta, \phi_1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$.

$$\widehat{\mathbf{A}}^{\mathrm{u}}_n = 10^2 \begin{array}{rrrrrrrl}
\overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} & \overline{\eta} & \overline{\phi}_1 & \\
0.479 & 0.049 & -0.288 & 0.068 & 0.154 & 0.011 & -0.042 & \overline{r}_{\mathrm{u,max}} \\
0.049 & 0.144 & -0.045 & 0.066 & 0.070 & 0.035 & -0.012 & \overline{a} \\
-0.288 & -0.045 & 2.213 & -0.243 & 0.195 & 0.007 & 0.040 & \overline{b} \\
0.068 & 0.066 & -0.243 & 0.213 & -0.072 & 0.018 & 0.003 & \overline{k}_{\mathrm{s}} \\
0.154 & 0.070 & 0.195 & -0.072 & 0.613 & 0.026 & -0.064 & \overline{k}_{\mathrm{f}} \\
0.011 & 0.035 & 0.007 & 0.018 & 0.026 & 0.067 & 0.031 & \overline{\eta} \\
-0.042 & -0.012 & 0.040 & 0.003 & -0.064 & 0.031 & 0.104 & \overline{\phi}_1
\end{array}$$

$$\widehat{\boldsymbol{\beta}}^{\mathrm{u}}_n = 10^2 \begin{array}{rrrrrrrl}
6.684 & 0.300 & -3.084 & 0.379 & 4.079 & -0.052 & -1.100 & \overline{r}_{\mathrm{u,max}} \\
0.300 & 1.941 & 0.983 & 1.487 & 1.203 & 0.069 & 0.128 & \overline{a} \\
-3.084 & 0.983 & 9.112 & -1.720 & 2.497 & 0.624 & -1.169 & \overline{b} \\
0.379 & 1.487 & -1.720 & 4.321 & -1.600 & -0.123 & 1.830 & \overline{k}_{\mathrm{s}} \\
4.079 & 1.203 & 2.497 & -1.600 & 9.556 & -0.040 & -3.084 & \overline{k}_{\mathrm{f}} \\
-0.052 & 0.069 & 0.624 & -0.123 & -0.040 & 0.220 & 0.221 & \overline{\eta} \\
-1.100 & 0.128 & -1.169 & 1.830 & -3.084 & 0.221 & 2.636 & \overline{\phi}_1
\end{array}$$

Table 5: Naive and sandwich variance matrices for MAP values of the HYMOD parameters using the Generalized Likelihood plus $\mathcal{L}_n^{\mathrm{g}+}(\boldsymbol{\theta},\xi,\phi_1 \mid s_0 = 10^{-3}, \beta = 0)$, Student Likelihood $\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta},\xi,\phi_1 \mid s_0 = 10^{-3}, \nu = n-d)$, Universal Likelihood $\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta},\eta,\phi_1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$ and Normal Likelihood $\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta},\phi_1 \mid s_0 = 10^{-3})$ functions. Nuisance parameters are estimated jointly but omitted from display.

	$\widehat{\boldsymbol{\Sigma}}_n^{\mathrm{naive}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}$	$\widehat{\boldsymbol{\Sigma}}_n^{\mathrm{sand}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}$
GL$^+$	$\begin{array}{cc} & \begin{array}{rrrrr} \overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} \end{array} \\ 10^{-5} & \left[\begin{array}{rrrrr} 1.425 & -0.130 & 0.184 & -0.110 & -0.376 \\ -0.130 & 5.684 & -0.092 & -1.846 & -0.574 \\ 0.184 & -0.092 & 0.329 & 0.390 & -0.108 \\ -0.110 & -1.846 & 0.390 & 3.905 & 0.578 \\ -0.376 & -0.574 & -0.108 & 0.578 & 1.262 \end{array}\right] \end{array}$	$\begin{array}{ccc} & \begin{array}{rrrrr} \overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} \end{array} & \\ 10^{-4} & \left[\begin{array}{rrrrr} 1.554 & -0.836 & 0.016 & 0.350 & 0.085 \\ -0.836 & 6.619 & 0.237 & 0.595 & 0.019 \\ 0.016 & 0.237 & 0.147 & 0.610 & 0.052 \\ 0.350 & 0.595 & 0.610 & 7.226 & 1.117 \\ 0.085 & 0.019 & 0.052 & 1.117 & 1.252 \end{array}\right] & \begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \end{array} \end{array}$
SL	$10^{-5}\left[\begin{array}{rrrrr} 1.499 & 0.009 & 0.183 & -0.405 & -0.445 \\ 0.009 & 5.617 & -0.041 & -1.789 & -0.614 \\ 0.183 & -0.041 & 0.319 & 0.287 & -0.122 \\ -0.405 & -1.789 & 0.287 & 3.843 & 0.668 \\ -0.445 & -0.614 & -0.122 & 0.668 & 1.299 \end{array}\right]$	$10^{-4}\left[\begin{array}{rrrrr} 1.633 & -0.763 & -0.023 & -0.495 & -0.058 \\ -0.763 & 6.510 & 0.222 & 0.724 & 0.062 \\ -0.023 & 0.222 & 0.118 & 0.383 & 0.030 \\ -0.495 & 0.724 & 0.383 & 6.920 & 1.256 \\ -0.058 & 0.062 & 0.030 & 1.256 & 1.302 \end{array}\right]\begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \end{array}$
UL	$10^{-5}\left[\begin{array}{rrrrr} 1.482 & 0.016 & 0.182 & -0.410 & -0.441 \\ 0.016 & 5.649 & -0.037 & -1.802 & -0.618 \\ 0.182 & -0.037 & 0.314 & 0.274 & -0.122 \\ -0.410 & -1.802 & 0.274 & 3.833 & 0.673 \\ -0.441 & -0.618 & -0.122 & 0.673 & 1.298 \end{array}\right]$	$10^{-4}\left[\begin{array}{rrrrr} 1.610 & -0.745 & -0.021 & -0.467 & -0.063 \\ -0.745 & 6.517 & 0.236 & 0.765 & 0.066 \\ -0.021 & 0.236 & 0.115 & 0.366 & 0.027 \\ -0.467 & 0.765 & 0.366 & 6.871 & 1.263 \\ -0.063 & 0.066 & 0.027 & 1.263 & 1.307 \end{array}\right]\begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \end{array}$
NL	$10^{-3}\left[\begin{array}{rrrrr} 2.507 & -0.022 & 1.464 & 0.215 & -0.047 \\ -0.022 & 0.051 & 0.002 & -0.039 & -0.008 \\ 1.464 & 0.002 & 1.074 & 0.147 & -0.023 \\ 0.215 & -0.039 & 0.147 & 0.114 & 0.005 \\ -0.047 & -0.008 & -0.023 & 0.005 & 0.010 \end{array}\right]$	$10^{-2}\left[\begin{array}{rrrrr} 6.852 & -0.067 & 4.393 & 0.451 & 0.039 \\ -0.067 & 0.037 & -0.001 & -0.022 & -0.002 \\ 4.393 & -0.001 & 3.071 & 0.315 & 0.033 \\ 0.451 & -0.022 & 0.315 & 0.089 & 0.008 \\ 0.039 & -0.002 & 0.033 & 0.008 & 0.007 \end{array}\right]\begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \end{array}$

The tabulated results of the different likelihoods highlight several interesting findings.

1. The naive variance matrices of the GL$^+$, SL and UL are almost perfectly aligned. The diagonal and off-diagonal entries of $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}$ are in remarkably close agreement.
2. The naive variance estimates of the GL$^+$, SL and UL are about 100× smaller on average than their counterparts from NL. This showcases the adaptive capabilities of the GL$^+$, SL and UL functions and demonstrates the strong control the likelihood exerts on data informativeness. The NL function does not adequately characterize the distribution of the discharge residuals [159]. Consequently, the naive variance estimates of HYMOD should be interpreted with caution.
3. The sandwich variance estimates of the GL$^+$, SL, UL, and NL are, on average, an order of magnitude larger than their naive counterparts, highlighting the substantial underestimation of parameter uncertainty by the naive estimator. This finding is consistent with our earlier results for the ABC model.
4. The sandwich variance matrices $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$ of the GL$^+$, SL and UL are in very close agreement and about two orders of magnitude larger than their estimates of the NL function.

When observations are dependent and the model is correctly specified, the second Bartlett identity continues to hold provided the *score* variance $\widehat{\mathbf{B}}_n$ is interpreted as the long-run variance $\boldsymbol{\beta}_n$; i.e., $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_n(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\beta}_n(\boldsymbol{\theta})]$. HAC estimator $\widehat{\boldsymbol{\beta}}_n$ in Equation (33) is consistent under standard mixing conditions with $b_n \to \infty$ and $b_n/n \to 0$. Under strict stationarity and ergodicity, $\widehat{\mathbf{A}}_n \xrightarrow{\text{p}} \mathbf{A}_* = \mathbb{E}_q[-\nabla^2 \mathcal{L}_\Omega(\boldsymbol{\theta}_*)]$, and the asymptotic sandwich covariance estimator $\frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}$ remains valid under weak dependence even though the naive curvature $\frac{1}{n}\widehat{\mathbf{A}}_n^{-1}$ does not account for serial correlation. In short, serial dependence inflates uncertainty through $\widehat{\boldsymbol{\beta}}_n$. The curvature term $\widehat{\mathbf{A}}_n$ is comparatively insensitive to that dependence under a law-of-large-numbers.

The parameters of conceptual hydrologic models are much less well defined than commonly thought and expressed by the sensitivity (= negative Hessian) matrix of the MAP parameter estimates $\widehat{\boldsymbol{\theta}}_n$ and/or posterior parameter distribution sampled by MCMC methods. Under model misspecification, there is (much) greater uncertainty in the statistical analysis than supposed by this theoretical estimate. As a result, we should draw inferences based on the sandwich variance matrix. Altogether, the tabulated results confirm our thesis that the MAP sensitivity matrix $\widehat{\mathbf{A}}_n$ provides overly optimistic estimates of the informativeness of streamflow measurements. As a result, the naive variance estimator underestimates HYMOD parameter uncertainty and its associated $100(1-\alpha)\%$ confidence regions will depart from asymptotic $100(1-\alpha)\%$ confidence regions [127]. In a Bayesian context, this is equivalent to overly optimistic estimates of parameter [14, 16, 162] and predictive [125, 164] uncertainty.

At this point, we remind the reader that the naive and sandwich variance matrices of Table 5 provide only a first-order approximation of the actual HYMOD parameter uncertainty. These two estimators are exact for models such as Philip's infiltration equation, whose output y depends linearly on $\boldsymbol{\theta}$. For all other models, the naive and sandwich variance estimators of Equation (31) approximate, but do not equal, the *true* parameter uncertainty. Figure 10 displays pdfs of the HYMOD parameters derived from the frequentist naive (blue) and sandwich (green) estimators under $\mathcal{L}_n^{\text{u}}(\boldsymbol{\theta}, \eta, \phi_1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$. For comparison, we also show histograms of the empirical densities of the HYMOD parameters obtained from MCMC simulation under the UL using the DREAM$_{\text{(ZS)}}$ algorithm. Blue histograms correspond to naive posterior samples, whereas the green histograms result from so-called open-faced sandwich adjustment (OFSA). A full account of this method is provided by Shaby [134], and Vrugt and Diks [160] compare OFSA against other sandwich-adjusted sampling approaches.

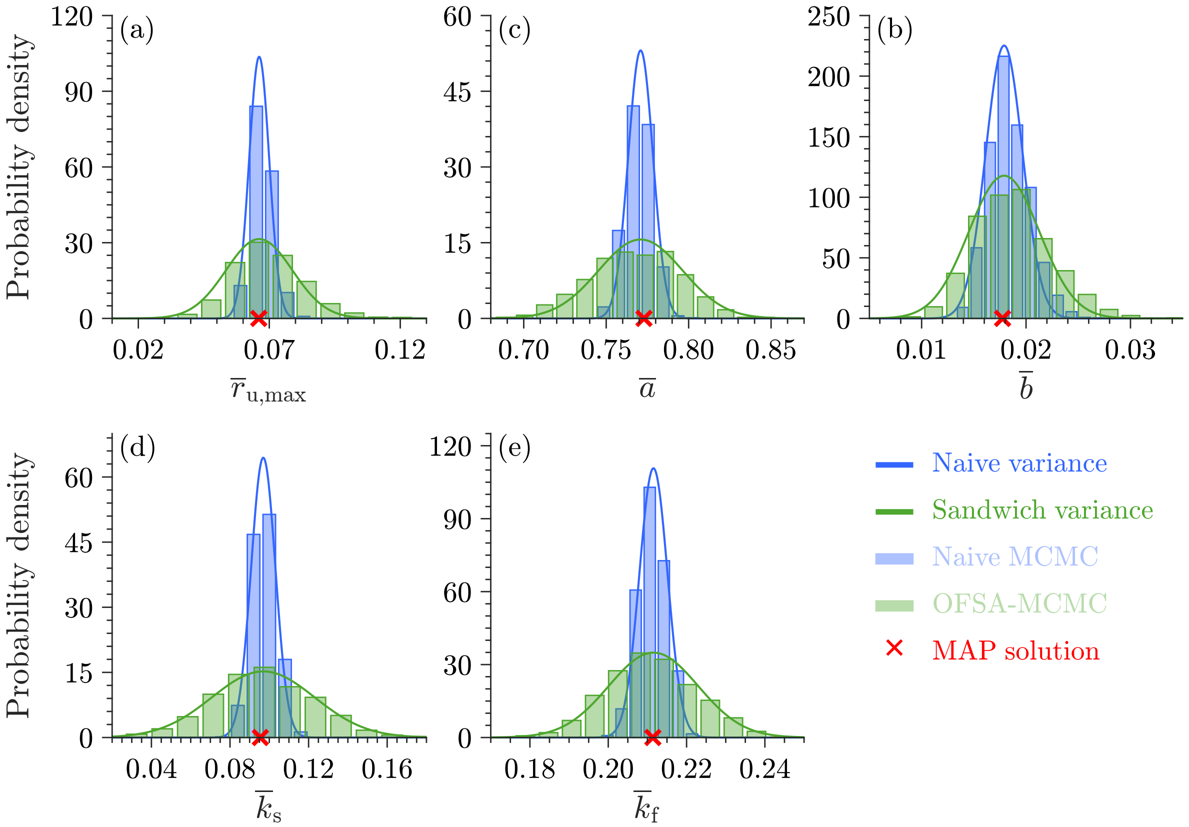

Figure 10: Probability density functions of the normalized HYMOD parameters (a) $\overline{r}_{\mathrm{u,max}}$, $\overline{a}$, $\overline{b}$, $\overline{k}_{\mathrm{s}}$ and $\overline{k}_{\mathrm{f}}$ under the Universal Likelihood $L_n^{\mathrm{u}}(\boldsymbol{\theta}, \eta, \phi_1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$ using naive (blue line) and sandwich (green line) estimators. Blue histograms are empirical densities of the HYMOD parameters derived from MCMC simulation with the DREAM$_{(\mathrm{ZS})}$ algorithm. Green histograms correspond to OFSA posterior samples. Red cross marks MAP estimate.

The density functions of the frequentist variance estimators confirm our previous conclusions for the ABC model. The naive variance $\widehat{\boldsymbol{\Sigma}}_n^{\mathrm{naive}}$ (blue line) provides an erroneous description of HYMOD parameter uncertainty. The MAP sensitivity matrix $\widehat{\mathbf{A}}_n^{\mathrm{u}}$ overestimates discharge data informativeness, and as result, the naive pdfs are overly concentrated and peaked. The normal marginal distributions of the sandwich estimator (green lines) extend over a much larger region of the parameter space, reflecting a greater dispersion and substantially larger uncertainty in the HYMOD parameters.

As expected, MCMC simulation with the DREAM$_{(\mathrm{ZS})}$ algorithm provides only a naive description of parameter uncertainty. The empirical density functions of the HYMOD parameters (blue histograms), obtained from the posterior realizations of the sampled chains, closely match the normal marginal distributions implied by the naive variance estimator of Equation (31a). Moreover, the MAP solution of the sampled Markov chains is indistinguishable from the ML parameter values obtained independently by maximizing $\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, \eta, \phi 1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$ with the simplex algorithm of Nelder and Mead [111].

But how can MCMC be modified so that the sampled chains target the asymptotically valid sandwich distribution? The OFSA method of Shaby [134] offers a practical route. In our application, the OFSA samples (green histograms) closely match the normal marginal distributions (green lines) implied by the frequentist sandwich estimator in Equation (31b). OFSA applies direction-specific dilations along the principal axes of the naive posterior samples to align their local curvature around the MAP solution $\widehat{\boldsymbol{\theta}}_n$ with the sandwich estimator

$\widehat{\mathbf{\Sigma}}_n^{\text{sand}}$. The post-hoc adjustment closely matches the frequentist sandwich distribution.

Further evidence for this claim is presented in Figure 11, which displays a scatter-plot matrix of the bivariate 95% parameter confidence regions sampled by the DREAM$_{(\text{ZS})}$ algorithm. This figure reiterates several of our earlier findings. First, the naive variance estimator substantially underestimates HYMOD parameter uncertainty. The culprit is the sensitivity (meat) matrix $\widehat{\mathbf{A}}_n^{\text{np}}$, which overstates streamflow data informativeness under model misspecification and, consequently, produces overly narrow confidence regions. The observed Godambe information $\widehat{\mathcal{G}}_n$ provides the correct measure of data informativeness under misspecification and substantially enlarges the confidence intervals and regions of the HYMOD parameters. Second, the DREAM$_{(\text{ZS})}$-sampled bivariate scatterplots of the posterior realizations align closely with the 95% confidence regions (blue ellipsoids) derived from the naive estimator.

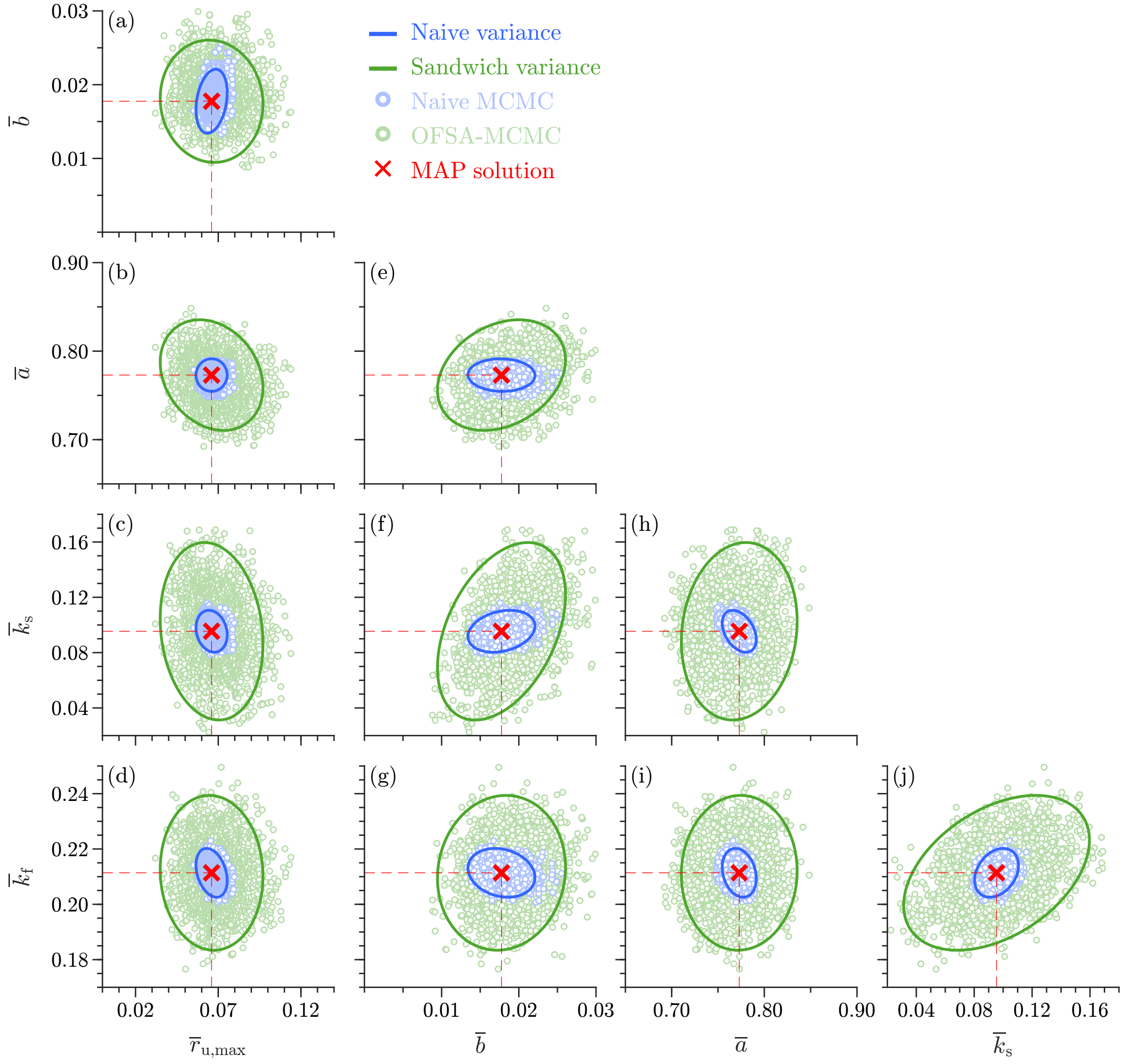

Figure 11: Scatter-plot matrix of bivariate 95% confidence regions of the normalized HYMOD parameters: (a) $\bar{r}_{\text{u,max}}$-$\bar{b}$, (b) $\bar{r}_{\text{u,max}}$-$\bar{a}$, (c) $\bar{r}_{\text{u,max}}$-$\bar{k}_{\text{s}}$, (d) $\bar{r}_{\text{u,max}}$-$\bar{k}_{\text{f}}$, (e) $\bar{b}$-$\bar{a}$, (f) $\bar{b}$-$\bar{k}_{\text{s}}$, (g) $\bar{b}$-$\bar{k}_{\text{f}}$, (h) $\bar{a}$-$\bar{k}_{\text{s}}$, (i) $\bar{a}$-$\bar{k}_{\text{f}}$, and (j) $\bar{k}_{\text{s}}$-$\bar{k}_{\text{f}}$. Blue dots denote the naive posterior samples from DREAM$_{(\text{ZS})}$ using $\mathcal{L}_n^{\text{u}}(\boldsymbol{\theta}, \eta, \phi_1 \mid s_0 = 10^{-3}, p = 2, q = \infty)$ with a uniform prior. Green dots are the resulting OFSA samples. Ellipses are 95% confidence regions of Equation (22) according to the naive (blue) and sandwich (green) estimators. The red cross marks the ML/MAP solution.

This confirms once again that MCMC simulation methods yield only a naive description of parameter uncertainty when the likelihood function is misspecified. Third, the strong similarities between the linear (blue ellipsoids) and nonlinear (blue dots) confidence regions validate the accuracy of the ML meat matrix $\widehat{\mathbf{A}}_n^{\text{np}}$ in quantifying naive HYMOD uncertainty. Finally, the 95% credible regions of the sandwich-adjusted posterior samples (green dots) match closely with the frequentist 95% confidence regions obtained from $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$. In Section 6, we briefly discuss approaches for modifying the posterior target or post-processing posterior draws so that MCMC sampling yields uncertainty consistent with the asymptotically valid sandwich distribution.

Then, Figure 12 shows posterior predictive bands for HYMOD-simulated streamflow over a representative segment of the five-year training period, obtained by propagating the DREAM$_{\text{(ZS)}}$-derived (a) naive and (b) OFSA posterior samples through the model. The larger sandwich variances translate into substantially wider streamflow intervals, most evident in the right-hand insets. The 95% band covers about 36.29% of the observations under the sandwich estimator, compared with about 10.18% under the naive estimator. This behavior is consistent with M-estimation theory and reflects a more realistic accounting of parameter uncertainty under model and data limitations.

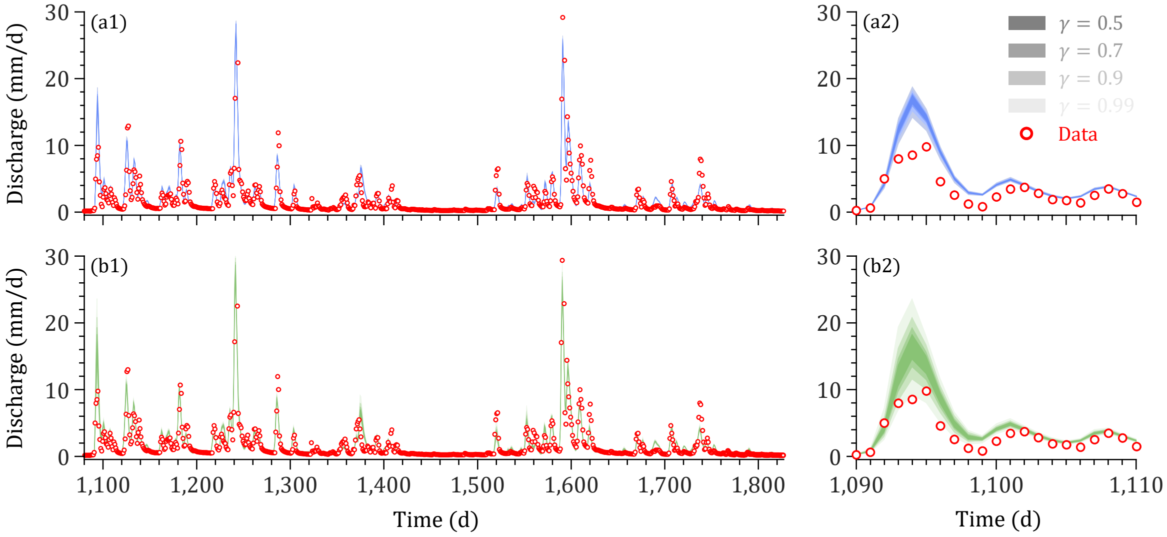

Figure 12: HYMOD intervals for simulated streamflow due to (a) the naive and (b) the sandwich variance estimators. Shaded bands denote the 68%, 90%, 95%, and 99% parameter-induced discharge intervals. Red dots are observed discharge.

6 Sandwich-Adjusted MCMC Simulation

The case studies have surfaced a fundamental limitation of Bayesian methods in the face of model misspecification. The asymptotic covariance matrix of the posterior distribution sampled by MCMC methods is the matrix inverse of a single slice of bread $\mathbf{A}_n^{-1}$ and not the asymptotically valid sandwich matrix. This warrants a deeper look into MCMC theory, specifically how must we adapt the acceptance probability of candidate points so that the Markov chains sample the sandwich distribution?

In this paper, we will not delve into MCMC theory and only briefly discuss one simple practical remedy that helps adjust the random walk of the Metropolis-Hastings algorithm [73, 107] to the sandwich posterior distribution. This so-called magnitude adjustment raises the likelihood to a constant, and helps the MCMC method into generating samples from the sandwich posterior distribution. In a companion paper by Vrugt and Diks [160] we review and evaluate existing approaches to robust Bayesian estimation of the sandwich distribution. This includes the OFSA method of Shaby [134], and magnitude- [41, 127] and curvature- [30, 127] adjusted MCMC simulation. This paper also presents theory and results of a more convenient and rigorous approach, which we coined *kernel-amendment* for sandwich-adjusted MCMC simulation. Unlike existing approaches that rely on arbitrary matrix square roots, eigendecompositions or a single scaling factor applied uniformly across the parameter space, *kernel-amendment* employs a power likelihood with a parameter-dependent learning rate that enables direction-specific tempering of the likelihood. This allows the sampler to capture directional asymmetries in the sandwich distribution, particularly under model misspecification or in small-sample regimes, and yields credible regions that remain valid when the naive Bayesian variance-covariance estimator underestimates uncertainty.

6.1 Magnitude-Adjusted MCMC Simulation

It is not too difficult to adjust the magnitude of the log-likelihood function so as to enforce the MAP sandwich variance matrix $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$ upon the posterior realizations of the sampled Markov chain(s). If $\mathcal{L}_n(\boldsymbol{\theta})$ is quadratic in the neighborhood of the MAP parameter values $\widehat{\boldsymbol{\theta}}_n$, then posterior exploration of the product $k\mathcal{L}_n(\boldsymbol{\theta})$ may yield a reasonable approximation of the sandwich posterior. The so-called *omnibus* magnitude adjustment, proposed by Pauli et al. [119], enforces the second Bartlett identity (18) by estimating scalar k as follows [41, 127]

$$k = \frac{d}{\operatorname{tr}\big((\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}})^{-1}\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}\big)} = \frac{d}{\operatorname{tr}(\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1})}, \tag{48}$$

where the unary trace operator $\operatorname{tr}(\mathbf{X})$ returns the sum of the diagonal elements of the square matrix $\mathbf{X} = \widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}$ and $\widehat{\mathbf{A}}_n = \mathbf{A}_n(\widehat{\boldsymbol{\theta}}_n)$ and $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_n(\widehat{\boldsymbol{\theta}}_n)$. The trace of $\mathbf{X}$ is equal to the sum of the eigenvalues of the matrix-matrix product $\widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n$.

Application of Equation (48) to our soil water infiltration study yields

$$k = \frac{d}{\operatorname{tr}\big(\lambda^2\sigma_e^{-2}(\mathbf{D}_n^\top\mathbf{D}_n)\lambda^{-1}\sigma_e^2(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}\big)} = \frac{d}{\lambda\operatorname{tr}(\mathbf{I}_d)} = \frac{d}{\lambda d} = \lambda^{-1}.$$

This choice of $k = \lambda^{-1}$ cancels the learning rate from the product $k\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ and results in the natural logarithm of the normal likelihood function $L_n^{\text{n}}(\boldsymbol{\theta})$ in Equation (37). Therefore, we yield $\widehat{\boldsymbol{\Sigma}}_n^{\text{np}} = \sigma_\epsilon^2(\mathbf{D}_n^\top\mathbf{D}_n)^{-1}$ and it should not require a practical demonstration that MCMC simulation with the magnitude-adjusted normal power log-likelihood function $k\mathcal{L}_n^{\text{np}}(\boldsymbol{\theta})$ yields the sandwich posterior distribution of Fig. 3b.

In our second study with synthetic streamflow data from the ABC model, we obtain $k = 0.974$ for the ideal case (Table 2) and $k = 0.406$ for the non-ideal case with error-corrupted rainfall (Table 3). The use of $k\mathcal{L}_n^{\text{n}}(\boldsymbol{\theta})$ with $k < 1$ decreases the curvature of $\mathcal{L}_n^{\text{n}}(\boldsymbol{\theta})$, thereby enhancing the dispersion of the MCMC-sampled posterior realizations in a manner consistent with the sandwich variance.

In our third and last study with HYMOD and measured streamflow data, we yield values of $k = 0.074$, $k = 0.074$, $k = 0.074$ and $k = 0.064$ for the bread and meat matrices of the UL, GL$^+$, SL and NL functions documented in Tables 4, E.1, E.2 and E.3, respectively. The better the agreement between the MAP sensitivity matrix $\widehat{\mathbf{A}}_n$ and HAC variability matrix $\widehat{\boldsymbol{\beta}}_n$, the closer k will be to unity and the less information biased the log-likelihood function is. The omnibus scalar favors the three distribution-adaptive likelihood functions, although their listed k-values of 0.074 are only a marginal improvement over $k = 0.064$ under $L_n^{\text{n}}(\boldsymbol{\theta}, \phi_1 \mid s_0 = 10^{-3})$ with AR(1) models of the studentized discharge residuals. Potentially larger values of k may be obtained when the GL$^+$, SL, and UL are fully parameterized, that is, when all nuisance parameters of their respective SEP, SST, and SGT densities are estimated jointly. Table 6 presents the outcome of this analysis. The most important results are as follows.

Table 6: Log-likelihood maximum $\mathcal{L}_n(\{\widehat{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\delta}}_n\})$ and `omnibus` scalar k for different formulations of the GL$^+$, SL, UL and NL functions. Inactive nuisance variables are set to default values. The second column lists the unique identifier of each likelihood function used in Table 8 of Vrugt et al. [166].

#	ID	Formulation	$\mathcal{L}_n(\{\widehat{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\delta}}_n\})$	k
1	1	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, q, \eta, \phi_1, \phi_2 \mid s_0 = 10^{-3})$	1504.6	0.310
2	8	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, q, \eta, \phi_1 \mid s_0 = 10^{-3},\ \phi_2 = 0)$	1505.2	0.342
3	6	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, q, \phi_1 \mid s_0 = 10^{-3},\ \eta = 0,\ \phi_2 = 0)$	1441.5	0.202
4	5	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, q, \eta \mid s_0 = 10^{-3},\ \phi_1 = 0,\ \phi_2 = 0)$	105.90	0.057
5	2	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, \eta, \phi_1 \mid s_0 = 10^{-3},\ p = 2,\ q = \infty,\ \phi_2 = 0)$	671.40	0.074
6	4	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, q, \phi_1 \mid s_0 = 10^{-3},\ p = 2,\ \eta = 0,\ \phi_2 = 0)$	1410.5	0.156
7	3	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, \phi_1 \mid s_0 = 10^{-3},\ q = \infty,\ \eta = 0,\ \phi_2 = 0)$	1416.6	0.593
8	7	$\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p \mid s_0 = 10^{-3}, q = \infty,\ \eta = 0,\ \phi_1 = 0,\ \phi_2 = 0)$	-150.62	0.076
9	10	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \beta, \xi, \phi_1, \phi_2 \mid s_0 = 10^{-3})$	1312.3	0.224
10	14	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \beta, \xi, \phi_1 \mid s_0 = 10^{-3},\ \phi_2 = 0)$	1312.5	0.219
11	12	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3},\ \beta = 0,\ \phi_2 = 0)$	671.42	0.074
12	11	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \beta, \phi_1 \mid s_0 = 10^{-3},\ \xi = 1,\ \phi_2 = 0)$	1280.4	0.112
13	13	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \beta, \xi \mid s_0 = 10^{-3},\ \phi_1 = 0,\ \phi_2 = 0)$	84.756	0.107
14	15	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \beta \mid s_0 = 10^{-3},\ \xi = 1,\ \phi_1 = 0,\ \phi_2 = 0)$	-150.66	0.081
15	16	$\mathcal{L}_n^{\mathrm{g+}}(\boldsymbol{\theta}, \xi \mid s_0 = 10^{-3},\ \beta = 0,\ \phi_1 = 0,\ \phi_2 = 0)$	-67.539	0.047
16	18	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \nu, \xi, \phi_1, \phi_2 \mid s_0 = 10^{-3})$	1424.2	0.136
17	22	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \nu, \xi, \phi_1 \mid s_0 = 10^{-3},\ \phi_2 = 0)$	1424.2	0.166
18	20	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3},\ \nu = n - d,\ \phi_2 = 0)$	671.42	0.074
19	19	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \nu, \phi_1 \mid s_0 = 10^{-3},\ \xi = 1,\ \phi_2 = 0)$	1410.6	0.157
20	21	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \nu, \xi \mid s_0 = 10^{-3},\ \phi_1 = 0,\ \phi_2 = 0)$	102.09	0.057
21	23	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \nu \mid s_0 = 10^{-3},\ \xi = 1,\ \phi_1 = 0,\ \phi_2 = 0)$	-92.299	0.096
22	24	$\mathcal{L}_n^{\mathrm{s}}(\boldsymbol{\theta}, \xi \mid s_0 = 10^{-3},\ \nu = n - d,\ \phi_1 = 0,\ \phi_2 = 0)$	-67.535	0.048
23	34	$\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta},\ \phi_1 \mid s_0 = 10^{-3}, \phi_2 = 0)$	635.70	0.064

1. None of the tabulated likelihood functions yield a unit k-value. This only confirms that HYMOD is misspecified and testifies to the need of a sandwich description for the posterior parameter distribution.

2. Distribution-adaptive likelihoods deliver on their promise and help reduce information bias when confronted with structural model errors and/or errors in the controlling variables. Most of the tabulated formulations of the UL, GL$^+$ and SL improve upon the value of $k = 0.064$ for the NL #23 with exception of formulations #4, #15 and #20, #22. These "outlier" likelihoods yield `omnibus` scalars smaller than $k = 0.064$ and share in common their lack of an AR(1)-treatment of the studentized discharge residuals. UL yields, on average, the largest k values followed by GL$^+$ and SL.

3. Formulation #7 of UL, $\mathcal{L}_n^{\mathrm{u}}(\boldsymbol{\theta}, p, \phi_1 \mid s_0 = 10^{-3},\ q = \infty,\ \eta = 0,\ \phi_2 = 0)$, maximizes the omnibus scalar. Its value of $k = 0.593$ is largest among all listed likelihood functions and is a significant improvement over the value of $k = 0.064$ derived from NL.

4. The omnibus adjustment k and log-likelihood maximum $\mathcal{L}_n(\{\widehat{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\delta}}_n\})$ do not exhibit a simple 1:1 relationship. The `omnibus` scalar k is positively associated with the MAP log-likelihood (Pearson's correlation coefficient $r = 0.64$), but larger $\mathcal{L}_n(\{\widehat{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\delta}}_n\})$ does not necessarily imply lower information bias or a better-specified model. For example, UL formulations #6 and #7 have nearly equal MAP log-likelihoods but yield markedly different k values. Hence, the log-likelihood alone is not a sufficient criterion for model adequacy.

5. The omnibus scalar k provides a simple, interpretable diagnostic of misspecification. A unit-optimal measure of conflict between the sensitivity $\widehat{\mathbf{A}}_n$ and variability $\widehat{\mathbf{B}}_n$ matrices that can inform model ranking and selection. However, because k is an *isotropic* adjustment, it is reliable only when the sandwich distribution is well approximated by a multivariate Gaussian, that is, when the posterior is nearly symmetric and locally quadratic around $\widehat{\boldsymbol{\theta}}_n$. Preserving local geometry and asymmetry typically requires a dimension-wise scaling, especially in small sample sizes [160].

The tabulated results for certain likelihoods should not give a false sense of HYMOD adequacy. Modeling the studentized discharge residuals with an AR(1) process tends to inflate k by attenuating serial correlation. By contrast, likelihoods that do not model residual correlation (ϕ_1, $\phi_2 = 0$: #4, #8, #13, #14, #15, #20, #21 and #22) yield substantially lower values of k on average. Consequently, enforcing unit variance for the studentized residuals leads to slopes s_1 that are markedly larger than those suggested by nonparametric measurement error estimators [37]. There is nothing unusual here: it is common practice to replace the measurement error variance σ_ϵ^2 in Equation (40b) with the sample residual variance s_e^2 when seeking well-calibrated parameter credible regions. Lastly, the tabulated k-values assume linearity of HYMOD in the neighborhood surrounding the ML or MAP solution. The validity of this assumption has been verified in Figs. 10 and 11, nevertheless, we should develop a nonlinear equivalent of the `omnibus` scalar.

6.2 A Formal Measure for the Degree of Model Misspecification

The misalignment of the naive and sandwich variance estimators can be formalized as a measure of model misspecification. The Kullback and Leibler [95] divergence between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^d$ with densities p_1 and p_2 (w.r.t. Lebesgue measure) is

$$\begin{aligned} d_{\text{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) &= \int p_1(\omega) \log\left(\frac{p_1(\omega)}{p_2(\omega)}\right) \mathrm{d}\omega \\ &= \mathbb{E}_{\mathbb{P}_1}\big[\log\big(p_1(\Omega)\big) - \log\big(p_2(\Omega)\big)\big]. \end{aligned}$$

The KL divergence is a statistical distance and not a metric. It is strictly positive and zero only iff $\mathbb{P}_1 = \mathbb{P}_2$, is asymmetric, $d_{\text{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) \neq d_{\text{KL}}(\mathbb{P}_2 \,\|\, \mathbb{P}_1)$, and violates the triangle inequality.

Let $\mathbb{P}_{\text{naive}} = \mathcal{N}_d\big(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}\big)$ and $\mathbb{P}_{\text{sand}} = \mathcal{N}_d\big(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}\big)$, be two d-variate Normal measures with a common mean $\widehat{\boldsymbol{\theta}}_n$ but different covariances. Appendix B of Vrugt [157] demonstrates

$$d_{\text{KL}}\big(\mathbb{P}_{\text{naive}} \,\|\, \mathbb{P}_{\text{sand}}\big) = \tfrac{1}{2}\big[\log\big(\big|\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}\big|\big) - \log\big(\big|\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}\big|\big) + \text{tr}\,\big((\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}})^{-1}\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}\big) - d\big].$$

This equals the multivariate divergence score of Dawid and Sebastiani [35] for equal means. Substituting $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}$ gives a convenient form in terms of the bread and meat matrices

$$d_{\text{XX}}\big(\mathbb{P}_{\text{naive}} \,\|\, \mathbb{P}_{\text{sand}}\big) = \tfrac{1}{2}\big[\log\big(\big|\widehat{\boldsymbol{\beta}}_n\widehat{\mathbf{A}}_n^{-1}\big|\big) + \text{tr}\,\big(\widehat{\mathbf{A}}_n\widehat{\boldsymbol{\beta}}_n^{-1}\big) - d\big]. \tag{49}$$

This strictly proper divergence is nonnegative and equals zero iff $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}} = \widehat{\boldsymbol{\Sigma}}_n^{\text{sand}}$ (equivalently, $\widehat{\mathbf{A}}_n = \widehat{\boldsymbol{\beta}}_n$). The larger the discord between the bread and meat matrices, the larger d_{XX}. This provides a convenient *information-bias* score, especially practical in machine-learning workflows where $\widehat{\mathbf{A}}_n$ and the *score*-based $\widehat{\boldsymbol{\beta}}_n$ arise naturally via automatic differentiation. We deliberately use the neutral subscript XX as a placeholder and leave formal nomenclature to future users and the research community. A symmetric alternative is the Jeffreys divergence $\frac{1}{2}\{d_{\text{XX}}(\mathbb{P}_{\text{naive}} \,\|\, \mathbb{P}_{\text{sand}}) + d_{\text{XX}}(\mathbb{P}_{\text{sand}} \,\|\, \mathbb{P}_{\text{naive}})\}$.

7 Conclusions

This paper explored the consequences of model misspecification on data informativeness of a streamflow record $\boldsymbol{\omega}_n = (\omega_1, \ldots, \omega_n)^\top$ for the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ of a vector-valued watershed model $\mathbf{y}_n = \boldsymbol{\mathcal{M}}(\boldsymbol{\theta}; \cdot)$ and associated confidence intervals and regions of $\boldsymbol{\theta}$. We reviewed basic theory of frequentist inference, specifically ML and M-estimation, and addressed the implications of model misspecification on the expected Fisher $\mathcal{I}_n$ and Godambe $\mathcal{G}_n$ information, the two currencies of data informativeness in ML theory, and the asymptotic distribution of the model parameters. The conclusions and insights developed here from the theoretical treatise should be familiar to most statisticians, yet they have not entered mainstream practice, particularly in Bayesian estimation. This is both remarkable and puzzling, given the far-reaching implications of the sandwich variance estimator for hydrologic model training and uncertainty quantification. The main consequences of M-estimation theory, as illustrated here, may be summarized as follows

1. Under the *true* but unknown distribution Q of the data-generating process, the ML (or MAP with a regular, locally flat prior) estimator $\widehat{\boldsymbol{\theta}}_n$ is consistent and converges in probability to the parameter values $\boldsymbol{\theta}_0$ of the data-generating process. Then, sample means of the sensitivity matrix $\widehat{\mathbf{A}}_n = -\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)$ and variability matrix $\widehat{\mathbf{B}}_n = \frac{1}{n}\sum_{t=1}^{n}\{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\omega_t}(\widehat{\boldsymbol{\theta}}_n)\nabla^\top_{\boldsymbol{\theta}}\mathcal{L}_{\omega_t}(\widehat{\boldsymbol{\theta}}_n)\}$ of ML/MAP estimator $\widehat{\boldsymbol{\theta}}_n$ will suggest an approximately equal observed Fisher information $\widehat{\mathcal{I}}_n = n\,\widehat{\mathbf{A}}_n = n\,\widehat{\mathbf{B}}_n$ for data $\omega_1, \ldots, \omega_n$. The parameter covariance matrix $\text{Var}(\widehat{\boldsymbol{\theta}}_n) \equiv \widehat{\boldsymbol{\Sigma}}_n$ simplifies to $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{I}}_n^{-1}$, which corresponds to the inverse of the Fisher information. This ML theory is widely practiced in hydrologic modeling, but under structural error (model misspecification) it yields an *erroneous* (naive) description of parameter uncertainty.

2. Under misspecification, the *true* parameter values $\boldsymbol{\theta}_0$ of the data-generating process are not in the model parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$. The best attainable values of the parameters or so-called *pseudo-true* parameter values $\boldsymbol{\theta}_*$ minimize the Kullback and Leibler [95] divergence between the *true* probability density function $q_\Omega(\omega; \boldsymbol{\theta}_0)$ of random variable Ω of interest and the incorrect family of densities $f(\omega; \boldsymbol{\theta})$ defined by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The ML (or MAP with a regular, locally flat prior) estimator $\widehat{\boldsymbol{\theta}}_*$ converges in probability to the *pseudo-true* parameter values $\boldsymbol{\theta}_*$.

3. Under model misspecification, the second Bartlett condition typically does not hold, that is, $\widehat{\mathbf{A}}_n \neq \widehat{\mathbf{B}}_n$. Then, the variance-covariance matrix of the ML or MAP parameter estimates $\widehat{\boldsymbol{\theta}}_*$ is $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{G}}_n^{-1}$, which corresponds to the inverse of the Godambe information. This *sandwich* (co)variance matrix is a metaphor for a *ham* or *meat* matrix $\widehat{\mathbf{B}}_n$ between two *bread* matrices $\widehat{\mathbf{A}}_n$ and delivers an asymptotically valid description of ML/MAP parameter uncertainty even when the likelihood is misspecified. The square roots of the diagonal elements of $\widehat{\boldsymbol{\Sigma}}_n$ are known as “robust standard errors” or “Eicker-Huber-White standard errors”.

4. When the entries of the successive *score* vectors $\nabla\mathcal{L}_{\omega_1}(\widehat{\boldsymbol{\theta}}_n),\ldots,\nabla\mathcal{L}_{\omega_n}(\widehat{\boldsymbol{\theta}}_n)$ exhibit serial dependence, matrix $\widehat{\mathbf{B}}_n$ must be replaced by its long-run variance $\widehat{\boldsymbol{\beta}}_n$, which can be estimated using a so-called heteroskedasticity- and autocorrelation-consistent or HAC estimator. Under independent and identically distributed residuals, $\widehat{\boldsymbol{\beta}}_n = \widehat{\mathbf{B}}_n$.

5. It is incorrect to claim that hydrologic parameter uncertainty can be inflated by raising $L_n(\boldsymbol{\theta})$ to an arbitrary power $0 < \lambda < 1$, which yields the Fisher-based (naive) variance $\widehat{\boldsymbol{\Sigma}}_n^{\text{naive}}(\lambda) = \frac{1}{n}\lambda^{-1}\,\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{I}}_n(\lambda)^{-1}$. Under misspecification, this merely rescales the *wrong* information. Tempering via the power likelihood $L_n^{\text{p}}(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta})^{\lambda}$, equivalently $\mathcal{L}_n^{\text{p}}(\boldsymbol{\theta}) = \lambda\,\mathcal{L}_n(\boldsymbol{\theta})$, implies $\widehat{\mathbf{A}}_n(\lambda) = \lambda\,\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n(\lambda) = \lambda^2\,\widehat{\mathbf{B}}_n$, so the asymptotically correct sandwich variance-covariance matrix, $\widehat{\boldsymbol{\Sigma}}_n^{\text{sand}} = \frac{1}{n}\lambda^{-1}\widehat{\mathbf{A}}_n^{-1}\,\lambda^2\,\widehat{\mathbf{B}}_n\,\lambda^{-1}\widehat{\mathbf{A}}_n^{-1} = \widehat{\mathcal{G}}_n^{-1}$ is *invariant* to λ. Consequently, credible intervals constructed from $\widehat{\mathcal{I}}_n(\lambda)^{-1}$ (posterior curvature) need not attain their nominal frequentist coverage under misspecification. The correct coverage requires the sandwich covariance.

6. Markov chain Monte Carlo (MCMC) sampling methods that target an unadjusted likelihood combined with flat or weakly informative priors provide a naive characterization of parameter uncertainty under model misspecification. In this common setting, the asymptotic covariance matrix $\boldsymbol{\Sigma}_n$ of the resulting posterior distribution is governed by the inverse curvature $\widehat{\mathbf{A}}_n^{-1}$ rather than the asymptotically valid sandwich matrix $\frac{1}{n}\mathbf{A}^{-1}n\mathbf{B}n\mathbf{A}_n^{-1}$. This finding calls into question the calibration of many published posterior *credible regions* in hydrology including our own when likelihood misspecification is present, as curvature-based (naive) variances generally underestimate uncertainty.

The implications of the sandwich variance estimator were demonstrated in three case studies of increasing complexity involving the modeling of soil water infiltration, watershed hydrologic fluxes and the rainfall-discharge transformation. The mathematical-physical description of the first two studies facilitated an analytic demonstration of the consequences of model misspecification on parameter uncertainty estimates derived from ML and Bayesian methods. The third and last study illustrated the sandwich variance estimator by application to measured streamflow data using the class of distribution-adaptive likelihood functions described by Vrugt et al. [166]. Our analytic and numerical results confirm the earlier take home messages from ML theory and can be summarized as follows

1. The maximum a-posteriori (MAP) sensitivity $\widehat{\mathbf{A}}_n$ and variability $\widehat{\mathbf{B}}_n$ matrices provide conflicting information about streamflow data informativeness in the face of model structural and/or input data errors. These differences will be small when model misspecification is benign and much larger when epistemic errors are significant. This disagreement of the bread and meat matrices gives rise to the sandwich variance.

2. The naive variance estimator $\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}$ was shown to provide overly tight confidence regions and intervals of the watershed model parameters. The culprit is the sensitivity matrix $\widehat{\mathbf{A}}_n$, which overestimates the Fisher information of the streamflow data. This is synonymous with overly optimistic estimates of model parameter and predictive uncertainty in a Bayesian context and poorly calibrated credible regions that do not correspond with frequentist coverage probabilities.

3. The sandwich variance estimator inflates considerably the confidence intervals and regions of the watershed model parameters. Sandwich intervals are wider (often substantially) and achieve better frequentist coverage for hydrologic parameters and predictions than naive (Fisher-based) intervals. So, the consequences of model misspecification are that the watershed model parameters are much less well defined than previously thought

as articulated by the sensitivity (Hessian) matrix and/or MCMC-sampled posterior parameter distribution.

4. Open-face sandwich adjustment of the naive posterior samples and post-hoc magnitude adjustments of the likelihood function $L_n(\boldsymbol{\theta})$ assist MCMC methods in recovering the asymptotically valid sandwich distribution. These fixes lack a solid theoretical foundation, but are critical for a robust quantification of hydrologic model uncertainty pending more rigorous advances to MCMC theory and the Metropolis-Hastings algorithm.

5. Distribution-adaptive likelihood functions deliver on their premise and offer some protection against model misspecification and/or imperfect knowledge of the controlling variables by harmonizing the MAP sensitivity and variability matrices. This reduces information bias.

6. Discrepancies between the sensitivity and variability matrices quantify "information bias." We formalized this via a misalignment score $d_{\mathrm{XX}}(P\|F)$ that measures the concordance of matrices $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$. This strictly proper divergence is invariant to smooth one-to-one reparameterizations (congruence transforms) and to common scaling.

Then, elastic stretching of the likelihood function by application of an arbitrary power λ to $L_n(\boldsymbol{\theta})$ is a widely practiced technique in the GLUE method to suppress problems of overconditioning and artificially inflate parameter uncertainty in the presence of epistemic and input data errors. The rationale is that such annealing counteracts the chronic effects of flawed residual assumptions and model inadequacy on the posterior distribution. While this approach is supported in theory by the naive variance estimator it rests on a fundamental misconception. Under model misspecification, the true currency of data informativeness is not $n\,\widehat{\mathbf{A}}_n$ (so-called observed information), but the observed Godambe information $\widehat{\mathcal{G}}_n = n\,\widehat{\mathbf{A}}_n\widehat{\mathbf{B}}_n^{-1}\widehat{\mathbf{A}}_n$, which accounts for both sensitivity and variability. Crucially, the Godambe information is invariant to the exponent λ in the power likelihood function $L_n^{\lambda}(\boldsymbol{\theta})$. As such, stretching the likelihood has no lasting effect on the posterior parameter distribution. Given this, the Godambe information should be the principal object of concern in GLUE-type methodologies and will help to produce estimates of uncertainty that are more consistent with model and data limitations. In fact, the sandwich variance estimator already offers a principled and statistically coherent route toward achieving this goal.

8 Limitations and Outlook

By now it should be evident that the sandwich estimator comes at a cost. First, common computational methods must be adjusted to deliver the asymptotically valid sandwich distribution under misspecification. This increases the computational burden and algorithmic complexity. Second, the sandwich variance is typically larger than the naive Fisher-based variance. This efficiency trade-off (sandwich vs. Fisher) is well known in the statistical literature [24, 46, 87, 173].

A limitation of this study is the use of uniform priors for the parameters of the soil-water infiltration functions and the ABC and HYMOD models. This choice enables a clean, apples-to-apples comparison between sandwich-based uncertainty quantification from frequentist and Bayesian analyses. An informative prior could, however, further accentuate discrepancies between the naive (Fisher-based) and sandwich covariance estimators.

Both naive and sandwich variances require the sensitivity matrix $\widehat{\mathbf{A}}_n$ and the variability matrix $\widehat{\mathbf{B}}_n$, evaluated at the ML/MAP estimate $\widehat{\boldsymbol{\theta}}_n$. The variability matrix, $\widehat{\mathbf{B}}_n = \frac{1}{n}\sum_{t=1}^{n}\nabla\mathcal{L}_{\omega_t}(\widehat{\boldsymbol{\theta}}_n)\nabla^{\top}\mathcal{L}_{\omega_t}(\widehat{\boldsymbol{\theta}}_n)$, is a sum of outer products of *score* vectors and is therefore positive semidefinite (typically positive definite under regularity). By contrast, the sensitivity matrix $\widehat{\mathbf{A}}_n = -\frac{1}{n}\nabla^2\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)$ involves second derivatives, which are more delicate numerically, especially when parameters lie near bounds, and may be ill-conditioned or non-positive definite in finite samples. Since invertibility of $\widehat{\mathbf{A}}_n$ is required for both naive and sandwich estimators, common remedies include model re-specification or collecting additional data. Fixing selected parameters can restore invertibility but reduces flexibility and risks bias if the fixed values are misspecified [98]. A practical alternative, requiring little additional computation, is to estimate $\widehat{\mathbf{A}}_n$ from the naive posterior. Under standard regularity, the naive posterior is locally normal around $\widehat{\boldsymbol{\theta}}_n$ (or $\widehat{\boldsymbol{\theta}}_*$ under misspecification) with covariance $\frac{1}{n}\mathbf{A}_n^{-1}$ (or $\frac{1}{n}\mathbf{A}_*^{-1}$). Hence, the inverse of the post-burn-in posterior samples provides a plug-in estimate of the bread matrix [134].

While reporting robust variances for an approximate model may be debated, robust inference remains essential for hypothesis testing, parameter estimation, and uncertainty quantification. To this end, the `omnibus` scalar k and the strictly proper misalignment score proposed here offer simple, interpretable measures of misspecification. The misalignment score supports a comparison across models with different number of parameters and helps guide model selection. In a companion paper, Vrugt and Diks [160] provide an information-theoretic interpretation of the misalignment score and introduce several other scalar diagnostics of misspecification. That same paper also presents a new recipe for sandwich-adjusted MCMC simulation.

Acknowledgments

We are grateful to the three anonymous referees for their constructive comments, which ultimately led to substantial improvements in the manuscript. The first author also acknowledges stimulating discussions with Prof. Keith Beven on model structural errors and their implications, and thanks Prof. Rich Vogel for pointing out the analytic value of the ABC model of Fiering [50] in demonstrating the consequences of model misspecification.

Data and Software Availability

The theory, methodology, and case studies presented in this paper are part of DREAM-Suite, a MATLAB-Python software package for Bayesian model training, evaluation, and diagnostics [156]. This software is available at https://github.com/jaspervrugt/dream-suite

Author Contributions

Conceptualization, J.V. and C.D.
Methodology, J.V., C.D. and P.G.
Software, J.V. and R.D.
Validation, J.V. and C.D.
Formal analysis, J.V. and C.D.
Investigation, J.V., C.D., R.D. and P.G.
Resources, J.V.
Data curation, J.V.
Writing—original draft preparation, J.V.
Writing—review and editing, J.V., C.D. and R.D.
Visualization, J.V.

Appendix A: M-Estimation: Loss, Influence & Weight Functions and Robust Efficiency

This Appendix provides a concise primer on M-estimation. We introduce main theory, illustrate the general features of an influence function and then tabulate, visualize, and discuss the loss ρ, influence ψ, and weight w functions of thirteen commonly used M-estimators. Last but not least, we suggest a robust alternative to the Nash and Sutcliffe [110] efficiency for hydrologic model training. We hope this primer encourages readers to explore M-estimation in greater depth.

A.1 Main Theory

Following Huber [80], an M-estimator is the solution to the following minimization problem

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^{n} \rho\big(\underline{e}_t(\boldsymbol{\theta})\big),$$

where $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,d})^\top$ is the d-vector of optimal parameters and the loss function $\rho(\cdot)$ admits standardized residuals

$$\underline{e}_t(\boldsymbol{\theta}) = \frac{\omega_t - y_t(\boldsymbol{\theta})}{\kappa\,\sigma_n}, \qquad t = 1, \ldots, n,$$

with $\kappa > 0$ a fixed constant and $\sigma_n > 0$ treated as a consistent estimator of the scale of the underlying distribution of the data. The denominator $S = \kappa\,\sigma_n$ renders the loss invariant to measurement units and controls M-estimator efficiency. A common choice of standardization is the median absolute deviation from the median [54].

If the function $\rho(\underline{e})$ is differentiable then the M-estimator of $\boldsymbol{\theta}$ is the solution of the following d estimating equations

$$\mathbf{g}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \nabla_{\boldsymbol{\theta}} \rho\big(\underline{e}_t(\boldsymbol{\theta})\big) = \mathbf{0}_d. \tag{A.1}$$

The ML method is a prototypical M-estimator, since the optimum solution $\widehat{\boldsymbol{\theta}}_n$ is a root (zero-point) of the $d \times 1$ *score* $\mathbf{g}_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$, where $\mathcal{L}_n(\boldsymbol{\theta})$ is the log-likelihood of $\boldsymbol{\theta}$ given data $\omega_1, \ldots, \omega_n$. We can use the chain rule to examine the jth element of the *score*

$$\begin{aligned} g_j(\boldsymbol{\theta}) &= \sum_{t=1}^{n} \frac{\partial \rho\big(\underline{e}_t(\boldsymbol{\theta})\big)}{\partial \underline{e}_t(\boldsymbol{\theta})} \frac{\partial \underline{e}_t(\boldsymbol{\theta})}{\partial \theta_j} \\ &= \sum_{t=1}^{n} \psi\big(\underline{e}_t(\boldsymbol{\theta})\big) \frac{\partial \underline{e}_t(\boldsymbol{\theta})}{\partial \theta_j}, \qquad j = 1, \ldots, d, \end{aligned} \tag{A.2}$$

where the first derivative $\psi(\underline{e}) = \mathrm{d}\rho(\underline{e})/\mathrm{d}\underline{e}$ is the influence function of Hampel [66]. This function, whose general features are illustrated in Figure A.1, measures the effect of a standardized residual on the parameter estimates and is often of greater interest than $\rho(\underline{e})$ itself [67, 68, 81].

Properly chosen ρ and ψ yield estimators less sensitive to outliers, heavy tails, and modest departures from distributional assumptions than least squares or ML methods. A robust estimator should have a bounded influence function so that large residuals (spurious data) do not corrupt the estimates $\widehat{\boldsymbol{\theta}}_n$. If we define the *weight function* $w(\underline{e}) = \psi(\underline{e})/\underline{e}$, then Equation (A.2) can be written as

$$g_j(\boldsymbol{\theta}) = \sum_{t=1}^{n} w(\underline{e}_t)\underline{e}_t \frac{\partial \underline{e}_t}{\partial \theta_j},$$

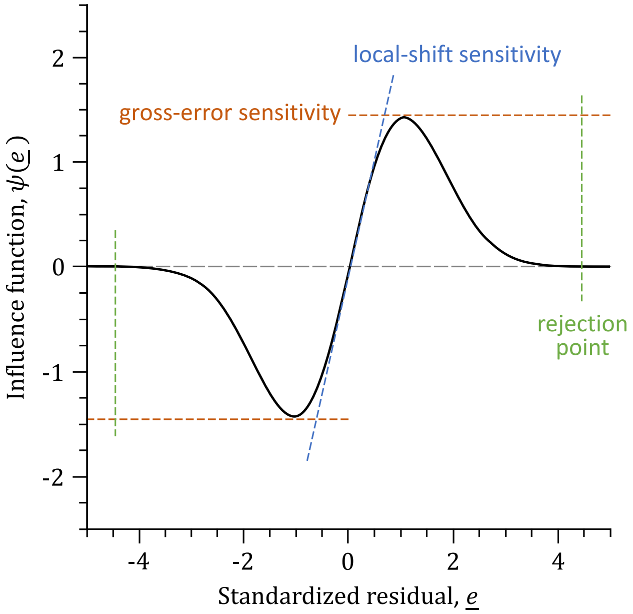

Figure A.1: Influence function $\psi(\underline{e})$ of a robust M-estimator. The rejection point is defined as the point where the influence function attains a zero value. Standardized residuals whose magnitude exceeds the rejection point do not impact the parameter estimates. The gross-error sensitivity is equal to the maximum value of the influence function and measures the largest impact a spurious residual can have on the estimator [75]. The local-shift sensitivity is equal to $\psi'(\underline{e})$ the derivative of the influence function at the respective standardized residual, $\underline{e}$.

which is reminiscent of iteratively reweighted least squares. Finally, differentiating the influence function yields $\psi'(\underline{e}) = \mathrm{d}\psi(\underline{e})/\mathrm{d}\underline{e}$, which corresponds to the second derivative $\partial^2\rho(\underline{e})/\partial\underline{e}^2$, that is, the Hessian of the loss function.

Table A.1 presents mathematical expressions of the loss, influence and weight functions of 13 commonly used M-estimators. These loss functions share several common properties: they are nonnegative, continuous, infinitely differentiable, symmetric and monotonically increasing away from their minimum at $\underline{e} = 0$. Strict convexity, however, is not a necessary requirement for successful application of an M-estimator [36]. All tabulated M-estimators, except for the ML and least squares estimators, are considered robust, even if their influence functions do not have a null (rejection) point but return a non-zero value instead for large $|\underline{e}|$.

Chen and Yin [32] classify M-estimators into three groups including those with influence functions that (i) strictly increase away from the origin with magnitude of the standardized residual (e.g. Least absolute, Least power, Fair [48], Charbonnier et al. [31] and Huber [78]), (ii) redescend soft (asymptotically) to zero with increasing value of $|\underline{e}|$ (e.g. Poisson [122]) and (iii) redescend hard to zero at some finite value of $|e|$ (e.g. Smith [139], Beaton and Tukey [10] and Andrews et al. [4]). This third and last class of M-estimators is particularly robust to outliers, yet remains largely unknown and seldom used in the hydrologic literature.

Figure A.2 plots the loss ρ, influence ψ, and weight w functions for the M-estimators in Table A.1. To maintain legibility, we display only a representative subset of the estimators, labeled 1, 2, 3, 4, 6, 7, 9, 10, 11, and 13. The graphs clarify the behavior of the various M-estimators. Their loss functions $\rho(\underline{e})$ are nonnegative, symmetric in $\underline{e}$ (even), and satisfy $\rho(0) = 0$. For convex choices (e.g., least squares (2), and Huber (6) on $|\underline{e}| \leq c$) the loss is non-decreasing in $|\underline{e}|$. For redescending losses (e.g., Welsch (9), Tukey (10), Andrews (12) and Hampel (13)) the loss increases on $[0, c]$ and then plateaus (becomes bounded). The

Table A.1: Loss $\rho(\underline{e})$, influence $\psi(\underline{e}) = \mathrm{d}\rho(\underline{e})/\mathrm{d}\underline{e}$, and weight $w(\underline{e}) = \psi(\underline{e})/\underline{e}$ functions for commonly used M-estimators. $\mathcal{L}^{\mathrm{n}}_{\omega}(\boldsymbol{\theta})$ is the normal log-likelihood for a single datum. Scalars ν, c and k are tuning constants.

#	Cond.	Loss: $\rho(\underline{e})$	Infl: $\psi(\underline{e})$	Wght: $w(\underline{e})$	Notes
1.		$\lvert\underline{e}\rvert$	$\mathrm{sign}(\underline{e})$	$1/\lvert\underline{e}\rvert$	Least absolute
2.		$\frac{1}{2}\underline{e}^2$	$\underline{e}$	1	Least squares
3.		$\lvert\underline{e}\rvert^{\nu}/\nu$	$\mathrm{sign}(\underline{e})\lvert\underline{e}\rvert^{\nu-1}$	$\lvert\underline{e}\rvert^{\nu-2}$	Least power, $\nu = 1.2$
4.		$2(1+\frac{1}{2}\underline{e}^2)^{\frac{1}{2}} - 2$	$\underline{e}/(1+\frac{1}{2}\underline{e}^2)^{\frac{1}{2}}$	$1/(1+\frac{1}{2}\underline{e}^2)^{\frac{1}{2}}$	Charbonnier et al. [31]
5.		$c\lvert\underline{e}\rvert - c^2\log(1+\lvert\underline{e}\rvert/c)$	$c\underline{e}/(c+\lvert\underline{e}\rvert)$	$c/(c+\lvert\underline{e}\rvert)$	Fair [48], $c = 1.3998$ Özyurt and Pike [116]
6.	$\lvert\underline{e}\rvert \leq c$	$\frac{1}{2}\underline{e}^2$	$\underline{e}$	1	Huber [78], $c = 1.345$
	$\lvert\underline{e}\rvert > c$	$c(\lvert\underline{e}\rvert - \frac{1}{2}c)$	$c\,\mathrm{sign}(\underline{e})$	$c/\lvert\underline{e}\rvert$	Holland and Welsch [76]
7.		$\frac{1}{2}c^2\log(1+\underline{e}^2/c^2)$	$c^2\underline{e}/(\underline{e}^2+c^2)$	$c^2/(\underline{e}^2+c^2)$	Poisson [122], $c = 2.3849$ apud Stigler [142]
8.		$\underline{e}^2/(2+2\underline{e}^2)$	$\underline{e}/(1+\underline{e}^2)^2$	$1/(1+\underline{e}^2)^2$	Geman and McClure [57]
9.		$\frac{1}{2}c^2[1-\exp(-\underline{e}^2/c^2)]$	$\underline{e}\exp(-\underline{e}^2/c^2)$	$\exp(-\underline{e}^2/c^2)$	Dennis and Welsch [39] Rey [126], $c = 2.9846$
10.	$\lvert\underline{e}\rvert \leq c$	$\frac{1}{6}c^2[1-(1-\underline{e}^2/c^2)^3]$	$\underline{e}(1-\underline{e}^2/c^2)^2$	$(1-\underline{e}^2/c^2)^2$	Beaton and Tukey [10]
	$\lvert\underline{e}\rvert > c$	$\frac{1}{6}c^2$	0	0	Pennacchi [120], $c = 4.6851$
11.	$\lvert\underline{e}\rvert \leq c$	$\frac{1}{4}c^2[1-(1-\underline{e}^2/c^2)^2]$	$\underline{e}(1-\underline{e}^2/c^2)$	$1-\underline{e}^2/c^2$	Smith [139], $c = 3.6732$
	$\lvert\underline{e}\rvert > c$	$\frac{1}{4}c^2$	0	0	de Menezes et al. [36]
12.	$\lvert\underline{e}\rvert \leq \pi c$	$c^2[1-\cos(\underline{e}/c)]$	$c\sin(\underline{e}/c)$	$c\sin(\underline{e}/c)/\underline{e}$	Andrews et al. [4], $c = 1.338$
	$\lvert\underline{e}\rvert > \pi c$	$2c^2$	0	0	Holland and Welsch [76]
13.	$\lvert\underline{e}\rvert \leq a$	$\frac{1}{2}\underline{e}^2$	$\underline{e}$	1	Hampel [66, 69], $k = 0.902$
	$\lvert\underline{e}\rvert \in (a,b]$	$\frac{1}{2}a^2 + a(\lvert\underline{e}\rvert - a)$	$a\,\mathrm{sign}(\underline{e})$	$a/\lvert\underline{e}\rvert$	$f(\underline{e}) = b-a+\lvert\underline{e}\rvert+g(\underline{e})(\lvert\underline{e}\rvert-b)$
	$\lvert\underline{e}\rvert \in (b,c]$	$\frac{1}{2}af(\underline{e})$	$a\,\mathrm{sign}(\underline{e})g(\underline{e})$	$a\,g(\underline{e})/\lvert\underline{e}\rvert$	$g(\underline{e}) = (c-\lvert\underline{e}\rvert)/(c-b)$
	$\lvert\underline{e}\rvert > c$	$\frac{1}{2}a(b-a+c)$	0	0	$a = \frac{3}{2}k$, $b = \frac{7}{2}k$ and $c = 8k$
14.		$\frac{1}{2}\underline{e}^2$	$\underline{e}$	1	$-\mathcal{L}^{\mathrm{n}}_{\omega}(\boldsymbol{\theta})$; $\underline{e} = e/\sigma$

When the weight function $w(\underline{e}) = \psi(\underline{e})/\underline{e}$ has a *removable* singularity at $\underline{e} = 0$, define $w(0)$ by continuity (e.g., 12 yields $w(0) = 1$). For *non-removable* singularities (e.g., 1), the limit diverges and we must use an ε-safeguard, e.g., $w(\underline{e}) = 1/\max(\lvert\underline{e}\rvert, \varepsilon)$. For piecewise forms such as Hampel (13), the branch active at $\underline{e} = 0$ gives $w(0) = 1$ directly.

smoothness of the functions varies. The least squares estimator is C^{∞} (derivatives of all orders exist and are continuous), and the least absolute (difference) estimator is not differentiable at 0. The Huber (6) loss function is continuously differentiable once (C^1) but not twice at $\lvert\underline{e}\rvert = c$. Redescending functions are piecewise-smooth with kinks at their tuning points. The quadratic loss corresponding to the Normal negative log-likelihood has $\rho(\underline{e}) = \frac{1}{2}\underline{e}^2$ for standardized residuals (i.e., $\sigma_{\underline{e}}^2 = 1$), and therefore also satisfies $\rho(0) = 0$. Following Hoaglin et al. [75, p. 366], tuning constants are typically chosen so that this normalization is met across estimators.

The various M-estimators share a common structure but differ markedly in their handling of large (standardized) residuals, as revealed by their influence $\psi(\underline{e})$ and weight $w(\underline{e})$ functions. Least squares and least-power (L_p, $p > 1$) estimators are not robust as their influence functions grow without bound as $\lvert\underline{e}\rvert$ increases. As a result, the parameter estimates derived from these

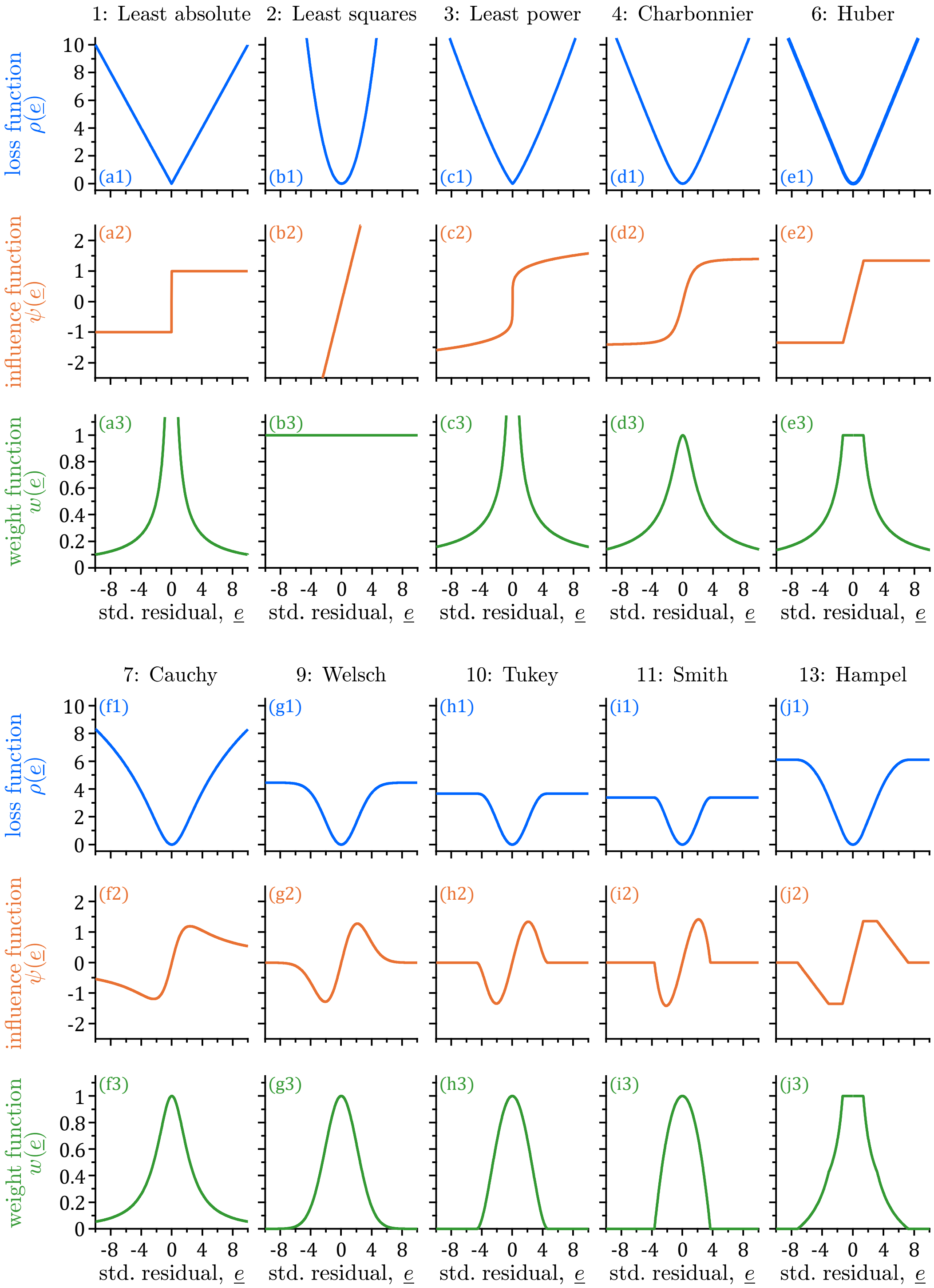

Figure A.2: Overview of M-estimators: Loss function, $\rho(\underline{e})$, influence function, $\psi(\underline{e})$, and weight function, $w(\underline{e})$, of ten of the M-estimators of Table A.1.

loss functions are sensitive to outliers. The Huber [78] estimator has bounded influence

$$\psi(\underline{e}) = \begin{cases} \underline{e} & \text{if } |\underline{e}| \leq c \\ c\,\mathrm{sign}(\underline{e}) & \text{otherwise,} \end{cases}$$

so $|\psi(\underline{e})| \leq c$, but it is not redescending. Indeed, Huber (6) clips at $\pm c$ rather than return to 0 as $|\underline{e}| \to \infty$. This bounded influence provides robustness without discarding very large residuals entirely. Equivalently

$$w(\underline{e}) = \frac{\psi(\underline{e})}{\underline{e}} = \begin{cases} 1 & \text{if } |\underline{e}| \leq c \\ c/|\underline{e}| & \text{otherwise.} \end{cases}$$

By contrast, the Welsch (9), Tukey (10), Smith (11), and Hampel (13) estimators are *redescending*. Their influence functions either decay smoothly to zero (Welsch: soft-redescending) or attain zero at a cutoff (Tukey, Smith, Hampel: hard-redescending). These M-estimators ignore large residuals altogether.

M-estimators have received little attention in hydrology, even though robust parameter estimation methods are of clear importance in watershed model calibration and evaluation. Notably, several objective functions already used in the field are *instances* of M-estimators even if they were not framed that way. For example, the least absolute M-estimator (#1 in Table A.1) is exactly the "sum of absolute residuals" measure in Table 1 of Beven and Binley [15]. Furthermore, the "inverse residual" measure of Beven [12] is closely related to the Welsch M-estimator (#9 in Table A.1). In other words, these works employed ρ-type criteria with bounded influence, albeit without the M-estimation terminology.

A.2 Robust Efficiency Metrics for Hydrologic Models

The principles of M-estimation can be used to improve the robustness of hydrologic performance measures such as the Nash and Sutcliffe [110] efficiency (NSE). In machine-learning-oriented hydrology, the Huber loss is sometimes used during training [168], however, model evaluation is typically still carried out with NSE or the Kling-Gupta efficiency [63].

Let $\boldsymbol{\omega}_n = (\omega_1, \dots, \omega_n)^\top$ and $\mathbf{y}_n = (y_1, \dots, y_n)^\top$ be n-vectors of observed and modeled streamflows, respectively, under $\boldsymbol{\theta}$. A robust NSE, $\text{RNSE}_{\rho_c} : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, 1]$, is

$$\text{RNSE}_{\rho_c}(\boldsymbol{\omega}_n, \mathbf{y}_n) = 1 - \frac{\sum_{t=1}^n \rho_c\big(S_\omega^{-1}(\omega_t - y_t)\big)}{\sum_{t=1}^n \rho_c\big(S_\omega^{-1}(\omega_t - \zeta_\omega)\big)}, \tag{A.3}$$

where ρ_c is the robust loss function with tuning constant c, and ζ_ω and $S_\omega = \kappa\,\sigma_n$ are robust location and scale functionals of the discharge data $\{\omega_t\}_{t=1}^n$. We can admit the Huber [78] loss (#6 in Table A.1 and Fig. A.2(e2))

$$\rho_c(\underline{e}) = \begin{cases} \frac{1}{2}\underline{e}^2 & \text{if } |\underline{e}| \leq c, \\ c|\underline{e}| - \frac{1}{2}c^2 & \text{otherwise,} \end{cases}$$

with $c = 1.345$ (about 95% asymptotic Normal efficiency). A common and convenient choice for σ_n is the median absolute deviation (MAD) from the median [54]

$$\begin{aligned} \sigma_n &= \text{MAD}(\omega_1, \dots, \omega_n) \\ &= \text{median}_{t=1,\dots,n}(|\omega_t - \zeta_\omega|), \end{aligned}$$

which centers the deviations at the sample median $\zeta_\omega = \text{median}(\omega_1, \dots, \omega_n)$. With $\kappa = 1.4826$, the scale $S_\omega = \kappa\,\sigma_n$ is a consistent estimator of the standard deviation of the ω's under Normality.

The RNSE inherits the key affine invariances of NSE. For any $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$, it holds that $\text{RNSE}_{\rho_c}(a\,\omega + b,\, a\,y + b) = \text{RNSE}_{\rho_c}(\omega, y)$ provided the location-scale functionals are translation/scale equivariant. Consequently RNSE preserves the canonical NSE scale $(-\infty, 1]$. Specifically, $\text{RNSE}_{\rho_c}(\omega, y = \omega) = 1$, $\text{RNSE}_{\rho_c}(\omega, y = \zeta_\omega) = 0$ and $\text{RNSE}_{\rho_c}(\omega, y) <$

0 if the model y is worse than the location functional ζ_ω. Practically, RNSE will exceed NSE on datasets where a few outliers dominate the sum of squares, because the robust loss downweights extreme residuals. Conversely, when apparent "outliers" are informative, the robust loss can downweight genuine signal. Finally, with quadratic loss $\rho(\underline{e}) = \frac{1}{2}\underline{e}^2$ and ζ_ω equal to the sample mean m_ω, RNSE reduces to the standard $\text{NSE}: \mathbb{R}^n \times \mathbb{R}^n \rightarrow (-\infty, 1]$ or

$$\text{NSE}(\boldsymbol{\omega}_n, \mathbf{y}_n) = 1 - \frac{\sum_{t=1}^{n}(\omega_t - y_t)^2}{\sum_{t=1}^{n}(\omega_t - m_\omega)^2}.$$

When the y_t's denote the in-sample OLS fitted values from a regression that includes an intercept, NSE $= R^2$. Thus, RNSE is a robust analogue of the well-known coefficient of determination R^2: it replaces (i) the squared loss by a robust loss and (ii) the mean/variance by robust location/scale functionals.

Appendix B: Information Identity for a Normal Distribution Model

In this Appendix we use a simple illustrative example to demonstrate that under correct specification the information equality $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_n(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{B}_n(\boldsymbol{\theta})]$ holds, where $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{A}_n(\boldsymbol{\theta})]$ is the expected curvature of the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$ under the data-generating process for a typical observation ω and $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{B}_n(\boldsymbol{\theta})]$ is the expected outer product of the *score* for a datum ω.

Suppose the observations $\omega_1, \ldots, \omega_n$ originate from a normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$ with unknown mean μ_0 and variance σ_0^2. We would like to estimate the mean and variance of random variable Ω under the Normal density $f(\omega; \mu, \sigma^2)$ using the n data points. We resort to maximum likelihood (ML) estimation and write out the total likelihood $L_n^{\mathrm{n}}(\mu, \sigma^2)$ of $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ given $\omega_1, \ldots, \omega_n$

$$L_n^{\mathrm{n}}(\mu, \sigma^2) = \prod_{t=1}^{n} \bigl(2\pi\sigma^2\bigr)^{-1/2} \exp\bigl(-\tfrac{1}{2}\sigma^{-2}(\omega_t - \mu)^2\bigr).$$

and the log-likelihood $\mathcal{L}_n(\mu, \sigma^2) = \log\bigl(L_n(\mu, \sigma^2)\bigr)$ becomes

$$\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2) = -\tfrac{1}{2}n\log(2\pi) - \tfrac{1}{2}n\log(\sigma^2) - \tfrac{1}{2}\sigma^{-2}\sum_{t=1}^{n}(\omega_t - \mu)^2.$$

where the superscript "n" in $L_n^{\mathrm{n}}(\mu, \sigma^2)$ and $\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2)$ stands for <u>n</u>ormal. Under standard regularity conditions, the ML estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)$ is consistent, that is, $\widehat{\mu}_n \xrightarrow{\mathrm{p}} \mu_0$ and $\widehat{\sigma}_n^2 \xrightarrow{\mathrm{p}} \sigma_0^2$ as the sample size $n \to \infty$.

We now want to determine the ML estimates of μ and σ^2. All we have to do is maximize $\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2)$. We can do this with an iterative (search) method. But given the relative simple data-generating process we follow an analytic procedure instead.

B.1 Maximum Likelihood Estimation

The ML estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)$ is a root of the *score function* or gradient vector, $\mathbf{g}_n(\mu, \sigma^2) = \nabla_{(\mu,\sigma^2)}\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2)$. We differentiate $\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2)$ with respect to μ

$$\begin{aligned}
\frac{\partial}{\partial\mu}\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2) &= -\tfrac{1}{2}\sigma^{-2}\sum_{t=1}^{n}\frac{\partial}{\partial\mu}(\omega_t - \mu)^2 \\
&= -\tfrac{1}{2}\sigma^{-2}\sum_{t=1}^{n}2(\omega_t - \mu)(-1) \\
&= \sigma^{-2}\sum_{t=1}^{n}(\omega_t - \mu).
\end{aligned}$$

Next, we differentiate the log-likelihood $\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2)$ with respect to σ^2

$$\frac{\partial}{\partial\sigma^2}\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2) = -\frac{1}{2}n\frac{\partial}{\partial\sigma^2}\bigl(\log(\sigma^2)\bigr) - \frac{1}{2}\frac{\partial}{\partial\sigma^2}\bigl(\sigma^{-2}\bigr)\sum_{t=1}^{n}(\omega_t - \mu)^2.$$

As a hint, substitute $x = \sigma^2$, and determine $\partial\mathcal{L}_n^{\mathrm{n}}(\mu, x)/\partial x$ to yield

$$\frac{\partial}{\partial\sigma^2}\mathcal{L}_n^{\mathrm{n}}(\mu, \sigma^2) = -\tfrac{1}{2}n\,\sigma^{-2} + \tfrac{1}{2}\sigma^{-4}\sum_{t=1}^{n}(\omega_t - \mu)^2.$$

Thus, the *score* or gradient vector is equal to

$$\mathbf{g}_n(\mu,\sigma^2) = \begin{bmatrix} \dfrac{\partial}{\partial\mu}\mathcal{L}_n(\mu,\sigma^2) \\[2ex] \dfrac{\partial}{\partial\sigma^2}\mathcal{L}_n(\mu,\sigma^2) \end{bmatrix} = \begin{bmatrix} \sigma^{-2}\sum\limits_{t=1}^{n}(\omega_t-\mu) \\[2ex] -\tfrac{1}{2}n\,\sigma^{-2} + \tfrac{1}{2}\sigma^{-4}\sum\limits_{t=1}^{n}(\omega_t-\mu)^2 \end{bmatrix}.$$

At the log-likelihood maximum $(\widehat{\mu}_n, \widehat{\sigma}^2_n)$, the *score* will be zero. Thus, we must solve

$$\begin{bmatrix} \widehat{\sigma}_n^{-2}\sum\limits_{t=1}^{n}(\omega_t-\widehat{\mu}_n) \\[2ex] -\tfrac{1}{2}n\,\widehat{\sigma}_n^{-2} + \tfrac{1}{2}\widehat{\sigma}_n^{-4}\sum\limits_{t=1}^{n}(\omega_t-\widehat{\mu}_n)^2 \end{bmatrix} = \begin{bmatrix} 0 \\[2ex] 0 \end{bmatrix},$$

and this results in

$$(1)\quad \widehat{\sigma}_n^{-2}\sum_{t=1}^{n}(\omega_t-\widehat{\mu}_n) = 0 \qquad \text{and} \qquad (2)\quad -\tfrac{1}{2}n\widehat{\sigma}_n^{-2} + \tfrac{1}{2}\widehat{\sigma}_n^{-4}\sum_{t=1}^{n}(\omega_t-\widehat{\mu}_n)^2 = 0.$$

To satisfy the first condition

$$\sum_{t=1}^{n}(\omega_t-\widehat{\mu}_n) = 0 \quad\Rightarrow\quad \sum_{t=1}^{n}\omega_t - n\,\widehat{\mu}_n = 0 \quad\Rightarrow\quad \widehat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n}\omega_t.$$

The second condition implies that

$$\sum_{t=1}^{n}(\omega_t-\widehat{\mu}_n)^2 = n\,\widehat{\sigma}^2_n \quad\Rightarrow\quad \widehat{\sigma}^2_n = \frac{1}{n}\sum_{t=1}^{n}(\omega_t-\widehat{\mu}_n)^2.$$

Thus, the ML estimate $(\widehat{\mu}_n, \widehat{\sigma}^2_n)$ of the mean and variance of the normal data-generating process $\mathcal{N}(\mu,\sigma^2)$ is given by the sample mean and the biased sample variance of data $\omega_1,\ldots,\omega_n$, respectively. The unbiased estimator replaces n by $n-1$.

B.2 Information Matrices

We confirm the information identity $\mathbb{E}_{(\mu,\sigma^2)}\big[\mathbf{A}_n(\mu,\sigma^2)\big] = \mathbb{E}_{(\mu,\sigma^2)}\big[\mathbf{B}_n(\mu,\sigma^2)\big]$ for all (μ,σ^2) under the Normal density $f(\cdot;\mu,\sigma^2)$ for i.i.d. realizations $\omega_1,\ldots,\omega_n$ of Ω. The superscript "n" clarifies that this sensitivity matrix corresponds to the likelihood of the normal distribution.

B.2.1 Sensitivity (Bread) Matrix, $\mathbf{A}^{\mathrm{n}}_n(\mu,\sigma^2)$

The per-observation Hessian matrix

$$\mathbf{H}_\omega(\mu,\sigma^2) = \nabla^2_{(\mu,\sigma^2)}\mathcal{L}^{\mathrm{n}}_\omega(\mu,\sigma^2),$$

is obtained as the second derivative of the log-likelihood function

$$\mathcal{L}^{\mathrm{n}}_\omega(\mu,\sigma^2) = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(\sigma^2) - \tfrac{1}{2}\sigma^{-2}(\omega-\mu)^2.$$

with respect to the parameters μ and σ^2. The first derivative of $\mathcal{L}^{\mathrm{n}}_\omega(\mu,\sigma^2)$, the per-observation *score*, is equal to

$$\begin{aligned}\mathbf{g}_\omega(\mu,\sigma^2) &= \nabla_{(\mu,\sigma^2)}\mathcal{L}^{\mathrm{n}}_\omega(\mu,\sigma^2) \\ &= \begin{bmatrix} \sigma^{-2}(\omega-\mu) \\[2ex] -\tfrac{1}{2}\sigma^{-2} + \tfrac{1}{2}\sigma^{-4}(\omega-\mu)^2 \end{bmatrix}.\end{aligned}$$

Differentiating the first entry of the *score* vector one more time with respect to μ gives

$$\begin{aligned}\frac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \mu^2} &= \frac{\partial}{\partial \mu}\biggl(\frac{\partial}{\partial \mu}\mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)\biggr)\\ &= \frac{\partial}{\partial \mu}\bigl(\sigma^{-2}(\omega-\mu)\bigr) = -\sigma^{-2}.\end{aligned}$$

We repeat this same step for the second entry of the *score* vector but with respect to σ^2

$$\begin{aligned}\frac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\biggl(\frac{\partial}{\partial \sigma^2}\mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)\biggr)\\ &= \frac{\partial}{\partial \sigma^2}\bigl(-\tfrac{1}{2}\sigma^{-2}+\tfrac{1}{2}\sigma^{-4}(\omega-\mu)^2\bigr) = \tfrac{1}{2}\sigma^{-4}-\sigma^{-6}(\omega-\mu)^2.\end{aligned}$$

Last, we look at the mixed partials

$$\begin{aligned}\frac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \mu\, \partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\biggl(\frac{\partial}{\partial \mu}\mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)\biggr)\\ &= \frac{\partial}{\partial \sigma^2}\bigl(\sigma^{-2}(\omega-\mu)\bigr) = -\sigma^{-4}(\omega-\mu),\end{aligned}$$

and by symmetry

$$\frac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \sigma^2\, \partial \mu} = \frac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \mu\, \partial \sigma^2}.$$

Hence, the per-observation Hessian matrix is

$$\begin{aligned}\mathbf{H}_{\omega}(\mu,\sigma^2) &= \nabla^2_{(\mu,\sigma^2)}\mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)\\ &= \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \mu^2} & \dfrac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \mu\, \partial \sigma^2} \\ \dfrac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \sigma^2\, \partial \mu} & \dfrac{\partial^2 \mathcal{L}^{\mathrm{n}}_{\omega}(\mu,\sigma^2)}{\partial \sigma^2} \end{bmatrix}\\ &= \begin{bmatrix} -\sigma^{-2} & -\sigma^{-4}(\omega-\mu) \\ -\sigma^{-4}(\omega-\mu) & \tfrac{1}{2}\sigma^{-4}-\sigma^{-6}(\omega-\mu)^2 \end{bmatrix}.\end{aligned}$$

For i.i.d. realizations $\omega_1,\ldots,\omega_n$, the total Hessian $\mathbf{H}_n(\mu,\sigma^2)$ is (see Definition 4, p. 7)

$$\begin{aligned}\mathbf{H}_n(\mu,\sigma^2) &= \sum_{t=1}^{n}\mathbf{H}_{\omega_t}(\mu,\sigma^2)\\ &= -\begin{bmatrix} n\,\sigma^{-2} & \sigma^{-4}\sum\limits_{t=1}^{n}(\omega_t-\mu) \\ \sigma^{-4}\sum\limits_{t=1}^{n}(\omega_t-\mu) & -\tfrac{1}{2}n\,\sigma^{-4}+\sigma^{-6}\sum\limits_{t=1}^{n}(\omega_t-\mu)^2 \end{bmatrix}.\end{aligned}$$

Then, Definition 5 on p. 7 states that

$$\begin{aligned}\mathbf{A}^{\mathrm{n}}_n(\mu,\sigma^2) &= -\tfrac{1}{n}\mathbf{H}_n(\mu,\sigma^2)\\ &= \begin{bmatrix} \sigma^{-2} & \tfrac{1}{n}\sigma^{-4}\sum\limits_{t=1}^{n}(\omega_t-\mu) \\ \tfrac{1}{n}\sigma^{-4}\sum\limits_{t=1}^{n}(\omega_t-\mu) & -\tfrac{1}{2}\sigma^{-4}+\tfrac{1}{n}\sigma^{-6}\sum\limits_{t=1}^{n}(\omega_t-\mu)^2 \end{bmatrix},\end{aligned}$$

and we yield the following expression for the expectation of the sensitivity matrix

$$\begin{aligned}\mathbb{E}\big[\mathbf{A}_n^{\mathrm{n}}(\mu,\sigma^2)\big] &= \mathbb{E}\left[\begin{bmatrix} \sigma^{-2} & \frac{1}{n}\sigma^{-4}\sum\limits_{t=1}^{n}(\Omega_t-\mu) \\ \frac{1}{n}\sigma^{-4}\sum\limits_{t=1}^{n}(\Omega_t-\mu) & -\frac{1}{2}\sigma^{-4}+\frac{1}{n}\sigma^{-6}\sum\limits_{t=1}^{n}(\Omega_t-\mu)^2 \end{bmatrix}\right] \\ &= \begin{bmatrix} \mathbb{E}[\sigma^{-2}] & \sigma^{-4}\mathbb{E}[\Omega-\mu] \\ \sigma^{-4}\mathbb{E}[\Omega-\mu] & -\frac{1}{2}\sigma^{-4}+\sigma^{-6}\mathbb{E}\big[(\Omega-\mu)^2\big] \end{bmatrix},\end{aligned}$$

where $\mathbb{E}[\cdot]$ is the expectation under the Normal density $f(\cdot \mid \mu,\sigma^2)$. For (μ,σ^2) we have

$$\mathbb{E}[\Omega-\mu]=0 \qquad \text{and} \qquad \mathbb{E}\big[(\Omega-\mu)^2\big]=\sigma^2,$$

and this leaves us with

$$\mathbb{E}\big[\mathbf{A}_n^{\mathrm{n}}(\mu,\sigma^2)\big] = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{bmatrix}.$$

B.2.2 The Variability (Meat) Matrix, $\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)$

According to Definition 6 on P. 7, the variability matrix is equal to

$$\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2) = \tfrac{1}{n}\sum_{t=1}^{n}\nabla_{(\mu,\sigma^2)}\mathcal{L}_{\omega_t}^{\mathrm{n}}(\mu,\sigma^2)\nabla_{(\mu,\sigma^2)}^{\top}\mathcal{L}_{\omega_t}^{\mathrm{n}}(\mu,\sigma^2).$$

We can write this as an expectation instead

$$\begin{aligned}\mathbb{E}\big[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)\big] &= \mathbb{E}\big[\nabla_{(\mu,\sigma^2)}\mathcal{L}_{\Omega}^{\mathrm{n}}(\mu,\sigma^2)\nabla_{(\mu,\sigma^2)}^{\top}\mathcal{L}_{\Omega}^{\mathrm{n}}(\mu,\sigma^2)\big] \\ &= \mathbb{E}\big[\mathbf{g}_{\Omega}(\mu,\sigma^2)\,\mathbf{g}_{\Omega}^{\top}(\mu,\sigma^2)\big] \\ &= \mathbb{E}\left[\begin{bmatrix} \sigma^{-2}(\Omega-\mu) \\ -\frac{1}{2}\sigma^{-2}+\frac{1}{2}\sigma^{-4}(\Omega-\mu)^2 \end{bmatrix}\begin{bmatrix} \sigma^{-2}(\Omega-\mu) & -\frac{1}{2}\sigma^{-2}+\frac{1}{2}\sigma^{-4}(\Omega-\mu)^2 \end{bmatrix}\right].\end{aligned}$$

We can work out the vector outer product

$$\mathbb{E}\big[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)\big] = \mathbb{E}\begin{bmatrix} \sigma^{-4}(\Omega-\mu)^2 & -\frac{1}{2}\sigma^{-4}(\Omega-\mu)+\frac{1}{2}\sigma^{-6}(\Omega-\mu)^3 \\ -\frac{1}{2}\sigma^{-4}(\Omega-\mu)+\frac{1}{2}\sigma^{-6}(\Omega-\mu)^3 & \frac{1}{4}\sigma^{-4}-\frac{1}{2}\sigma^{-6}(\Omega-\mu)^2+\frac{1}{4}\sigma^{-8}(\Omega-\mu)^4 \end{bmatrix},$$

and determine entry-wise each expectation. We end up with

$$\mathbb{E}\big[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)\big] = \begin{bmatrix} \sigma^{-4}\mathbb{E}\big[(\Omega-\mu)^2\big] & -\frac{1}{2}\sigma^{-4}\mathbb{E}[\Omega-\mu]+\frac{1}{2}\sigma^{-6}\mathbb{E}\big[(\Omega-\mu)^3\big] \\ -\frac{1}{2}\sigma^{-4}\mathbb{E}[\Omega-\mu]+\frac{1}{2}\sigma^{-6}\mathbb{E}\big[(\Omega-\mu)^3\big] & \frac{1}{4}\sigma^{-4}-\frac{1}{2}\sigma^{-6}\mathbb{E}\big[(\Omega-\mu)^2\big]+\frac{1}{4}\sigma^{-8}\mathbb{E}\big[(\Omega-\mu)^4\big] \end{bmatrix}.$$

For (μ,σ^2) we have

$$\mathbb{E}[\Omega-\mu]=0, \quad \mathbb{E}\big[(\Omega-\mu)^2\big]=\sigma^2, \quad \mathbb{E}\big[(\Omega-\mu)^3\big]=\mu_3, \quad \text{and} \quad \mathbb{E}\big[(\Omega-\mu)^4\big]=\mu_4.$$

For i.i.d. data the third and fourth central moments equal $\mu_3=0$ and $\mu_4=3\sigma^4$, respectively. Then

$$\mathbb{E}\big[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)\big] = \begin{bmatrix} \sigma^{-4}\sigma^2 & -\frac{1}{2}\sigma^{-4}\cdot 0+\frac{1}{2}\sigma^{-6}\cdot 0 \\ -\frac{1}{2}\sigma^{-4}\cdot 0+\frac{1}{2}\sigma^{-6}\cdot 0 & \frac{1}{4}\sigma^{-4}-\frac{1}{2}\sigma^{-6}\sigma^2+\frac{3}{4}\sigma^{-8}\sigma^4 \end{bmatrix},$$

and the expectation of the variability matrix is

$$\mathbb{E}\big[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)\big] = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{bmatrix}.$$

B.3 Information Identity

The expressions for the sensitivity and variability matrices confirm the second Bartlett identity (18), i.e., $\mathbb{E}[\mathbf{A}_n^{\mathrm{n}}(\mu,\sigma^2)] = \mathbb{E}[\mathbf{B}_n^{\mathrm{n}}(\mu,\sigma^2)]$. Thus, under correct specification, the Fisher information equals

$$\mathcal{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\big[-\nabla^2\mathcal{L}_n(\boldsymbol{\theta})\big] = \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla\mathcal{L}_n(\boldsymbol{\theta})\,\nabla^\top\mathcal{L}_n(\boldsymbol{\theta})\big],$$

and, thus, the expected curvature of the log-likelihood at $\boldsymbol{\theta}$ equals the expected outer product of the *score* at the same point. For i.i.d. data of size n, we can write $\mathcal{I}_n(\mu,\sigma^2) = n\,\mathcal{I}_\omega(\mu,\sigma^2)$.

In practice, we only observe a finite sample $\omega_1,\ldots,\omega_n$ from the data-generating process, and therefore we cannot compute expectations of the sensitivity and variability matrices. Let $\mathbf{A}_0$ and $\mathbf{B}_0$ denote the corresponding matrices under the true process $\mathcal{N}(\mu_0,\sigma_0^2)$. Then, under correct specification

$$\widehat{\mathbf{A}}_n \xrightarrow{\mathrm{p}} \mathbf{A}_0, \quad \text{and} \quad \widehat{\mathbf{B}}_n \xrightarrow{\mathrm{p}} \mathbf{B}_0,$$

as the sample size $n \to \infty$. With a finite sample, $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ are typically close (up to sampling variability) when the model is correctly specified (see Figure 1). Under misspecification, however, the two matrices generally do not coincide. Analytic examples are provided in Vrugt and Diks [160].

Appendix C: Bread and Meat Matrices for the ABC-Model

This Appendix derives analytic expressions for the $d \times d$ variability and sensitivity matrices, $\mathbf{A}_\omega(\boldsymbol{\theta})$ and $\mathbf{B}_\omega(\boldsymbol{\theta})$ respectively, of the ABC model of Fiering [50], for a typical discharge measurement $\omega \in \Omega$. As a reminder, we use lowercase letters for fluxes such as rainfall, p (L/T), infiltration, i (L/T), and evaporation, e (L/T).

To derive algebraic expressions for the sensitivity and variability matrices we must write the discharge of the ABC model in a convenient analytic form. If precipitation p at time t is written as p_t (L/T) then ABC simulated streamflow q_t (L/T) is equal to (see Fig. 6)

$$q_t = (1 - a - b)p_t + cr_{t-1}, \tag{C.1}$$

where groundwater storage r_{t-1} at the previous time $t-1$ satisfies the following relationship

$$r_{t-1} = c'r_{t-2} + ap_{t-1}, \tag{C.2}$$

where $c' = 1 - c$ is the complement of parameter c. We can use the recurrence relationship of Equation (C.2) to find the relationship between $r_{t-1-\Delta t}$ and r_{t-1}. For example, for $\Delta t = 5$ we yield the following expression

$$\begin{aligned} r_{t-1} &= c'[c'r_{t-3} + a\,p_{t-2}] + a\,p_{t-1} \\ &= c'[c'\{c'r_{t-4} + a\,p_{t-3}\} + a\,p_{t-2}] + a\,p_{t-1} \\ &= c'[c'\{c'[c'r_{t-5} + a\,p_{t-4}] + a\,p_{t-3}\} + a\,p_{t-2}] + a\,p_{t-1} \\ &= c'[c'\{c'[c'\{c'r_{t-6} + a\,p_{t-5}\} + a\,p_{t-4}] + a\,p_{t-3}\} + a\,p_{t-2}] + a\,p_{t-1}. \end{aligned} \tag{C.3}$$

Choosing $\Delta t > 5$ merely increases the number of antecedent-precipitation terms and is not required to accurately compute the first- and second-order derivatives of the ABC model, provided groundwater residence times are sufficiently short.

If we insert Equation (C.3) into Equation (C.1) we yield a simple algebraic expression for the relationship between the discharge q_t at time t and present and antecedant precipitation and groundwater storage

$$q_t = (1 - a - b)p_t + c\{c'[c'\{c'[c'\{c'r_{t-6} + a\,p_{t-5}\} + a\,p_{t-4}] + a\,p_{t-3}\} + a\,p_{t-2}] + a\,p_{t-1}\}.$$

Now let us consider the normal power likelihood function

$$\begin{aligned} L_n^{\text{np}}(\boldsymbol{\theta}) &= f^{\lambda}_{\mathcal{N}_n}(\mathbf{y}_n; \boldsymbol{\omega}_n, \boldsymbol{\Sigma}_e) \\ &= (2\pi)^{-\frac{1}{2}n\lambda}|\boldsymbol{\Sigma}_e|^{-\frac{1}{2}\lambda}\exp\Big(-\tfrac{1}{2}\big(\boldsymbol{\omega}_n - \boldsymbol{\mathcal{M}}_{\text{abc}}(\boldsymbol{\theta};\mathbf{p}_n)\big)^{\top}\boldsymbol{\Sigma}_e^{-1}\big(\boldsymbol{\omega}_n - \boldsymbol{\mathcal{M}}_{\text{abc}}(\boldsymbol{\theta};\mathbf{p}_n)\big)\Big)^{\lambda}, \end{aligned}$$

where $\mathbf{y}_n = \boldsymbol{\mathcal{M}}_{\text{abc}}(\boldsymbol{\theta};\mathbf{p}_n)$ is the vector-valued discharge $\mathbf{y}_n = (y_1, \ldots, y_n)^{\top}$ simulated by the ABC model for parameters $\boldsymbol{\theta} = (a, b, c)^{\top}$ and the $n \times 1$ rainfall record $\mathbf{p}_n = (p_1, \ldots, p_n)^{\top}$. The normal power log-likelihood for a single ABC simulated discharge value y_t becomes

$$\begin{aligned} \mathcal{L}^{\text{np}}_{\omega_t}(\boldsymbol{\theta}) &= -\tfrac{1}{2}\lambda\log(2\pi) - \tfrac{1}{2}\lambda\log(\sigma_e^2) - \tfrac{1}{2}\lambda\,\sigma_e^{-2}\big(\omega_t - \mathcal{M}_t(\boldsymbol{\theta};\mathbf{p}_n)\big)^2 \\ &= -\tfrac{1}{2}\lambda\log(2\pi) - \tfrac{1}{2}\lambda\log(\sigma_e^2) - \tfrac{1}{2}\lambda\,\sigma_e^{-2}(\omega_t - [(1-a-b)p_t + c\{c'[c'\{c'[c'\{c'r_{t-6} + a\,p_{t-5}\} \\ &\qquad + a\,p_{t-4}] + a\,p_{t-3}\} + a\,p_{t-2}] + a\,p_{t-1}\}])^2. \end{aligned}$$

We can now determine the first-order derivatives of the normal log-likelihood with respect to

a, b and c

$$\begin{aligned}
\frac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} &= -\lambda\,\sigma_e^{-2}(\,p_t - c[p_{t-1} + c'\{p_{t-2} + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\}]) \\
&\quad \times [\omega_t - c(a\,p_{t-1} + c'[a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\}]) \\
&\quad + p_t(a + b - 1)]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} &= -\lambda\,\sigma_e^{-2}\,p_t\{\omega_t - c[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})] \\
&\quad + p_t(a + b - 1)\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} &= \lambda\,\sigma_e^{-2}\{\omega_t - c[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})] \\
&\quad + p_t(a + b - 1)\} \\
&\quad \times [a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\}) \\
&\quad + c(-a\,p_{t-2} - c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\} \\
&\quad - c'\{a\,p_{t-3} - c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c'r_{t-6}) - c'(a\,p_{t-5} + 2c'r_{t-6})] \\
&\quad + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})],
\end{aligned}$$

where the exogeneous variables (rainfall and groundwater storage) are conveniently omitted as input arguments of the likelihood. The 3×1 gradient vector or *score function* $\mathbf{g}^{\mathrm{np}}_{\omega_t}(a,b,c)$ now becomes

$$\begin{aligned}
\mathbf{g}^{\mathrm{np}}_{\omega_t}(a,b,c) \equiv \nabla_{(a,b,c)}\mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c) &\equiv \frac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial(a,b,c)} \\
&= \begin{bmatrix} \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} \\ \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} \\ \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} \end{bmatrix}.
\end{aligned}$$

The analytic expressions of the first-order derivatives demonstrate a linear dependence on the learning rate. The *score* of $L^{\mathrm{np}}_{\omega_t}(\boldsymbol{\theta})$ may thus also be written as $\mathbf{g}^{\mathrm{np}}_{\omega_t}(a,b,c) = \lambda\,\mathbf{g}^{\mathrm{n}}_{\omega_t}(a,b,c)$, where $\mathbf{g}^{\mathrm{n}}_{\omega_t}(a,b,c)$ is the *score* of the normal log-likelihood for datum ω.

The 3×3 variability matrix $\mathbf{B}^{\mathrm{np}}_{\omega_t}(a,b,c)$ is now equal to

$$\begin{aligned}
\mathbf{B}^{\mathrm{np}}_{\omega_t}(a,b,c) &= \mathbf{g}^{\mathrm{np}}_{\omega_t}(a,b,c)\mathbf{g}^{\mathrm{np}}_{\omega_t}(a,b,c)^{\top} \\
&= \begin{bmatrix} \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} \\ \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} \\ \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} \end{bmatrix} \begin{bmatrix} \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} \end{bmatrix} \\
&= \begin{bmatrix}
\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} \\
\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c} \\
\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial a} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial b} & \dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c}\dfrac{\partial \mathcal{L}^{\mathrm{np}}_{\omega_t}(a,b,c)}{\partial c}
\end{bmatrix},
\end{aligned}$$

and we yield the total variability matrix $\mathbf{B}_n^{\mathrm{np}}(a,b,c) = \frac{1}{n}\big(\mathbf{B}_{\omega_1}^{\mathrm{np}}(a,b,c) + \ldots + \mathbf{B}_{\omega_n}^{\mathrm{np}}(a,b,c)\big)$. The variability matrix is symmetric by construction, $b_{n,ij}^{\mathrm{np}}(a,b,c) = b_{n,ji}^{\mathrm{np}}(a,b,c)$ for all $i \neq j$. Equivalently, we can stack the per-observation *score* vectors $\mathbf{g}_{\omega_t} = \nabla\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)$ for $t = 1,\ldots,n$ into an $n \times 3$ Jacobian matrix

$$\mathbf{J}_n^{\mathrm{np}}(a,b,c) = \begin{bmatrix} \dfrac{\partial\mathcal{L}_{\omega_1}^{\mathrm{np}}(a,b,c)}{\partial a} & \dfrac{\partial\mathcal{L}_{\omega_1}^{\mathrm{np}}(a,b,c)}{\partial b} & \dfrac{\partial\mathcal{L}_{\omega_1}^{\mathrm{np}}(a,b,c)}{\partial c} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial\mathcal{L}_{\omega_n}^{\mathrm{np}}(a,b,c)}{\partial a} & \dfrac{\partial\mathcal{L}_{\omega_n}^{\mathrm{np}}(a,b,c)}{\partial b} & \dfrac{\partial\mathcal{L}_{\omega_n}^{\mathrm{np}}(a,b,c)}{\partial c} \end{bmatrix},$$

and compute the total variability matrix at once using

$$\mathbf{B}_n^{\mathrm{np}}(a,b,c) = \tfrac{1}{n}\mathbf{J}_n^{\mathrm{np}}(a,b,c)^\top\mathbf{J}_n^{\mathrm{np}}(a,b,c).$$

The sum of each column of $\mathbf{J}_n^{\mathrm{np}}(a,b,c)$ yields the entries of the *total score* $\mathbf{g}_n^{\mathrm{np}}(a,b,c)$. The vector outer product implies that the variability matrix $\mathbf{B}_n^{\mathrm{np}}(a,b,c)$ depends linearly on the square of the learning rate. Thus, we can write $\mathbf{B}_n^{\mathrm{np}}(a,b,c) = \lambda^2\,\mathbf{B}_n^{\mathrm{n}}(a,b,c)$, where $\mathbf{B}_n^{\mathrm{n}}(a,b,c)$ is the variability matrix of the log-likelihood $\mathcal{L}_n^{\mathrm{n}}(\boldsymbol{\theta})$.

For the 3×3 sensitivity or Hessian matrix $\mathbf{H}_{\omega_t}(a,b,c)$ we must differentiate one more time the first-order derivatives of the log-likelihood. This yields the second-order partial derivatives of $\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)$

$$\begin{aligned}
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial a^2} &= -\lambda\,\sigma_e^{-2}[p_t - c(p_{t-1} + c'\{p_{t-2} + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\})]^2 \\
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial a\,\partial b} &= -\lambda\,\sigma_e^{-2}\,p_t[p_t - c(p_{t-1} + c'\{p_{t-2} + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\})] \\
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial a\,\partial c} &= -\lambda\,\sigma_e^{-2}[\omega_t - c[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})] \\
&\qquad + p_t(a+b-1)](-p_{t-1} + c\{p_{t-2} - c'[-p_{t-3} - c'(p_{t-4} + c'p_{t-5}) \\
&\qquad - c'(p_{t-4} + 2c'p_{t-5})] + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\} \\
&\qquad - c'\{p_{t-2} + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\}) \\
&\quad + \lambda\,\sigma_e^{-2}[p_t - c(p_{t-1} + c'\{p_{t-2} + c'[p_{t-3} + c'(p_{t-4} + c'p_{t-5})]\})] \\
&\qquad \times [a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\}) \\
&\qquad + c(-a\,p_{t-2} - c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\} \\
&\qquad - c'\{a\,p_{t-3} - c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c'r_{t-6}) - c'(a\,p_{t-5} + 2c'r_{t-6})] \\
&\qquad + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial b^2} &= -\lambda\,\sigma_e^{-2}\,p_t^2 \\
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial b\,\partial c} &= \lambda\,\sigma_e^{-2}\,p_t[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\}) \\
&\qquad + c(-a\,p_{t-2} - c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\} \\
&\qquad - c'\{a\,p_{t-3} - c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c'r_{t-6}) - c'(a\,p_{t-5} + 2c'r_{t-6})] \\
&\qquad + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\})] \\
\frac{\partial^2\mathcal{L}_{\omega_t}^{\mathrm{np}}(a,b,c)}{\partial c^2} &= -\lambda\,\sigma_e^{-2}[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\}) \\
&\qquad + c(-a\,p_{t-2} - c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c'r_{t-6})]\} \\
&\qquad - c'\{a\,p_{t-3} - c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c'r_{t-6}) - c'(a\,p_{t-5} + 2c'r_{t-6})]
\end{aligned}$$

$$
\begin{aligned}
&\quad + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c' r_{t-6})]\})]^2 \\
&+ \lambda\,\sigma_e^{-2}\{\omega_t - c[a\,p_{t-1} + c'(a\,p_{t-2} + c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c' r_{t-6})]\})] \\
&\quad + p_t(a + b - 1)\}(c\{2a\,p_{t-3} - 2c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c' r_{t-6}) \\
&\quad - c'(a\,p_{t-5} + 2c' r_{t-6})] - c'[-2a\,p_{t-4} - 2c'(a\,p_{t-5} + c' r_{t-6}) \\
&\quad - 2c'(a\,p_{t-5} + 2c' r_{t-6}) - c'(2a\,p_{t-5} + 6c' r_{t-6})] \\
&\quad + 2c'[a\,p_{t-4} + c'(a\,p_{t-5} + c' r_{t-6})]\} - 2a\,p_{t-2} \\
&\quad - 2c'\{a\,p_{t-3} + c'[a\,p_{t-4} + c'(a\,p_{t-5} + c' r_{t-6})]\} \\
&\quad - c'\{a\,p_{t-3} - c'[-a\,p_{t-4} - c'(a\,p_{t-5} + c' r_{t-6}) - c'(a\,p_{t-5} + 2c' r_{t-6})] \\
&\quad + 2c'[a\,p_{t-4} + c'(a\,p_{t-5} + c' r_{t-6})]\}),
\end{aligned}
$$

and we obtain the Hessian matrix

$$
\begin{aligned}
\mathbf{H}^{\text{np}}_{\omega_t}(a,b,c) \equiv \nabla^2_{(a,b,c)} \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c) &\equiv \frac{\partial}{\partial(a,b,c)}\left(\frac{\partial \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial(a,b,c)}\right) \\
&= \begin{bmatrix}
\dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial a^2} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial a\,\partial b} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial a\,\partial c} \\
\dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial b\,\partial a} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial b^2} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial b\,\partial c} \\
\dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial c\,\partial a} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial c\,\partial b} & \dfrac{\partial^2 \mathcal{L}^{\text{np}}_{\omega_t}(a,b,c)}{\partial c^2}
\end{bmatrix}.
\end{aligned}
$$

The order of partial differentiation does not matter; $\partial^2\mathcal{L}_\omega(\boldsymbol{\theta})/\partial a\,\partial b = \partial^2\mathcal{L}_\omega(\boldsymbol{\theta})/\partial b\,\partial a$ for all parameter pairs. Through matrix addition we yield the total Hessian, $\mathbf{H}^{\text{np}}_n(a,b,c)$ in Equation (6). Finally, we obtain the 3×3 sensitivity matrix, $\mathbf{A}^{\text{np}}_n(a,b,c) = -\frac{1}{n}\mathbf{H}^{\text{np}}_n(a,b,c)$, of the ABC model parameters.

The analytic expressions of the second-order partial derivatives demonstrate that the sensitivity matrix $\mathbf{A}^{\text{np}}_n(a,b,c)$ is linearly dependent on λ. As a result, we can write $\mathbf{A}^{\text{np}}_n(a,b,c) = \lambda\mathbf{A}^{\text{n}}_n(a,b,c)$, where $\mathbf{A}^{\text{n}}_n(a,b,c)$ is the sensitivity matrix of the normal log-likelihood $\mathcal{L}^{\text{n}}_n(\boldsymbol{\theta})$. This concludes the derivation of the sensitivity and variability matrices of the ABC model.

Appendix D: Description of the HYdrologic MODEL

The HYdrologic MODel (HYMOD) originates from the PhD thesis of Boyle [23] and describes the rainfall-discharge relationship using five fictitious control volumes. These reservoirs simulate processes such as evaporation, percolation, river inflow and baseflow (see Figure D.1). HYMOD is coded in MATLAB and C++ and uses a mass-conservative second-order integration method. Adaptive time stepping guarantees a robust and accurate numerical solution of the simulated fluxes and state variables.

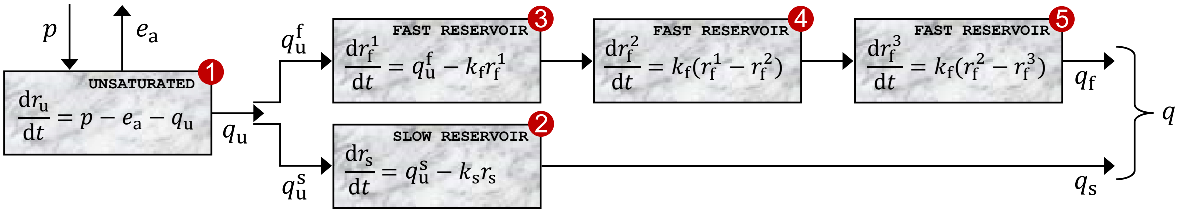
p
e_{a}
UNSATURATED
1
$\frac{\mathrm{d}r_{\mathrm{u}}}{\mathrm{d}t} = p - e_{\mathrm{a}} - q_{\mathrm{u}}$
q_{u}
$q_{\mathrm{u}}^{\mathrm{f}}$
$q_{\mathrm{u}}^{\mathrm{s}}$
FAST RESERVOIR
3
$\frac{\mathrm{d}r_{\mathrm{f}}^{1}}{\mathrm{d}t} = q_{\mathrm{u}}^{\mathrm{f}} - k_{\mathrm{f}} r_{\mathrm{f}}^{1}$
FAST RESERVOIR
4
$\frac{\mathrm{d}r_{\mathrm{f}}^{2}}{\mathrm{d}t} = k_{\mathrm{f}}(r_{\mathrm{f}}^{1} - r_{\mathrm{f}}^{2})$
FAST RESERVOIR
5
$\frac{\mathrm{d}r_{\mathrm{f}}^{3}}{\mathrm{d}t} = k_{\mathrm{f}}(r_{\mathrm{f}}^{2} - r_{\mathrm{f}}^{3})$
SLOW RESERVOIR
2
$\frac{\mathrm{d}r_{\mathrm{s}}}{\mathrm{d}t} = q_{\mathrm{u}}^{\mathrm{s}} - k_{\mathrm{s}} r_{\mathrm{s}}$
q_{f}
q_{s}
q

Figure D.1: Schematic illustration of the HYdrologic MODel. The gray boxes are fictitious control volumes of the watershed which govern the rainfall-runoff transformation. The water storage in each reservoir is measured by the state variables r_{u}, r_{s}, r_{f}^{1}, r_{f}^{2} and r_{f}^{3} with units of length (mm). Arrows signify water fluxes into and out of the compartments, including precipitation, p, evaporation, e_{a}, precipitation converted into flow, q_{u}, fast flow, q_{f}, and baseflow, q_{s}. These fluxes have units of length per time (mm/d) and are computed as follows, $q_{\mathrm{u}} = p\{1 - (1 - \overline{r}_{\mathrm{u}})^{b}\}$, $e_{\mathrm{a}} = e_{\mathrm{p}}\overline{r}_{\mathrm{u}}(1 + c)/(\overline{r}_{\mathrm{u}} + c)$, $q_{\mathrm{u}}^{\mathrm{f}} = aq_{\mathrm{u}}$, $q_{\mathrm{u}}^{\mathrm{s}} = (1 - a)q_{\mathrm{u}}$, $q_{\mathrm{f}} = k_{\mathrm{f}} r_{\mathrm{f}}^{3}$ and $q_{\mathrm{s}} = k_{\mathrm{s}} r_{\mathrm{s}}$, where e_{p} (mm/d) signifies the potential evapotranspiration rate, $\overline{r}_{\mathrm{u}} = r_{\mathrm{u}}/r_{\mathrm{u,max}}$, $c = 10^{-2}$ and $r_{\mathrm{u,max}}$ (mm), a (-), b (-), k_{s} (d^{-1}) and k_{f} (d^{-1}) are unknown parameters.

Table D.1 discusses the five HYMOD parameters with their corresponding symbols, units, and lower and upper bounds. This concludes the description of HYMOD.

Table D.1: Description of HYMOD parameters and their symbols, units, and lower and upper bounds.

Parameter	Description	Units	Min.	Max.
$r_{\mathrm{u,max}}$	Maximum storage unsaturated zone	mm	50	1000
a	Flow partitioning coefficient	–	10^{-3}	1
b	Spatial variability of soil moisture capacity	–	10^{-1}	10
k_{s}	Recession constant of slow reservoir	1/d	10^{-4}	1
k_{f}	Recession constant of fast reservoir	1/d	10^{-1}	5

Appendix E: Bread and Meat Matrices

In this Appendix, we present the sensitivity and variability matrices of the MAP solution of the HYMOD parameters $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^{\top}$ for a few of the likelihood functions analyzed by Vrugt et al. [166].

Tables E.1 and E.2 present the HYMOD bread and meat matrices for the GL^{+}: $L_n^{\mathrm{g+}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3}, \beta = 0)$ and SL: $L_n^{\mathrm{s}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3}, \nu = n-d)$ functions with nuisance variables ξ and ϕ_1. For benchmark purposes, Table E.3 present the results for the Normal Likelihood $L_n^{\mathrm{n}}(\boldsymbol{\theta}, \phi_1 \mid s_0 = 10^{-3})$, with nuisance variable ϕ_1 for the first-order autoregressive scheme of the studentized discharge residuals.

Table E.1: Sensitivity $\widehat{\mathbf{A}}_n^{\mathrm{g+}}$ and HAC variability $\widehat{\boldsymbol{\beta}}_n^{\mathrm{g+}}$ matrices for the MAP solution of HYMOD parameters $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^{\top}$ and nuisance variables ξ and ϕ_1 under the Generalized Likelihood plus $L_n^{\mathrm{g+}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3}, \beta = 0)$.

$$\widehat{\mathbf{A}}_n^{\mathrm{g+}} = 10^2 \begin{array}{rrrrrrrl} \overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} & \overline{\xi} & \overline{\phi}_1 & \\ 0.477 & 0.049 & -0.268 & 0.041 & 0.154 & 0.045 & -0.042 & \overline{r}_{\mathrm{u,max}} \\ 0.049 & 0.144 & -0.045 & 0.066 & 0.070 & 0.140 & -0.012 & \overline{a} \\ -0.268 & -0.045 & 2.213 & -0.277 & 0.195 & 0.031 & 0.041 & \overline{b} \\ 0.041 & 0.066 & -0.277 & 0.213 & -0.072 & 0.075 & 0.003 & \overline{k}_{\mathrm{s}} \\ 0.154 & 0.070 & 0.195 & -0.072 & 0.612 & 0.103 & -0.064 & \overline{k}_{\mathrm{f}} \\ 0.045 & 0.140 & 0.031 & 0.075 & 0.103 & 1.097 & 0.125 & \overline{\xi} \\ -0.042 & -0.012 & 0.041 & 0.003 & -0.064 & 0.125 & 0.104 & \overline{\phi}_1 \end{array}$$

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{g+}} = 10^2 \begin{array}{rrrrrrrl} 6.674 & 0.293 & -3.084 & 0.366 & 4.083 & -0.203 & -1.098 & \overline{r}_{\mathrm{u,max}} \\ 0.293 & 1.941 & 0.974 & 1.489 & 1.197 & 0.281 & 0.129 & \overline{a} \\ -3.084 & 0.974 & 9.154 & -1.717 & 2.440 & 2.558 & -1.149 & \overline{b} \\ 0.366 & 1.489 & -1.717 & 4.322 & -1.607 & -0.496 & 1.833 & \overline{k}_{\mathrm{s}} \\ 4.083 & 1.197 & 2.440 & -1.607 & 9.561 & -0.160 & -3.090 & \overline{k}_{\mathrm{f}} \\ -0.203 & 0.281 & 2.558 & -0.496 & -0.160 & 3.591 & 0.898 & \overline{\xi} \\ -1.098 & 0.129 & -1.149 & 1.833 & -3.090 & 0.898 & 2.645 & \overline{\phi}_1 \end{array}$$

Table E.2: Sensitivity $\widehat{\mathbf{A}}_n^{\mathrm{s}}$ and HAC variability $\widehat{\boldsymbol{\beta}}_n^{\mathrm{s}}$ matrices for the MAP solution of HYMOD parameters $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^\top$ and nuisance variables ξ and ϕ_1 under the Student likelihood $L_n^{\mathrm{s}}(\boldsymbol{\theta}, \xi, \phi_1 \mid s_0 = 10^{-3}, \nu = n - d)$.

$$\begin{array}{cc}
 & \begin{array}{ccccccc} \overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} & \overline{\xi} & \overline{\phi}_1 \end{array} \\
\widehat{\mathbf{A}}_n^{\mathrm{s}} = 10^2 & \left[\begin{array}{rrrrrrr}
0.473 & 0.049 & -0.285 & 0.067 & 0.153 & 0.044 & -0.042 \\
0.049 & 0.144 & -0.045 & 0.066 & 0.070 & 0.140 & -0.012 \\
-0.285 & -0.045 & 2.192 & -0.246 & 0.195 & 0.034 & 0.041 \\
0.067 & 0.066 & -0.246 & 0.213 & -0.072 & 0.074 & 0.004 \\
0.153 & 0.070 & 0.195 & -0.072 & 0.612 & 0.103 & -0.064 \\
0.044 & 0.140 & 0.034 & 0.074 & 0.103 & 1.099 & 0.125 \\
-0.042 & -0.012 & 0.041 & 0.004 & -0.064 & 0.125 & 0.105
\end{array}\right] \begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \\ \overline{\xi} \\ \overline{\phi}_1 \end{array}
\end{array}$$

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{s}} = 10^2 \left[\begin{array}{rrrrrrr}
6.603 & 0.288 & -2.990 & 0.311 & 4.107 & -0.194 & -1.112 \\
0.288 & 1.945 & 0.930 & 1.470 & 1.201 & 0.284 & 0.126 \\
-2.990 & 0.930 & 9.051 & -1.751 & 2.492 & 2.498 & -1.166 \\
0.311 & 1.470 & -1.751 & 4.314 & -1.626 & -0.491 & 1.838 \\
4.107 & 1.201 & 2.492 & -1.626 & 9.556 & -0.157 & -3.085 \\
-0.194 & 0.284 & 2.498 & -0.491 & -0.157 & 3.584 & 0.906 \\
-1.112 & 0.126 & -1.166 & 1.838 & -3.085 & 0.906 & 2.642
\end{array}\right] \begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \\ \overline{\xi} \\ \overline{\phi}_1 \end{array}$$

Table E.3: Sensitivity $\widehat{\mathbf{A}}_n^{\mathrm{n}}$ and HAC variability $\widehat{\boldsymbol{\beta}}_n^{\mathrm{n}}$ matrices for the MAP solution of HYMOD parameters $\boldsymbol{\theta} = (r_{\mathrm{u,max}}, a, b, k_{\mathrm{s}}, k_{\mathrm{f}})^\top$ under the Normal Likelihood $L_n^{\mathrm{n}}(\boldsymbol{\theta}, \phi_1 \mid s_0 = 10^{-3})$ with AR(1)-process of the studentized discharge residuals.

$$\begin{array}{cc}
 & \begin{array}{cccccc} \overline{r}_{\mathrm{u,max}} & \overline{a} & \overline{b} & \overline{k}_{\mathrm{s}} & \overline{k}_{\mathrm{f}} & \overline{\phi}_1 \end{array} \\
\widehat{\mathbf{A}}_n^{\mathrm{n}} = 10^1 & \left[\begin{array}{rrrrrr}
0.131 & 0.154 & -0.170 & 0.060 & 0.311 & -0.171 \\
0.154 & 2.047 & -0.261 & 0.919 & 1.273 & -0.797 \\
-0.170 & -0.261 & 0.292 & -0.189 & -0.223 & 0.180 \\
0.060 & 0.919 & -0.189 & 1.194 & -0.041 & -0.907 \\
0.311 & 1.273 & -0.223 & -0.041 & 7.369 & -0.587 \\
-0.171 & -0.797 & 0.180 & -0.907 & -0.587 & 2.553
\end{array}\right] \begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \\ \overline{\phi}_1 \end{array}
\end{array}$$

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{n}} = 10^3 \left[\begin{array}{rrrrrr}
0.014 & 0.006 & -0.009 & -0.005 & 0.063 & -0.018 \\
0.006 & 0.234 & 0.007 & 0.133 & 0.224 & -0.241 \\
-0.009 & 0.007 & 0.023 & -0.001 & 0.030 & -0.017 \\
-0.005 & 0.133 & -0.001 & 0.146 & 0.009 & -0.210 \\
0.063 & 0.224 & 0.030 & 0.009 & 1.025 & -0.234 \\
-0.018 & -0.241 & -0.017 & -0.210 & -0.234 & 0.566
\end{array}\right] \begin{array}{l} \overline{r}_{\mathrm{u,max}} \\ \overline{a} \\ \overline{b} \\ \overline{k}_{\mathrm{s}} \\ \overline{k}_{\mathrm{f}} \\ \overline{\phi}_1 \end{array}$$

References

[1] A. C. Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936. doi: 10.1017/S0370164600014346.

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

[3] T. W. T. W. Anderson. *The statistical analysis of time series / [by] T.W. Anderson*. Wiley series in probability and mathematical statistics. Wiley, New York, 1971. ISBN 0471029009.

[4] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972. URL http://www.jstor.org/stable/j.ctt13x12sw.

[5] D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991. ISSN 0012-9682. doi: 10.2307/2938229. URL https://doi.org/10.2307/2938229.

[6] D. W. K. Andrews and J. C. Monahan. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–966, 1992.

[7] A. Bárdossy and S. K. Singh. Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, 12(6):1273–1283, 2008. doi: 10.5194/hess-12-1273-2008. URL https://hess.copernicus.org/articles/12/1273/2008/.

[8] M. S. Bartlett. Approximate confidence intervals. *Biometrika*, 40(1/2):12–19, 1953. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2333091.

[9] M. S. Bartlett. Approximate confidence intervals. ii. more than one unknown parameter. *Biometrika*, 40(3/4):306–317, 1953. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2333349.

[10] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974. ISSN 00401706. URL http://www.jstor.org/stable/1267936.

[11] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 2000. ISBN 9780471494645. doi: 10.1002/9780470316870.

[12] K. Beven. Changing ideas in hydrology — the case of physically-based models. *Journal of Hydrology*, 105(1):157–172, 1989. ISSN 0022-1694. doi: https://doi.org/10.1016/0022-1694(89)90101-7. URL https://www.sciencedirect.com/science/article/pii/0022169489901017.

[13] K. Beven. A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1):18–36, 2006. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2005.07.007. URL http://www.sciencedirect.com/science/article/pii/S002216940500332X.

[14] K. Beven. On doing better hydrological science. *Hydrological Processes*, 22(17):3549–3553, 2008. doi: 10.1002/hyp.7108. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.7108.

[15] K. Beven and A. Binley. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298, 1992. doi: 10.1002/hyp.3360060305. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.3360060305.

[16] K. Beven and P. Smith. Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *Journal of Hydrologic Engineering*, 20(1):A4014010, 2015. doi: 10.1061/(ASCE)HE.1943-5584.0000991. URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29HE.1943-5584.0000991.

[17] K. J. Beven. *Environmental Modelling: An Uncertain Future?* Routledge, London, 2009.

[18] V. P. Bhapkar. On a measure of efficiency of an estimating equation. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 34(4):467–472, 1972. ISSN 0581572X. URL http://www.jstor.org/stable/25049833.

[19] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

[20] G. Box, G. Jenkins, G. Reinsel, and G. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN 9781118674925. URL https://books.google.com/books?id=rNt5CgAAQBAJ.

[21] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953. ISSN 00063444. URL http://www.jstor.org/stable/2333350.

[22] G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, New York, NY, USA, 1992. ISBN 9780471574286. doi: 10.1002/9781118033197.

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

[24] N. Breslow. Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85(410):565–571, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289799.

[25] K. L. Bristow and M. J. Savage. Estimation of parameters for the Philip two-term infiltration equation applied to field soil experiments. *Australian Journal of Soil Research*, 25:369–375, 1987. doi: 10.1071/SR9870369.

[26] N. Bulygina and H. Gupta. Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. *Water Resources Research*, 47(5), 2011. doi: 10.1029/2010WR009614. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR009614.

[27] A. Cameron and P. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005. ISBN 9780521848053. URL https://books.google.nl/books?id=Zf0gCwxC9ocC.

[28] G. Capkun, A. C. Davison, and A. Musy. A robust rainfall-runoff transfser model. *Water Resources Research*, 37(12):3207–3216, 2001. doi: 10.1029/2001WR000295. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001WR000295.

[29] K. C. Chanda. A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, 41(1/2):56–61, 1954. ISSN 00063444. URL http://www.jstor.org/stable/2333005.

[30] R. E. Chandler and S. Bate. Inference for clustered data using the independence log-likelihood. *Biometrika*, 94(1):167–183, 03 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm015. URL https://doi.org/10.1093/biomet/asm015.

[31] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, 1997. doi: 10.1109/83.551699.

[32] C. Chen and G. Yin. Computing the efficiency and tuning constants for m-estimation. In *Proceedings of the 2002 Joint Statistical Meetings*. American Statistical Association, 2002.

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

[34] P. Dawid and S. L. Lauritzen. The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pages 22–28, Tokyo, Japan, 2006.

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

[36] D. Q. F. de Menezes, D. M. Prata, A. R. Secchi, and J. C. Pinto. A review on robust m-estimators for regression analysis. *Computers & Chemical Engineering*, 147:107254, 2021. ISSN 0098-1354. doi: 10.1016/j.compchemeng.2021.107254. URL https://www.sciencedirect.com/science/article/pii/S0098135421000326.

[37] D. Y. de Oliveira and J. A. Vrugt. The treatment of uncertainty in hydrometric observations: A probabilistic description of streamflow records. *Water Resources Research*, 58(11):e2022WR032263, 2022. doi: 10.1029/2022WR032263. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022WR032263. e2022WR032263 2022WR032263.

[38] W. J. den Haan and A. T. Levin. A practitioner's guide to robust covariance matrix estimation. In *Robust Inference*, volume 15 of *Handbook of Statistics*, pages 299–342. Elsevier, 1997. doi: https://doi.org/10.1016/S0169-7161(97)15014-3. URL https://www.sciencedirect.com/science/article/pii/S0169716197150143.

[39] J. E. Dennis, Jr. and R. E. Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics - Simulation and Computation*, 7(4):345–359, 1978. doi: 10.1080/03610917808812083. URL https://doi.org/10.1080/03610917808812083.

[40] J. D'Errico. Adaptive robust numerical differentiation, 2024. URL https://www.mathworks.com/matlabcentral/fileexchange/13490-adaptive-robust-numerical-differentiation. MATLAB Central File Exchange. Retrieved March 29, 2024.

[41] M. L. di San Miniato and N. Sartori. Adjusted composite likelihood for robust Bayesian meta-analysis. 2021. doi: 10.48550/arXiv.2104.01920.

[42] P. Diggle. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. OUP Oxford, 2002. ISBN 9780198524847. URL https://books.google.com/books?id=kKLbyWycRwcC.

[43] N. R. Draper and I. Guttman. Confidence intervals versus regions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(3):399–403, 1995. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2348711.

[44] J. Durbin. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1):139–153, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1960.tb00361.x. URL https://doi.org/10.1111/j.2517-6161.1960.tb00361.x.

[45] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2 edition, 2012. ISBN 9780199641178.

[46] B. Efron. Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1301–1304, 1986. ISSN 00905364. URL http://www.jstor.org/stable/2241457.

[47] F. Eicker. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, California, 1965/66), Vol. I: Statistics*, pages 59–82. Univ. California Press, Berkeley, CA, 1967.

[48] R. C. Fair. On the Robust Estimation of Econometric Models. In *Annals of Economic and Social Measurement, Volume 3, number 4*, NBER Chapters, pages 667–677. National Bureau of Economic Research, Inc, June 1974. URL https://ideas.repec.org/h/nbr/nberch/10207.html.

[49] P. E. Ferreira. Multiparametric estimating equations. *Annals of the Institute of Statistical Mathematics*, 34:423–431, 1982.

[50] M. B. Fiering. *Streamflow Synthesis*. PhD thesis, Harvard University, 1967.

[51] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1922.0009.

[52] D. Frazier, R. Kohn, C. Drovandi, and D. Gunawan. Reliable Bayesian inference in misspecified models. Technical Report arXiv:2302.06031, Monash University, 2023. URL http://arxiv.org/abs/2302.06031.

[53] D. A. Freedman. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302, 2006. doi: 10.1198/000313006X152207. URL https://doi.org/10.1198/000313006X152207.

[54] D. A. Freedman and P. Diaconis. On inconsistent m-estimators. *The Annals of Statistics*, 10(2):454–461, 1982. ISSN 00905364. URL http://www.jstor.org/stable/2240679.

[55] J. E. Freer, T. Krueger, and K. J. Beven. Limits of acceptability: A framework for combining data errors and modelling uncertainty to benchmark our predictive capability. In *Paper number H31L-03. American Geophysical Union Meeting, San Francisco, USA, 13-17th December, 2010*, dec 2010.

[56] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, London, 1995.

[57] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the ISI, Bulletin of the ISI*, volume 52, 1987.

[58] F. Giummolè, V. Mameli, E. Ruli, and L. Ventura. Objective Bayesian inference with proper scoring rules. *TEST*, 28(3):728–755, jul 2018. doi: 10.1007/s11749-018-0597-z.

[59] V. Godambe. *Estimating Functions*. Oxford science publications. Clarendon Press, 1991. ISBN 9780198522287. URL https://books.google.nl/books?id=V7P2wAEACAAJ.

[60] V. P. Godambe. An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960. doi: 10.1214/aoms/1177705693. URL https://doi.org/10.1214/aoms/1177705693.

[61] H. V. Gupta, S. Sorooshian, and P. O. Yapo. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998. doi: 10.1029/97WR03495. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97WR03495.

[62] H. V. Gupta, T. Wagener, and Y. Liu. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22(18):3802–3813, 2008. doi: 10.1002/hyp.6989. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.6989.

[63] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1):80–91, 2009. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2009.08.003. URL http://www.sciencedirect.com/science/article/pii/S0022169409004843.

[64] H. V. Gupta, M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48(8), 2012. doi: https://doi.org/10.1029/2011WR011044. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011WR011044.

[65] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994. URL http://www.jstor.org/stable/j.ctv14jx6sm.

[66] F. R. Hampel. *Contribution to the theory of robust estimation*. PhD thesis, University of California, Berkeley, 1968.

[67] F. R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971. doi: 10.1214/aoms/1177693054. URL https://doi.org/10.1214/aoms/1177693054.

[68] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962. URL https://app.dimensions.ai/details/publication/pub.1058301356.

[69] F. R. Hampel. *Robust Statistics: The Approach Based on Influence Functions*. Probability and Statistics Series. Wiley, 1986. ISBN 9780471632382. URL https://books.google.com/books?id=KXWMNAAACAAJ.

[70] B. E. Hansen. Autoregressive conditional density estimation. *International Economic Review*, 35:705–730, 1994.

[71] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

[72] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. doi: 10.1109/PROC.1978.10837.

[73] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. URL http://www.jstor.org/stable/2334940.

[74] M. R. Hernández-López and F. Francés. Bayesian joint inference of hydrological and generalized error models with the enforcement of total laws. *Hydrology and Earth System Sciences Discussions*, pages 1–40, 01 2017. doi: 10.5194/hess-2017-9. URL https://hess.copernicus.org/preprints/hess-2017-9/.

[75] D. C. Hoaglin, F. Mosteller, and J. W. Tukey. *Understanding Robust and Exploratory Data Analysis*. Wiley series in probability and statistics: Probability and Statistics. Wiley, 1983. ISBN 9780471097778. URL https://books.google.nl/books?id=HRrvAAAAMAAJ.

[76] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, 1977. doi: 10.1080/03610927708827533. URL https://doi.org/10.1080/03610927708827533.

[77] D. Huard and A. Mailhot. Calibration of hydrological model gr2m using Bayesian uncertainty analysis. *Water Resources Research*, 44(2), 2008. doi: 10.1029/2007WR005949. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007WR005949.

[78] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

[79] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 221–233. University of California Press, Berkeley, CA, 1967.

[80] P. J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973. doi: 10.1214/aos/1176342503. URL https://doi.org/10.1214/aos/1176342503.

[81] P. J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1981. doi: 10.1002/0471725250.

[82] A. G. Hunt, R. Holtzman, and B. Ghanbarian. A percolation-based approach to scaling infiltration and evapotranspiration. *Water*, 9(2), 2017.

[83] P. Jaiswal, Y. Gao, M. Rahmati, J. Vanderborght, J. Šimůnek, H. Vereecken, and J. A. Vrugt. Parasite inversion for determining the coefficients and time-validity of Philip's two-term infiltration equation. *Vadose Zone Journal*, 21(1):e20166, 2022. doi: 10.1002/vzj2.20166. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/vzj2.20166.

[84] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(none):175–193, 1906. doi: 10.1007/BF02418571. URL https://doi.org/10.1007/BF02418571.

[85] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, 2nd Edition*, volume 2 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, NY, USA, 1995. ISBN 978-0-471-58494-0.

[86] N. F. Kahal Musakkal and D. Gabda. The sandwich estimator approach counting for inter-site dependence of extreme river flow in sabah. In *Journal of Physics: Conference Series*, volume 890, page 012148. IOP Publishing, 2017.

[87] G. Kauermann and R. J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001. ISSN 01621459. URL http://www.jstor.org/stable/3085907.

[88] D. Kavetski, G. Kuczera, and S. W. Franks. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, 42(3), 2006. doi: 10.1029/2005WR004368. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004368.

[89] B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none):354–381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

[90] J. Knight and P. Raats. The contributions of lewis fry richardson to drainage theory, soil physics, and the soil-plant-atmosphere continuum. *Geophysical Research Abstracts*, 18(EGU2016-10980-1), 2016. URL https://meetingorganizer.copernicus.org/EGU2016/EGU2016-10980-1.pdf.

[91] D. Koutsoyiannis and A. Montanari. Bluecat: A local uncertainty estimator for deterministic simulations and predictions. *Water Resources Research*, 58(1):e2021WR031215, 2022. doi: https://doi.org/10.1029/2021WR031215. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021WR031215. e2021WR031215 2021WR031215.

[92] R. Krzysztofowicz. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, 35(9):2739–2750, 1999. doi: 10.1029/1999WR900099. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999WR900099.

[93] R. Krzysztofowicz. The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249(1):2–9, 2001. ISSN 0022-1694. doi: 10.1016/S0022-1694(01)00420-6. URL https://www.sciencedirect.com/science/article/pii/S0022169401004206.

[94] G. Kuczera. On the validity of first-order prediction limits for conceptual hydrologic models. *Journal of Hydrology*, 103(3):229–247, 1988. ISSN 0022-1694. doi: 10.1016/0022-1694(88)90136-9. URL https://www.sciencedirect.com/science/article/pii/0022169488901369.

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

[96] R. Lamb, K. J. Beven, S. Myrabø, and K. Mørken. Overcoming the limitations of glue uncertainty estimation in hydrological modelling: a comparison of formal and informal bayesian approaches. *Hydrological Processes*, 23(14):1873–1886, 2009. doi: 10.1002/hyp.7249.

[97] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Classics in Applied Mathematics. Society of Industrial and Applied Mathematics, 1995. ISBN 978-0-89871-356-5. doi: 10.1137/1.9781611971217.

[98] E. E. Leamer. Multicollinearity: A bayesian interpretation. *The Review of Economics and Statistics*, 55(3):371–380, 1973. doi: 10.2307/1927962. URL https://doi.org/10.2307/1927962.

[99] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, 1998. ISBN 0-387-98502-6. URL http://www.math.nagoya-u.ac.jp/~richard/teaching/s2019/LC.pdf.

[100] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.1.13. URL https://doi.org/10.1093/biomet/73.1.13.

[101] K.-Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):3–40, 1992. ISSN 00359246. URL http://www.jstor.org/stable/2345947.

[102] B. Lindsay. Conditional score functions: Some optimality results. *Biometrika*, 69(3):503–512, 12 1982. ISSN 0006-3444. doi: 10.1093/biomet/69.3.503. URL https://doi.org/10.1093/biomet/69.3.503.

[103] D. Lu, M. Ye, and M. C. Hill. Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. *Water Resources Research*, 48(9):1087–1096, 2012. doi: 10.1029/2011WR011289. URL https://doi.org/10.1029/2011WR011289.

[104] V. Mameli and L. Ventura. Higher-order asymptotics for scoring rules. *Journal of Statistical Planning and Inference*, 165:13–26, 2015. ISSN 0378-3758. doi: 10.1016/j.jspi.2015.03.005. URL https://www.sciencedirect.com/science/article/pii/S0378375815000567.

[105] P. Mantovan and E. Todini. Hydrological forecasting uncertainty assessment: Incoherence of the glue methodology. *Journal of Hydrology*, 330(1):368–381, 2006. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2006.04.046. URL https://www.sciencedirect.com/science/article/pii/S0022169406002162. Hydro-ecological functioning of the Pang and Lambourn catchments, UK.

[106] J. B. McDonald and W. K. Newey. Partially adaptive estimation of regression models via the generalized t distribution. *Econometric Theory*, 4:428–457, 1988.

[107] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 06 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

[108] A. Montanari and E. Toth. Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resources Research*, 43(5), 2007. doi: 10.1029/2006WR005184. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005184.

[109] R. Morton. Efficiency of estimating equations and the use of pivots. *Biometrika*, 68(1): 227–233, 1981. ISSN 00063444. URL http://www.jstor.org/stable/2335823.

[110] J. Nash and J. Sutcliffe. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. ISSN 0022-1694. doi: 10.1016/0022-1694(70)90255-6. URL https://www.sciencedirect.com/science/article/pii/0022169470902556.

[111] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965.

[112] W. K. Newey and D. L. McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics, Volume 4*, chapter 36, pages 2111–2245. Elsevier, Amsterdam, 1994.

[113] W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913610.

[114] W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61(4):631–653, 1994.

[115] R. A. Olinda, J. Blanchet, C. A. C. dos Santos, V. A. Ozaki, and P. J. Ribeiro Jr. Spatial extremes modeling applied to extreme precipitation data in the state of paraná. *Hydrology and Earth System Sciences Discussions*, 11(11):12731–12764, 2014.

[116] D. B. Özyurt and R. W. Pike. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, 28(3):381–402, 2004. ISSN 0098-1354. doi: 10.1016/j.compchemeng.2003.07.001. URL https://www.sciencedirect.com/science/article/pii/S0098135403001960.

[117] S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277, 2010. doi: 10.1198/jasa.2009.tm08577. URL https://doi.org/10.1198/jasa.2009.tm08577.

[118] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. doi: 10.1214/aoms/1177704472.

[119] F. Pauli, W. Racugno, and L. Ventura. Bayesian composite marginal likelihoods. *Statistica Sinica*, 21(1):149–164, 2011. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24309266.

[120] P. Pennacchi. Robust estimate of excitations in mechanical systems using m-estimators—theoretical background and numerical applications. *Journal of Sound and Vibration*, 310(4):923–946, 2008. ISSN 0022-460X. doi: 10.1016/j.jsv.2007.08.007. URL https://www.sciencedirect.com/science/article/pii/S0022460X07006694.

[121] J. R. Philip. The theory of infiltration: 1. the infiltration equation and its solution. *Soil Science*, 83(5):345–358, 1957.

[122] S. D. Poisson. Sur la probabilite des resultats moyens des observations. *Connaissance des Tems pour l'an 1827*, pages 273–302, 1824.

[123] M. Rahmati, J. Vanderborght, J. Šimůnek, J. A. Vrugt, D. Moret-Fernández, B. Latorre, L. Lassabatere, and H. Vereecken. Soil hydraulic properties estimation from one-dimensional infiltration experiments using characteristic time concept. *Vadose Zone Journal*, 19:1–22, 2020. doi: 10.1002/vzj2.20068.

[124] N. I. Ramesh and A. C. Davison. Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology*, 256(1-2):106–119, 2002.

[125] B. Renard, D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5), 2010. doi: 10.1029/2009WR008328. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009WR008328.

[126] W. J. J. Rey. *Introduction to Robust and Quasi-Robust Statistical Methods*. Universitext. Springer Berlin Heidelberg, 2012. ISBN 9783642693892. URL https://books.google.nl/books?id=UAzpCAAAQBAJ.

[127] M. Ribatet, D. Cooley, and A. C. Davison. Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845, 2012. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24310036.

[128] L. A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1(5): 318–333, 1931. doi: 10.1063/1.1745010. URL https://doi.org/10.1063/1.1745010.

[129] L. F. Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations. *Philosophical Transactions of the Royal Society of London. Series A*, 210:307–357, 1911.

[130] L. F. Richardson. *Weather prediction by numerical process*. Cambridge Mathematical Library. Cambridge University Press, 1922. ISBN 978-0-51161-829-1. doi: 10.1017/CBO9780511618291.

[131] Z. Rulfová, A. Buishand, M. Roth, and J. Kyselý. A two-component generalized extreme value distribution for precipitation frequency analysis. *Journal of hydrology*, 534:659–668, 2016.

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

[133] G. Schoups and J. A. Vrugt. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46(10), 2010. doi: 10.1029/2009WR008933. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009WR008933.

[134] B. A. Shaby. The open-faced sandwich adjustment for mcmc using estimating functions. *Journal of Computational and Graphical Statistics*, 23(3):853–876, 2014. ISSN 10618600. URL http://www.jstor.org/stable/43304925.

[135] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1948.tb01338.x.

[136] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb00917.x. URL https://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1948.tb00917.x.

[137] S. C. Sian and D. Gabda. Modeling the extreme rainfall data of several sites in sabah using sandwich estimator. *Journal of Applied Science and Engineering*, 25(3):517–520, 2021.

[138] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992. doi: 10.2307/2684188.

[139] R. H. Smith. True average of observations? *Nature*, 37, 1888. doi: 10.1038/037464a0. URL https://doi.org/10.1038/037464a0.

[140] P. B. Stark and R. L. Parker. Bounded-variable least-squares: An algorithm and applications. *Computational Statistics*, 10:129–141, 1995.

[141] J. R. Stedinger, R. M. Vogel, S. U. Lee, and R. Batchelder. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, 44(12), 2008. doi: 10.1029/2008WR006822. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR006822.

[142] S. M. Stigler. Studies in the history of probability and statistics. xxxiii cauchy and the witch of agnesi: An historical note on the cauchy distribution. *Biometrika*, 61(2):375–380, 1974. ISSN 00063444. URL http://www.jstor.org/stable/2334368.

[143] M. T. Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31:296–301, 1923.

[144] SymPy Development Team. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. doi: 10.7717/peerj-cs.103. URL doi.org/10.7717/peerj-cs.103.

[145] The MathWorks Inc. MATLAB version: 9.13.0 (r2022b), 2022. URL https://www.mathworks.com.

[146] P. Theodossiou. Financial data and the skewed generalized t distribution. *Management Science*, 44(12):1650–1661, 1998. doi: 10.1287/mnsc.44.12.1650. URL https://www.jstor.org/stable/2634700.

[147] P. Theodossiou. Skewed generalized error distribution of financial assets and options pricing. *Multinational Finance Journal*, 19:223–266, 2015. doi: 10.17578/19-4-1.

[148] M. Thyer, H. Gupta, S. Westra, D. McInerney, H. R. Maier, D. Kavetski, A. Jakeman, B. Croke, C. Simmons, D. Partington, M. Shanafield, and C. Tague. Virtual hydrological laboratories: Developing the next generation of conceptual models to support decision making under change. *Water Resources Research*, 60(4):e2022WR034234, 2024. doi: 10.1029/2022WR034234. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022WR034234. e2022WR034234 2022WR034234.

[149] J. W. Tukey. Some elementary problems of importance to small sample practice. *Human Biology*, 20(4):205–214, 1948.

[150] J. W. Tukey. A survey of sampling from contaminated distributions. In *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, 1960.

[151] J. W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. doi: 10.1214/aoms/1177704711. URL https://doi.org/10.1214/aoms/1177704711.

[152] J. W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2:523–531, 1975. URL https://cir.nii.ac.jp/crid/1573950399770196096.

[153] S. M. Vandeskog, S. Martino, and R. Huser. An efficient workflow for modelling high-dimensional spatial extremes, 2022.

[154] L. Ventura and W. Racugno. Pseudo-likelihoods for Bayesian inference. In T. Di Battista, E. Moreno, and W. Racugno, editors, *Topics on Methodological and Applied Statistical Inference*, pages 205–220, Cham, 2016. Springer International Publishing. ISBN 978-3-319-44093-4.

[155] R. M. Vogel and A. Sankarasubramanian. Validation of a watershed model without calibration. *Water Resources Research*, 39(10), 2003. doi: https://doi.org/10.1029/2002WR001940. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002WR001940.

[156] J. A. Vrugt. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and matlab implementation. *Environmental Modelling & Software*, 75:273–316, 2016. ISSN 1364-8152. doi: 10.1016/j.envsoft.2015.08.013. URL http://www.sciencedirect.com/science/article/pii/S1364815215300396.

[157] J. A. Vrugt. Distribution-based model evaluation and diagnostics: Elicitability, propriety, and scoring rules for hydrograph functionals. *Water Resources Research*, 60(6):e2023WR036710, 2024. doi: 10.1029/2023WR036710. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR036710. e2023WR036710 2023WR036710.

[158] J. A. Vrugt and W. Bouten. Validity of first-order approximations to describe parameter uncertainty in soil hydrologic models. *Soil Science Society of America Journal*, 66(6):1740–1751, 2002. doi: 10.2136/sssaj2002.1740. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.2136/sssaj2002.1740.

[159] J. A. Vrugt and D. Y. de Oliveira. Confidence intervals of the kling-gupta efficiency. *Journal of Hydrology*, 612:127968, 2022. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2022.127968. URL https://www.sciencedirect.com/science/article/pii/S0022169422005431.

[160] J. A. Vrugt and C. G. Diks. The learning rate is not a constant: Sandwich-adjusted Markov chain Monte Carlo simulation. *Entropy*, 27, 2025. doi: 10.3390/e27100999. URL https://doi.org/10.3390/e27100999.

[161] J. A. Vrugt and M. Sadegh. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, 49(7):4335–4345, 2013a. doi: 10.1002/wrcr.20354. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/wrcr.20354.

[162] J. A. Vrugt, H. V. Gupta, W. Bouten, and S. Sorooshian. A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8), 2003. doi: 10.1029/2002WR001642. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002WR001642.

[163] J. A. Vrugt, C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, 41(1), 2005. doi: 10.1029/2004WR003059. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004WR003059.

[164] J. A. Vrugt, C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44(12), 2008. doi: 10.1029/2007WR006720. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007WR006720.

[165] J. A. Vrugt, C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, Oct 2009. ISSN 1436-3259. doi: 10.1007/s00477-008-0274-y. URL https://doi.org/10.1007/s00477-008-0274-y.

[166] J. A. Vrugt, D. Yumi de Oliveira, G. Schoups, and C. G. H. Diks. On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. *Journal of Hydrology*, 615:128542, 2022. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2022.128542. URL https://www.sciencedirect.com/science/article/pii/S002216942201112X.

[167] J. A. Vrugt, J. W. Hopmans, Y. Gao, M. Rahmati, J. Vanderborght, and H. Vereecken. The time validity of Philip's two-term infiltration equation: An elusive theoretical quantity? *Vadose Zone Journal*, page e20309, 2024. doi: 10.1002/vzj2.20309. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/vzj2.20309.

[168] J. A. Vrugt, J. M. Frame, and E. Bollman. Reclaiming first principles: A differentiable framework for conceptual hydrologic models. *Water Resources Research*, 2026. Manuscript under review.

[169] J. Watson and C. Holmes. Approximate Models and Robust Decisions. *Statistical Science*, 31(4):465–489, 2016. doi: 10.1214/16-STS592. URL https://doi.org/10.1214/16-STS592.

[170] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912934.

[171] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

[172] H. White and I. Domowitz. Nonlinear regression with dependent observations. *Econometrica*, 52(1):143–161, 1984. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911465.

[173] C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986. doi: 10.1214/aos/1176350142. URL https://doi.org/10.1214/aos/1176350142.

[174] J. Yang, P. Reichert, and K. C. Abbaspour. Bayesian uncertainty analysis in distributed hydrologic modeling: A case-study in the Thur river basin (Switzerland). *Water Resources Research*, 43(W10401), 2007. doi: 10.1029/2006WR005497.

[175] F. Zheng, E. Thibaud, M. Leonard, and S. Westra. Assessing the performance of the independence method in modeling spatial extreme rainfall. *Water Resources Research*, 51(9):7744–7758, 2015.